T.C.
MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN
PURE AND APPLIED SCIENCES

MODEL BASED FAULT MONITORING SYSTEM IN MACHINING: AN APPLIED STUDY

Necla YEL

THESIS

FOR THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

SUPERVISORS
Prof. Dr. S. Ümit OKTAY FIRAT
Associate Prof. Dr. Farid NOUREDDINE

İSTANBUL 2011

T.C.
MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN
PURE AND APPLIED SCIENCES

MODEL BASED FAULT MONITORING SYSTEM IN MACHINING: AN APPLIED STUDY

Necla YEL
(524408007)

THESIS

FOR THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

SUPERVISORS
Prof. Dr. S. Ümit OKTAY FIRAT
Associate Prof. Dr. Farid NOUREDDINE

İSTANBUL 2011

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor, Prof. Dr. S. Ümit OKTAY FIRAT, for supporting and encouraging to me during the preparation of this thesis. I would like to mention her invaluable experience and knowledge that helped me complete my study.

I would like to express my sincere gratitude to my thesis co-advisor Associate Professor Dr. Farid NOUREDDINE, for supporting me during the preparation of the application of this thesis in France.

I would like to express my sincere gratitude to Dr. Rachid NOUREDDINE, for providing me data during the preparation of application of this thesis in France.

I would like to thank, Prof. Dr. Abbas Azimli, from Yildiz Technical University for encouraging to me to begin master education. I would like to express my sincere gratitude to him.

Finally, I would like to thank my family and my friends for supporting and encouraging to me throughout all my life.

CONTENTS

PAGE

ABSTRACT

Model Based Fault Monitoring System in Machining:
An Applied Study

Technological evolutions, spreading into our life on an increasing scale with each passing day, are making human life simpler in many aspects. With technology, product and service quality has increased, the error rate has decreased and processing speed has increased. Technological improvements in computer are currently being considered as the most essential revolution in history of civilization . Despite all these positive results, it has caused many types of fraud which is a significant source of lost revenue to the companies.

Anomaly Detection and Fraud Detection, which are one of Data Mining application areas, are two of techniques which used for making detection. The aim of these two disciplines is to find object that are different from most other object .The importance of anomaly detection is due to the fact that anomalies in data translate to significant (and often critical) actionable information in a wide variety of computer science and engineering.

In this study, in order to predict future possible faults, models were built by analyzing the data which are taken from the cutting machine Latent by using Taguchi method. The study contains two parts: in the first part artificial neural networks and response surface analysis used to find optimum parameters and in the second part, by assuming other inputs are constraint, regression model were built between dependent variable and time.

The aim of this study is to help companies not to lose income by using Fault Detection Techniques, primarily. As a result of comprehensive literature research, resources which found were benchmarked, made comparison to their usage and advantages, disadvantages and made suggestion to improve these methods.

May 2011 **Necla YEL**

ÖZET

Makinelerde Model Tabanlı Arıza Yönetim Sistemi: Bir Uygulama

Her geçen gün, teknolojide meydana gelen gelişmeler, hayatımızın hemen hemen her alanında kendini göstermenin yanında, birçok konuda insan hayatını oldukça kolaylaştırmaktadır. Teknolojiyle beraber, ürün ve hizmet kalitesi artmış, hata oranı azalmış ve işlem hızı artmıştır. Bilgisayarın icadı insan hayatı açısından devrim niteliğindedir. Tüm bu olumlu sonuçlara rağmen, meydana gelen gelişmeler beraberinde bir takım dolandırıcılık türlerinin ortaya çıkmasına neden olmuş , bu da firmaların önemli ölçüde gelir kaybetmelerine mal olmuştur.

Veri madenciliği yöntemlerinden olan Anomali belirleme ve Sahtekarlık tespiti bu dolandırıcılık türlerini ortaya çıkarmak için kullanılan yöntemlerden ikisidir. Her iki prensipte de amaç; genel işlem hacminden farklı olarak meydana gelen bir anormalliği belirlemektir .Anomali belirlemenin önemi, çeşitli bilgisayar bilimleri ve mühendislikte , verideki anomalileri anlamlı (sık sık kritik) dava edilebilir veriye dönüştürebilmektir.

Bu çalışmada, Latent adlı kesici bir makineden Taguchi Methoduyla alınan veriler analiz edilerek, gelecekte meydana gelebilecek muhtemel arızaları tahmin etmeye yönelik modeller kurulmustur. Çalışmanın uygulama aşaması iki kısımdan oluşmaktadır; ilk kısımda, yapay sinir ağları ve yanıt yüzey analizi yöntemi kullanılarak, optimum parametreler belirlenmiş, ikinci aşamada ise diğer girdilerin sabit oldugu varsayılarak zamanla ilişkili regresyon modelleri kurulmustur.

Bu çalışmanın amacı; öncelikle, güncel olarak kullanılan ariza tespiti yöntemlerini kullanarak , firmaların gelir kaybetmemesine yardımcı olmaktır. Bunun için yapılan kapsamlı literatür taraması sonucundaulaşılan yayınlar arasında karşılaştırma yapılmış, birbirlerine göre kullanım açısından avantaj ve dezavantajları karşılaştırılmış ve geliştirilmesi açısından öneriler getirilmiştir.

Mayıs 2011 **Necla YEL**

LIST OF ABBREVIATIONS

2-D	**:** Two Dimensional
SVM	**:** Support Vector Machine
LOF	**:** Local Outlier Factor
COF	**:** Connectivity Based Outlier Factor
ODIN	**:** Outlier Detection Using In-degree Number
MDEF	**:** Multi Granularity Deviation Factor
SOM	**:** Self Organizing Maps
EM	**:** Expectation Maximization
CBLOF	**:** Cluster-Based Local Outlier Factor
MLE	**:** Maximum Likelihood Estimates
ARIMA	**:** Autoregressive Integrated Moving Average
ARMA	**:** Autoregressive Moving Average
Pdf	**:** Probability density function
PCA	**:** Principal Component Analysis
CMD	**:** Compact Matrix Decomposition
SLOM	**:** Spatial Local Outlier Factor
FSA	**:** Finite State Automation
MCMC	**:** Markov Chain Monte Carlo
ART	**:** Adaptive Resonance Theory
HMM	**:** Hidden Markov Model
PST	**:** Probabilistic Suffix Tree
SMT	**:** Sparse Markov Trees
WCAD	**:** Window Comparison Anomaly Detection
CRF	**:** Conditional Random Fields
IMM	**:** Interpolated Markov Models
MGMRF	**:** Multivariate Gaussian Random Markov Fields

LIST OF FIGURES

LIST OF TABLES

CHAPTER I:

INTRODUCTION AND AIM

Anomaly detection is described as a problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Anomalies and outliers are two terms that are used most commonly in the anomaly detection framework equivalently. Many techniques that are detecting outliers are fundamentally identical however, different names are chosen by the authors. For example, most of the authors describe their various approaches as outlier detection, novelty detection, anomaly detection, noise detection, deviation detection or exception mining.

Detecting outliers or anomalies in data has been studied in the statistics community as early as the 19*th* century (Edgeworth, 1887). Over time, a variety of anomaly detection techniques have been developed in several research communities. Many of these techniques have been specifically developed for certain application domains, while others are more inclusive.

The importance of anomaly detection is due to anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination (Kumar 2005). Anomalies in credit card transaction data could indicate credit card or identity theft (Aleskerov et al. 1997) or anomalous readings from a space craft sensor could signify a fault in some component of the space craft (Fujimaki et al. 2005).

What are Anomalies?

In the literature several anomaly definitions have been proposed. According to the definition of Grubbs (1969);

“An outlying observation, or outlier, is one that appears to deviate markedly from

other members of the sample in which it occurs.”

Another definition of anomalies;

“An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.”

Generally, Anomalies are patterns in data that do not adapt to a well defined notion of normal behavior. Figure 1 (Kumar et al.,2005) illustrates anomalies in a simple 2-dimensional data set. The data has two normal regions, N1 and N2, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., point’s *o*1 and *o*2, and points in region *O*3, are anomalies.

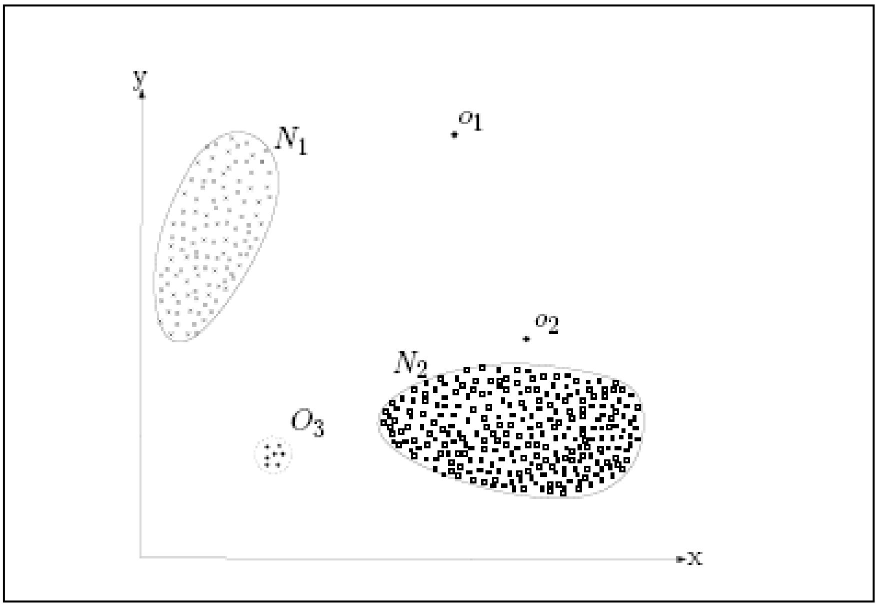


Figure 1: A simple example of anomalies in a 2D data set.(Kumar,2005)

Recently, in aerospace, aviation, transportation, and other complex industrial systems, Predictive Maintenance (PM) strategy has been increasing interest and investment. PM has been proved to be an effective way to control the cost of services to engineering systems while maintaining their desirable operational reliability. PM is a maintenance system where an equipment repair or replacement decisions are based on the current and projected future health of the equipment.

Furthermore it is a program that recommends maintenance decisions with support systems provided by sensors and signal processing techniques. Data mining (DM) algorithms involve monitoring changes of the system state to detect any abnormal behavior. The goal is to detect faults early to provide lead-time for maintenance actions and planning based on the collected faults. Moreover, data mining can discover hard to find incident affinities that can reduce down time resulting from failures. Neural networks are the preferred tool for many predictive data mining applications because of their power, flexibility and ease of use.

A neural network can approximate a wide range of statistical models without requiring that one hypothesizes in advance some relationships between dependent and independent variables. Instead, the form of the relationship is determined during the learning process. In this study, the aim is to show the use of DM algorithms on PM, especially the use of a neural networks on a monitoring surface roughness of machined products. The validity of the models of the surface roughness will be controlled with residual tests, then, the best model will be chosen.

This study contains five main chapters. The first chapter is introduction and aim of this study. In the second chapter, the literature related to subjects were presented to focus on the researches. Aspects of anomaly detection problem, anomaly detection techniques, relative strenghts and weaknesses of anomaly detection techniques, applications of anomaly detection and fraud detection are the related topics of this section.

The third section contains the main related topics of case study. It was focused on the meintenance systems, artificial neural networks, response surface methodologym regression analysis and statistical process control. The forth section is application, it contains all of the analyses related to subject and the results. Finally. The fifth chapter is conclusion that it is summarized the problem and results together.

CHAPTER II:
LITERATURE REVIEW

The literature rewiew was classified as five main groups which are:

1. Aspects of Anomaly Detection Problem
2. Anomaly detection Techniques
3. Relative strength an Weaknesses of Anomaly Detection Techniques
4. Applications of Anomaly Detection
5. Fraud Detection

II.1 ASPECTS OF ANOMALY DETECTION PROBLEM

Aspects of anomaly detection problem are related to the type of data structure. These are can be classified as the following titles:

1. Nature of Input Data
2. Types of Anomalies
3. Availability of Class Labels
4. Output of Anomaly Detection

II.1.1 Nature Of Input Data

One of the key attitudes of any anomaly detection technique is the nature of the input data. Input is generally a collection of data instances (also referred as object, record, point, vector, pattern, event, case, sample, observation, entity) (Tan et al. 2005, Chapter 2). Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, field, and dimension). The attributes can be of different types such as binary, categorical or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate). In the case of multivariate data instances, all attributes might be of same type or might be a mixture of different data types. The nature of attributes determines the applicability of anomaly detection techniques. For example, for statistical techniques different statistical models have to be used for continuous and categorical data. Similarly, for nearest neighbor based techniques, the nature of attributes would determine the distance measure to be used. Often, instead of the actual data, the pair wise distance between instances might be provided in the form of a distance (or similarity) matrix. In such cases, techniques that require original data instances are not applicable, e.g., many statistical and classification based techniques. Input data can also be categorized based on the relationship present among data instances (Tan et al. 2005). Most of the existing anomaly detection techniques deal with record data (or point data), in which no relationship is assumed among the data instances.

In general, data instances can be related to each other. Some examples are sequence data, spatial data, and graph data. In sequence data, the data instances are linearly ordered, e.g., time-series data, genome sequences, protein sequences. In spatial data, each data instance is related to its neighboring instances, e.g., vehicular traffic data, ecological data. When the spatial data has a temporal (sequential) component it is referred to as spatiotemporal data, e.g., climate data. In graph data, data instances are represented as vertices in a graph and are connected to other vertices with edges. Later in this section we will discuss situations where such relationship among data instances becomes relevant for anomaly detection.

II.1.2 Types Of Anomalies

An important aspect of an anomaly detection technique is the nature of the desired anomaly. Anomalies can be classified into following three categories which are shown in figure 2:

1. Point Anomalies
2. Contextual anomalies
3. Collective anomalies

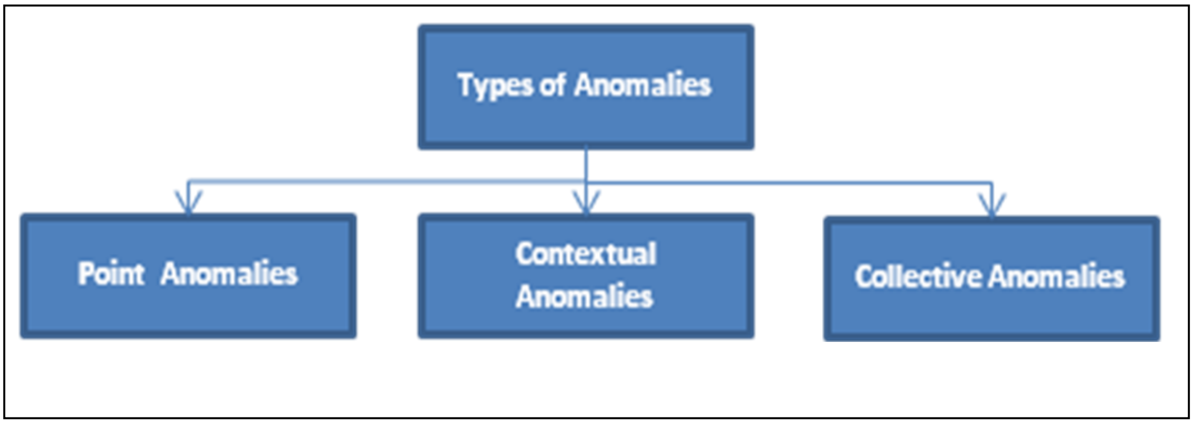


Figure 2: Types of Anomalies

II.1.2.1 Point Anomalies

If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point anomaly. This is the simplest type of anomaly and is the focus of majority of research on anomaly detection. For example, in Figure 1, points *o*1 and *o*2 as well as points in region *O*3 lie outside the boundary of the normal regions, and hence are point anomalies since they are different from normal data points. As a real life example, consider credit card fraud detection. Let the data set correspond to an individual's credit card transactions. For the sake of simplicity, let us assume that the data is defined using only one feature: *amount spent*. A transaction for which the amount spent is very high compared to the normal range of expenditure for that person will be a point anomaly.

II.1.2.2 Contextual Anomalies (Conditional Anomalies)

If a data occasion is not normal in a specific context (but not otherwise), then it is termed as a contextual anomaly. The notion of a context is induced by the structure in the data set and has to be specified as a part of the problem formulation. Each data instance is defined using following two sets of attributes:

(1) *Contextual attributes*. The contextual attributes are used to determine the context (or neighborhood) for that occurrence. For instance, in spatial data sets, the longitude and latitude of a location are the contextual attributes. In time- series data, time is a contextual attribute which determines the position of an instance on the entire sequence.

(2) *Behavioral attributes*. The behavioral attributes define the non-contextual characteristics of an instance. For example, in a spatial data set describing the average rainfall of the entire world, the amount of rainfall at any location is a behavioral attribute.

The anomalous behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual anomaly in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. This property is key in identifying contextual and behavioral attributes for a contextual anomaly detection technique**.**

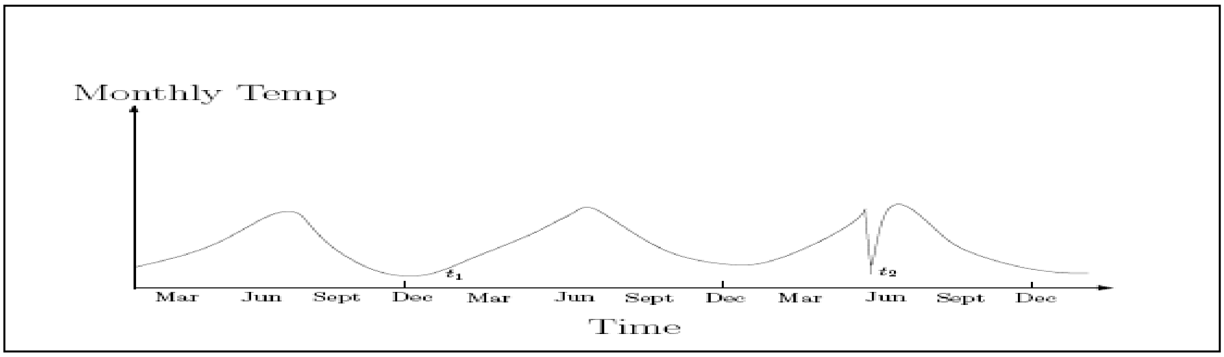


Figure 3: Contextual anomaly *t*2 in a temperature time series. Note that the temperature at time *t*1 is same as that at time *t*2 but occurs in a different context and hence is not considered as an anomaly. (Kumar,2005)

Contextual anomalies have been most commonly explored in time-series data (Weigend et al. 1995; Salvador and Chan 2003) and spatial data. Figure 3 (Kumar at all., 2005) shows such an example for a temperature time series which shows the monthly temperature of an area over last few years. A temperature of 35F might be normal during the winter (at time t1) at that place, but the same value during summer (at time t2) would be an anomaly.

The choice of applying a contextual anomaly detection technique is determined by the meaningfulness of the contextual anomalies in the target application domain. Another key factor is the availability of contextual attributes. In several cases defining a context is straightforward, and hence applying a contextual anomaly detection technique makes sense. In other cases, defining a context is not easy, making it difficult to apply such techniques.

II.1.2.3 Collective Anomalies

If a collection of related data instances is anomalous with respect to the entire data set, it is termed as a collective anomaly. The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous. Figure 4 (Kumar 2005) illustrates an example which shows a human electrocardiogram output. The highlighted region denotes an anomaly because the same low value exists for an abnormally long time (corresponding to an Atrial Premature Contraction). Note that that low value by itself is not an anomaly.

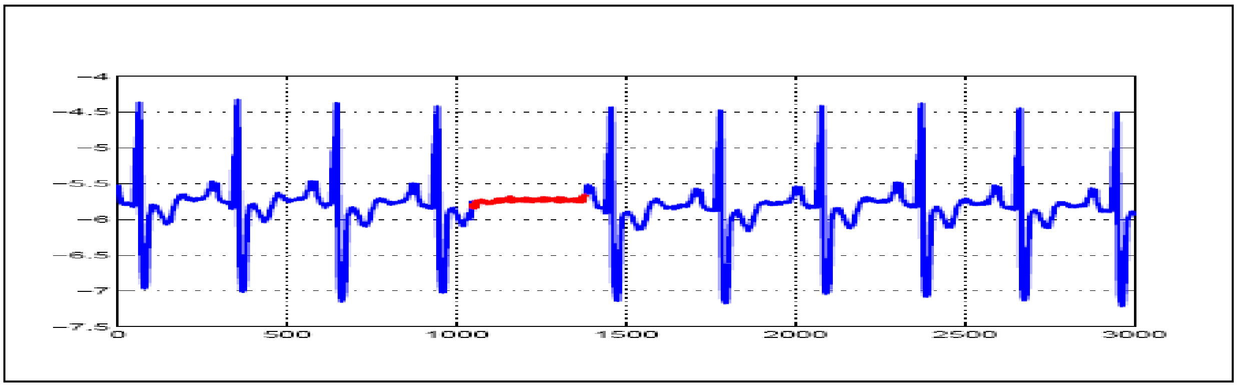


Figure 4: Collective anomaly corresponding to an *Atrial Premature Contraction* in an human electrocardiogram output (Kumar, 2005)

Collective anomalies have been explored for sequence data (Sun et al. 2006), graph data (Noble and Cook 2003), and spatial data (Shekhar et al.2001).

It should be noted that while point anomalies can occur in any data set, collective anomalies can occur only in data sets in which data instances are related. In contrast, occurrence of contextual anomalies depends on the availability of context attributes in the data. A point anomaly or a collective anomaly can also be a contextual anomaly if analyzed with respect to a context. Thus a point anomaly detection problem or collective anomaly detection problem can be transformed to a contextual anomaly detection problem by incorporating the context information.

II.1.3 Availability Of Class Labels

Based on the extent to which the labels are available, anomaly detection techniques can operate in one of the following three modes:

1. Supervised Anomaly Detection Techniques
2. Semi-Supervised Anomaly Detection Techniques
3. Unsupervised Anomaly detection Techniques

II.1.3.1 Supervised Anomaly Detection Techniques

Supervised anomaly detection techniques trained in supervised mode assume that both anomaly and normal behavior labels are available. Typical approach in such cases is to build a predictive model for normal *vs.* anomaly classes. Any hidden data illustration is compared against the model to determine which class it belongs to. There are two main affairs that appear in supervised anomaly detection techniques. First, the anomalous observations are far fewer compared to the normal observation in the data. Issues that arise due to imbalanced class distributions have been addressed in the data mining and machine learning literature (Chawla et al. 2004; Weiss and Hirsh 1998). Second, obtaining accurate and representative labels, especially for the anomaly class is usually challenging.

A number of techniques have been submitted that insert artificial anomalies in a normal data set to obtain a labeled training data set (Abe et al. 2006). Other than these two affairs, the supervised anomaly detection problem is similar to building predictive models.

II.1.3.2 Semi-Supervised Anomaly Detection Techniques

Semi-Supervised anomaly detection techniques that operate in a semi-supervised mode, assume that the training data set has been labeled illustrations for only the normal class behavior. Due to the fact that they do not require labels for the anomaly class, they are more widely preferred than supervised techniques. For instance, in space craft fault detection (Fujimaki et al. 2005), an anomaly scenario would signify an accident, which is not easy to model. The typical approach used in such techniques is to formulate a model for the class identical to normal behavior, and use this model in order to identify anomalies in the test dataset. A limited set of anomaly detection techniques exist that assume availability of only the anomaly instances for training (Forrest et al. 1996). Such techniques are not generally used, primarily because it is difficult to obtain a training data set which covers every possible anomalous behavior that can occur in the data.

II.1.3.3 Unsupervised Anomaly Detection Techniques

Unsupervised anomaly detection techniques that operate in unsupervised mode do not require training data and any kinds of class labels thus are most widely applicable. The techniques in this category make the implied assumption that normal instances are far more frequent than anomalies in the test data. If this assumption is not true then such techniques suffer from high false alarm rate. Many semi-supervised techniques can be adapted to operate in an unsupervised mode by using a sample of the unlabeled data set as training data. Such adaptation assumes that the test data contains very few anomalies and the model learnt during training is robust to these few anomalies.

II.1.4 Output Of Anomaly Detection

An important aspect for any anomaly detection technique is the manner in which the anomalies are reported. Typically, there are two types outputs produced by anomaly detection techniques:

1. Scores
2. Labels

II.1.4.1 Scores

Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Therefore the output of such techniques is a sorted list of anomalies. Two ways exist to select the anomalies;The first one is analyzing top few anomalies and the second one is using cut-off threshold.

II.1.4.2 Labels

Techniques in this category assign a label (normal or anomalous) to each test instance. Scoring based anomaly detection techniques allow the analyst to use a domain- specific threshold to select the most relevant anomalies. Techniques that provide binary labels to the test instances do not directly allow the analysts to make such a choice, though this can be controlled indirectly through parameter choices within each technique.

II.2 ANOMALY DETECTION TECHNIQUES

According to the type of data, availability of class labels and programs, different anomaly detection techniques are exists. A taxonomy about anomaly detection techniaues is shown in figure 5:

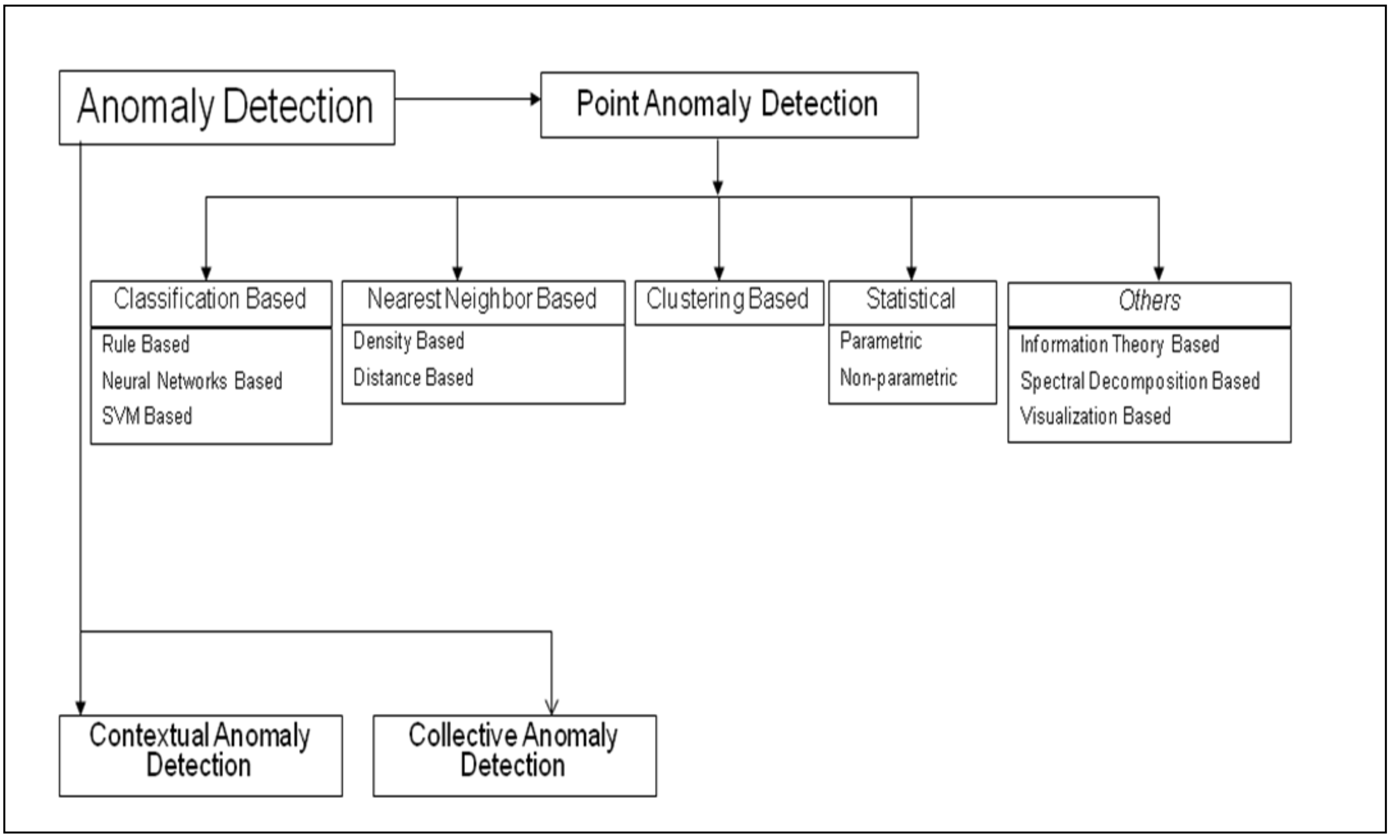


Figure 5: Anomaly Detection Techniques

II.2.1. Classification Based Anomaly Detection:

Classification is used to build a model from a set of labeled data instances (*training*) and then, classify a test data into one of the classes using the learnt model (*testing*) (Tan et al. 2005). Classification based anomaly detection techniques act in a similar two-phase fashion. The training phase learns a classifier using the available labeled training data. The testing phase classifies a test instance as normal or anomalous using the classifier. Classification based anomaly detection techniques operate under the following general assumption that a classifier that can determine between normal and anomaly classes can be learnt in the given feature space.

Based on the labels available for training phase, classification based anomaly detection techniques can be grouped into two broad categories: multi-class and one-class anomaly detection techniques. Multi-class classification based anomaly detection techniques assume that the training data contains labeled instances belonging to multiple normal. Such anomaly detection techniques learn a classifier to distinguish between each normal class against the rest of the classes.

The illustration of this technique is shown in the figure 6(a) (Kumar, 2005). A test instance is considered anomalous if it is not classified as normal by any of the classifiers. Some techniques in this sub-category associate a confidence score with the prediction made by the classifier. If none of the classifiers are confident in classifying the test instance as normal, the instance is declared to be anomalous. One-class classification based anomaly detection techniques assume that all training instances have only one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm, e.g., one-class SVMs, one-class Kernel Fisher Discriminants as shown in Figure 6(b) (Kumar, 2005). Any test instance that does not fall within the learnt boundary is declared as anomalous.

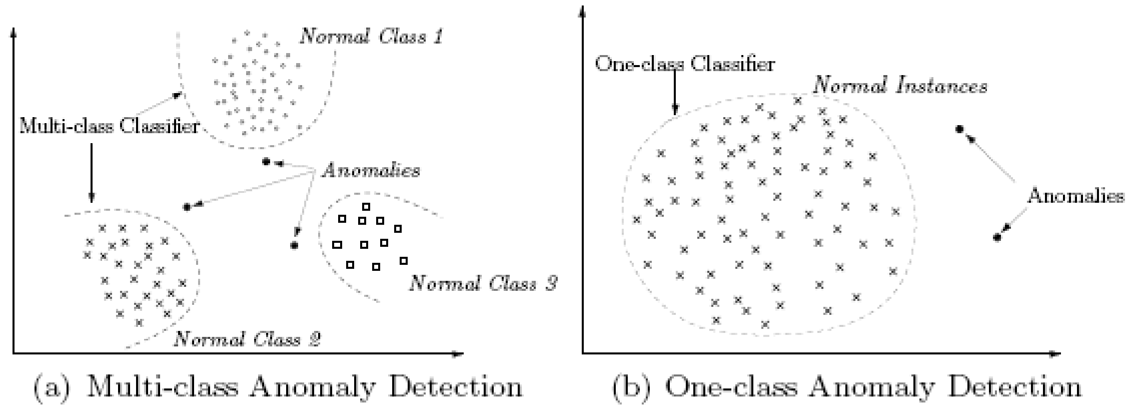


Figure 6: Using Classification for Anomaly Detection (Kumar,2005)

A variety of anomaly detection techniques that use different classification algorithms to build classifiers:

II.2.1.1Rule Based Techniques

Rule based anomaly detection techniques used for learning rules to distinguish the normal behaviors from anomalous class. A test instance that is not covered by any such rule is considered as an anomaly. Rule based techniques can be applied to both multi-class and one-class setting.

A basic multi-class rule based technique includes two steps. First step is to learn rules from the training data using a rule learning algorithm. Each rule has an associated confidence value which is proportional to ratio between the number of training instances correctly classified by the rule and the total number of training instances covered by the rule. Second step is to find, for each test instance, the rule that best captures the test instance. The inverse of the confidence associated with the best rule is the anomaly score of the test instance. Several minor variants of the basic rule based technique have been proposed (Lee et al. 2000; Salvador and Chan 2003).

Association rule mining has been used for one-class anomaly detection by generating rules from the data in an unsupervised fashion. Association rules are generated from a categorical data set. To ensure that the rules correspond to strong patterns, a support threshold is used to prune out rules with low support (Tan et al. 2005). Association rule mining based techniques have been used for network intrusion detection (Tandon and Chan 2007), system call intrusion detection (Lee et al. 2000), credit card fraud detection, and fraud detection in spacecraft housekeeping data (Yairi et al. 2001). Frequent item sets are generated in the intermediate step of association rule mining algorithms. (Yairi et al. 2001) propose an anomaly detection algorithm for categorical data sets in which the anomaly score of a test instance is equal to the number of frequent item sets it occurs in.

II.2.1.2 Neural Networks Based Techniques

Neural networks have been applied to anomaly detection in multi-class as well as one-class setting.

A basic multi-class anomaly detection technique using neural networks operates in two steps. First, a neural network is trained on the normal training data to learn the different normal classes. Second, each test instance is provided as an input to the neural network. If the network accepts the test input, it is normal and if the network rejects a test input, it is an anomaly. Several variants of the basic neural network technique have been proposed that use different types of neural networks. Replicator Neural Networks have been used for one-class anomaly detection. A multi-layer feed forward neural network is constructed that has the same number of input and output neurons (corresponding to the features in the data). The training involves compressing data into three hidden layers. The testing phase involves reconstructing each data instance *xi* using the learnt network to obtain the reconstructed output o_i. The reconstruction error δ_i *for* the test instance x_i is then computed as:

$$\delta_i = \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - o_{ij})^2 \tag{1}$$

Where n is the number of features over which the data is defined. The reconstruction error δ_i is directly used as an anomaly score for the test instance.

II.2.1.3 Support Vector Machines Based Techniques

Support Vector Machines (SVMs) have been applied to anomaly detection in the one-class setting. Such techniques use one class learning techniques for SVM and learn a region that contains the training data instances (a boundary). Kernels, such as radial basis function (RBF) kernel, can be used to learn complex regions. For each test instance, the basic technique determines if the test instance falls within the learnt region. If a test instance falls within the learnt region, it is determined as normal; else it is declared as anomalous. Variants of the basic technique have been proposed for anomaly detection in audio signal, novelty detection in power generation plants and system call intrusion detection. The basic technique also been extended to detect anomalies in temporal. A variant of the basic technique finds the smallest hyper-sphere in the kernel space, which contains all training instances, and then determines which side of that hyper-sphere does a test instance lie. If a test instance lies outside the hyper-sphere, it is declared to be anomalous. Hodge et al. (2004) use Robust Support Vector Machines (RSVM) which is robust to the presence of anomalies in the training data. RSVM have been applied to system call intrusion detection.

II.2.1.4. Bayesian Networks Based Techniques

Bayesian networks have been used for anomaly detection in the multi-class setting. A basic technique for a univariate categorical data set using a naïve Bayesian network estimates the posterior probability of observing a class label (from a set of normal class labels and the anomaly class label), given a test data instance.

The class label with largest posterior is chosen as the predicted class for the given test instance. The likelihood of observing the test instance given a class, and the prior on the class probabilities, are estimated from the training data set. The zero probabilities, especially for the anomaly class, are smoothed using Laplace Smoothing.

The basic technique can be generalized to multivariate categorical data set by aggregating the per-attribute posterior probabilities for each test instance and using the aggregated value to assign a class label to the test instance. Several variants of the basic technique has been proposed for network intrusion, for novelty detection in video surveillance, for anomaly detection in text data, and for disease outbreak detection. The basic technique described above assumes independence between the different attributes. Several variations of the basic technique have been proposed that capture the conditional dependencies between the different attributes using more complex Bayesian networks.

II.2.1.5.Computational Complexity

The computational complexity of classification based techniques depends on the classification algorithm being used. Generally, training decision trees tends to be faster while techniques that involve quadratic optimization, such as SVMs, are more expensive, though linear time SVMs have been proposed that have linear training time. The testing phase of classification techniques is usually very fast since the testing phase uses a learnt model for classification.

II.2.1.6.Advantages and Disadvantages

The advantages of classification based techniques are as follows:

(1) Classification based techniques, especially the multi-class techniques, can make use of powerful algorithms that can distinguish between instances belonging to different classes.

(2) The testing phase of classification based techniques is fast since each test instance needs to be compared against the pre-computed model.

The disadvantages of classification based techniques are as follows:

(1) Multi-class classification based techniques rely on availability of accurate labels for various normal classes, which is often not possible.

(2) Classification based techniques assign a label to each test instance, which can also become a disadvantage when a meaningful anomaly score is desired for the test instances. Some classification techniques that obtain a probabilistic prediction score from the output of a classifier can be used to address this issue.

II.2.2. Nearest Neighbor Based Anomaly Detection Techniques

The concept of nearest neighbor analysis has been used in several anomaly detection techniques. Such techniques are based on the assumption that, normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors.

Nearest neighbor based anomaly detection techniques require a distance or similarity measure defined between two data instances. Distance (or similarity) between two data instances can be computed in different ways. For continuous attributes, Euclidean distance is a popular choice but other measures can be used (Tan et al. 2005, Chapter 2). For categorical attributes, simple matching coefficient is often used but more complex distance measures can be used. For multivariate data instances, distance or similarity is usually computed for each attribute and then combined (Tan et al. 2005, Chapter 2). Most of the techniques that will be discussed in this section, as well as the clustering based techniques do not require the distance measure to be strictly metric. The measures are typically required to be positive-definite and symmetric, but they are not required to satisfy the triangle inequality.

Nearest neighbor based anomaly detection techniques can be broadly grouped into two categories:

(1) Techniques that use the distance of a data instance to its kth nearest neighbor as the anomaly score.

(2) Techniques that compute the relative density of each data instance to compute its anomaly score.

Additionally there are some techniques that use the distance between data instances in a different manner to detect anomalies and will be briefly discussed later.

II.2.2.1.Using Distance to *kth* Nearest Neighbor

A basic nearest neighbor anomaly detection technique is based on the assumption that; the anomaly score of a data instance is defined as its distance to its kth nearest neighbor in a given data set. This basic technique has been applied to detect land mines from satellite ground images and to detect shorted turns (anomalies) in the DC field windings of large synchronous turbine-generators. In the latter paper the authors use k = 1. Usually, a threshold is then be applied on the anomaly score to determine if a test instance is anomalous or not, on the other hand, select n instances with the largest anomaly scores as the anomalies. The basic technique has been extended by researchers in three different ways. The first set of variants modifies the above definition to obtain the anomaly score of a data instance. The second set of variants use different distance/similarity measures to handle different data types. The third set of variants focus on improving the efficiency of the basic technique (the complexity of the basic technique is O(N2), where N is the data size) in different ways.

Zhang and Wang (2006) compute the anomaly score of a data instance as the sum of its distances from its *k* nearest neighbors. A similar technique has been applied to detect credit card frauds by called *Peer Group Analysis*. A different way to compute the anomaly score of a data instance is to count the number of nearest neighbors (*n*) that are not more than *d* distance apart from the given data. This method can also be viewed as estimating the global density for each data instance since it involves counting the number of neighbors in a hyper-sphere of radius *d*. For example, in a 2-D data set, the density of a data instance = $n/\pi d^2$. The inverse of the density is the anomaly score for the data instance. Instead of computing the actual density, several techniques fix the radius d and use 1 /n as the anomaly score, while several techniques fix n and use 1/d as the anomaly score.

While most techniques discussed in this category so far have been proposed to handle continuous attributes, several variants have been proposed to handle other data types. A hyper-graph based technique is proposed by called HOT where the authors model the categorical values using a hyper-graph, and measure distance between two data instances by analyzing the connectivity of the graph. A distance measure for data containing a mix of categorical and continuous attributes has been proposed for anomaly detection. The authors define links between two instances by adding distance for categorical and continuous attributes separately. For categorical attributes, the number of attributes for which the two instances have same values defines the distance between them. For continuous attributes, a covariance matrix is maintained to capture the dependencies between the continuous values. (Yu et al. 2006) adapts the technique proposed to continuous sequences. (Zeevi et al.1997) extend the technique proposed to spatial data. Several variants of the basic technique have been proposed to improve the efficiency. Some techniques prune the search space by either ignoring instances that cannot be anomalous or by focusing on instances that are most likely to be anomalous. It is shown that for a sufficiently randomized data, a simple pruning step could result in the average complexity of the nearest neighbor search to be nearly linear. After calculating the nearest neighbors for a data instance, the algorithm sets the anomaly threshold for any data instance to the score of the weakest anomaly found so far. Using this pruning procedure, the technique discards instances that are close, and hence not interesting. It is proposed a *partition* based technique, which first clusters the instances and computes lower and upper bounds on distance of a instance from its *kth* nearest neighbor for instances in each partition. This information is then used to identify the partitions that cannot possibly contain the top *k* anomalies; such partitions are pruned. Anomalies are then computed from the remaining instances (belonging to unpruned partitions) in a final phase. (Wu and Jermaine 2006) use sampling to improve the efficiency of the nearest neighbor based technique. The authors compute the nearest neighbor of every instance within a smaller sample from the data set. Thus the complexity of the proposed technique is reduced to $O(MN)$ where M is the sample size chosen. To prune the search space for nearest neighbors, several techniques partition the attribute space into a hyper-grid consisting of hyper-cubes of fixed sizes. The intuition behind such techniques is that if a hypercube contains many instances, such instances are likely to be normal. Moreover, if for a given instance, the hypercube that contains the instance and its adjoining hyper-cubes contain very few instances, the given instance is likely to be anomalous. Techniques based on this intuition have been proposed. It is extended by linearizing the search space through the Hilbert space filling curve. The *d*-dimensional data set is fitted in a hypercube D = (0*;* 1)d. This hypercube is then mapped to the interval I = (0*;* 1) using the Hilbert Space Filling Curve and the *k*-nearest neighbors of a data instance are obtained by examining its successors and predecessors in *I*.

II.2.2.2.Using Relative Density

Density based anomaly detection techniques estimate the density of the neighborhood of each data instance. An instance that lies in a neighborhood with low density is declared to be anomalous while an instance that lies in a dense neighborhood is declared to be normal. For a given data instance, the distance to its kth nearest neighbor is equivalent to the radius of a hyper-sphere, centered at the given data instance, which contains *k* other instances. Thus the distance to the kth nearest neighbor for a given data instance can be viewed as an estimate of the inverse of the density of the instance in the data set and the basic nearest neighbor based technique described in the previous subsection can be considered as a density based anomaly detection technique.

Density based techniques perform poorly if the data has regions of varying densities. For example, consider a 2 dimensional data set shown in Figure 7. Due to the low density of the cluster *C*1 it is apparent that for every instance q inside the cluster *C*1, the distance between the instance q and its nearest neighbor is greater than the distance between the instance *p*2 and the nearest neighbor from the cluster *C*2, and the instance p2 will not be considered as anomaly. Hence, the basic technique will fail to distinguish between *p*2 and instances in *C*1. However, the instance p1 may be detected. To handle the issue of varying densities in the data set, a set of techniques have been proposed to compute density of instances relative to the density of their neighbors.

(Vapnik et al 1995) assign an anomaly score to a given data instance, known as Local Outlier Factor (LOF). For any given data instance, the LOF score is equal to ratio of average local density of the *k* nearest neighbors of the instance and the local density of the data instance itself. To find the local density for a data instance, the authors first find the radius of the smallest hyper-sphere centered at the data instance that contains its *k* nearest neighbors. The local density is then computed by dividing *k* by the volume of this hyper-sphere. For a normal instance lying in a dense region, its local density will be similar to that of its neighbors, while for an anomalous instance, its local density will be lower than that of its nearest neighbors. Hence the anomalous instance will get a higher LOF score.

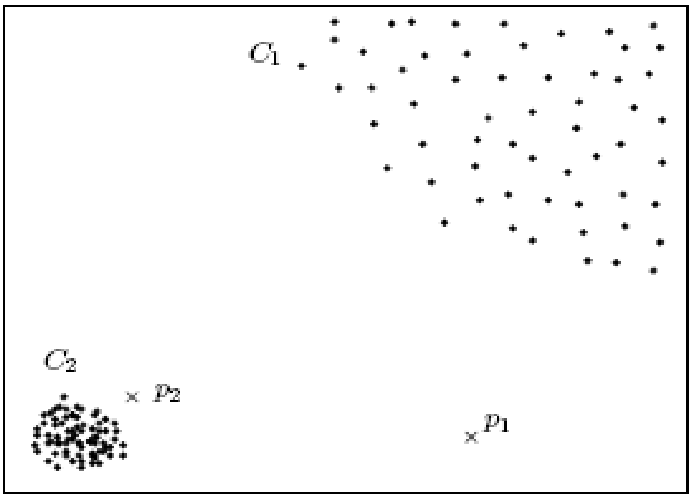


Figure 7: Advantage of local density based techniques over global density based techniques. (Kumar at al., 2005)

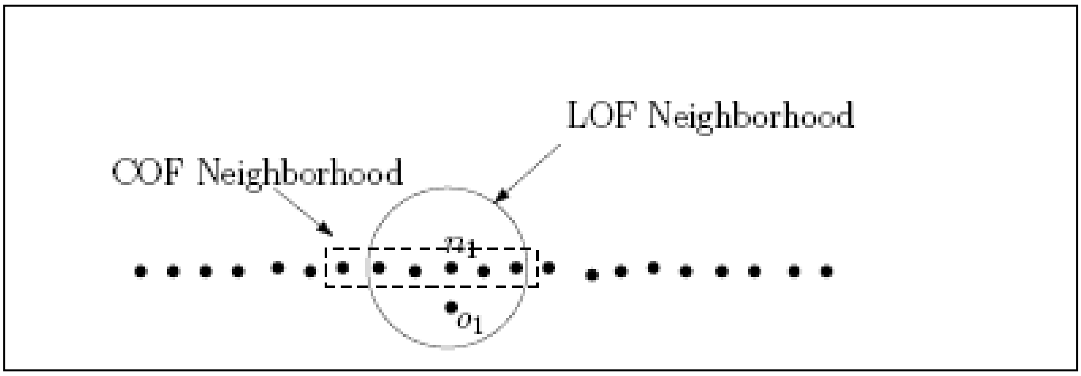


Figure 8 : Difference between the neighborhoods computed by LOF and COF (Kumar at al., 2005)

Several researchers have proposed variants of LOF technique. Some of these variants estimate the local density of an instance in a different way. Some variants have adapted the original technique to more complex data types. Since the original LOF technique is O(N2) (N is the data size), several techniques have been proposed that improve the efficiency of LOF.(Tang et al. 2002) discuss a variation of the LOF, which they call Connectivity- based Outlier Factor (COF). The difference between LOF and COF is the manner in which the *k* neighborhood for an instance is computed. In COF, the neighborhood for an instance is computed in an incremental mode. To start, the closest instance to the given instance is added to the neighborhood set. The next instance added to the neighborhood set is such that its distance to the existing neighborhood set is minimum among all remaining data instances. The distance between an instance and a set of instances is defined as the minimum distance between the given instance and any instance belonging to the given set. The neighborhood is grown in this manner until it reaches size *k*. Once the neighborhood is computed, the anomaly score (COF) is computed in the same manner as LOF. COF is able to capture regions such as straight lines, as shown in Figure 8 (Kumar 2005).

For a given data instance, ODIN is equal to the number of *k* nearest neighbors of the data instance which have the given data instance in their *k* nearest neighbor list. The inverse of ODIN is the anomaly score for the data instance propose a measure called Multi-granularity Deviation Factor (MDEF) which is a variation of LOF. MDEF for a given data instance is equal to the standard deviation of the local densities of the nearest neighbors of the given data instance (including the data instance itself). The inverse of the standard deviation is the anomaly score for the data instance. The anomaly detection technique presented in the paper is called LOCI, which not only finds anomalous instances but also anomalous micro-clusters.

Several variants of LOF have been proposed to handle different data types. A variant of LOF is applied for detecting spatial anomalies in climate data by (Chawla 2004). Yu et al. (2006) use a similarity measure instead of distance to handle categorical attributes. Similar technique has been proposed to detect sequential anomalies in protein sequences by Sun et al. (2006). This technique uses Probabilistic Suffix Trees (PST) to find the nearest neighbors for a given sequence.

Some variants of the LOF technique have been proposed to improve its efficiency. It is proposed a variant, in which only the top n anomalies are found instead of finding LOF score for every data instance. The technique includes finding micro-clusters in the data and then finding upper and lower bound on LOF for each of the micro-clusters proposed three variants of LOF which enhance its performance by making certain assumptions about the problem to prune all those clusters which definitely do not contain instances which will figure in the top n “anomaly list”. For the remaining clusters a detailed analysis is done to find the LOF score for each instance in these clusters.

II.2.2.3.Computational Complexity

A drawback of the basic nearest neighbor based technique and the LOF technique, is the O (N2) complexity required. Since these techniques involve finding nearest neighbors for each instance efficient data structures such as k-d trees (Bentley 1975) and R-trees (Roussopoulos et al. 1995) can be used. But such techniques do not scale well as the number of attributes increases. Several techniques have directly optimized the anomaly detection technique under the assumption that only top few anomalies are interesting. If an anomaly score is required for every test instance, such techniques are not applicable. Techniques that partition the attribute space into a hyper-grid, are linear in data size but are exponential in the number of attributes, and hence are not well suited for large number of attributes. Sampling techniques try to address the O (N2) complexity issue by determining the nearest neighbors within a small sample of the data set. But sampling might result in incorrect anomaly scores if the size of the sample is very small.

II.2.2.4. Advantages and Disadvantages of Nearest Neighbor Based Techniques

The advantages of nearest neighbor based techniques are as follows:

(1) A key advantage of nearest neighbor based techniques is that they are unsupervised in nature and do not make any assumptions regarding the generative distribution for the data. Instead, they are purely data driven.

(2) Semi-supervised techniques perform better than unsupervised techniques in terms of missed anomalies, since the likelihood of an anomaly to form a close neighborhood in the training data set is very low.

(3) Adapting nearest neighbor based techniques to a different data type is straight-forward, and primarily requires defining an appropriate distance measure for the given data.

The disadvantages of nearest neighbor based techniques are as follows:

(1) For unsupervised techniques, if the data has normal instances that do not have enough close neighbors or if the data has anomalies that have enough close neighbors, the technique fails to label them correctly, resulting in missed anomalies.

(2) For semi-supervised techniques, if the normal instances in test data do not have enough similar normal instances in the training data, the false positive rate for such techniques is high.

(3) The computational complexity of the testing phase is also a significant challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data, to compute the nearest neighbors.

(4) Performance of a nearest neighbor based technique greatly relies on a distance measure, defined between a pair of data instances that can effectively distinguish between normal and anomalous instances. Defining distance measures between instances can be challenging when the data is complex, e.g. graphs, sequences, etc.

II.2.3.Clustering Based Anomaly Detection Techniques

Clustering (Tan et al. 2005) is used to group similar data instances into clusters in an unlabeled dataset. Clustering is primarily an unsupervised technique though semi-supervised clustering (Basu et al. 2004) has also been explored lately. Even though clustering and anomaly detection appear to be fundamentally different from each other, several clustering based anomaly detection techniques have been developed. Clustering based anomaly detection techniques can be grouped into three categories.

First category of clustering based techniques relies on the assumption that:

Normal data instances belong to a cluster in the data, while anomalies either do not belong to any cluster.

Techniques based on the above assumption apply a known clustering based algorithm to the data set and declare any data instance that does not belong to any cluster as anomalous. Several clustering algorithms that do not force every data instance to belong to a cluster, such as DBSCAN (Ester et al. 1996), ROCK (Guha et al. 2000), and SNN clustering (Ertïoz et al. 2003) can be used. The Find Out algorithm (Yu et al. 2002) is an extension of the Wave Cluster algorithm (Sheikholeslami et al. 1998) in which the detected clusters are removed from the data and the residual instances are declared as anomalies. A disadvantage of such techniques is that they are not optimized to find anomalies, since the main aim of the underlying clustering algorithm is to find clusters.

Second category of clustering based techniques relies on the following assumption that ;

Normal data instances lie close to their closest cluster centroid, while anomalies are far away from their closest cluster centroid.

Techniques based on the above assumption consist of two steps. In the first step, the data is clustered using a clustering algorithm. In the second step, for each data instance, its distance to its closest cluster centroid is calculated as its anomaly score.

A number of anomaly detection techniques that follow this two step approach have been proposed using different clustering algorithms. Song et al. (2007) studied Self-Organizing Maps (SOM), K-means Clustering, and Expectation Maximization (EM) to cluster training data and then use the clusters to classify test data. In particular, SOM has been widely used to detect anomalies in a semi-supervised mode in several applications such as intrusion detection, fault, and fraud detection. Song et al. (2002) propose a technique is robust to anomalies in the training data. The authors first separate normal instances from anomalies in the training data, using frequent item-set mining, and then use the clustering based technique to detect anomalies. Several techniques have also been proposed to handle sequence data.

Techniques based on the second assumption can also operate in semi-supervised mode, in which the training data is clustered and instances belonging to the test data are compared against the clusters to obtain an anomaly score for the test data instance (Allan et al. 1998). If the training data has instances belonging to multiple classes, semi-supervised clustering can be applied to improve the clusters. Harris (2003) incorporate the knowledge of labels to improve on their unsupervised clustering based anomaly detection technique (Harris 2003) by calculating a measure called semantic anomaly factor which is high if the class label of an object in a cluster is different from the majority of the class labels in that cluster.

Note that if the anomalies in the data form clusters by themselves, the above discussed techniques will not be able to detect such anomalies. To address this issue a third category of clustering based techniques have been proposed that rely on the following assumption that; Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters.

Techniques based on the above assumption declare instances belonging to clusters whose size and/or density is below a threshold as anomalous. Several variations of the third category of techniques have been proposed (Harris 2003). The technique proposed by (Harris 2003), called Find CBLOF, assigns an anomaly score known as Cluster-Based Local Outlier Factor (CBLOF) for each data instance. The CBLOF score captures the size of the cluster to which the data instance belongs, as well as the distance of the data instance to its cluster centroid.

Several clustering based techniques have been proposed to improve the efficiency of the existing techniques discussed above. Fixed width clustering is a linear time

(O(Nd)) approximation algorithm used by various anomaly detection techniques (Harris 2003). An instance is assigned to a cluster whose center is within a pre-specified distance to the given instance. If no such cluster exists then a new cluster with the instance as the center is created. Then they determine which clusters are anomalies based on their density and distance from other clusters. The width can either be a user-specified parameter or can be derived from the data. Chen et al. (2005) propose an anomaly detection technique using k-d trees which provide a partitioning of the data in linear time. They apply their technique to detect anomalies in astronomical data sets where computational efficiency is an important requirement. Another technique which addresses this issue is proposed by Sun et al. (2004). The authors propose an indexing technique called CD-trees to efficiently partition data into clusters. The data instances which belong to sparse clusters are declared as anomalies.

II.2.3.1.Distinction between Clustering Based and Nearest Neighbor Based Techniques

Several clustering based techniques require distance computation between a pair of instances. Thus, in that respect, they are similar to nearest neighbor based techniques. The choice of the distance measure is critical to the performance of the technique; hence the discussion in the previous section regarding the distance measures hold for clustering based techniques also. The key difference between the two techniques, however, is that clustering based techniques evaluate each instance with respect to the cluster it belongs to, while nearest neighbor based techniques analyze each instance with respect to its local neighborhood.

II.2.3.2Computational Complexity

The computational complexity of training a clustering based anomaly detection technique depends on the clustering algorithm used to generate clusters from the data. Thus such techniques can have quadratic complexity if the clustering technique requires computation of pair wise distances for all data instances, or linear when using heuristic based techniques such as *k*-means or approximate clustering techniques. The test phase of clustering based techniques is fast, since it involves comparing a test instance with a small number of clusters.

II.2.3.3.Advantages and Disadvantages

The advantages of clustering based techniques are as follows:

(1) Clustering based techniques can operate in an unsupervised mode.

(2) Such techniques can often be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type.

(3) The testing phase for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.

The disadvantages of clustering based techniques are as follows:

(1) Performance of clustering based techniques is highly dependent on the effectiveness of clustering algorithm in capturing the cluster structure of normal instances.

(2) Many techniques detect anomalies as a by-product of clustering, and hence are not optimized for anomaly detection.

(3) Several clustering algorithms force every instance to be assigned to some cluster. This might result in anomalies getting assigned to a large cluster, thereby being considered as normal instances by techniques that operate under the assumption that anomalies do not belong to any cluster.

(4) Several clustering based techniques are effective only when the anomalies do not form significant clusters among themselves.

(5) The computational complexity for clustering the data is often a bottleneck, especially if O(N2d) clustering algorithms are used.

II.2.4.Statistical Anomaly Detection Techniques

The underlying principle of any statistical anomaly detection technique is: An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed. Statistical anomaly detection techniques are based on the following key assumption:

Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Statistical techniques fit a statistical model (usually for normal behavior) to the given data and then apply a statistical inference test to determine if an unexpected instance belongs to this model or not. Instances that have a low probability to be generated from the learnt model, based on the applied test statistic, are declared as anomalies. Both parametric as well as non-parametric techniques have been applied to fit a statistical model. While parametric techniques assume the knowledge of underlying distribution and estimate the parameters from the given data, non-parametric techniques do not generally assume knowledge of underlying distribution.

II.2.4.1Parametric Techniques

Parametric techniques assume that the normal data is generated by a parametric distribution with parameters Θ and probability density function f(x; Θ), where x is an observation. The anomaly score of a test instance (or observation) x is the inverse of the probability density function, f(x; Θ). The parameters are estimated from the given data. Alternatively, a statistical hypothesis test (also referred to as discordancy test in statistical outlier detection literature maybe used. The null hypothesis (H_0) for such tests is that the data instance x has been generated using the estimated distribution (with parameters Θ). If the statistical test rejects H_0, x is declared to be anomaly. A statistical hypothesis test is associated with a test statistic, which can be used to obtain a probabilistic anomaly score for the data instance x.

Based on the type of distribution assumed, parametric techniques can be further categorized as follows:

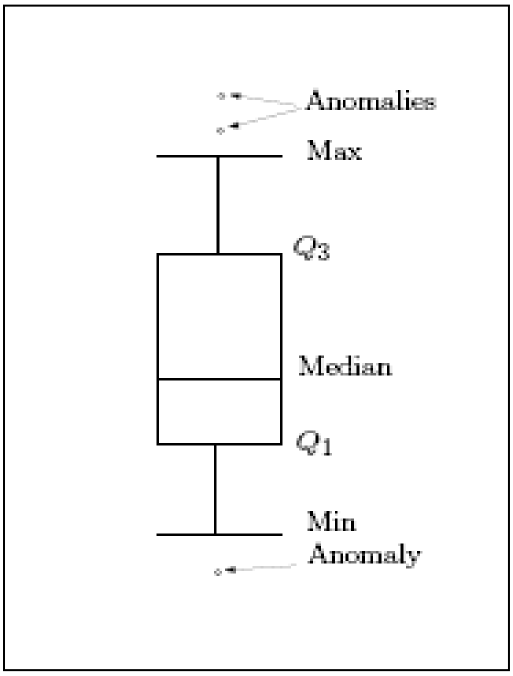


Figure 9: A box plot for a univariate data set.(Kumar et al., 2005)

Gaussian Model Based : Such techniques assume that the data is generated from a Gaussian distribution. The parameters are estimated using Maximum Likelihood Estimates (MLE). The distance of a data instance to the estimated mean is the anomaly score for that instance. A threshold is applied to the anomaly scores to determine the anomalies. Different techniques in this category calculate the distance to the mean and the threshold in different ways. A simple outlier detection technique, often used in process quality control domain, is to declare all data instances that are more than 3σ distance away from the distribution mean , where σ is the standard deviation for the distribution. The ± 3 σ region contains 99:7% of the data instances.

More sophisticated statistical tests have also been used to detect anomalies. The box plot rule (Figure 9) is the simplest statistical technique that has been applied to detect univariate and multivariate anomalies in medical domain data and turbine rotors data. A box-plot graphically depicts the data using summary attributes such as smallest non-anomaly observation (min), lower quartile (Q1), median, upper quartile (Q3), and largest non-anomaly observation (max). The quantity Q3-Q1 is called the Inter Quartile Range (IQR). The box plots also indicate the limits beyond which any observation will be treated as an anomaly. A data instance that lies more than 1.5 *IQR lower than Q1 or 1.5 *IQR higher than Q3 is declared as an anomaly. The region between Q1 -1.5IQR and Q3 +1.5IQR contains 99:3% of observations, and hence the choice of 1:5IQR boundary makes the box plot rule equivalent to the 3σ technique for Gaussian data.

Grubb's test (also known as the maximum normed residual test) is used to detect anomalies in a univariate data set (Grubbs 1969) under the assumption that the data is generated by a Gaussian distribution. For each test instance x, its z score is computed as follows:

$$\mathrm{z} = \frac{|\mathrm{x} - \bar{\mathrm{x}}|}{s}$$

(3)

where $\bar{\mathrm{x}}$ and s are the mean and standard deviation of the data sample, respectively. A test instance is declared to be anomalous if:

$$\mathrm{z} > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

(4)

Where N is the data size and $t_{\alpha/(2N),N-2}$ is a threshold used to declare an instance to be anomalous or normal. This threshold is the value taken by a t-distribution at a significance level of α/2N. The significance level rejects the confidence associated with the threshold and indirectly controls the number of instances declared as anomalous.

A variant of the Grubb's test for multivariate data was proposed by Song et al. (2007), which uses the Mahalanobis distance of a test instance x to the sample mean ¹x, to reduce multivariate observations to univariate scalars.

$$y^2 = (\mathrm{x} - \bar{\mathrm{x}})' S^{-1} (\mathrm{x} - \bar{\mathrm{x}}),$$

(5)

where S is the sample covariance matrix. The univariate Grubb's test is applied to y to determine if the instance x is anomalous or not. Several other variants of Grubb's test have been proposed to handle multivariate data sets (Aggarwal and Yu 2001; Eskin et al. 2000), graph structured data (Sun et al. 2005), and Online Analytical Processing (OLAP) data cubes (Sarawagi et al. 1998).The student's t-test has also been applied for anomaly detection to detect damages in structural beams. A normal sample, N1 is compared with a test sample, N2 using the t-test. If the test shows significant difference between them, it signifies the presence of an anomaly in N2.

The multivariate version of student's t-test called the Hotelling t^2-test is also used as an anomaly detection test statistic in (Wu et al. 2003) to detect anomalies in bioavailability/ bioequivalence studies.

Yu et al., (2006) use a χ^2 statistic to determine anomalies in operating system call data. The training phase assumes that the normal data has a multivariate normal distribution. The value of the χ^2 statistic is determined as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i}$$

(6)

Where X_i is the observed value of the ith variable, E_i is the expected value of the ith variable (obtained from the training data) and n is the number of variables. A large value of X^2 denotes that the observed sample contains anomalies.

Several other statistical anomaly detection techniques that assume that the data follows a Gaussian distribution have been proposed that use other statistical tests, such as: Rosner test, Slippage Detection test (Otey et al., 2006).

Regression Model Based: Anomaly detection using regression has been extensively investigated for time-series data. The basic regression model based anomaly detection technique consists of two steps. In the first step, a regression model is fitted on the data. In the second step, for each test instance, the residual for the test instance is used to determine the anomaly score. The residual is the part of the instance which is not explained by the regression model. The magnitude of the residual can be used as the anomaly score for the test instance, though statistical tests have been proposed to determine anomalies with certain confidence. Certain techniques detect the presence of anomalies in a data set by analyzing the Akaike Information Content (AIC) during model fitting.

Presence of anomalies in the training data can influence the regression parameters and hence the regression model might not produce accurate results. A popular technique to handle such anomalies, while fitting regression models is called robust regression (Rousseeuw and Leroy 1987) (estimation of regression parameters while accommodating anomalies). The authors argue that the robust regression techniques not only hide the anomalies, but can also detect the anomalies, because the anomalies tend to have larger residuals from the robust fit. A similar robust anomaly detection approach has been applied in Autoregressive Integrated Moving Average (ARIMA) models (Chen et al. 2002). Variants of the basic regression models based technique have been proposed to handle multivariate time-series data. Salvador et al. (2003) discuss the additional complexity in multivariate time-series over the univariate time-series and come up with statistics that can be applied to detect anomalies in multivariate ARIMA models. This is a generalization of statistics proposed earlier.

Another variant that detect anomalies in multivariate time-series data generated by an Autoregressive Moving Average (ARMA) model, was proposed by Wei et al. (2003). In this technique the authors transform the multivariate time-series to univariate time-series by linearly combining the components of the multivariate time-series. The interesting linear combinations (projections in 1-d space) are obtained using a projection pursuit technique (Torr 1993) that maximizes the Kurtosis coefficient (a measure for the degree of peakedness /flatness in the variable distribution) of the time-series data. The anomaly detection in each projection is done by using univariate test statistics as proposed by (Tandon et al., 2007).

Mixture of Parametric Distributions Based: Such techniques use a mixture of parametric statistical distributions to model the data. Techniques in this category can be grouped into two sub-categories. The first sub-category of techniques model the normal instances and anomalies as separate parametric distributions, while the second sub-category of techniques model only the normal instances as a mixture of parametric distributions.

For the first sub-category of techniques, the testing phase involves determining which distribution| normal or anomalous |the test instance belongs to. Abraham and Box (1979) assume that the normal data is generated from a Gaussian distribution ($N(0,\sigma^2)$) and the anomalies are also generated from a Gaussian distribution with same mean but with larger variance, $N(0,k^2\sigma^2)$. A test instance is tested using the Grubb's test on both distributions, and accordingly labeled as normal or anomalous. Similar techniques have been proposed in (Eskin 2000; Agarwal 2005). Eskin (2000) use Expectation Maximization (EM) algorithm to develop a mixture of models for the two classes, assuming that each data point is an anomaly with apriori probability

λ, and distribution of the entire data, and M and A represent the distributions of the normal and anomalous data respectively, then $D = \lambda A + (1 - \lambda) M$. M is learnt using any distribution estimation technique, while A is assumed to be uniform.

Initially all points are considered to be in M. The anomaly score is assigned to a point based on how much the distributions change if that point is removed from M and added to A.

The second sub-category of techniques models the normal instances as a mixture of parametric distributions. A test instance which does not belong to any of the learnt models is declared to be anomaly. Gaussian mixture models have been mostly used for such techniques and have been used to detect strains in airframe data to detect anomalies in mammographic image analysis and for network intrusion. Similar techniques have been applied to detecting anomalies in biomedical signal data, where extreme value statistics are used to determine if a test point is an anomaly with respect to the learnt mixture of models or not.

II.2.4.2 Non-parametric Techniques

The anomaly detection techniques in this category use non-parametric statistical models, such that the model structure is not defined a priori, but is instead determined from given data. Such techniques typically make fewer assumptions regarding the data, such as smoothness of density, when compared to parametric techniques.

Histogram Based: The simplest non-parametric statistical technique is to use histograms to maintain a profile of the normal data. Such techniques are also referred to as frequency based or counting based. Histogram based techniques are particularly popular in intrusion detection community (Eskin 2000) and fraud, since the behavior of the data is governed by certain profiles (user or software or system) that can be efficiently captured using the histogram model.

A basic histogram based anomaly detection technique for univariate data consists of two steps. The first step involves building a histogram based on the different values taken by that feature in the training data. In the second step, the technique checks if a test instance falls in any one of the bins of the histogram. If it does, the test instance is normal, otherwise it is anomalous. A variant of the basic histogram based technique is to assign an anomaly score to each test instance based on the height (frequency) of the bin in which it falls.

The size of the bin used when building the histogram is key for anomaly detection. If the bins are small, many normal test instances will fall in empty or rare bins, resulting in a high false alarm rate. If the bins are large, many anomalous test instances will fall in frequent bins, resulting in a high false negative rate. Thus a key challenge for histogram based techniques is to determine an optimal size of the bins to construct the histogram which maintains low false alarm rate and low false negative rate.

Histogram based techniques require normal data to build the histograms. Some techniques even construct histograms for the anomalies, if labeled anomalous instances are available.

For multivariate data, a basic technique is to construct attribute-wise histograms. During testing, for each test instance, the anomaly score for each attribute value of the test instance is calculated as the height of the bin that contains the attribute value. The per-attribute anomaly scores are aggregated to obtain an overall anomaly score for the test instance.

The basic histogram based technique for multivariate data has been applied to system call intrusion detection, network intrusion detection, fraud detection, damage detection in, detecting web-based attacks, and anomalous topic detection in text data. A variant of the simple technique is used in Packet Header Anomaly Detection (PHAD) and Application Layer Anomaly Detection (ALAD), applied to network intrusion detection.

The SRI International's real-time Network Intrusion Detection System (NIDES has a sub-system that maintains long-term statistical profiles to capture the normal behavior of a computer system. The authors propose a Q statistic to compare a long-term profile with a short term profile (observation). The statistic is used to determine another measure called S statistic which rejects the extent to which the behavior in a particular feature is anomaly with respect to the historical profile. The feature-wise S statistics are combined to get a single value called IS statistic which determines if a test instance is anomalous or not. A variant has been proposed for anomaly detection in link-state routing protocols.

Kernel Function Based: A non-parametric technique for probability density estimation is parzen windows estimation. This involves using kernel functions to approximate the actual density. Anomaly detection techniques based on kernel functions are similar to parametric methods described earlier. The only difference is the density estimation technique used. It is proposed a semi-supervised statistical technique to detect anomalies which uses kernel functions to estimate the probability distribution function (pdf) for the normal in- stances. A new instance which lies in the low probability area of this pdf is declared to be anomalous.

Similar application of parzen windows is proposed for network intrusion detection, for novelty detection in oil flow data, and for mammographic image analysis (Sarawagi 1998).

II.2.4.3 Computational Complexity

The computational complexity of statistical anomaly detection techniques depends on the nature of statistical model that is required to be fitted on the data. Fitting single parametric distributions from the exponential family, e.g., Gaussian, Poisson, Multinomial, etc., is typically linear in data size as well as number of attributes. Fitting complex distributions (such as mixture models, HMM, etc.) using iterative estimation techniques such as Expectation Maximization (EM), are also typically linear per iteration, though they might be slow in converging depending on the problem and/or convergence criterion. Kernel based techniques can potentially have quadratic time complexity in terms of the data size.

II.2.4.4.Advantages and Disadvantages

The advantages of statistical techniques are:

(1) If the assumptions regarding the underlying data distribution hold true, statistical techniques provide a statistically justifiable solution for anomaly detection.

(2) The anomaly score provided by a statistical technique is associated with a confidence interval, which can be used as additional information while making a decision regarding any test instance.

(3) If the distribution estimation step is robust to anomalies in data, statistical techniques can operate in an unsupervised setting without any need for labeled training data.

The disadvantages of statistical techniques are:

(1) The key disadvantage of statistical techniques is that they rely on the assumption that the data is generated from a particular distribution. This assumption often does not hold true, especially for high dimensional real data sets.

(2) Even when the statistical assumption can be reasonably justified, there are several hypothesis test statistics that can be applied to detect anomalies; choosing the best statistic is often not a straightforward task. In particular, constructing hypothesis tests for complex distributions that are required to fit high dimensional data sets is nontrivial.

(3) Histogram based techniques are relatively simple to implement, but a key short-coming of such techniques for multivariate data is that they are not able to capture the interactions between different attributes. An anomaly might have attribute values that are individually very frequent, but their combination is very rare, but an attribute-wise histogram based technique would not be able to detect such anomalies.

II.2.5. Information Theoretic Anomaly Detection Techniques

Information theoretic techniques analyze the information content of a data set using different information theoretic measures such as Kolmogorov Complexity, entropy, relative entropy, etc. Such techniques are based on the following key assumption: Anomalies in data induce irregularities in the information content of the data set.

Let C(D) denote the complexity of a given data set, D. A basic information theoretic technique can be described as follows. Given a data set D, find the minimal subset of instances, I, such that C(D)¡C(D¡I) is maximum. All instances in the subset thus obtained, are deemed as anomalous. The problem addressed by this basic technique is to find a pareto-optimal solution, which does not have single optima, since there are two different objectives that need to be optimized.

In the above described technique, the complexity of a data set (C) can be measured in different ways. Kolmogorov complexity has been used by several techniques. Arning et al. (1996) use the size of the regular expression to measure the Kolmogorov Complexity of data (represented as a string) for anomaly detection. Barbara et al. (2001a) use the size of the compressed data file (using any standard compression algorithm), as a measure of the data set's Kolmogorov Complexity. Other information theoretic measures such as entropy, relative uncertainty, etc., have also been used to measure the complexity of a categorical data.

The basic technique described above, involves dual optimization to minimize the subset size while maximizing the reduction in the complexity of the data set. Thus an exhaustive approach in which every possible subset of the data set is considered would run in exponential time. Several techniques have been proposed that perform approximate search for the most anomalous subset. Chiu et al. (2003) use an approximate algorithm called Local Search Algorithm (LSA) (Song et al. 2005) to approximately determine such a subset in a linear fashion, using entropy as the complexity measure. A similar technique that uses an information bottleneck measure was proposed by (Abe 2006).

Information theoretic techniques have also been used in data sets in which data instances are naturally ordered, e.g., sequential data, spatial data. In such cases, the data is broken into substructures (segments for sequences, sub-graphs for graphs, etc.), and the anomaly detection technique finds the substructure, I, such that C(D) - C(D - I) is maximum. This technique has been applied to sequences, graph data (Noble and Cook 2003), and spatial data. A key challenge of such techniques is to find the optimal size of the substructure which would result in detecting anomalies.

II.2.5.1.Computational Complexity

The basic information theoretic anomaly detection technique has exponential time complexity, though approximate techniques have been proposed that have linear time complexity.

II.2.5.2.Advantages and Disadvantages

The advantages of information theoretic techniques are as follows:

(1) They can operate in an unsupervised setting.

(2) They do not make any assumptions about the underlying statistical distribution for the data.

The disadvantages of information theoretic techniques are as follows:

(1) The performance of such techniques is highly dependent on the choice of the information theoretic measure. Often, such measures can detect the presence of anomalies only when there is significantly large number of anomalies present in the data.

(2) Information theoretic techniques applied to sequences and spatial data sets rely on the size of the substructure, which is often nontrivial to obtain.

(3) It is difficult to associate an anomaly score with a test instance using an information theoretic technique.

II.2.6. Spectral Anomaly Detection Techniques

Spectral techniques try to find an approximation of the data using a combination of attributes that capture the bulk of variability in the data. Such techniques are based on the following key assumption:

Assumption: Data can be embedded into a lower dimensional subspace in which normal instances and anomalies appear significantly different.

Thus the general approach adopted by spectral anomaly detection techniques is to determine such subspaces (embeddings, projections, etc.) in which the anomalous instances can be easily identified (Aeyels 1991). Such techniques can work in an unsupervised as well as semi-supervised setting.

Several techniques use Principal Component Analysis (PCA) for projecting data into a lower dimensional space. One such technique analyzes the projection of each data instance along the principal components with low variance. A normal instance that satisfies the correlation structure of the data will have a low value for such projections while anomalous instances that deviate from the correlation structure will have a large value. Ando (2007) adopt this approach to detect anomalies in astronomy catalogs.

It is proposed a spectral technique to detect anomalies in a time series of graphs. Each graph is represented as an adjacency matrix for a given time. At every time instance, the principle component of the matrix is chosen as the activity vector for the given graph. The time-series of the activity vectors is considered as a matrix and the principal left singular vector is obtained to capture the normal dependencies over time in the data. For a new (test) graph, then angle between its activity vector and the principal left singular vector obtained from the previous graphs is computed and used to determine the anomaly score of the test graph. In a similar approach, Sun et al. (2007) propose an anomaly detection technique on a sequence of graphs by performing Compact Matrix Decomposition (CMD) on the adjacency matrix for each graph and thus obtaining an approximation of the original matrix. For each graph in the sequence, the authors perform CMD and compute the approximation error between the original adjacency matrix and the approximate matrix. The authors construct a time series of the approximation errors and detect anomalies in the time series of errors; the graph corresponding to anomalous approximation error is declared to be anomalous.

Sun et al. (2007) present an anomaly detection technique where the authors perform robust PCA (Hodge 2004) to estimate the principal components from the covariance matrix of the normal training data. The testing phase involves comparing each point with the components and assigning an anomaly score based on the point's distance from the principal components. Thus if the projection of x on the principal components are y_1, y_2,…, y_p and the corresponding eigenvalues are λ_1,λ_2,…, λ_p, then

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \ldots + \frac{y_q^2}{\lambda_q}, q \leq p$$

(8)

has a chi-square distribution (Hawkins 2009). Using this result, the authors propose that, for a given significance level α, observation x is an anomaly if

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} > \chi_q^2(\alpha)$$

(9)

It can be shown that the quantity calculated in Equation 5 is equal to the Mahalanobis distance of the instance x from the sample mean (See Equation 3) when q = p (Sun et al. 2007). Thus the robust PCA based technique is same as a statistical technique discussed in Section 7.1.1 in a smaller subspace.

The robust PCA based technique has been applied to the network intrusion detection domain and for detecting anomalies in space craft components.

II.2.6.1.Computational Complexity

Standard PCA based techniques are typically linear in data size but often quadratic in the number of dimensions. Non linear techniques can improve the time complexity to be linear in the number of dimensions but polynomial in the number of principal components (Hautamaki et al. 2004). Techniques that perform SVD on the data typically quadratic in data size.

II.2.6.2.Advantages and Disadvantages of Spectral Techniques

The advantages of spectral anomaly detection techniques are as follows:

(1) Spectral techniques automatically perform dimensionality reduction and hence are suitable for handling high dimensional data sets. Moreover, they can also be used as a pre-processing step followed by application of any existing anomaly detection technique in the transformed space.

(2) Spectral techniques can be used in an unsupervised setting.

The disadvantages of spectral anomaly detection techniques are as follows:

(1) Spectral techniques are useful only if the normal and anomalous instances are separable in the lower dimensional embedding of the data.

(2) Spectral techniques typically have high computational complexity.

II.2.7.Handling Contextual Anomalies

The anomaly detection techniques discussed in the previous sections primarily focus on detecting point anomalies. In this section, we will discuss anomaly detection techniques that handle contextual anomalies.

Contextual anomalies require that the data has a set of contextual attributes (to define a context), and a set of behavioral attributes (to detect anomalies within a context). Song et al. (2002) use the terms environmental and indicator attributes which are analogous to our terminology. Some of the ways in which contextual attributes can be defined are:

(1) Spatial: The data has spatial attributes, which define the location of a data instance and hence a spatial neighborhood. A number of context based anomaly detection techniques (Sun and Chawla 2004) have been proposed for data with spatial data.

(2) Graphs: The edges that connect nodes (data instances) define neighborhood for each node. Contextual anomaly detection techniques have been applied to graph based data by Sun et al. (2005).

(3) Sequential: The data is sequential, i.e., the contextual attributes of a data instance is its position in the sequence.

Time-series data has been extensively explored in the contextual anomaly detection.

Another form of sequential data for which anomaly detection techniques have been developed is event data, in which each event has a timestamp. The difference between time-series data and event sequences is that for the latter, the inter-arrival time between consecutive events is uneven.

(4) Profile: Often times the data might not have an explicit spatial or sequential structure, but can still be segmented or clustered into components using a set of contextual attributes. These attributes are typically used to profile and group users in activity monitoring systems, such as cell-phone fraud detection, CRM databases and credit-card fraud detection. The users are then analyzed within their group for anomalies. In comparison to the rich literature on point anomaly detection techniques, the research on contextual anomaly detection has been limited. Broadly, such techniques can be classified in two categories. The first category of techniques

reduce a contextual anomaly detection problem to a point anomaly detection problem while the second category of techniques model the structure in the data and use the model to detect anomalies.

II.2.7.1.Reduction to Point Anomaly Detection Problem

Since contextual anomalies are individual data instances (like point anomalies), but are anomalous only with respect to a context, one approach is to apply a known point anomaly detection technique within a context.

A generic reduction based technique consists of two steps. First, identify a context for each test instance using the contextual attributes. Second, compute anomaly score for the test instance within the context using a known point anomaly detection technique.

An example of the generic reduction based technique has been proposed for the scenario where identifying the context is not straightforward (Song et al. 2002).

The authors assume that the attributes are already partitioned into contextual and behavioral attributes. Thus each data instance d can be represented as (x; y).

The contextual data is partitioned using a mixture of Gaussian model, say U. The behavioral data is also partitioned using another mixture of Gaussian model, say V . A mapping function, $p(V_j|U_i)$ is also learnt. This mapping indicates the probability of the indicator part of a data point y to be generated from a mixture component V_j , when the environmental part x is generated by U_i. Thus for a given test instance d = (x, y), the anomaly score is given by:

$$Anomaly\ Score = \sum_{i=1}^{n_U} p(x \in U_i) \sum_{j=1}^{n_V} p(y \in V_j) p(V_j|U_i)$$

(10)

Where n_U is the number of mixture components in U and NV is the number of mixture components in V . $p(x \in U_i)$ indicates the probability of a sample point x to be generated from the mixture component U_i while $p(y \in U_j)$ indicates the probability of a sample point y to be generated from the mixture component V_j .

Another example of the generic technique is applied to cell-phone fraud detection. The data in this case consists of cell-phone usage records. One of the attributes in the data is the cell-phone user which is used as the contextual attribute. The activity of each user is then monitored to detect anomalies using other attributes. A similar technique is adopted for computer security, where the contextual attributes are: user id, time of the day. The remaining attributes are compared with existing rules representing normal behavior to detect anomalies. Peer group analysis is another similar technique where users are grouped together as peers and analyzed within a group for fraud.

For spatial data, neighborhoods are intuitive and straightforward to detect (Wei et al., 2003) by using the location coordinates. Graph-based anomaly detection use Grubb's score or similar statistical point anomaly detection techniques to detect anomalies within a spatial neighborhood. Sun and Chawla (2004) use a distance based measure called SLOM (Spatial Local Outlier Measure (Sun and Chawla 2006)) to detect spatial anomalies within a neighborhood. Another example of the generic technique applied to time-series data is proposed by Yu et al., (2007). For a given instance in a time-series the authors compare the observed value to the median of the neighborhood values. A transformation technique for time-series data has been proposed by using phase spaces. This technique converts a time-series into a set of vectors by unfolding the time-series into a phase space using a time-delay embedding process.

The temporal relations at any time instance are embedded in the phase vector for that instance. The authors use this technique to transform a time-series into feature space and then use one-class SVMs to detect anomalies. Each anomaly can be translated to a value at certain time instance in the original time-series.

II.2.7.2. Utilizing the Structure in Data

In several scenarios, breaking up data into contexts is not straightforward. This is typically true for time-series data and event sequence data. In such cases, time series modeling and sequence modeling techniques are extended to detect contextual anomalies in the data.

A generic technique in this category can be described as follows. A model is learnt from the training data which can predict the expected behavior with respect to a given context. If the expected behavior is significantly different from the observed behavior, an anomaly is declared. A simple example of this generic technique is regression in which the contextual attributes can be used to predict the behavioral attribute by fitting a regression line on the data. For time series data, several regression based techniques for time-series modeling such as robust regression, auto-regressive models , ARMA models (Zeevi et al. 1997), and ARIMA models (Eskin et al. 2000), have been developed for contextual anomaly detection. Regression based techniques have been extended to detect contextual anomalies in a set of co-evolving sequences by modeling the regression as well as correlation between the sequences.

One of the earliest works in time-series anomaly detection was proposed by Tang (2002), where a time-series was modeled as a stationary auto-regressive process. Any observation is tested to be anomaly by comparing it with the covariance matrix of the auto-regressive process. If the observation falls outside the modeled error for the process, it is declared to be an anomaly. An extension to this technique is made by using Support Vector Regression to estimate the regression parameters and then using the learnt model to detect novelties in the data (Ma and Perkins 2003a). A technique to detect a single anomaly (discord) in a sequence of alphabets was proposed by Manson et al. (2001). The technique adopts a divide and conquers approach. The sequence is divided into two parts and the Kolmogorov Complexity is calculated for each. The one with higher complexity contains the anomaly. The sequence is recursively divided until they are left with a single event which is declared to be the anomaly in the sequence.

Markou et al. (2003a) propose a technique to detect rare events in sequential data, where they use events occurring before a particular time to predict the event occurring at that time instance. If the prediction does not match the actual event, it is declared to be rare. This idea is extended in other areas, where the authors have used Frequent Itemset Mining (Markou et al. 2003b), Finite State Automaton (FSA) (Salvador and Chan 2003) and Markov Models to determine conditional probabilities for events based on the history of events.

Nairac et al., (1997) use FSA to predict the next event of a sequence based on the previous n events. They apply this technique to the domain of system call intrusion detection. HMM has been employed for cell phone fraud detection. The authors use a hierarchical regime switching call model to model the cell phone activity of a user.

The model predicts the probability of a fraud taking place for a call using the learnt model. The parameter estimation is done using the EM algorithm.

A model to detect intrusions in telephone networks was proposed and for modeling web clicks data. Both papers follow a technique in which they assume that the normal behavior in a time-series is generated by a non-stationary Poisson process while the anomalies are generated by a homogenous Poisson process. The transition between normal and anomalous behavior is modeled using a Markov process. The proposed techniques in each of these papers use Markov Chain Monte Carlo (MCMC) estimation technique to estimate the parameters for these processes. For testing, a time series is modeled using this process and the time instances for which the anomalous behavior was active are considered as anomalies.

Bipartite graph structure in P2P networks has been used to first identify a neighborhood for any node in the graph (Sun et al. 2004), and then detecting the relevance of that node within the neighborhood. A node with a low relevance score is treated as an anomaly. The authors also propose an approximate technique where the graph is first partitioned into non-overlapping sub-graphs using graph partitioning algorithm such as METIS (Kumar et al., 2005). The neighborhood of a node is then computed within its partition.

II.2.7.3.Computational Complexity

The computational complexity of the training phase in reduction based contextual anomaly detection techniques depends on the reduction technique as well as the point anomaly detection technique used within each context. While segmenting/partitioning techniques have a fast reduction step, techniques that use clustering, or mixture of models estimation, are relatively slower. Since the reduction simplifies the anomaly detection problem, fast point anomaly detection techniques can be used to speed up the second step. The testing phase is relatively expensive since for each test instance, its context is determined, and then an anomaly label or score is assigned using a point anomaly detection technique.

The computational complexity of training phase in contextual anomaly detection techniques that utilize the structure in the data to build models is typically higher that of techniques that reduce the problem to point anomaly detection. An advantage for such techniques is the testing phase is relatively fast, since each instance is just compared to the single model and assigned an anomaly score or an anomaly label.

II.2.7.4.Advantages and Disadvantages

The key advantage of contextual anomaly detection techniques is that they allow a natural definition of an anomaly in many real life applications where data instances tend to be similar within a context. Such techniques are able to detect anomalies that might not be detected by point anomaly detection techniques that take a global view of the data.

The disadvantage of contextual anomaly detection techniques is that they are applicable only when a context can be defined.

II.2.8.Handling Collective Anomalies

This section discusses the anomaly detection techniques which focus on detecting collective anomalies. As mentioned earlier, collective anomalies are a subset of instances that occur together as a collection and whose occurrence is not normal with respect to a normal behavior. The individual instances belonging to this collection are not necessarily anomalies by themselves, but it is their co-occurrence in a particular form that makes them anomalies. Collective anomaly detection problem is more challenging than point and contextual anomaly detection because it involves exploring structure in the data for anomalous regions.

A primary data requirement for collective anomaly detection is the presence of relationship between data instances. Three types of relations that have been exploited most frequently are sequential, spatial, and graphs:

1. -Sequential Anomaly Detection Techniques: These techniques work with sequential data and find subsequences as anomalies (also referred to as sequential anomalies). Typical data sets include event sequence data, such as system call data or numerical time-series data (Chan and Mahoney 2005).
2. -Spatial Anomaly Detection Techniques: These techniques work with spatial data and find connected sub-regions within the data as anomalies (also

referred to as spatial anomalies). Anomaly detection techniques have been applied to multi- spectral imagery data.

3. -Graph Anomaly Detection Techniques: These techniques work with graph data and find connected sub-graphs within the data as anomalies (also referred to as graph anomalies). Anomaly detection techniques have been applied to graph data (Noble and Cook 2003).

Substantial research has been done in the field of sequential anomaly detection; this can be attributed to the existence of sequential data in several important application domains. Spatial anomaly detection has been explored primarily in the domain of image processing. The following subsections discuss each of these categories in detail.

II.2.8.1. Handling Sequential Anomalies

As mentioned earlier, collective anomaly detection in sequence data involves detecting sequences that are anomalous with respect to a definition of normal behavior. Sequence data is very common in a wide range of domains where a natural ordering is imposed on data instances by either time or position. In anomaly detection literature, two types of sequences are dealt with. First type of sequences is symbolic, such as a sequence of operating system calls, or a sequence of biological entities. Second type of sequences is continuous or time series. Sequences can also be univariate, in which each event in the sequence is a univariate observation, or multivariate, in which each event in the sequence is a multivariate observation.

The anomaly detection problem for sequences can be defined in different ways and are discussed below.

Detecting anomalous sequence in a set of sequences; the objective of the techniques in this category is to detect anomalous sequences from a given set of unsupervised mode.

Key challenges faced by techniques in this category are:

1. The sequences might not be of equal length.
2. The test sequences may not be aligned with each other or with normal sequences.

For example, the first event in one sequence might correspond to the third event in another sequence. Comparing such sequences is a fundamental problem with biological sequences where different sequence alignment and sequence matching techniques are explored.

Techniques addressing this problem follow one of the following two approaches:

A general approach to solve the above problem would be to transform the sequences to a finite feature space and then use a point anomaly detection technique in the new space to detect anomalies.

Certain techniques assume that all sequences are of equal lengths. Thus they treat each sequence as a vector of attributes and employ a point anomaly detection technique to detect anomalies. For example, if a data set contains length 10 sequences, they can be treated as data records with 10 features. A similarity or distance measure can be defined between a pair of sequences and any point anomaly detection technique can be applied to such data sets. This approach has been adopted for time-series data sets.

In the former paper, the authors apply ART (Adaptive Resonance Theory) neural networks based anomaly detection technique to detect anomalies in a time-series data set, while the latter paper uses a clustering based anomaly detection technique to identify cyclone regimes (anomalies) in weather data.

As mentioned earlier, the given sequences may not be of equal length. Certain techniques address this issue by transforming each sequence into a record of equal number of attributes. A transformation technique has been proposed for multiple time-series data (Chan and Mahoney 2005), known as Box Modeling. In a box model, for each time-series, each instance of this time-series is assigned to a box depending on its value. These boxes are then treated as features (the number of boxes is the number of features in the transformed feature space). The authors then apply point anomaly detection techniques _ a Euclidean distance based technique and a classification based technique using RIPPER to detect anomalous time series in the data.

Several techniques address the issue of unequal length of sequences by using a similarity or distance metric that can compute similarity or distance between two unequal length sequences. For example, (Qin et al. 2004) employ length of longest common subsequence as the similarity measure for symbolic sequences. The transformations discussed in the previous section are appropriate when all the sequences are properly aligned. Often times the alignment assumption becomes too prohibitive. Research dealing with system call data, biological data, etc., explore other alternatives to detect collective anomalies.

Such techniques operate in a semi-supervised mode, and hence require a training set of normal sequences.

Sequential association modeling has been used to generate sequential rules from sequences. The authors use an approach called time-based inductive learning to generate rules from the set of normal sequences. The test sequence is compared to these rules and is declared an anomaly if it contains patterns for which no rules have been generated.

Markovian modeling of sequences has been the most popular approach in this category. The modeling techniques used in this category range from Finite State Automatons (FSA) to Markov models. FSA have been used to detect anomalies in network protocol data. Anomalies are detected when a given sequence of events does not result in reaching one of the final states.

The authors also apply their technique to operating system call intrusion detection. It is proposed a simple 1-order markov chain modeling approach to detect if a given sequence S is an anomaly. The author determines the likelihood of S, P(S) using the following equation

$$P(S) = q_{S_1} \prod_{t=2}^{|S|} p_{S_{t-1}S_t}$$

(11)

where q_{S1} is the probability of observing the symbol S_1 in the training set and p_{St-1St} is the probability of observing the symbol S_t after symbol S_{t-1} in the training set. The inverse of P(S) is the anomaly score for S. The drawback of this technique is that single order markov chain cannot model higher order dependencies in the sequences.

Palshikar et al. (2005) propose a Hidden Markov Model (HMM) based technique to detect anomalous program traces in operating system call data. The authors train an HMM using the training sequences. The authors propose two testing techniques.

In the first technique they compute the likelihood of a test sequence S to be generated by the learnt HMM using the Viterbi algorithm. The second technique is to use the underlying Finite State Automaton (FSA) of the HMM.

The state transitions and the outputs made by the HMM to produce the test sequence are recorded. The authors count the number of times the HMM had to make an unlikely state transition or output an unlikely symbol (using a user-defined threshold) as mismatches. The total number of mismatches denotes the anomaly score for that sequence.

A Probabilistic Suffix Trees (PST) is another modeling tool which has been applied to detect collective anomalies in sequential databases. A PST is a compact representation of a variable order markov chain. Yang and Wang (2003) use PST to cluster sequences and detect anomalous sequences as a by-product. Similarly Roussopoulos et al. (1995) use HMMs to cluster the set of sequences and detect any sequences which do not belong to any cluster as anomalies. Another modeling tool used for sequential anomaly detection is Sparse Markov Trees (SMT), which is similar to a PST with the difference that it allows wild card symbols within a path. This technique has been used by Eskin et al. (2000), who train a mixture of SMT using the training set. Each SMT has a different location of wildcards. Testing phase involves predicting the probability $P(S_n|S_{n-1} \ldots S_1)$ using the best SMT from the mixture. If this probability is below a certain threshold, the test sequence is declared as an anomaly.

Detecting anomalous subsequences in a long sequence; The objective of techniques belonging to this category is to detect a subsequence within a given sequence which is anomalous with respect to the rest of the sequence. This problem formulation occurs in event and time-series data sets where the data is in the form of a long sequence and contains regions that are anomalous. The techniques that address this problem, typically work in an unsupervised mode, due to the lack of any training data. The underlying assumption is that the normal behavior of the time-series follows a defined pattern. A subsequence within the long sequence which does not conform to this pattern is an anomaly.

Key challenges faced by techniques in this category are:

1. The length of the anomalous subsequence to be detected is not generally defined. A long sequence could contain anomalous regions of variable lengths. Thus fixed length segmenting of the sequence is often not useful.
2. Since the input sequence contains anomalous regions, it becomes challenging to create a robust model of normalcy.

It is proposed a surprise detection technique in market basket transactions. The data is a sequence of item sets, ordered by time. The authors propose to segment the sequence of item sets such that the sum of number of bits require to encode each segment (using Shannon's classical Information Theorem) is minimized. The authors show that an optimal solution exists to find such segmentation. The segments which require highest number of bits for encoding are treated as anomalies.

It is proposed an algorithm called Window Comparison Anomaly Detection (WCAD), where the authors extract subsequences out of a given sequence of continuous observations using a sliding window. The authors compare each sub-sequence with the entire sequence using a compression based dissimilarity measure. The anomaly score of each subsequence is its dissimilarity with the entire sequence.

It is proposed a related technique (HOT SAX) to solve the above problem for continuous time-series. The basic approach followed by the authors is to extract subsequences out of the given sequence using sliding window, and then computing the distance of each subsequence to its closest non-overlapping subsequence within the original sequence. The anomaly score of a subsequence is proportional to its distance from its nearest neighbors. Distance between two sequences is measured using Euclidean measure. Similar approach is also applied to the domain of medical data. The same authors propose the use of Haar Wavelet based transformation to make the previous technique more efficient.

The problem formulation there is to predict the most likely state sequence for a given observation sequence. Any anomalous segment within the observation sequence will have a low conditional probability for any state sequence. Determining if the frequency of a query pattern in a given sequence is anomalous with respect to its expected frequency. Such formulation of the anomaly detection problem is motivated from the case vs control type of data. The idea is to detect patterns whose occurrence in a given test data set (case) is different from its occurrence in a normal data set (control For each of these substrings they determine if this substring is anomalous with respect to a normal database of strings. The authors use suffix trees to estimate the expected frequency of a substring in the normal database of strings. In a similar approach, the authors use Interpolated Markov Models (IMM) to estimate the expected frequency.

II.2.8.2. Handling Spatial Anomalies

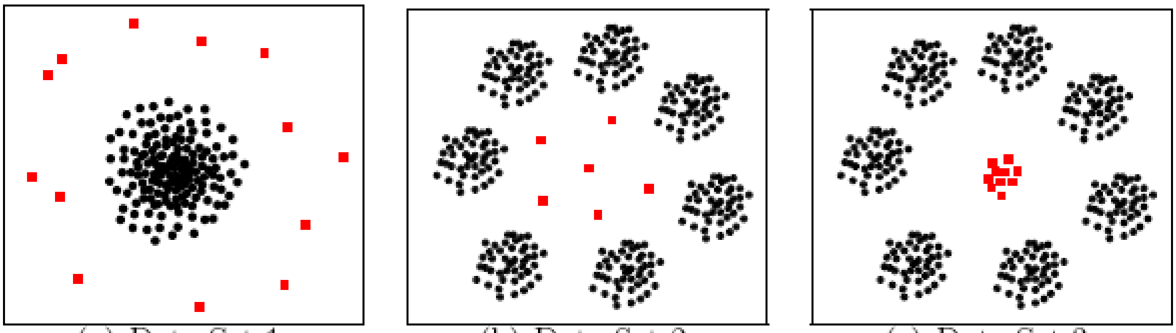
Collective anomaly detection in spatial data involves finding sub-graphs or sub-components in the data that are anomalous. A limited amount of research has been done in this category so we will discuss them individually.

Yairi (2001) propose a technique to detect regions in an image that are anomalous with respect to rest of the image. The proposed technique makes use of Multivariate Gaussian Random Markov Fields (MGMRF) to segment a given image. The authors make an assumption that each pixel belonging to an anomalous region of the image is also a contextual anomaly within its segment. These pixels are detected as contextual anomalies with respect to the segments (by estimating the conditional probability of each pixel), and then connected using a spatial structure available, to find the collective anomalies.

Anomaly detection for graphs has been explored in application domains where the data can be modeled as graphs. It is addressed two distinct collective anomaly detection problems for graph data. The first problem involves detecting anomalous sub-graphs in a given large graph. The authors use a bottom-up sub-graph enumeration technique and analyze the frequency of a sub-graph in the given graph to determine if it is an anomaly or not. The size of the sub-graph is also taken into account, since a large sub-graph (such as the graph itself) is bound to occur very rarely in the graph while a small sub-graph (such as an individual node) will be more frequent. The second problem involves detecting if a given sub-graph is an anomaly with respect to a large graph. The authors measure the regularity or entropy of the sub-graph in the context of the entire graph to determine its anomaly score.

II.3. RELATIVE STRENGTHS AND WEAKNESSES OF ANOMALY DETECTION TECHNIQUES

Each of the large number of anomaly detection techniques discussed in previous sections has their unique strengths and weaknesses. It is important to know which anomaly detection technique is best suited for a given anomaly detection problem. Given the complexity of the problem space, it is not feasible to provide such an understanding for every anomaly detection problem. In this section we analyze the relative strengths and weaknesses of different categories of techniques for a few simple problem settings.



(a) Data Set 1 (b) Data Set 2 (c) Data Set 3

Figure 10: 2-D data sets. Normal instances are shown as circles and anomalies are shown as squares (Kumar et al., 2005)

For example, let us consider the following anomaly detection problem: The normal data instances are generated from a Gaussian distribution and are located in a tight cluster in the 2D space. The anomalies are a very few instances generated from another Gaussian distribution whose mean is very far from the first distribution. A representative training data set that contains instances from the normal data set is also available. Thus the assumptions made by techniques in Sections 4{9 hold for this data set, and hence any anomaly detection techniques belonging to these categories will detect the anomalies in such a scenario.

Let the normal instances be such that they are generated by a large number of different Gaussian distributions with means arranged on a circle and very low variance. Thus the normal data will be a set of tight clusters arranged on a circle. A one-class classification based technique might learn a circular boundary around the entire data set and hence will not be able to detect the anomalies that lie within the circle of clusters. On the other hand if each cluster was labeled as a different class, a multi-class classification based technique might be able to learn boundaries around each cluster, and hence be able to detect the anomalies in the center. A statistical technique that uses a mixture model approach to model the data may be able to detect the anomalies.

Similarly, clustering based and nearest neighbor based techniques will be able to detect the anomalies since they are far from all other instances. In a similar example (Figure 10(c)), if the anomalous instances form a tight cluster of significant size at the center of the circle, both clustering based and nearest neighbor based techniques will treat these instances as normal, thus exhibiting poor performance. For more complex data sets, different types of techniques face different challenges.

Nearest neighbor and clustering based techniques suffer when the number of dimensions is high because the distance measures in high number of dimensions are not able to differentiate between normal and anomalous instances. Spectral techniques explicitly address high dimensionality problem by mapping data to a lower dimensional projection. But their performance is highly dependent on the assumption that the normal instances and anomalies are distinguishable in the projected space. Classification based techniques can be a better choice in such scenario. But to be most effective, classification based techniques require labels for both normal and anomalous instances, which are not often available. Even if the labels for both normal and anomalous instances are available, the imbalance in the distribution of the two labels often makes learning a classifier quite challenging. Semi-supervised nearest neighbor and clustering techniques, that only use the normal labels, can often be more effective than the classification based techniques. Statistical techniques, though unsupervised, are effective only when the dimensionality of data is low and statistical assumptions hold. Information theoretic techniques require a measure that is sensitive enough to detect the effects of even a single anomaly. Otherwise, such techniques can detect anomalies only when there are significantly enough number of anomalies.

Nearest neighbor and clustering based techniques require distance computation between a pair of data instances. Thus, such techniques assume that the distance measure can discriminate between the anomalies and normal instances well enough. In situations where identifying a good distance measure is difficult, classification based or statistical techniques might be a better choice.

The computational complexity of an anomaly detection technique is a key aspect, especially when the technique is applied to a real domain. While classification based, clustering based, and statistical techniques have expensive training times, testing is usually fast. Often this is acceptable, since models can be trained in an off-line fashion while testing is required to be in real time. In contrast, techniques such as nearest neighbor based, information theoretic, and spectral techniques which do not have a training phase, have expensive testing phase which can be a limitation in a real setting.

Anomaly detection techniques typically assume that anomalies in data are rare when compared to normal instances. Though this assumption is generally true, anomalies are not always rare. For example, when dealing with worm detection in computer networks, the anomalous (worm) traffic is actually more frequent than the normal traffic. Unsupervised techniques are not suited for such bulk anomaly detection. Techniques operating in supervised or semi-supervised modes can be applied to detect bulk anomalies. The comparison of techniques can be seen in Table 1:

Table 1: Comparison of Anomaly Detection Techniques

Techniques	Advantages	Disadvantages
Classification Based Techniques	(1) Classification based techniques, especially the multi-class techniques, can make use of powerful algorithms that can distinguish between instances belonging to different classes. (2) The testing phase of classification based techniques is fast since each test instance needs to be compared against the pre-computed model.	(1) Multi-class classification based techniques rely on availability of accurate labels for various normal classes, which is often not possible. (2) Classification based techniques assign a label to each test instance, which can also become a disadvantage when a meaningful anomaly score is desired for the test instances.
Nearest Neighbor Based Techniques	(1) A key advantage of nearest neighbor based techniques is that they are unsupervised in nature and do not make any assumptions regarding the generative distribution for the data. Instead, they are purely data driven. (2) Semi-supervised techniques perform better than unsupervised techniques in terms of missed anomalies, since the likelihood of an anomaly to form a close neighborhood in the training data set is very low. (3) Adapting nearest neighbor based techniques to a different data type is straight-forward, and primarily requires defining an appropriate distance measure for the given data.	(1) For unsupervised techniques, if the data has normal instances that do not have enough close neighbors or if the data has anomalies that have enough close neighbors, the technique fails to label them correctly, resulting in missed anomalies. (2) For semi-supervised techniques, if the normal instances in test data do not have enough similar normal instances in the training data, the false positive rate for such techniques is high. (3) The computational complexity of the testing phase is also a significant challenge since it involves computing the distance of each test instance with all instances belonging to either the test data itself, or to the training data, to compute the nearest neighbors.

		(4) Performance of a nearest neighbor based technique greatly relies on a distance measure, defined between a pair of data instances that can effectively distinguish between normal and anomalous instances. Defining distance measures between instances can be challenging when the data is complex, e.g. graphs, sequences, etc.
Clustering Based Techniques	(1) Clustering based techniques can operate in an unsupervised mode. (2) Such techniques can often be adapted to other complex data types by simply plugging in a clustering algorithm that can handle the particular data type. (3) The testing phase for clustering based techniques is fast since the number of clusters against which every test instance needs to be compared is a small constant.	(1) Performance of clustering based techniques is highly dependent on the effectiveness of clustering algorithm in capturing the cluster structure of normal instances. (2) Many techniques detect anomalies as a by-product of clustering, and hence are not optimized for anomaly detection. (3) Several clustering algorithms force every instance to be assigned to some cluster. This might result in anomalies getting assigned to a large cluster, thereby being considered as normal instances by techniques that operate under the
		assumption that anomalies do not belong to any cluster. (4) Several clustering based techniques are effective only when the anomalies do not form significant clusters among themselves. (5) The computational complexity for clustering the data is often a bottleneck, especially if O(N2d) clustering algorithms are used.
Statistical Techniques	(1) If the assumptions regarding the underlying data distribution hold true, statistical techniques provide a statistically justifiable solution for anomaly detection. (2) The anomaly score provided by a statistical technique is associated with a confidence interval, which can be used as additional information while making a decision regarding any test instance. (3) If the distribution estimation step is robust to anomalies in data, statistical techniques can operate in an unsupervised setting without any need for labeled training data.	(1) The key disadvantage of statistical techniques is that they rely on the assumption that the data is generated from a particular distribution. This assumption often does not hold true, especially for high dimensional real data sets. (2) Even when the statistical assumption can be reasonably justified, there are several hypothesis test statistics that can be applied to detect anomalies; choosing the best statistic is often not a straightforward task. In particular, constructing hypothesis tests for complex distributions that are required to fit high dimensional data sets is nontrivial. (3) Histogram based techniques are relatively simple to implement, but a key short- coming of such techniques for multivariate data is that they are not able to capture the interactions between different attributes.
Information Theory Based Techniques	(1) They can operate in an unsupervised setting. (2) They do not make any assumptions about the underlying statistical distribution for the data.	(1) The performance of such techniques is highly dependent on the choice of the information theoretic measure. Often, such measures can detect the presence of anomalies only when there is significantly large number of anomalies present in the data. (2) Information theoretic techniques applied to sequences and spatial data sets rely on the size of the substructure, which is often nontrivial to obtain. (3) It is difficult to associate an anomaly score with a test instance using an information theoretic technique.

Spectral Techniques	(1) Spectral techniques automatically perform dimensionality reduction and hence are suitable for handling high dimensional data sets. Moreover, they can also be used as a pre-processing step followed by application of any existing anomaly detection technique in the transformed space. (2) Spectral techniques can be used in an unsupervised setting.	(1) Spectral techniques are useful only if the normal and anomalous instances are separable in the lower dimensional embedding of the data. (2) Spectral techniques typically have high computational complexity.
Contextual Anomaly Detection	The key advantage of contextual anomaly detection techniques is that they allow a natural definition of an anomaly in many real life applications where data instances tend to be similar within a context. Such techniques are able to detect anomalies that might not be detected by point anomaly detection techniques that take a global view of the data.	The disadvantage of contextual anomaly detection techniques is that they are applicable only when a context can be defined.

II.4. APPLICATIONS IN THE LITERATURE

An exhaustive list of applications that utilize outlier detection is:

1. Loan application processing - to detect fraudulent applications or potentially problematical customers.
2. Intrusion detection - detecting unauthorized access in computer networks.
3. Activity monitoring - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
4. Network performance - monitoring the performance of computer networks, for example to detect network bottlenecks.
5. Fault diagnosis - monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for example.
6. Structural defect detection - monitoring manufacturing lines to detect faulty production runs for example cracked beams.
7. Satellite image analysis - identifying novel features or misclassified features.
8. Detecting novelties in images - for robot neo taxis or surveillance systems.
9. Motion segmentation - detecting image features moving independently of the background.
10. Time-series monitoring - monitoring safety critical applications such as drilling or high-speed milling.
11. Medical condition monitoring - such as heart-rate monitors.

12. Pharmaceutical research - identifying novel molecular structures.
13. Detecting novelty in text - to detect the onset of news stories, for topic detection and tracking or for traders to pinpoint equity, commodities, FX trading stories, outperforming or underperforming commodities.
14. Detecting unexpected entries in databases - for data mining to detect errors, frauds or valid but unexpected entries.
15. Detecting mislabeled data in a training data set.

The applications of anomaly detection problem and the references is shown in table 2:

Table 2: Applications in the literature

Application	**Techniques**	**References**
Intrusion Detection	Rough Set Theory, Classification Trees, Artificial Neural Networks, Support Vector Machines	Anderson et al.(1994), Barbara et al.(2001a,2001b,2003), Fan et al.(2001), Labib et al,(2002), Lee et al.(2000)
Fraud Detection	Taguchi Method, Regression and Classification Trees, Time Series, SVM,	Lin et al. (2000), Ghoting et al. (2006), Aleskerov et al. (2007), Harris (2003), Aggarwal et al.(2005), He et al.(2003),Baker et al. (1999)
Fault Detection	Artificial Neural Networks, Association Rules, Fuzzy Systems, Time Series, Regression	Gosh et al.(2007), Aggarwal (2001), Ihler et al.(2006), Manson et al.(2001), Zeevi et al.(2007)
Novelty Detection	Extreme Value Distribution, Clustering, Artificial Neural Networks	Augutejin et al. (2002), Ma et al.(2003a,2003b), Markou et al. (2003a,2003b), Roberts (2002)
Internet Anomaly Detection	Genetic Algorithm, Spatial Techniques, Radial Basis Function, Classification	Qin et al. (2004), Palshikar et al.(2005), Tandon et al.(2007)

II.5. FRAUD DETECTION

Fraud detection refers to detection of criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, etc. The malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft). The fraud occurs when these users consume the resources provided by the organization in an unauthorized way. The organizations are interested in immediate detection of such frauds to prevent economic losses. Ghoting et al. (2006) introduce the term activity monitoring as a general approach to fraud detection in these domains. The typical approach of anomaly detection techniques is to maintain a usage profile for each customer and monitor the profiles to detect any deviations. Some of the specific applications of fraud detection are discussed below:

1. Credit Card Fraud Detection
2. Mobile Phone Fraud detection
3. Insurance Claim Fraud Detection
4. Insider Trading Detection

II.5.1. Credit Card Fraud Detection.

In this domain, anomaly detection techniques are applied to detect fraudulent credit card applications or fraudulent credit card usage (associated with credit card thefts). Detecting fraudulent credit card applications is similar to detecting insurance fraud.

The data typically comprises of records defined over several dimensions such as the user ID, amount spent, time between consecutive card usage, etc. The frauds are typically rejected in transactional records (point anomalies) and correspond to high payments, purchase of items never purchased by the user before, high rate of purchase, etc. The credit companies have complete data available and also have labeled records. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain (Aleskerov et al.(2007).

The challenge associated with detecting unauthorized credit card usage is that it requires online detection of fraud as soon as the fraudulent transaction takes place. Anomaly detection techniques have been applied in two different ways to address this problem. The first one is known as by-owner in which each credit card user is profiled based on his/her credit card usage history. Any new transaction is compared to the user's profile and flagged as an anomaly if it does not match the profile. This approach is typically expensive since it requires querying a central data repository, every time a user makes a transaction. Another approach known as by-operation detects anomalies from among transactions taking place at a specific geographic location. Both by-user and by-operation techniques detect contextual anomalies. In the first case the context is a user, while in the second case the context is the geographic location.

II.5.2. Mobile Phone Fraud Detection.

Mobile/cellular fraud detection is a typical activity monitoring problem. The task is to scan a large set of accounts, examining the calling behavior of each, and to issue an alarm when an account appears to have been misused. Calling activity may be represented in various ways, but is usually described with call records. Each call record is a vector of features, both continuous (e.g., CALL-DURATION) and discrete (e.g., CALLING-CITY) (Lin et al., 2000). However, there is no inherent primitive representation in this domain. Calls can be aggregated by time, for example into call-hours or call-days or user or area depending on the granularity desired. The anomalies correspond to high volume of calls or calls made to unlikely destinations.

II.5.3.Insurance Claim Fraud Detection

An important problem in the property-casualty insurance industry is claims fraud, e.g. automobile insurance fraud. Individuals and conspiratorial rings of claimants and providers manipulate the claim processing system for unauthorized and illegal claims. Detection of such fraud has been very important for the associated companies to avoid financial losses.

The available data in this domain are the documents submitted by the claimants. The techniques extract different features (both categorical as well as continuous) from these documents. Typically, claim adjusters and investigators assess these claims for frauds. These manually investigated cases are used as labeled instances by supervised and semi-supervised techniques for insurance fraud detection.

Insurance claim fraud detection is quite often handled as a generic activity monitoring problem (Ghoting et al. 2006). Neural network based techniques have also been applied to identify anomalous insurance claims (He et al.2003; Baker et al. 1999).

II.5.4. Insider Trading Detection

Another recent application of anomaly detection techniques has been in early detection of Insider Trading. Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on (or leaking) inside information before the information is made public. The inside information can be of different forms (Harris 2003). It could refer to the knowledge of a pending merger/acquisition, a terrorist attack affecting a particular industry, a pending legislation affecting a particular industry or any information which would affect the stock prices in a particular industry. Insider trading can be detected by identifying anomalous trading activities in the market.

The available data is from several heterogeneous sources such as option trading data, stock trading data, news. The data has temporal associations since the data is collected continuously. The temporal and streaming nature has also been exploited in certain techniques (Aggarwal 2005). Anomaly detection techniques in this domain are required to detect fraud in an online manner and as early as possible, to prevent people/organizations from making illegal profits.

II.6. EXISTANCE OF ANOMALIES

A number of formal outlier tests have proposed in the literature. These can be grouped by the following characteristics:

1. What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution.
2. Is the test designed for a single outlier or is it designed for multiple outliers?
3. If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers?

The following are a few of the more commonly used outlier tests for normally distributed data. This list is not exhaustive (a large number of outlier tests have been proposed in the literature). The tests given here are essentially based on the criterion of "distance from the mean". This is not the only criterion that could be used. For example, the Dixon test, which is not discussed here, is based a value being too large (or small) compared to its nearest neighbor.

1. Grubbs' Test - this is the recommended test when testing for a single outlier.
2. Tietjen-Moore Test - this is a generalization of the Grubbs' test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.
3. Generalized Extreme Studentized Deviate (ESD) Test - this test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

The tests discussed here are specifically based on the assumption that the data follow an approximately normal distribution. If your data follow an approximately lognormal distribution, you can transform the data to normality by taking the logarithms of the data and then applying the outlier tests discussed here.

Iglewicz and Hoaglin provide an extensive discussion of the outlier tests given above (as well as some not given above) and also give a good tutorial on the subject of outliers. Barnett and Lewis provide a book length treatment of the subject.

In addition to discussing additional tests for data that follow an approximately normal distribution, these sources also discuss the case where the data are not normally distributed.

CHAPTER III:

OTHER RELATED TOPICS ABOUT APPLIED STUDY

The aim of this section is to focus on yhe main techniques which are widely use in the application in the literature. Related topic of this section is:

1. Maintenance Systems
2. Artificial Neural Networks
3. Response surface Methodology
4. Regression analysis
5. Statistical Process control

III.1. MAINTENANCE SYSTEMS

All tools, equipment and vehicles must be properly maintained so that workers are not endangered. Construction regulations require inspections of vehicles, tools, machines and equipment before use.

The degree of detail to include in the company's program regarding equipment maintenance will depend on the kinds of tools/equipment used. Some construction equipment (cranes) have very specific inspection and maintenance requirements. Mobile heavy equipment (dozers, loaders, scrapers) may have different maintenance requirements. Passenger Vehicles (company trucks, cars and vans) may require only basic maintenance. Power Tools should be maintained in good working order. This may be limited to ensuring that blades/bits are replaced when needed and that guards or other safety devices are operable and any damaged electrical cords/plugs are repaired or replaced. Damaged or defective equipment/tools should be tagged and removed from service (Gosh et al. 2005).

Most manufacturers can provide maintenance schedules for their equipment. Large companies with a fleet of vehicles/equipment typically have a comprehensive maintenance program due to the capital investment and/or leasing agreements. Smaller companies may lease equipment and maintenance services may be included in the leasing agreement. There are three main kinds of maintenance are shown in Figure 11:

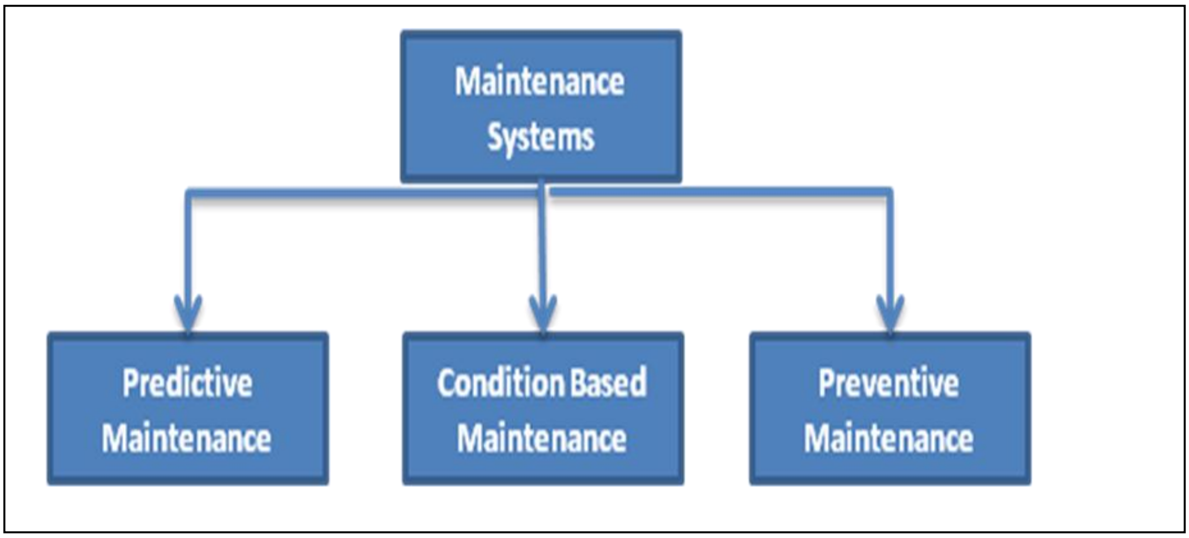


Figure 11: Classification of Maintenance systems

III.1.1 Predictive Maintenance (PdM)

Predictive maintenance techniques help determine the condition of in-service equipment in order to predict when maintenance should be performed. This approach offers cost savings over routine or time-based preventive maintenance, because tasks are performed only when warranted (Tang et al., 2002).

The main value of Predicted Maintenance is to allow convenient scheduling of corrective maintenance, and to prevent unexpected equipment failures. The key is "the right information in the right time". By knowing which equipment that needs maintenance, the maintenance work can be better planned (spare parts, people etc.) and what would had been "unplanned stops" are transformed to shorter and less "planned stops" thus increasing plant availability. Other values are increased equipment life time, increased plant safety, less accidents with negative impact on environment, an optimised spare parts handling, etc.

PdM, or condition-based maintenance, attempts to evaluate the condition of equipment by performing periodic or continuous (online) equipment condition monitoring. The ultimate goal of PdM is to perform maintenance at a scheduled point in time when the maintenance activity is most cost-effective and before the equipment loses performance within a threshold. This is in contrast to time- and/or operation count-based maintenance, where a piece of equipment gets maintained whether it needs it or not. Time-based maintenance is labor intensive, ineffective in identifying problems that develop between scheduled inspections, and is not cost-effective (Zeevi et al., 2007).

The "predictive" component of predictive maintenance stems from the goal of predicting the future trend of the equipment's condition. This approach uses principles of statistical process control to determine at what point in the future maintenance activities will be appropriate.

Most PdM inspections are performed while equipment is in service, thereby minimizing disruption of normal system operations. Adoption of PdM can result in substantial cost savings and higher system reliability.

Reliability-centered maintenance, or RCM, emphasizes the use of predictive maintenance (PdM) techniques in addition to traditional preventive measures. When properly implemented, RCM provides companies with a tool for achieving lowest asset Net Present Costs (NPC) for a given level of performance and risks.

One area that many times is overlooked is how to, in an efficient way, transfer the PdM data to a Computerized Maintenance Management System (CMMS) system so that the equipment condition data is sent to the right equipment object in the CMMS system in order to trigger maintenace planning, execution and reporting. Unless this is achieved, the PdM solution is of limited value, at least if the PdM solution is implemented on a medium to large size plant with tens of thousands pieces of equipment. In 2010, the mining company Boliden, as a first, implemented a combined Distributed Control System (DCS) and Pdm solution integrated with the plant CMMS system on an object to object level, transferring equipment data using protocols like Highway Addressable Remote Transducer Protocol (HART), IEC61850 and OLE for process control (OPC).

To evaluate equipment condition, predictive maintenance utilizes nondestructive testing technologies such as infrared, acoustic (partial discharge and airborne ultrasonic), corona detection, vibration analysis, sound level measurements, oil analysis, and other specific online tests. New methods in this area are to utilize measurements on the actual equipment in combination with measurement of process performance, measured by other devices, to trigger maintenance conditions. This is primarily available in Collaborative Process Automation Systems (CPAS). Site measurements are often supported by wireless sensor networks to reduce the wiring cost.

Vibration analysis is most productive on high-speed rotating equipment and can be the most expensive component of a PdM program to get up and running. Vibration analysis, when properly done, allows the user to evaluate the condition of equipment and avoid failures. The latest generation of vibration analyzers comprises more capabilities and automated functions than its predecessors. Many units display the full vibration spectrum of three axes simultaneously, providing a snapshot of what is going on with a particular machine. But despite such capabilities, not even the most sophisticated equipment successfully predicts developing problems unless the operator understands and applies the basics of vibration analysis.

Acoustical analysis can be done on a sonic or ultrasonic level. New ultrasonic techniques for condition monitoring make it possible to “hear” friction and stress in rotating machinery, which can predict deterioration earlier than conventional techniques. Ultrasonic technology is sensitive to high-frequency sounds that are inaudible to the human ear and distinguishes them from lower-frequency sounds and mechanical vibration. Machine friction and stress waves produce distinctive sounds in the upper ultrasonic range. Changes in these friction and stress waves can suggest deteriorating conditions much earlier than technologies such as vibration or oil analysis. With proper ultrasonic measurement and analysis, it’s possible to differentiate normal wear from abnormal wear, physical damage, imbalance conditions, and lubrication problems based on a direct relationship between asset and operating conditions.

Sonic monitoring equipment is less expensive, but it also has fewer uses than ultrasonic technologies. Sonic technology is useful only on mechanical equipment, while ultrasonic equipment can detect electrical problems and is more flexible and reliable in detecting mechanical problems.

Infrared monitoring and analysis has the widest range of application (from high- to low-speed equipment), and it can be effective for spotting both mechanical and electrical failures; some consider it to currently be the most cost-effective technology. Oil analysis is a long-term program that, where relevant, can eventually be more predictive than any of the other technologies. It can take years for a plant's oil program to reach this level of sophistication and effectiveness. Analytical techniques performed on oil samples can be classified in two categories: used oil analysis and wear particle analysis. Used oil analysis determines the condition of the lubricant itself, determines the quality of the lubricant, and checks its suitability for continued use. Wear particle analysis determines the mechanical condition of machine components that are lubricated. Through wear particle analysis, you can identify the composition of the solid material present and evaluate particle type, size, concentration, distribution, and morphology.

Because of the “predictive” part of these maintenance systems, this methodology allows to build a model based fault analisis by using data mining techniques. Therefore, it is preferred to use in our study, due to prediction of future outliers. Data mining and Predictive maintenance relationship is shown in figure12:

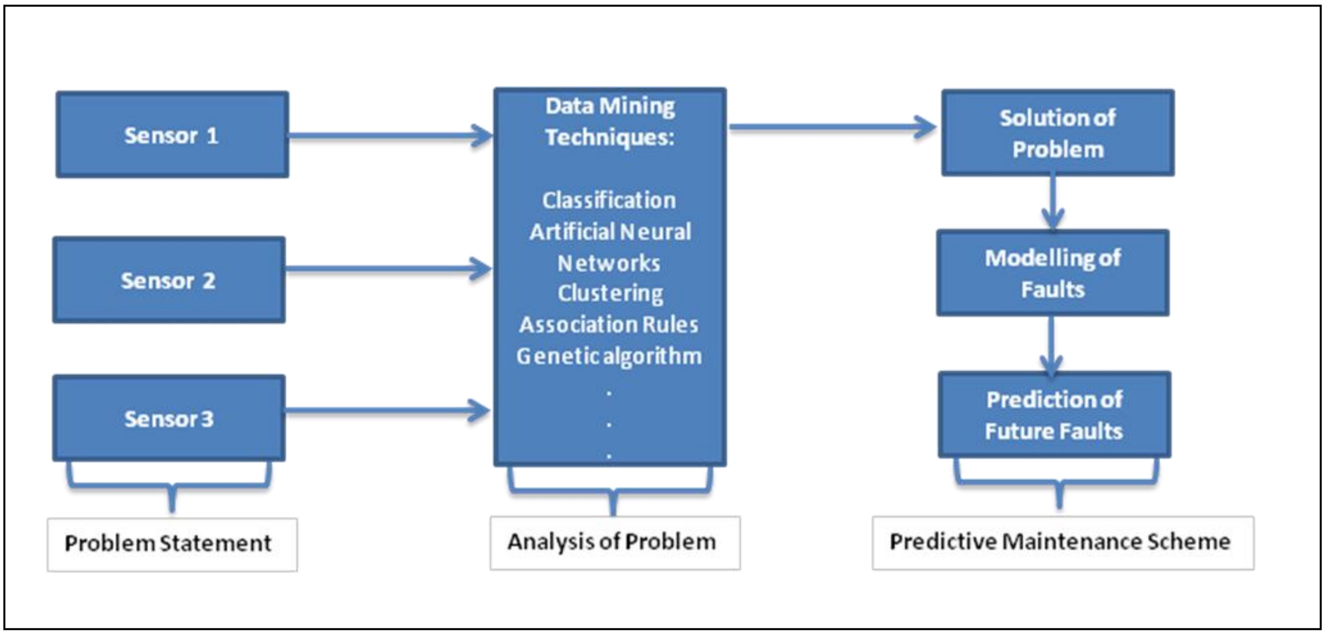


Figure 12: Data Mining and Predictive Maintenance Relationship

III.1.2 Condition-Based Maintenance (CBM)

Condition- based maintenance shortly described, is maintenance when need arises. This maintenance is performed after one or more indicators show that equipment is going to fail or that equipment performance is deteriorating (Aggarwal et. al, 2001).

Condition-based maintenance was introduced to try to maintain the correct equipment at the right time. CBM is based on using real-time data to prioritize and optimize maintenance resources. Observing the state of the system is known as condition monitoring. Such a system will determine the equipment's health, and act only when maintenance is actually necessary. Developments in recent years have allowed extensive instrumentation of equipment, and together with better tools for analyzing condition data, the maintenance personnel of today are more than ever able to decide what is the right time to perform maintenance on some piece of equipment. Ideally condition-based maintenance will allow the maintenance personnel to do only the right things, minimizing spare parts cost, system downtime and time spent on maintenance.

Despite its usefulness, there are two main challenges to the use of CBM. First and most important of all, the initial cost of CBM is high. It requires improved instrumentation of the equipment. Often the cost of sufficient instruments can be quite large, especially on equipment that is already installed. Therefore, it is important for the installer to decide the importance of the investment before adding CBM to all equipment. A result of this cost is that the first generation of CBM in the oil and gas industry has only focused on vibration in heavy rotating equipment. Secondly, introducing CBM will invoke a major change in how maintenance is performed, and potentially to the whole maintenance organization in a company. Organizational changes are in general difficult.

Also, the technical side of it is not always as simple. Even if some types of equipment can easily be observed by measuring simple values as vibration (displacement or acceleration), temperature or pressure, it is not trivial to turn this measured data into actionable knowledge about health of the equipment.

As systems get more costly, and instrumentation and information systems tend to become cheaper and more reliable, CBM becomes an important tool for running a plant or factory in an optimal manner. Better operations will lead to lower production cost and lower use of resources. And lower use of resources may be one of the most important differentiators in a future where environmental issues become more important by the day.
A more down to earth scenario where value can be created is by monitoring the health of your car motor. Rather than changing parts at predefined intervals, the car itself can tell you when something needs to be changed based on cheap and simple instrumentation.
It is Department of Defense policy that condition-based maintenance (CBM) be implemented to improve maintenance agility and responsiveness, increase operational availability, and reduce life cycle total ownership costs.

CBM has some advantages over planned maintenance (Ihler et al., 2006):

1. Improved system reliability
2. Decreased maintenance costs
3. Decreased number of maintenance operations causes decreasing of human error influence

Disadvantages are:

1. High installation costs, for minor equipment items more than value of equipment
2. Unpredictable maintenance periods are causing costs to be divided unequally
3. Increased number of parts (CBM installation) that need maintenance and checking

CBM today due to its costs is not used for less important parts of machinery despite obvious advantages. However it can be found everywhere where increased reliability and safety is required, and in future will be even more applied.

III.1.3 Preventive maintenance (PM)

Preventive Maintenance has the following meanings:

1. The care and servicing by personnel for the purpose of maintaining equipment and facilities in satisfactory operating condition by providing for systematic inspection, detection, and correction of incipient failures either before they occur or before they develop into major defects.
2. Maintenance, including tests, measurements, adjustments, and parts replacement, performed specifically to prevent faults from occurring.

The main difference of subgroups is determination of maintenance time, or determination of moment when maintenance should be performed (Ma et al., 2003b)

While preventive maintenance is generally considered to be worthwhile, there are risks such as equipment failure or human error involved when performing preventive maintenance, just as in any maintenance operation. Preventive maintenance as scheduled overhaul or scheduled replacement provides two of the three proactive failure management policies available to the maintenance engineer. Common methods of determining what Preventive (or other) failure management policies should be applied are; OEM recommendations, requirements of codes and legislation within a jurisdiction, what an "expert" thinks ought to be done, or the maintenance that's already done to similar equipment, and most important measured values and performance indications.

To make it simple:

1. PM is conducted to keep equipment working and/or extend the life of the equipment.
2. Corrective maintenance, sometimes called "repair," is conducted to get equipment working again.

The primary goal of maintenance is to avoid or mitigate the consequences of failure of equipment. This may be by preventing the failure before it actually occurs which Planned Maintenance and Condition Based Maintenance help to achieve. It is designed to preserve and restore equipment reliability by replacing worn components before they actually fail. Preventive maintenance activities include partial or complete overhauls at specified periods, oil changes, lubrication and so on. In addition, workers can record equipment deterioration so they know to replace or repair worn parts before they cause system failure. The ideal preventive maintenance program would prevent all equipment failure before it occurs.

There is a controversy of sorts regarding the propriety of the usage preventive. The consensus of internet entries concerning the respective usages seems to indicate that “preventive” is the preferred term (Manson et al., 2001).

III.2. ARTIFICIAL NEURAL NETWORKS

The term neural network applies to a loosely related family of models, characterized by a large parameter space and flexible structure, descending from studies of brain functioning. As the family grew, most of the new models were designed for non-biological applications, though much of the associated terminology reflects its origin.

Specific definitions of neural networks are as varied as the fields in which they are used. While no single definition properly covers the entire family of models, for now, considers the following description (Hawkins 2009):

A neural network is a massively parallel distributed processor that has a natural propensity for the storing experimental knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Inter neuron connection strengths known as synaptic weights are used to store knowledge.

The original inspiration for the term *Artificial Neural Network* came from examination of central nervous systems and their neurons, axons, dendrites, and synapses, which constitute the processing elements of biological neural networks investigated by neuroscience. In an artificial neural network, simple artificial nodes, variously called "neurons", "neurodes", "processing elements" (PEs) or "units", are connected together to form a network of nodes mimicking the biological neural networks — hence the term "artificial neural network".

Because neuroscience is still full of unanswered questions, and since there are many levels of abstraction and therefore many ways to take inspiration from the brain, there is no single formal definition of what an artificial neural network is. Generally, it involves a network of simple processing elements that exhibit complex global behavior determined by connections between processing elements and element parameters. While an artificial neural network does not have to be adaptive per se, its practical use comes with algorithms designed to alter the strength (weights) of the connections in the network to produce a desired signal flow.

These networks are also similar to the biological neural networks in the sense that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned (see also connectionism). Currently, the term Artificial Neural Network (ANN) tends to refer mostly to neural network models employed in statistics, cognitive psychology and artificial intelligence. Neural network models designed with emulation of the central nervous system (CNS) in mind are a subject of theoretical neuroscience and computational neuroscience.

In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on statistics and signal processing. In some of these systems, neural networks or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive elements. While the more general approach of such adaptive systems is more suitable for real-world problem solving, it has far less to do with the traditional artificial intelligence connectionist models. What they do have in common, however, is the principle of non-linear, distributed, parallel and local processing and adaptation

Although computing these days is truly advanced, there are certain tasks that a program made for a common microprocessor is unable to perform; even so a software implementation of a neural network can be made with their advantages and disadvantages (Tandon and Chan, 2007):

Advantages:

1. A neural network can perform tasks that a linear program cannot.

2. When an element of the neural network fails, it can continue without any problem by their parallel nature.
3. A neural network learns and does not need to be reprogrammed.
4. It can be implemented in any application.
5. It can be implemented without any problem.

Disadvantages:

1. The neural network needs training to operate.
2. The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated.
3. Requires high processing time for large neural networks.

Another aspect of the artificial neural networks is that there are different architectures, which consequently requires different types of algorithms, but despite to be an apparently complex system, a neural network is relatively simple.

III.3. RESPONSE SURFACE METHODOLOGY

In statistics, response surface methodology (RSM) explores the relationships between several explanatory variables and one or more response variables. The method was introduced by G. E. P. Boxand K. B. Wilson in 1951. The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response. Box and Wilson suggest using a second-degree polynomial model to do this. They acknowledge that this model is only an approximation, but use it because such a model is easy to estimate and apply, even when little is known about the process.

An easy way to estimate a first-degree polynomial model is to use a factorial experiment or a fractional factorial designs. This is sufficient to determine which explanatory variables have an impact on the response variable(s) of interest. Once it is suspected that only significant explanatory variables are left, then a more complicated design, such as a central composite design can be implemented to estimate a second-degree polynomial model, which is still only an approximation at best. However, the second-degree model can be used to optimize (maximize, minimize, or attain a specific target for) a response.

Response surface methodology uses statistical models, and therefore practitioners need to be aware that even the best statistical model is an approximation to reality. In practice, both the models and the parameter values are unknown, and subject to uncertainty on top of ignorance. Of course, an estimated optimum point need not be optimum in reality, because of the errors of the estimates and of the inadequacies of the model.

Nonetheless, response surface methodology has an effective track-record of helping researchers improve products and services: For example, Box's original response-surface modeling enabled chemical engineers to improve a process that had been stuck at a saddle-point for years. The engineers had not been able to afford to fit a cubic three-level design to estimate a quadratic model, and their biased linear-models estimated the gradient to be zero. Box's design reduced the costs of experimentation so that a quadratic model could be fit, which led to a (long-sought) ascent direction (Wei et al., 2003)

Application of response surface methodology can also be incorporated in matters of smaller proportions, such as video games. This is especially prevalent in golf games, where the use of response surface methodology can be applied to help the player achieve greater accomplishments than before. Once the important factors have been identified, the next step is to determine the settings for these factors that result in the optimum value of the response. The optimum value of the response may either be a maximum value or a minimum value, depending upon the product or process in question. For example, if the response in an experiment is the yield from a chemical process, then the objective might be to find the settings of the factors affecting the yield so that the yield is maximized. On the other hand, if the response in an experiment is the number of defects, then the goal would be to find the factor settings that minimize the number of defects. Methodologies that help the experimenter reach the goal of optimum response are referred to as *Response Surface Methods*. These methods are exclusively used to examine the "surface" or the relationship between the response and the factors affecting the response. Regression models are used for the analysis of the response, as the focus now is on the nature of the relationship between the response and the factors, rather than identification of the important factors.

Response surface methods usually involve the following steps (Baker et al., 1999):

1. The experimenter needs to move from the present operating conditions to the vicinity of the operating conditions where the response is optimum. This is done using the *method of steepest ascent* in the case of maximizing the response. The same method can be used to minimize the response and is then referred to as the *method of steepest descent.*
2. Once in the vicinity of the optimum response the experimenter needs to fit a more elaborate model between the response and the factors. Special experiment designs, referred to as *RSM designs,* are used to accomplish this. The fitted model is used to arrive at the best operating conditions that result in either a maximum or minimum response.
3. It is possible that a number of responses may have to be optimized at the same time. For example, an experimenter may want to maximize strength, while keeping the number of defects to a minimum. The optimum settings for each of the responses in such cases may lead to conflicting settings for the factors. A balanced setting has to be found that gives the most appropriate values for all the responses. *Desirability functions* are useful in these cases.

III.4 REGRESSION ANALYSİS

In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables that is, the average value of the dependent variable when the independent variables are held fixed. Less commonly, the focus is on a quartile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

A large body of techniques for carrying out regression analysis has been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

Regression models involve the following variables:

1. The unknown parameters denoted as β; this may be a scalar or a vector of length *k*.
2. The independent variables, X.
3. The dependent variable, *Y*.

A regression model relates *Y* to a function of X and β.

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta})$$

The approximation is usually formalized as $E(Y \mid X) = f(X, \beta)$. To carry out regression analysis, the form of the function *f* must be specified. Sometimes the form of this function is based on knowledge about the relationship between *Y* and X that does not rely on the data. If no such knowledge is available, a flexible or convenient form for *f* is chosen.

Assume now that the vector of unknown parameters β is of length *k*. In order to perform a regression analysis the user must provide information about the dependent variable *Y*:

1. If *N* data points of the form (*Y*,X) are observed, where $N < k$, most classical approaches to regression analysis cannot be performed: since the system of

equations defining the regression model is underdetermined, there is not enough data to recover β.

2. If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(X, \beta)$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of β), which has a unique solution as long as the X are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
3. The most common situation is where $N > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for β that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in β.

In the last case, the regression analysis provides the tools for:

1. Finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).
2. Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters β and predicted values of the dependent variable Y.

Two key assumptions are common to all estimation methods used in linear regression analysis:

1. The design matrix X must have full column rank p. For this property to hold, we must have $n > p$, where n is the sample size (this is a necessary but not a sufficient condition). If this condition fails this is called the multicollinearity in the regressors. In this case the parameter vector β will be not identifiable — at most we will be able to narrow down its value to some linear subspace of $\mathbf{R}^p$. Methods for fitting linear models with multicollinearity have been developed, but require additional assumptions such as “effect sparsity” — that a large fraction of the effects are exactly zero.
2. A simpler statement of this is that there must be enough data available compared to the number of parameters to be estimated. If there is too little data, then you end up with a system of equations with no unique solution. See partial least squares regression.

3. The regressors x_i are assumed to be error-free, that is they are not contaminated with measurement errors. Although not realistic in many settings, dropping this assumption leads to significantly more difficult errors-in-variables models.
4. Beyond these two assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:
5. Some estimation methods are based on a lack of correlation, among the *n* observations $(y_i,\, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n$. Statistical independence of the observations is not needed, although it can be exploited if it is known to hold.
6. The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
7. The variances of the error terms may be equal across the *n* units (termed *homoscedasticity*) or not (termed *heteroscedasticity*). Some linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present.
8. The arrangement, or probability distribution of the predictor variables *x* has a major influence on the precision of estimates of *β*. Sampling and design of experiments are highly-developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of *β*.

III.5. STATISTICAL PROCESS CONTROL

Statistical process control (SPC) is the application of statistical methods to the monitoring and control of a process to ensure that it operates at its full potential to produce conforming product. Under SPC, a process behaves predictably to produce as much conforming product as possible with the least possible waste. While SPC has been applied most frequently to controlling manufacturing lines, it applies equally well to any process with a measurable output. Key tools in SPC are control charts, a focus on continuous improvement and designed experiments.

Much of the power of SPC lies in the ability to examine a process and the sources of variation in that process using tools that give weight to objective analysis over subjective opinions and that allow the strength of each source to be determined numerically. Variations in the process that may affect the quality of the end product or service can be detected and corrected, thus reducing waste as well as the likelihood that problems will be passed on to the customer. With its emphasis on early detection and prevention of problems, SPC has a distinct advantage over other quality methods, such as inspection, that apply resources to detecting and correcting problems after they have occurred.

In addition to reducing waste, SPC can lead to a reduction in the time required to produce the product or service from end to end. This is partially due to a diminished likelihood that the final product will have to be reworked, but it may also result from using SPC data to identify bottlenecks, wait times, and other sources of delays within the process. Process cycle time reductions coupled with improvements in yield have made SPC a valuable tool from both a cost reduction and a customer satisfaction standpoint.

Freaks

A freak is a single point that is beyond a control limits. It signifies that something changed dramatically in the process for a short time or that a mistake was made. When a freak occurs;

1. Did a rare event occur?
2. Did something happen to cause the freak to occur?

The most causes of freaks in three types of charts are as follows:

R Chart

i. A sudden change in material
ii. A mistake in measurement
iii. An error in recording
iv. An error in arithmetic
v. An error in plotting
vi. An incomplete or omitted operation
vii. Demage in handling

X bar and R chart in control

i. Same as the R charts

P chart

i. Variation in sample size
ii. Sample taken from a different population
iii. An occasional very good or very poor lot

III.5.1. Freak Patterns

Test 1 : If four out of five consecutive points are in zone B or beyond on the same side of the centre line , a freak pattern is formed.

Test 2 : A freak pattern is formed if two out of three consecutive points fall in zone A or beyond on the same side of the centre line.

The most common process problems cause of freak pattern

R chart

i. Operator errors
ii. Poor technique owing to improper training or instruction
iii. Excessive vibration and increased variability caused inadequate fixtures
iv. Inconsistent materials or materials from different suppliers
v. Defective parts used in assembly
vi. Chasing the target measurement and excessive variability caused by over adjustment of the machine

X bar with the R chart in control

i. The preceding possible causes
ii. Planned rework resulting from deliberate crowding of one side of the specifications as a hedge against producing scrap

p chart

i. Poor maintenance of the machine
ii. Defective parts in an assembly
iii. Variations in sample size
iv. Nonrandom sampling

v. Sampling different distributions

III.5.2. Shifts

Shifts, sets of seven or more consecutive points that are all on one side of the centre line , indicate that the center of the distribution has changed.

Any special cause of the shift in the process is generally in one of six major categories:

i. Operator
ii. Method
iii. Machine
iv. Material
v. Tooling
vi. Environment

Shifts can indicate improvements as well as problems in the process

i. A shift away from the target line on the x bar chart signifies that the parts being measured and charted are now too large or too small ,depending on the direction of the shift.
ii. A shift up on the R chart is a trouble indicator because product variation has increased
iii. A shift down on the R charts indicates process improvement: Variation has decreased.
iv. A shift on a p chart may be an improvement or worsening of the defective proportion, but quick judgments are not advisable because a change in the inspection criteria

The chance of getting seven consecutive point above the line can be calculated with the “ and” rule:

P(7 Points above) = 0.5*0.5*0.5*0.5*0.5*0.5*0.5=(0.5)^7

=0.008

=0.8%

Possible Causes of Shifts

R chart

i. A careless or poorly trained operator
ii. Maintenance problems
iii. Change in material : poorer material will shift up, better material will shift down, a mixture of material will shift up
iv. Fixtures not holding the work in a place
v. Downshifts caused by better labor quality
vi. Downshifts owing to improved process quality

X bar chart R chart is in control

i. A change in the machine setting or speed
ii. A new operator or inspector
iii. A change in method
iv. A new lot of material
v. A new set up

p chart

i. Any changes listed for the x bar chart
ii. A change in standards

III.5.3. Runs And Trends

A chart shows a pattern of points that are steadily climbing or steadily falling is called a run.

Another pattern that is associated with a run is a trend

i. An increasing trend would be a sequence of points that are gradually climbing on the control chart with some decreases mixed in but an overall pattern showing that the measurements are increasing
ii. A decreasing trend shows the sequence of points gradually falling with some increases mixed in.

The distinction between the trend or run pattern and the shift pattern

i. The shift classification implies that something in process suddenly changed.
ii. The trend or run classification implies that there is gradual change occurring
iii. The following list summarizes the specifications for runs and trends
iv. Seven consecutive points that steadily increase or steadily decrease indicate run

v. Ten out of 11 points that steadily increase or steadily decrease form a run

vi. A long fluctuating pattern that is steadily decreasing is a trend

Runs and trends may stem from a variety of causes;

R chart, upward trend(process deteriorating)

i. Gradually increasing material variability
ii. Loosening fixtures
iii. Operator fatigue (check day-to-day time patterns)
iv. Machine wear

R chart, downward trend (process improving)

i. Improvements owing to SPC and the quality program
ii. Better training of operators
iii. Better maintenance program
iv. Gradual introduction of better material

X bar chart with the R chart in control

i. Gradual introduction of new material
ii. Tool wear
iii. Machine due for adjustment

p chart

i. Changing defective proportion
ii. Changing requirements or standards

III.5.4. Cycles

Cycles on a control chart are patterns that repeat on a regular basis.

To key to finding the problem that is causing the cycles is concentration on the factors that change the process periodically

There is no number rule for spotting cycles; it involves a recognition of repeating patterns

Cycles may be short term in duration and clearly show up on a control chart

It is also possible to have long range cycles in which several or many charts must be considered simultaneously

In some cases , trouble patterns of one type exist within another.

There are various causes of cycles:

R chart

i. Operator fatigue and shift changes
ii. Eccentric tooling wear, periodic changes
iii. Maintenance schedules
iv. Periodic speed changes

X bar chart the R chart in control

i. The listed R chart causes
ii. Seasonal or environmental changes
iii. Worn threads and locking devices
iv. Power fluctuations
v. Reliance on different suppliers

p chart

i. Material variations
ii. Different suppliers
iii. Change in sampling methods

III.5.5. Grouping

Grouping or bunching, occurs when the points on a chart occur in cluster.

Large fluctuations between clusters may qualify as one of the instability patterns: freaks or freak patterns.

But the overall pattern of grouping may be the pattern that provides the necessary information for solving problem

Grouping indicates that several different distributions are present.

The causes are following:

R chart

i. Differences work quality
ii. Inconsistent materials

x bar chart with the R chart in control

i. Differences in work quality
ii. Inconsistent materials
iii. Shifting fixtures
iv. Inconsistencies in method

v. Several upstream problems

III.5.6. Instability

An erratic pattern that has large fluctuations on a control chart is classified as instability, or unstable mixture.

Instability is generally hard to track because it usually has several causes

A pattern of instability often indicates that the trouble lies upstream from the chart that indicates the trouble.

To test unstable mixture, look for violations in that normal pattern. This process consists of a search for two characteristic:

1. More than one third of the points lie outside the center ± 1σ band or more than 4% of the points fall in or beyond the outer 1σ bands
2. The chart has a steep, zigzag pattern

The following are possible causes:

R charts, x bar charts, p charts:

i. Frequent breakdowns and startups (R and X bar)

ii. Over adjustment (x bar)

iii. A mixture of materials or parts from different machines or spindles(R, x bar and p)

iv. Loose fixtures (R and x bar)

v. Poor sampling procedure (x bar)

vi. Inconsistent material s(R and x bar)

vii. Inconsistent inspection standards (p)

III.5.7. Stable Mixtures

A mixture pattern that has erratic ups and downs similar to the instability pattern but has very few points in the middle of the chart is a stable mixture.

The point crowd or overlap the control limits

This pattern usually indicates a mixing of two different stable distributions: one for the upper set of the points and one for the lower set

Test for a stable mixture by looking for two characteristics:

i. Five or more consecutive points that are outside the center ± 1σ band
ii. Steep zigzag lines on the control chart with very few points in the middle of the chart

A stable mixture pattern on an R , x bar, or p chart may be caused mixing two stable distributions from the following list:

i. Different suppliers
ii. Different inspectors
iii. Different operators
iv. Different lines
v. Different lots

III.5.8. Stratification

In a stratification pattern, the points hug the averages line on a control chart.

The rule for spotting this pattern is to look for 14 or more points in two C zones

To the untrained eye, a stratification pattern seems to identify a smoothly running process. If the process has really improved a downward trend or shift on the range chart will accompany the stratification pattern on x bar chart

Possible causes of stratification:

R charts and x bar chart

i. Falsification of data
ii. Nonrandom, representative sampling
iii. Process improvement when x bar stratifies and R values decreas

CHAPTER IV:

APPLIED STUDY: FAULT MONITORING IN LATENT

IV.1 THE AIM OF THE APPLIED STUDY

Breakdowns in industrial manufacturing systems can have a significant impact on the profitability of a business. Expensive production equipment is idled, labor is no longer optimized, and the ratio of fixed costs to product output is negatively affected. As far as maintenance is concerned, beyond the preventive maintenance and fault-based unplanned trend, a new strategy is needed.

Recently, in aerospace, aviation, transportation, and other complex industrial systems, Predictive Maintenance (PM) strategy has been increasing interest and investment. PM has been proved to be an effective way to control the cost of services to engineering systems while maintaining their desirable operational reliability. PM is a maintenance system where an equipment repair or replacement decisions are based on the current and projected future health of the equipment.

Data mining (DM) algorithms involve monitoring changes of the system state to detect any abnormal behavior. The goal is to detect faults early to provide lead-time for maintenance actions and planning based on the collected faults. Moreover, data mining can discover hard to find incident affinities that can reduce down time resulting from failures. Neural networks are the preferred tool for many data mining applications.

A neural network can approximate a wide range of statistical models without requiring that one hypothesizes in advance some relationships between dependent and independent variables. Instead, the form of the relationship is determined during the learning process. In this study, the aim is to show the use of DM algorithms on PM, especially the use of a neural networks on a monitoring surface roughness of machined products.

VI.2. PROBLEM STATEMENT

The challenges of modern industries is mainly focused on the achievement of high quality, in terms of cost saving and increase the performance of the product with reduced environmental impact. The quality of the surface plays a very important role in the performance of the turning as a good quality. The greatest advantage of using finish hard turning is to reduced machining time and complexity required to manufacture metal parts and some other benefits.

In earlier studies, the surface finish prediction strategy has been developed using three main methods: the multiple regression, mathematical modeling based on the physics of the process, and neural network modeling (Vapniket al.,1995). In mathematical modeling approach, theoretical surface roughness achievable based on the tool geometry and feed rate is given approximately by the Formula

$R_a=0.0024f^t/r^{\varepsilon}$,(Worden et al., 2007) (1)

In this formula

Ra is the arithmetic mean of roughness mm,
f is feed rate mm/rev,
r is the tool nose radius mm,
t is time second and
ε is correction factor which is between $0<\varepsilon<1$

Nevertheless this model only considers geometrical conditions of variables. Furthermore, classical approach is not good enough for prediction. Hasan, (2004) mentioned about analysis of surface roughness by turning process. Worden et al. (2009) used multiple regression techniques to develop an in-process surface roughness prediction system. In this paper it is studied addressed the following question;

1. How do various parameters and vibration features relate to the surface roughness of the work piece being turned?
2. What are the minimum turning and vibration parameters that can produce a valid production model?
3. Based on the first two questions, what is the best prediction model and can it be repeated in a confirmation run?

According to this paper, predictions were made with errors of 9.96% on the experimental run and 10.77% based on the validation run. Wu and Jarmine (2009) made a comparison between multiple regression and artificial neural network on finish turning process. The results were more efficient with artificial neural network. Yu et al. (2006) used only neural networks to make prediction of surface roughness in lathe. developed fuzzy inference systems to present nonlinear dependence of surface finish on the process parameters. In this paper, Neural Networks are used to predict surface roughness. It is believed that, neural networks would model surface roughness and tool flank wear better than regression model, because they generally offer the ability to model more complex nonlinearities and interactions than linear and exponential regression models can offer. On the other hand, using a single neural network, it is possible to train and predicts many as performance measure desired.

Our application contains three part. The first part is to find optimum condition with given parameters. Artificial neural networks and response surface analysis are used to obtain the results. The second part is regression analysis which is used to built model to predict future faults related to time. This part implements the adaptation of predictive maintenance technique. The last part is statistical process control which is used to show process in a Shewart control chart. The flow chart about the applied study is shown in figure 13:

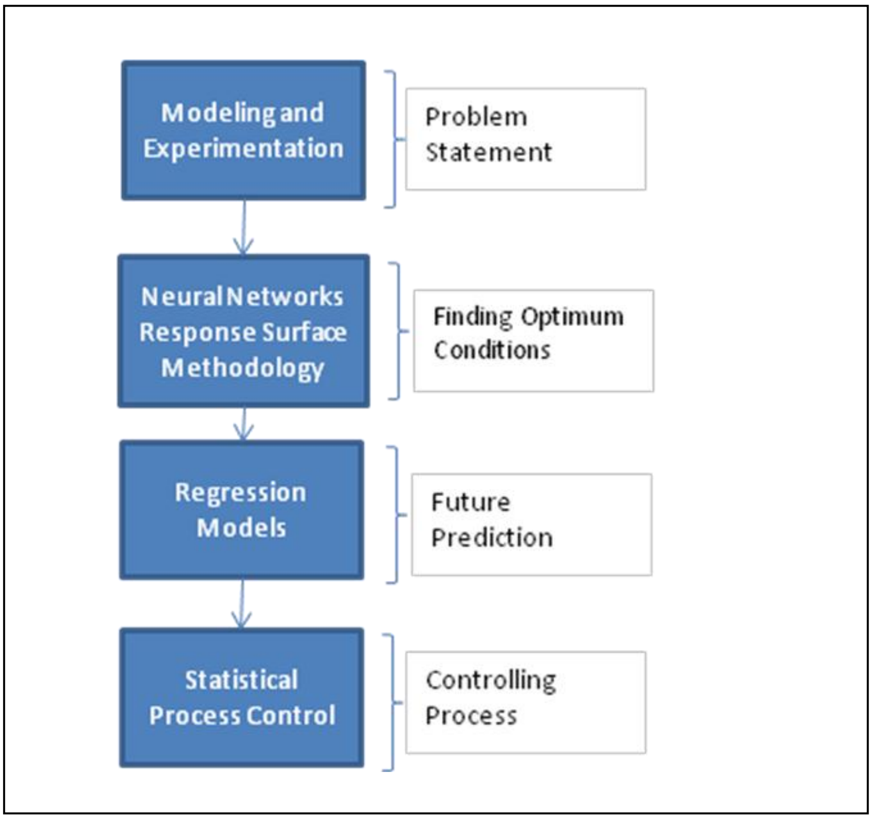

Modeling and Experimentation
Problem Statement
Neural Networks Response Surface Methodology
Finding Optimum Conditions
Regression Models
Future Prediction
Statistical Process Control
Controlling Process

Figure 13: Flowchart of applied study

IV.3 MODELING

As previously mentioned, neural network model, response surface methodology and regression analysis used as a model according to the problem statement, literature review and data type. Neural networks and response surface methodology used to obtain the optimum cutting conditions which are mentioned the experimentation section.

First of all, neural network is the first step of the applied study for modeling section. It has been mentioned in the previous chapter and also will be mentioned through the application, however it is necessary to explain something about the modeling of neurons and the importance of this technique about our study. Specific definitions of neural networks are as varied as the fields in which they are used. Three layer feed-forward neural networks have been used for modeling the surface roughness. The input parameters of the artificial neural network are the cutting conditions, namely cutting speed (V c), feed rate (S) and depth of cut (f). The output parameter is the surface roughness (Ra).After evaluating the results of radial basis function and multilayer perceptron approach for artificial neural networks, the multilayer perceptron has been chosen due to having higher correlation of predicted data with actual data.

Secondly, after neural network model, response surface analysis was used as an alternatif way for deciding the optimum and to visualize the relationship of input variables relevant to the output variable. Minitab 15 program was used to draw the graphs and the results can be seen in the next section.

After that, Regression analysis was used to predict future possible faults which are related to time. Moreover, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

Finally, Shewart chart was used to visualise trouble indicating pattern with statistical process control methods.

All in all, SPSS 17, Excel 2007 and Minitab 15 are the program which are chosen to analyse this data. After evaluating the results of analysis, the best ones are selected to present in applied study section.

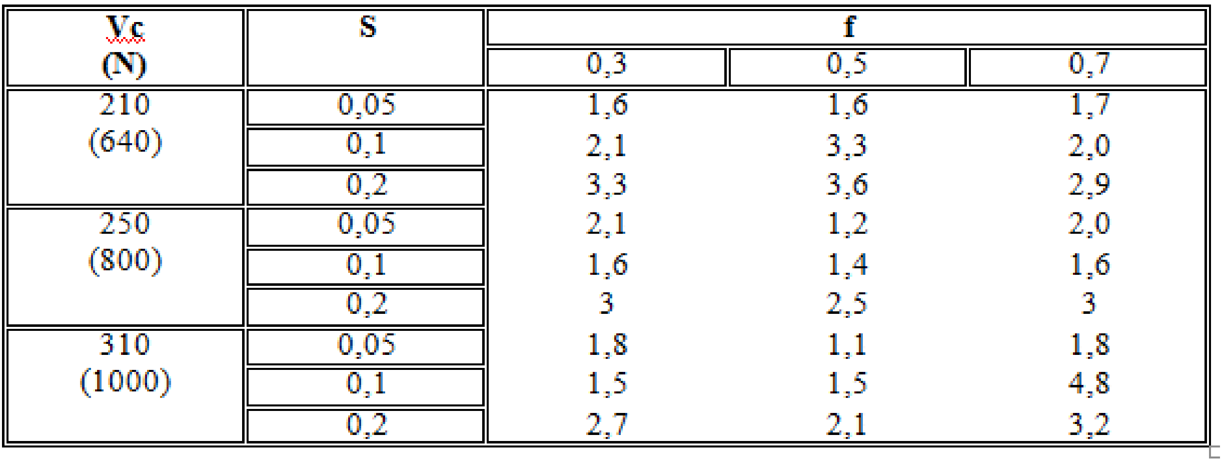
IV.4 EXPERIMENTATION

The data was taken from the finish turning operation process of Latent 201. The machining process deals with a tool material couple. The tool is made up of a carbide grade P125, and the material is heat-treated steel grade C69. The dimensions are:

V_c: Cutting speed
S: Feed rate
f: Depth of cut

The data collection made according to Taguchi method and the experiments were made randomly. Design of experimentm the dimensions and the results can be seen in table 3:

Table 3: Dataset which is used for modeling



Vc (N)	S	f		
		0,3	0,5	0,7
210 (640)	0,05	1,6	1,6	1,7
	0,1	2,1	3,3	2,0
	0,2	3,3	3,6	2,9
250 (800)	0,05	2,1	1,2	2,0
	0,1	1,6	1,4	1,6
	0,2	3	2,5	3
310 (1000)	0,05	1,8	1,1	1,8
	0,1	1,5	1,5	4,8
	0,2	2,7	2,1	3,2

There are three dimensions for each input variablem and totally 27 data observed to build models for this chapter.

In this study, independent variables were chosen according to Best Subset Algorithm which is commonly used in Statistics. It was assumed that input variables are independent and multicollinearity was not taken into account. Least Square

Algorithm was used to calculate coefficient. The list of symbols which are used for regression analysis is shown in table

Table 4: List of Symbols for Regression Analysis Outputs

R_a	Arithmetic Mean of Roughness	**P**	The probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
t	Time	**F**	F Distribution Value
t^2	Square of Time	**VIF**	Variance Inflation Factor
t^3	Cube of Time	**SS**	Sum of Square
Coef	Coefficients	**df**	Degrees of Freedom
SE	Standard Error	**MS**	Mean of Squares
T	Probability value for Student-t Distribution	**ANOVA**	Analysis of Variance

IV.5 ANALYSES

In this section the following titles have been discussed:

1. Neural network design
2. Response surface analysis
3. Regression analysis
4. Statistical process control

IV.5.1 Neural network design

All of the input data have been normalized in the range of 0-1 for ANN modeling by the following equation:

$$VN = \frac{\mathbf{V - Vmin}}{\mathbf{Vmax - Vmin}} \qquad (2)$$

Where VN is the normalized value of a variable V (real value in a parameter), Vmax and Vmin are the maximum and minimum values respectively. Furthermore, standardized roughness was used as dependent variable by the following
equation:

$$V_S = \frac{V - \bar{V}}{\sigma_V / N} \qquad (3)$$

Where V_S is the standardized value of V, $\bar{V}$ is the average of V and $\bar{V}$ is the standard deviation of V and N is the sample size. Multilayer perceptron were performed to make this analysis. Hyperbolic tangent function was used as activation function for hidden layers by the giving formula and identity function which is a function like that used by the original perceptron was used as activation function for output layer. The output is a certain value, A_1, if the input sum is above a certain threshold and A_0 if the input sum is below a certain threshold. The values used by the perceptron were $A_1 = 1$ and $A_0 = 0$. Sum of square was used for error function.

SPSS 17 software was used to perform training and test data graphics are shown in 1. The networks have 3 neurons in the inputs, corresponding to each of the three cutting parameters plus a bias which the rule allows you to shift the activation function to the left or right, which might be critical for successful learning, and one neuron in the output layer.

Networks with one hidden layer which has been chosen by the system automatically, is used.

27 measurement of roughness were used in the experiment, 21 of them were chosen for training and 6 of them were chosen for testing. ANN for roughness is shown in figure 9 and parameter estimates are shown in table 5:

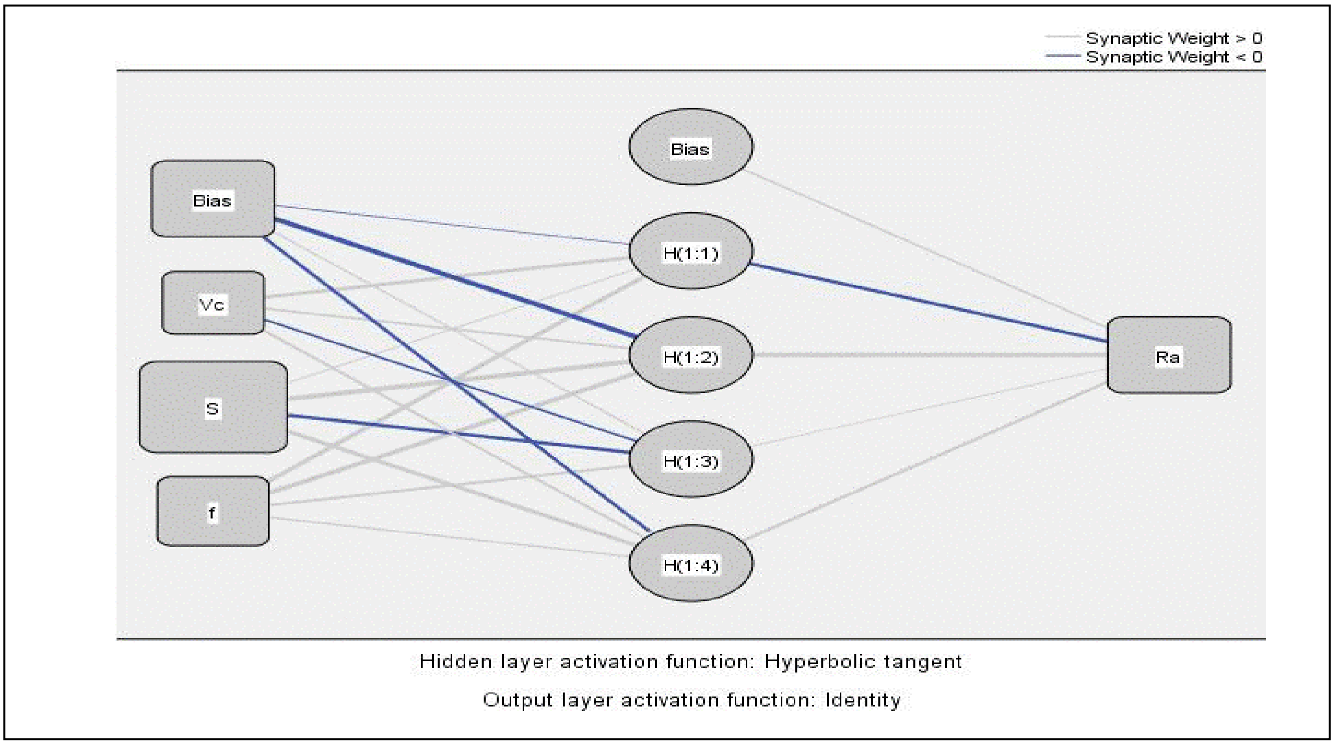
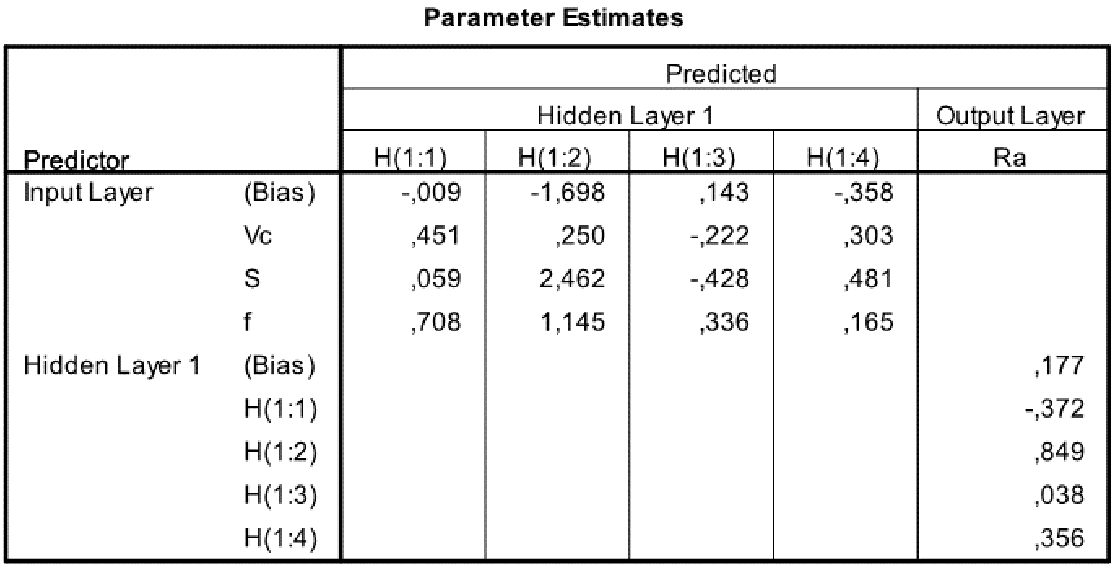

Synaptic Weight > 0
Synaptic Weight < 0
Bias
Vc
S
f
Bias
H(1:1)
H(1:2)
H(1:3)
H(1:4)
Ra
Hidden layer activation function: Hyperbolic tangent
Output layer activation function: Identity

Figure 14: Neural Network Design for Roughness according to SPSS 17

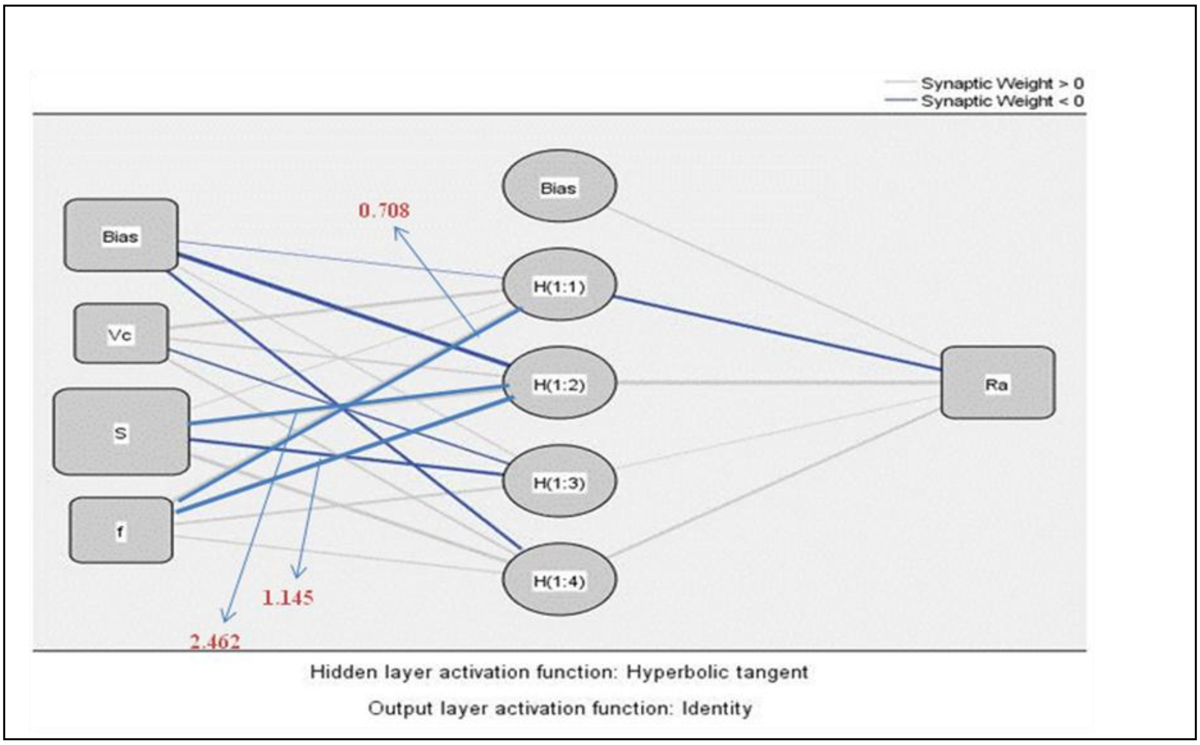
27 observation values analyzed for roughness of the machine. The optimum conditions for roughness are the minimum value for S which is approximately 0.05, medium value for f approximately 0.45 and highest value for V_c approximately 300. The neural network design is shown in figure 14. According to the model which is obtained as a result of ANN analysis the optimum roughness is approximately 1.02.The highest probabilities are between S and H(1:2), f and H(1:1) and f and H(1:2) which are shown in figure 15:

Table 5: Parameter Estimation of ANN



Parameter Estimates

Predictor		Predicted				
		Hidden Layer 1				Output Layer
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	Ra
Input Layer	(Bias)	-,009	-1,698	,143	-,358	
	Vc	,451	,250	-,222	,303	
	S	,059	2,462	-,428	,481	
	f	,708	1,145	,336	,165	
Hidden Layer 1	(Bias)					,177
	H(1:1)					-,372
	H(1:2)					,849
	H(1:3)					,038
	H(1:4)					,356



Synaptic Weight > 0
Synaptic Weight < 0
Bias
Vc
S
f
0.708
Bias
H(1:1)
H(1:2)
H(1:3)
H(1:4)
Ra
1.145
2.462
Hidden layer activation function: Hyperbolic tangent
Output layer activation function: Identity

Figure 15: The highest probabilities for layers

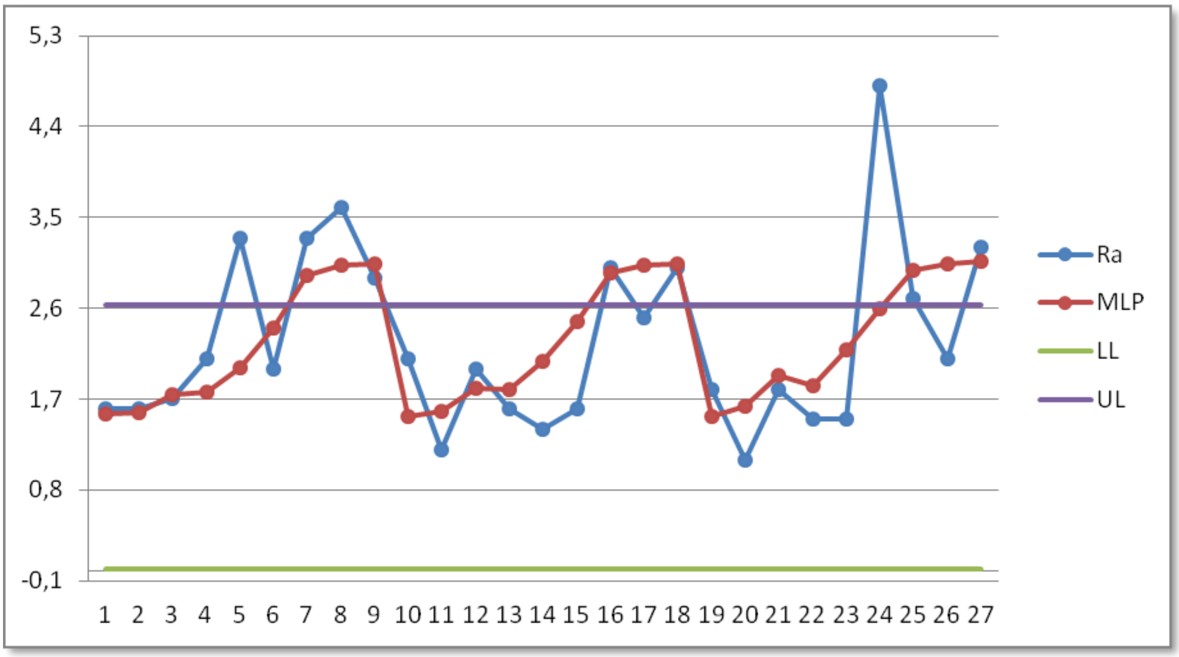


Figure 16: Multilayer Perceptron Algorithm Results versus Actual Data

Figure 16 shows the relationship between predicted and actual data. As previously mentioned, The MLP approach used to predict best fitting due to highest correlation value between the model and real measurements.

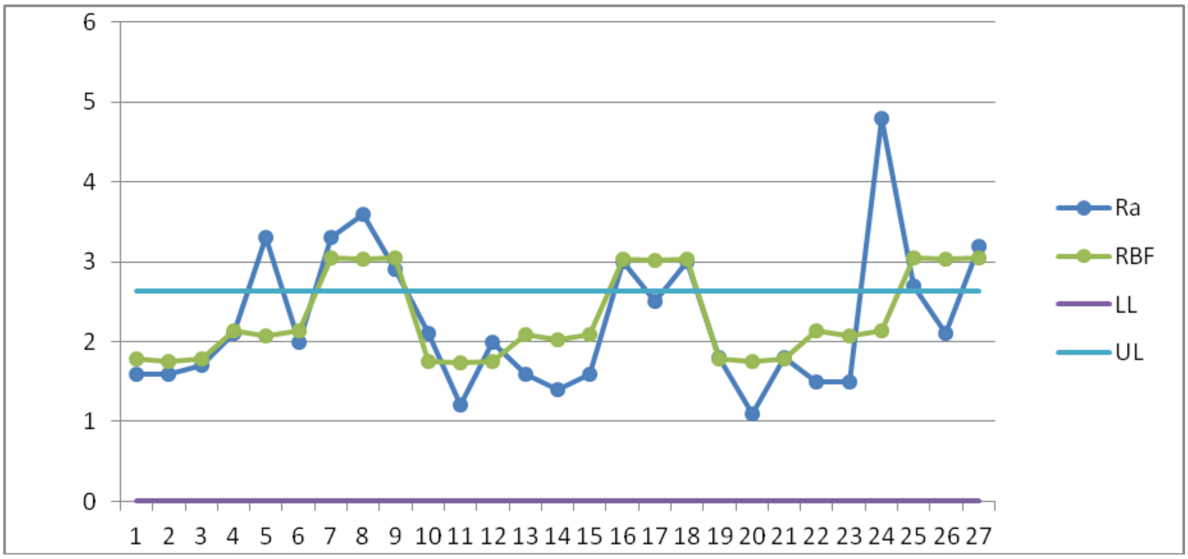
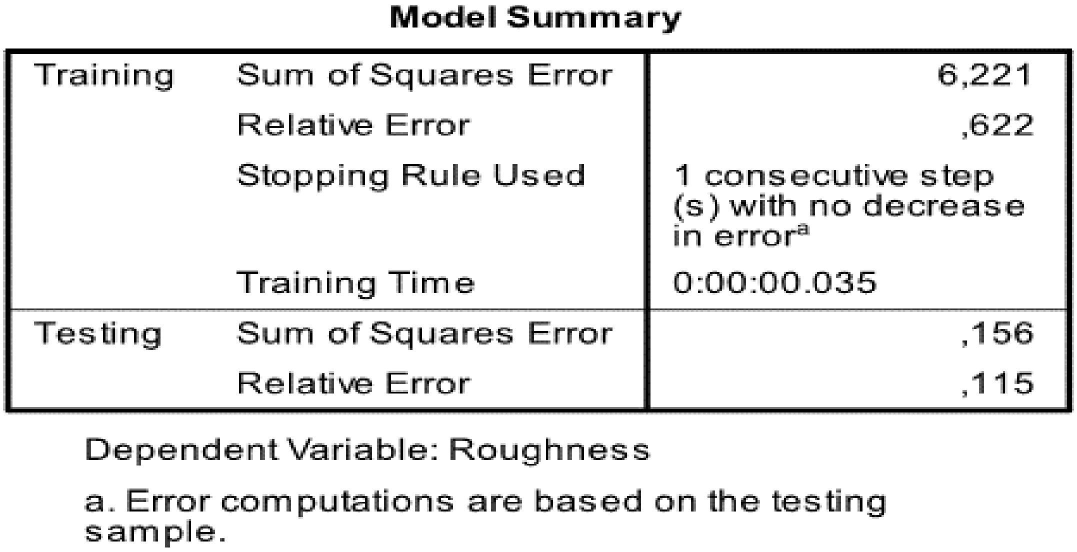


Figure 17: Radial Basis Function Results versus Actual Data

Figure 17 shows the comparison of the result of RBF results and actual data. Although the model fits quite well, because of higher correlation, the MLP approach has been chosen.

Table 6: Model Summary for ANN



Model Summary

Training	Sum of Squares Error	6,221
	Relative Error	,622
	Stopping Rule Used	1 consecutive step (s) with no decrease in error[a]
	Training Time	0:00:00.035
Testing	Sum of Squares Error	,156
	Relative Error	,115

Dependent Variable: Roughness

a. Error computations are based on the testing sample.

According to the table 6, sum of square error of training set is 6,221 and relative error is 0,622. However, sum of square error of testing sample is 0,156 and relative error of testing set is 0,115.

IV.5.2 Response Surface Analysis

Response surface analysis is used for an alternative way of neural networks to decide the optimum cutting conditions. The process is performed with MINITAB 15 program and the results are shown in Figure 18,19 and 20.

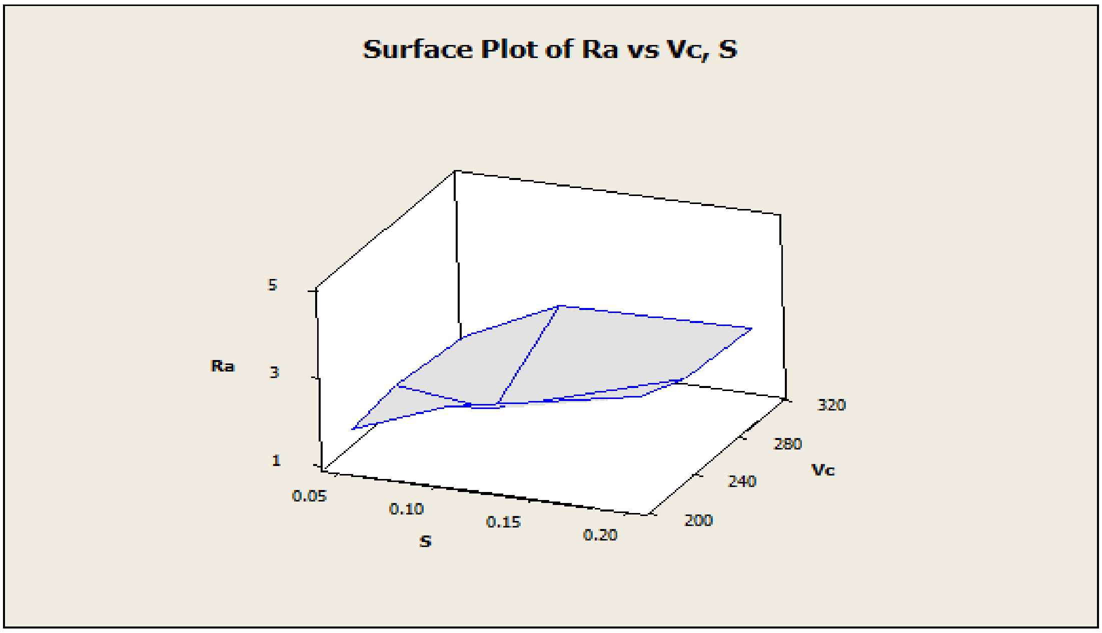


Figure 18: Surface Plot of Roughness, V_c versus S

Figure 18 indicates the surface roughness as a reaction of cutting speed (V_c) and feed rate (S). It is shown that, the optimum roughness has measured when the feed rate value is minimum and middle value of cutting speed.

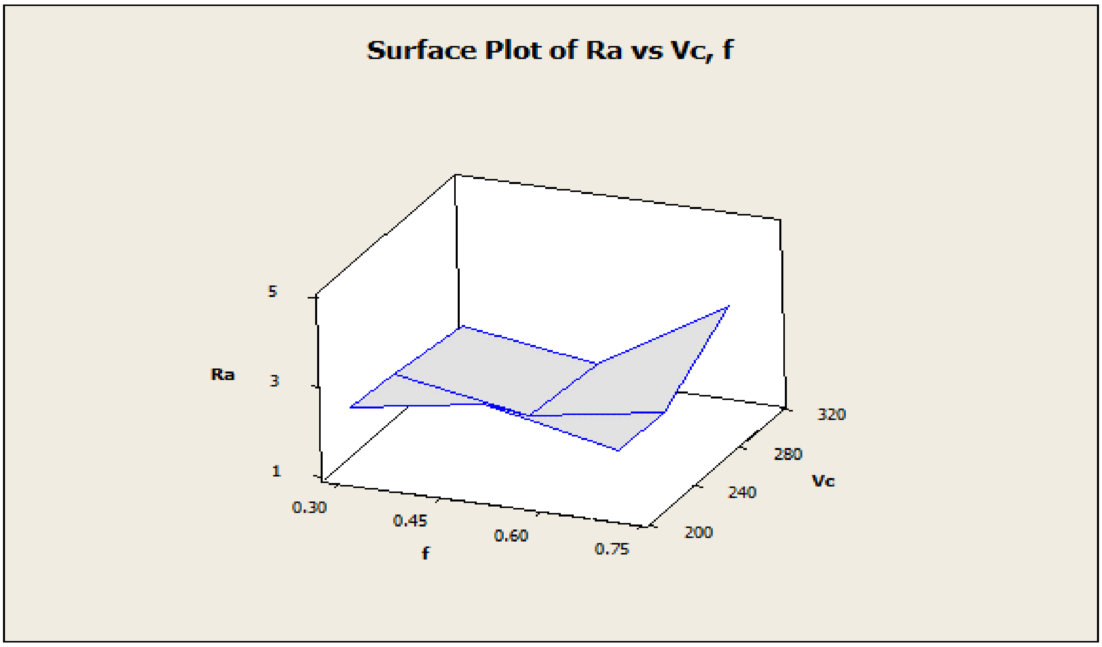


Figure 19: Surface Plot of Roughness, V_c versus f

Figure 19 indicates the surface of the roughness as a reaction of cutting speed and depth of cut. It can be claimed that the optimal conditions for minimum roughness is the highest value of depth of cut and minimum value of cutting speed.

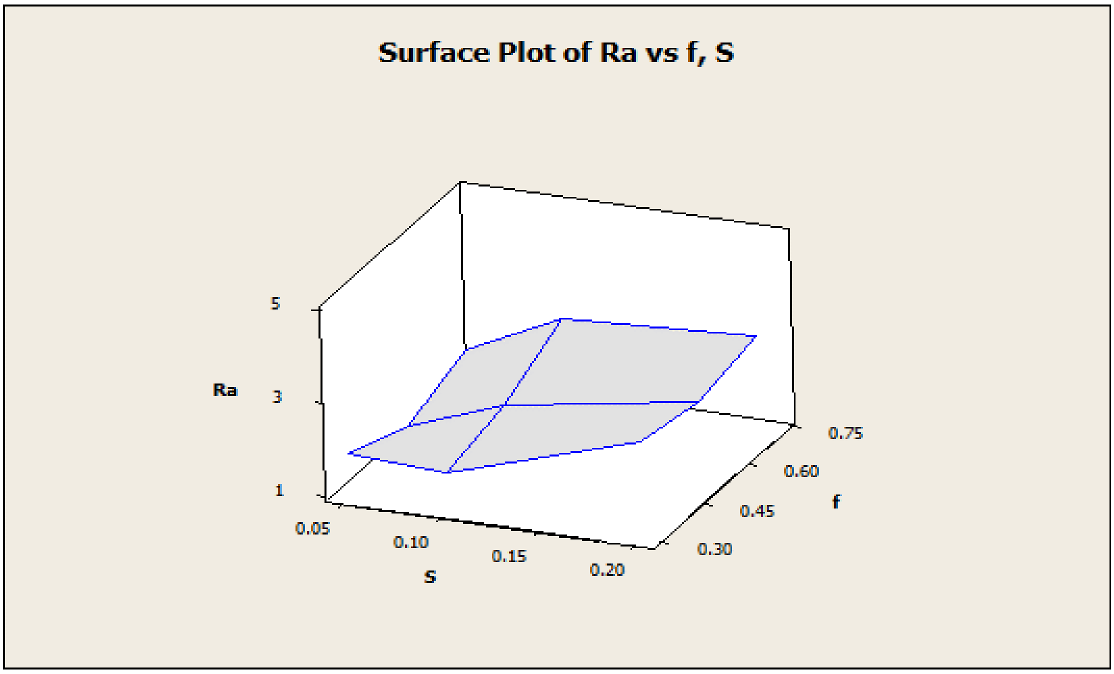


Figure 20: Surface Plot of Roughness, f versus S

Figure 20 shows the surface of the roughness as a reaction of depth of cut and feed rate. It is clear that the minimum roughness value occurs in the middle value of depth of cut and minimum value of feed rate.

IV.5.3 Regression Model for Roughness Analysis

In order to understand the relationship between dependent variable and time, scatter plot drawn in Excel which can be seen in figure 21:

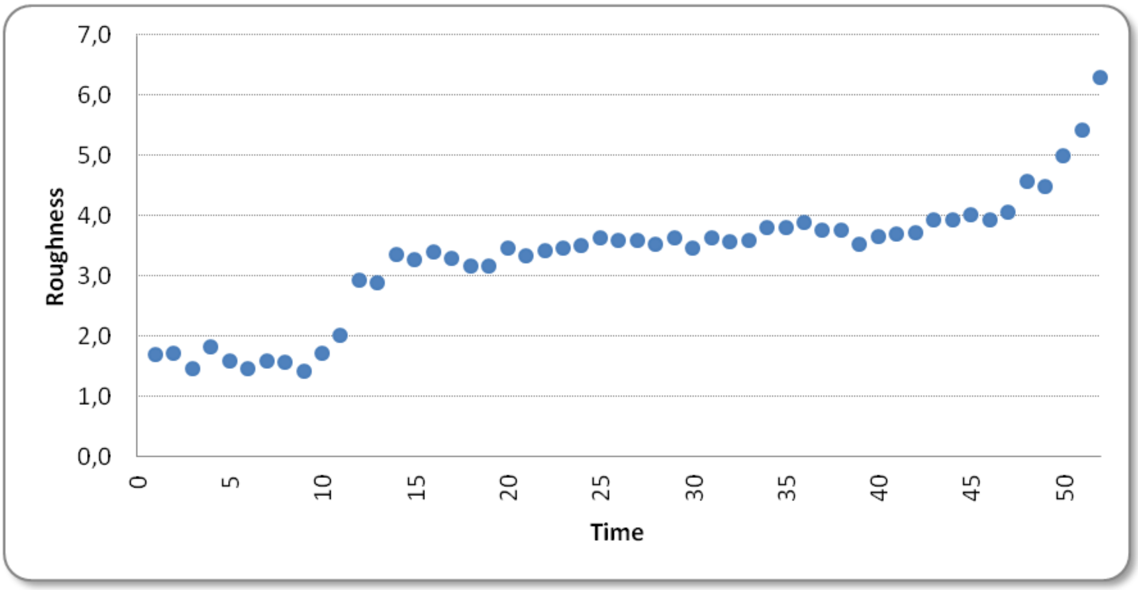


Figure 21: Scatter Plot of Roughness versus Time

According to the hyperbolic shape of the scatter plot (Figure 21) it can be claimed that the quality of roughness is related to time, square of time and cube of time. Several regression techniques have been used for these study :

IV.6.3.1. Ra versus t, t^2, t^3

$$\mathbf{Ra = 0.644 + 0.263\ t - 0.00866\ t^2 + 0.000102\ t^3} \qquad (4)$$

According to the Table 7, all of the input variables are statistically significant with 95% confidence level. It means, when other variables are fixed and T is increased one unit, Ra will increase 0.263 unit. By the same way, if t^2 is increased one unit, Ra will decrease 0.00866 unit. Furthermore, If t^3 is increased one unit, Ra will increase 0.000102 unit.

Table 7: Coefficients of Regression Analysis

Predictor	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	0.6438	0.2230	2.89	0.006	
t	0.26265	0.03609	7.28	0.000	109.442
t^2	-0.008657	0.001574	-5.50	0.000	622.385
t^3	0.00010249	0.00001953	5.25	0.000	239.826

In this model, Standard Error of Estimation (SEE) is 0.373, R^2 is 87.9 % and adjusted R^2 is 87.1%. Therefore, it can be claimed that , this model can explain 87.9% of variation in Roughness. Moreover, according to the ANOVA table (Table 8) this model is statistically significant with 5% error rate. Durbin-Watson Statistic is equal to 0.432; it means autocorrelation does not exist.

Table 8: Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	3	48.374	16.125	115.70	0.000
Residual Error	48	6.690	0.139		
Total	51	55.063			

In Figure 22, the first graph is Normal Probability Plot to test residuals' normality. It is clearly seen that, Residuals fit Normal Distribution quite well. The second graph is Residual versus Fits Plot to understand independence of residuals. According to the plot, because of lack of any pattern, It can be said that residuals are independent from each other. The third plot is Histogram; it can be claimed that, because of the shape of histogram, the distribution of residuals is Normal. According to the last plot which is called as Residuals versus Order; there is no pattern, therefore residuals are independent.

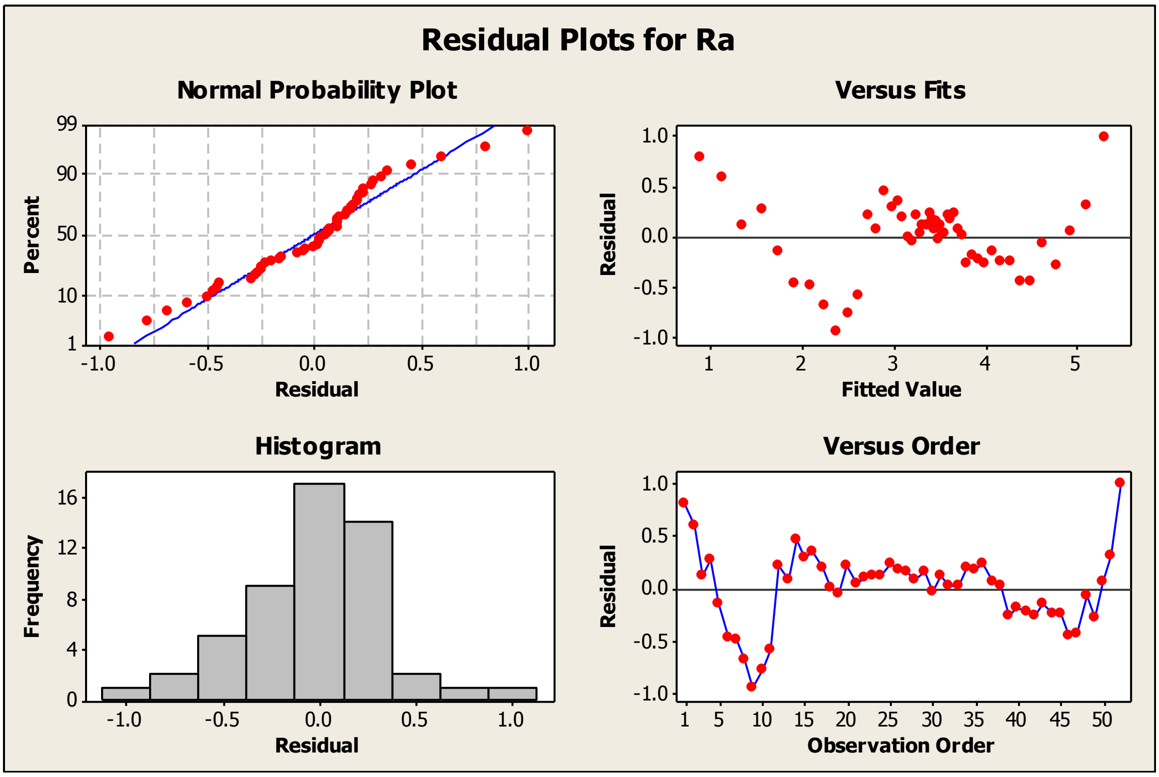


Figure 22: Residuals Plot for Regression Analysis

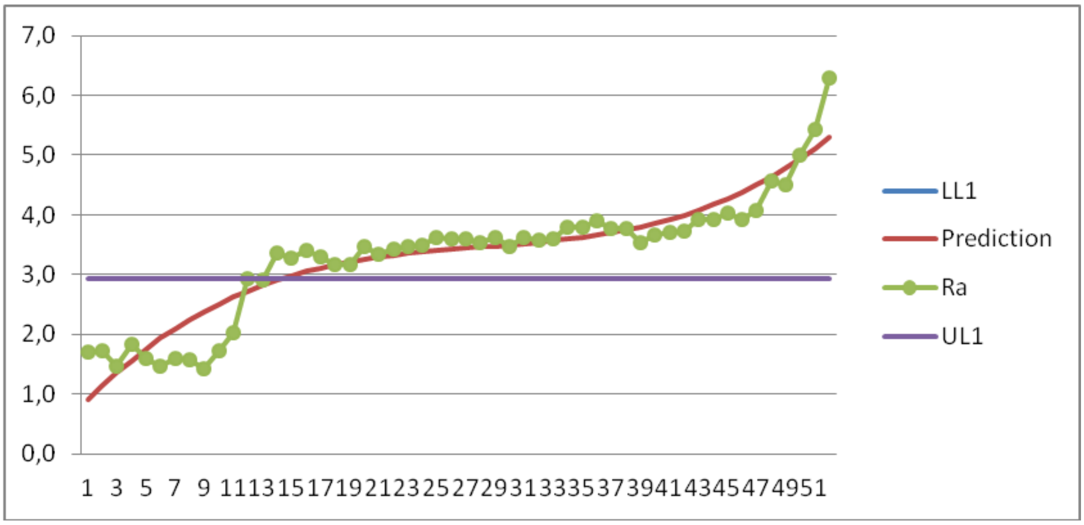


Figure 23: Fault Detection with Actual and Predicted Values

According to the Figure 23, both prediction and actual values are out-of-control at the time 14. Nonetheless, In Figure 24, Residuals seem in control.

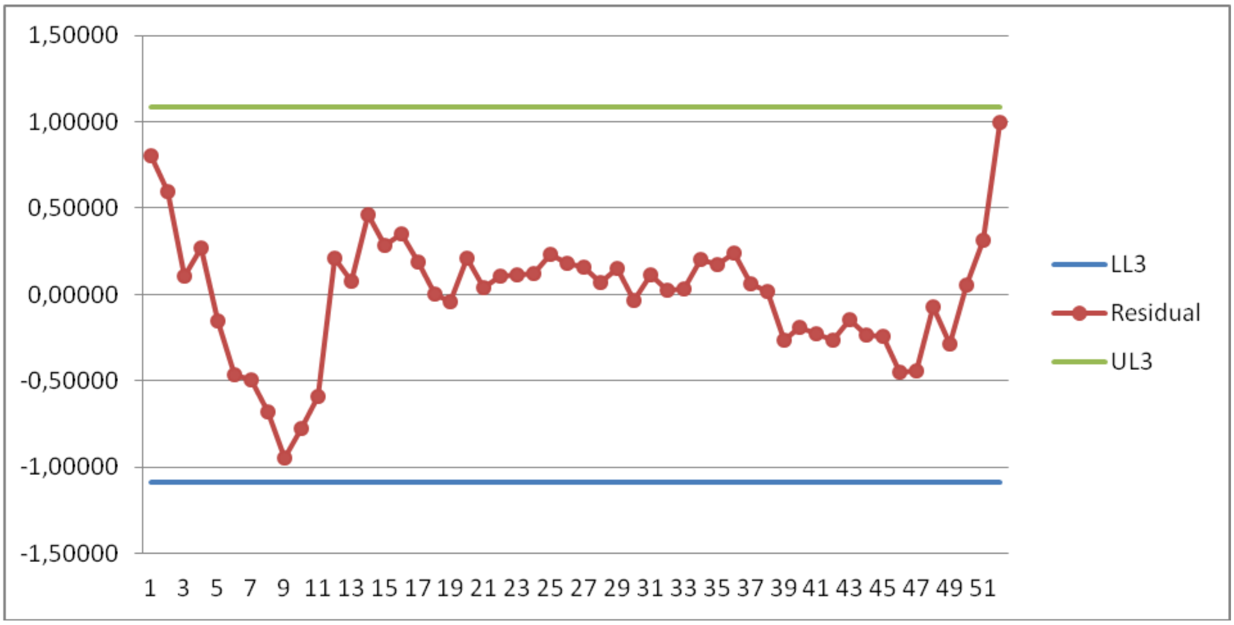


Figure 24: Residual Control Chart

IV.5.3.2. LnRa versus t, t^2, t^3

$$\mathbf{lnRa = 0.129 + 0.0962\ t - 0.00282\ t^2 + 0.000029\ t^3} \qquad (5)$$

According to the Table 9 , except constant all of the independent variables are statistically significant with 95% level of confidence. It means, if all other variables held constant every one point increase in the percentage t multiplies per Ra by $\mathbf{e^{0.0962}}$= 1.10097924. In other words, time increases per Roughness by 10.09%.

Table 9: Coefficients of Regression Analysis

Predictor	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	0.12895	0.08099	1.59	0.118	
t	0.09617	0.01311	7.34	0.000	109.442
$\mathbf{t^2}$	-0.0028162	0.0005717	-4.93	0.000	622.385
$\mathbf{t^3}$	0.00002945	0.00000710	4.15	0.000	239.826

In this model, SEE is 0.1356, R^2 is 87.0% and adjusted R^2 is 86.2%. It can be said that, this model can explain 87.0% of variation in Ra. Furthermore, According to the

Table 10, the model is statistically significant with 95% confidence level. Durbin-Watson Statistic is 0.461, so there is no autocorrelation between residuals.

Table 10: Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	3	5.9078	1.9693	107.09	0.000
Residual Error	48	0.8826	0.0184		
Total	51	6.7905			

When the Figure 25 is considered, in the first plot, residuals fit Normal Distribution quite well. In Residuals versus Fits Plot, it is seen that no pattern exist, so residuals are independent. However according to the histogram, the shape of the distribution of residuals is left skewed. Apart from this, there is no trend in Residuals versus Order Plot.

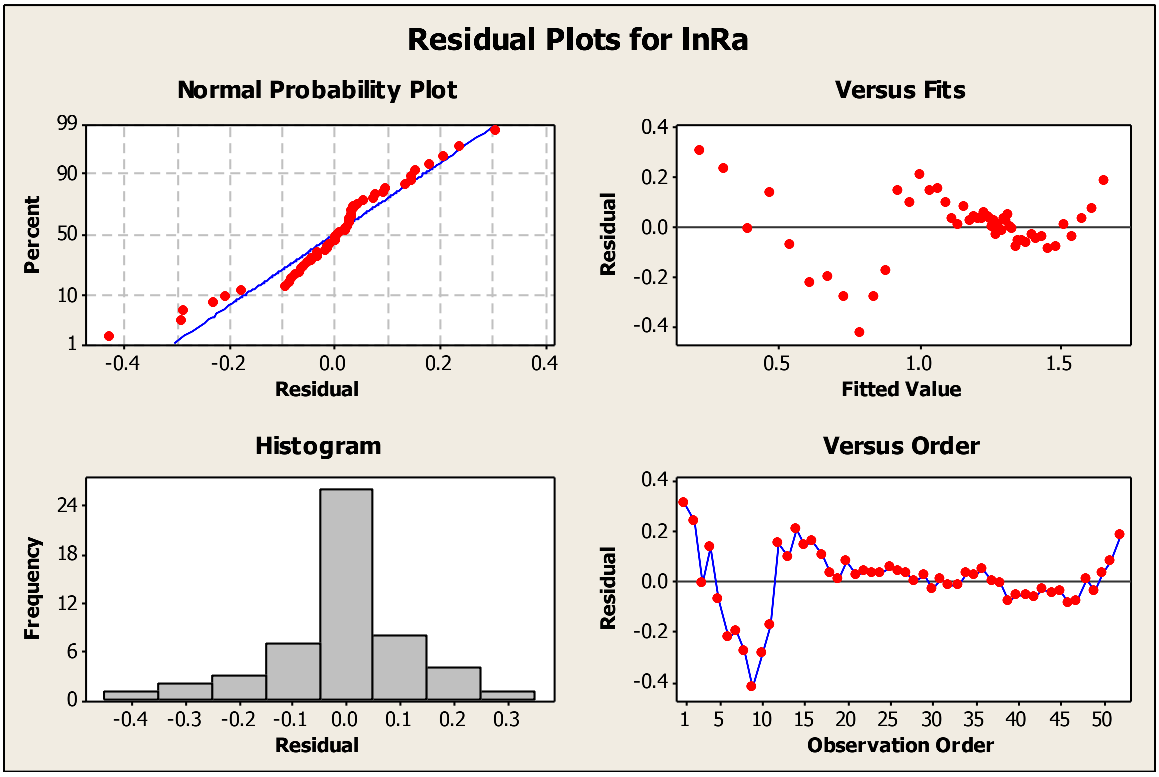


Figure 25: Residual Plots of Regression Analysis

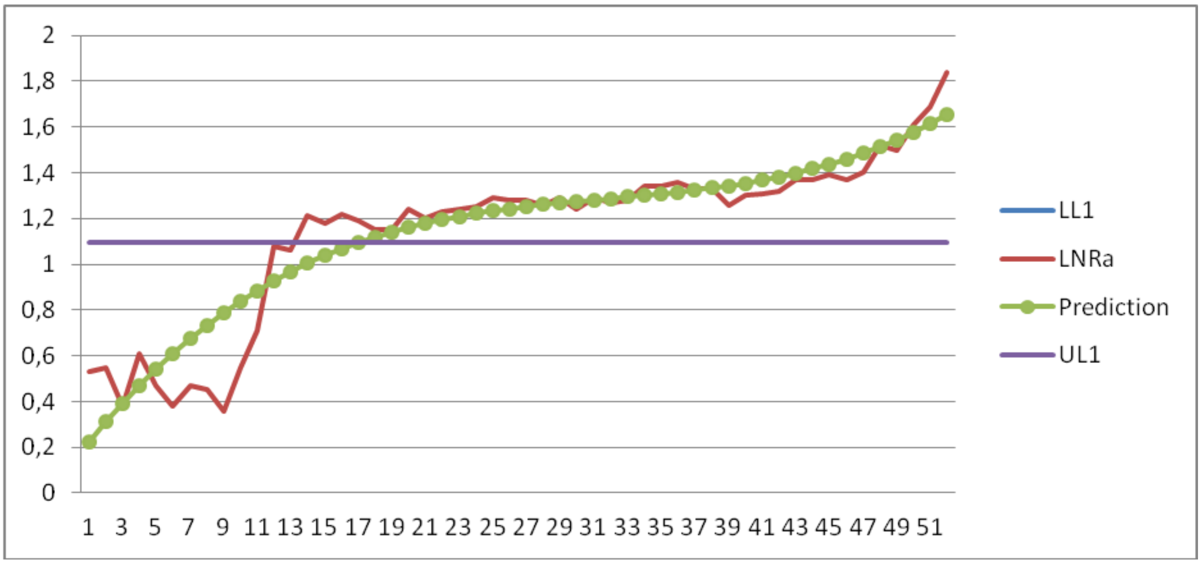


Figure 26: Fault Detection with Actual and Predicted Values

Although, the curve fits actual values very well, they are not out- of -control at the same time according to the Figure 26.

Furthermore, another interesting result is that, residual of this model is out-of-control at the time 9. It is shown in Figure 27;

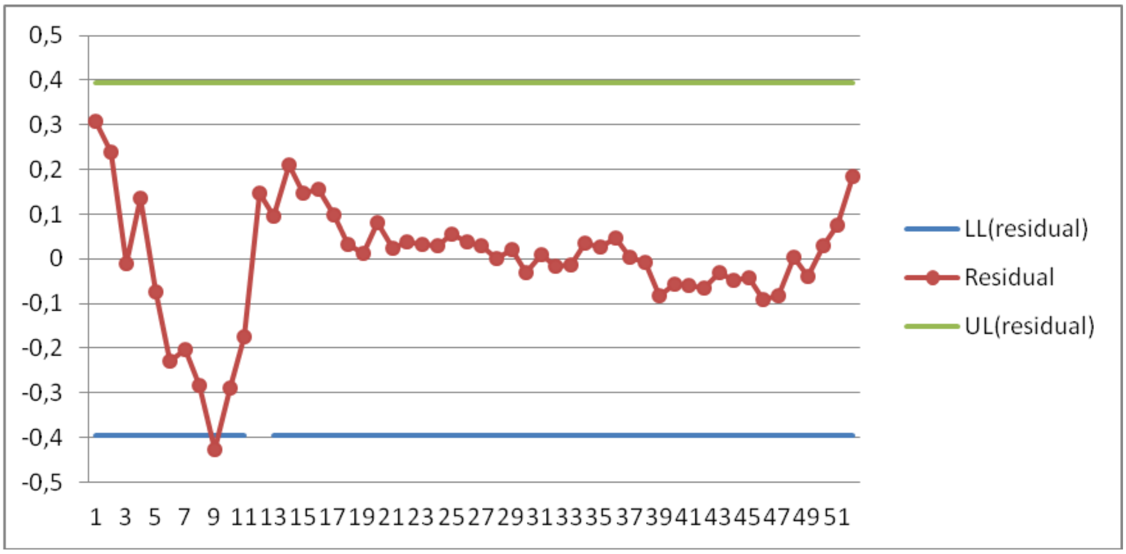


Figure 27: Residual Control Chart for Regression Analysis

IV.5.3.3. LnRa versus Lnt

lnRa = 0.0502 + 0.363 lnt (6)

According to the Table 11, the elasticity coefficient is statistically significant with 5% error rate. The meaning of the model is; Thus multiplying per time by 2.718 multiplies the roughness by $\mathbf{e^{0.363}}$= 1.43763586.

Table 11: Coefficients of Regression Analysis

Predictor	**Coef**	**SE Coef**	**T**	**P**	**VIF**
Constant	0.05020	0.08390	0.60	0.552	
lnt	0.36276	0.02677	13.55	0.000	1.000

In this model, SEE is equal to 0.170, R^2 is equal to 78.6% and adjusted R^2 is equal to 78.2%. 78.6%of variance in the dependent variable that can be explained by the model. According to the Table 12, This model is statistically significant with 5% error rate. Durbin Watson Statistic is 0.356, therefore autocorrelation does not exist.

Table 12: Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	1	5.3371	5.3371	183.61	0.000
Residual Error	50	1.4534	0.0291		
Total	51	6.7905			

In Figure 28, Normal Probability Plot indicates that, residuals fit Normal Distribution quite well. In Residuals versus Fits Plot, there is no significant pattern, therefore it can be claimed that, residuals are independent. The shape of the Histogram is quite symmetrical. Apart from these, in Residuals versus Order Plot, no trend exists.

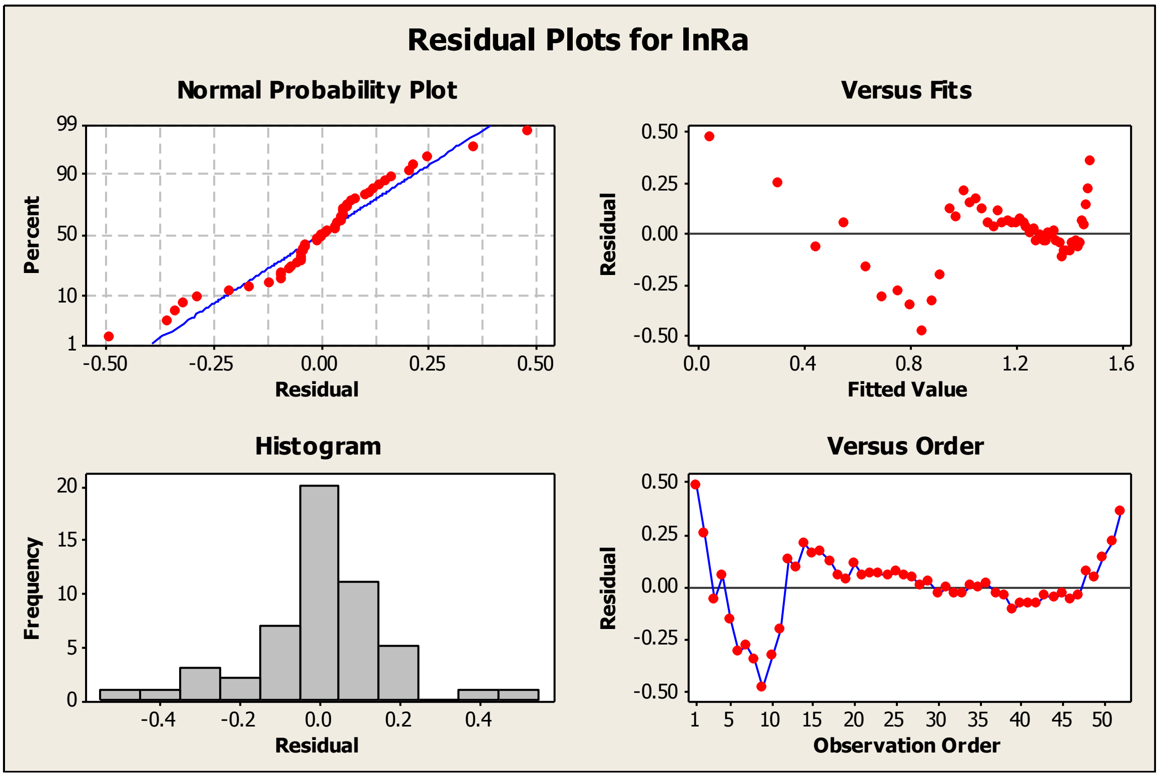


Figure 28: Residual Plots for Regression Analysis

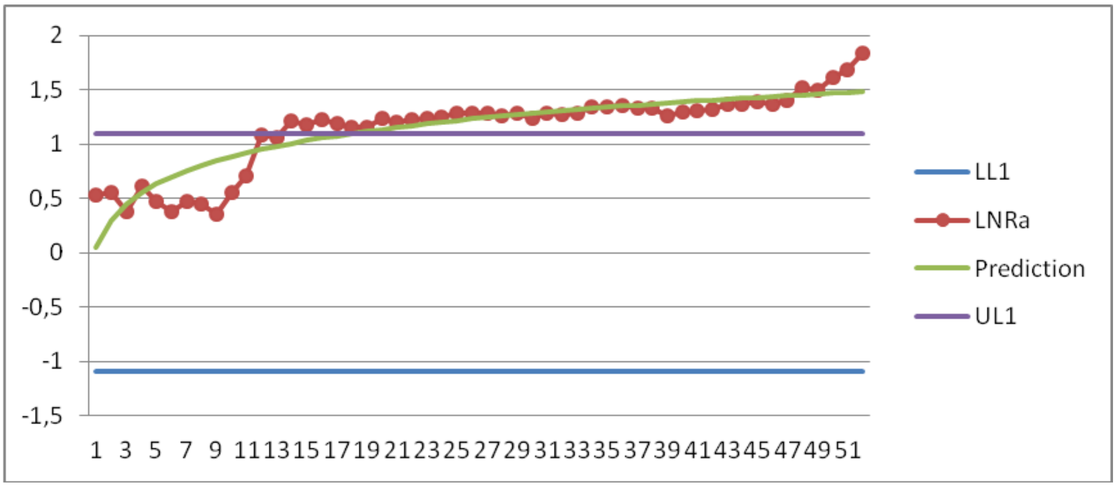


Figure 29: Fault Detection With Actual and Predicted Values

According to the Figure 29, actual and predicted values are not out-of-control at the same time. Nevertheless, in Figure 30, there is no trouble indicating pattern in residuals.

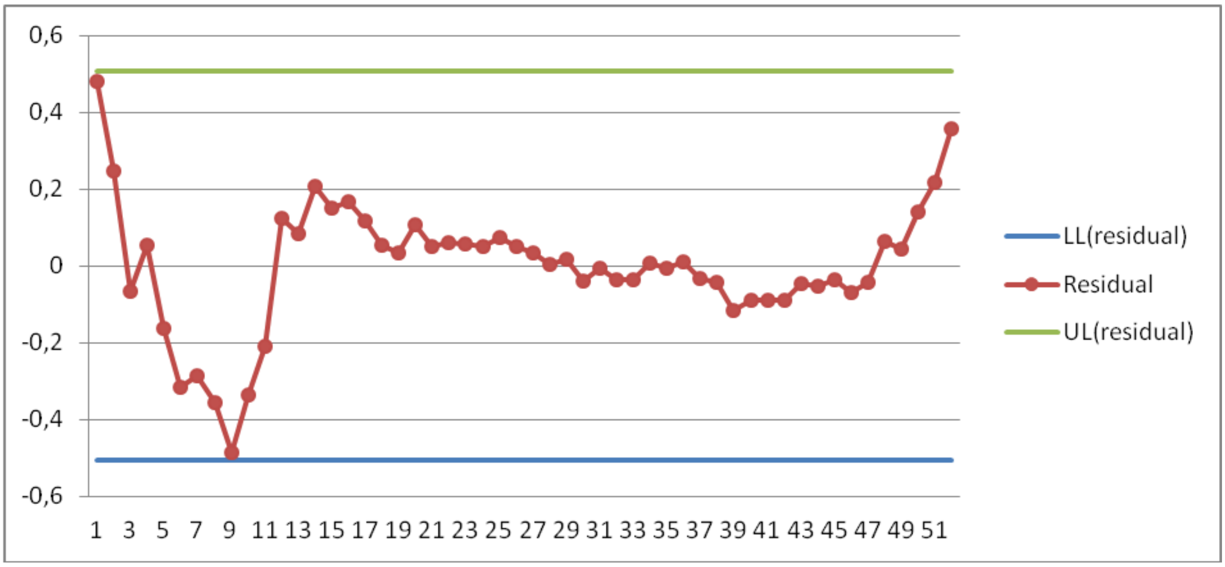


Figure 30: Residual Control Chart

IV.5.4 Statistical Process Control

Statistical process control was implemented to see trouble indicating pattern in our study. Upper and lower limits are calculated according to specifications which is given by company. The results are shown in figure 31:

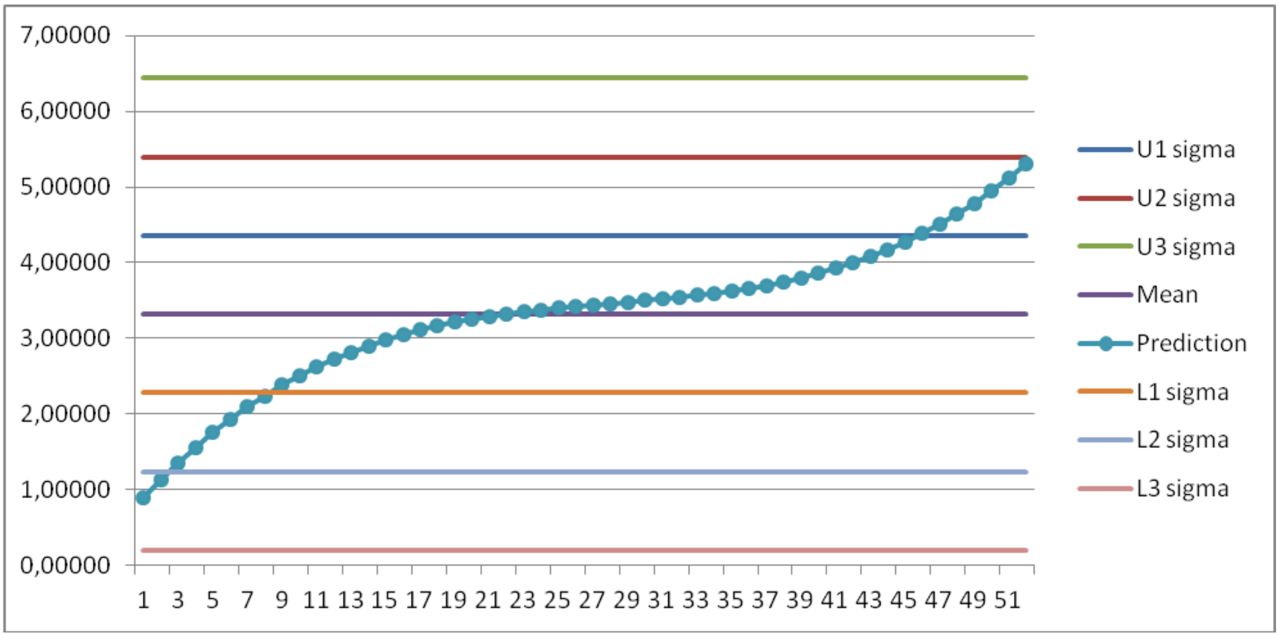


Figure 31: Outliers in Statistical Process Control

As it is seen from the graph the roughness at the beginning of the process approaches lower control limit. And at the and of the process it approaches upper control limit. Furthermore the halh of the process occurs under the mean and the rest half is on the mean, there fore trouble indicating patterns seem as a result of control chart. Nevertheless, the company's definition of faults limited with specification limits.

IV.5.5 Roughness monitoring

Fault detection scheme

The principle of the model-based fault detection approach consists of a comparison between the measured outputs of the system y_{measured} (t) and the same estimated outputs $y_{\text{estimated}}$ (t), as shown in Figure 32. The same input value u (t) is applied to the system and to its model. Without fault, these two outputs are identical (to more or less the noise and the modeling errors) and r (t), signal called residual, tends towards zero. Any occurrence of fault results in a deviation of the measured output and r (t) is different from zero. A decision logic reduced to a thresholding makes it possible to determine the occurrence of a fault. The threshold is selected to take into account noise level and modeling uncertainties.

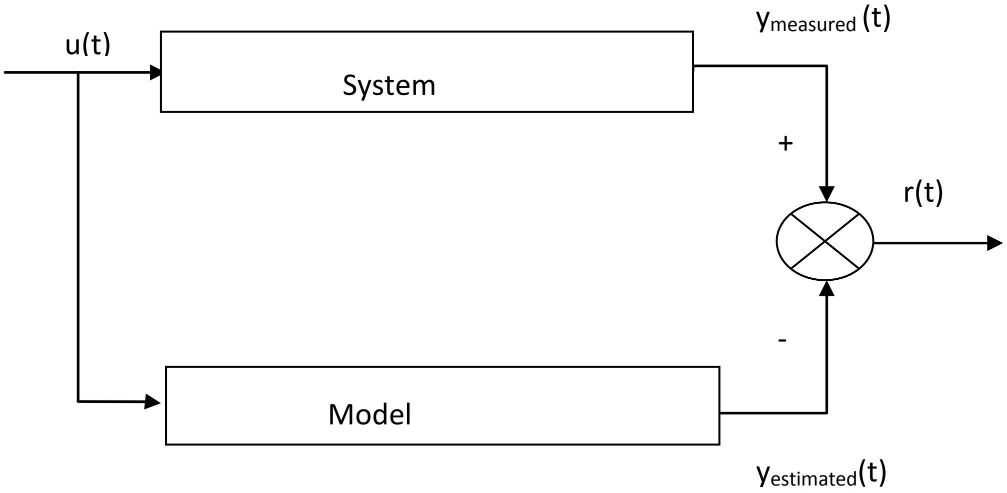


Figure 32: Residual Construction

This approach is often used for continuous system monitoring. Detection is achieved on line and in real time. The model is computed and works parallel to the system.

VI.5.6 Optimum Selection

The ANN approach has been applied as the first step in order to find optimal condition and importance of giving parameters. The methodology is found to be quite effective and utilizes fewer training and testing data. In order to understand there is significant difference between predicted and actual roughness, independent sample T test has been applied in Minitab 15. According to the test result there is no statistically significant difference between predicted and actual values with 95% confidence level.

The second step is fault monitoring with regression analysis for roughness as a function of time. Polynomial model has been built by applying least square estimator for model parameter. In this model, Standard Error of Estimation (SEE) is 0.373, R^2 is 87.9 % and adjusted R^2 is 87.1%. Therefore, it can be claimed that, this model can explain 87.9% of variation in Roughness. Moreover, according to the ANOVA table (Table 8) this model is statistically significant with 5% error rate. Durbin-Watson Statistic is equal to 0.432; it means autocorrelation does not exist. It can be claimed that this model is good enough to make prediction of faults.

CHAPTER V: CONCLUSIONS

Anomaly detection is described as a problem of finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains. Anomalies and outliers are two terms that are used most commonly in the anomaly detection framework equivalently. Many techniques that are detecting outliers are fundamentally identical however, different names are chosen by the authors.

The importance of anomaly detection is due to anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination (Kumar 2005). An anomalous MRI image may indicate presence of malignant. Anomalies in credit card transaction data could indicate credit card or identity theft (Aleskerov et al. 1997) or anomalous readings from a space craft sensor could signify a fault in some component of the space craft (Fujimaki et al. 2005).

Because of the “predictive” part of this maintenance systems, this methodology allows to built a model based fault analisis by using data mining techniques. Therefore, it is preferred to use in our study, due to prediction of future outliers**.**

The objective of this study is to prepare two stage fault monitoring system for surface roughness. The ANN approach has been applied as the first step in order to find optimal condition and importance of giving parameters. The biggest advantage of ANN is simplicity and speed of calculations. The methodology is found to be quite effective and utilizes fewer training and testing data. The experimental data and the developed system analyses showed that ANN reduces disadvantages such as time, material and economical losses to a minimum. In order to understand there is significant difference between predicted and actual roughness, independent sample T test has been applied in Minitab 15. According to the test result there is no statistically significant difference between predicted and actual values with 95% confidence level.

The second step is fault monitoring with regression analysis for roughness as a function of time. tandard error of estimation, R^2 and adjusted R^2 were used for accuracy of forecasting. All of these values and models are given in the Table 13.

Table 13: Comparison of Regression Models

Model type	R^2	Adj R^2	SEE
$Ra = 0.644 + 0.263\ t - 0.00866\ t^2 + 0.000102\ t^3$	87.90%	87.10%	0.373317
$\ln Ra = 0.129 + 0.0962\ t - 0.00282\ t^2 + 0.000029\ t^3$	87.00%	86.20%	0.135603
$\ln Ra = 0.0502 + 0.363 \ln t$	78.60%	78.20%	0.170493

According to the Table 13, the first model has the highest adjusted R^2 value. The second model has the second highest adjusted R^2 value. On the other hand the first model has the maximum standard error of estimation value. For a good forecast , it is necessary for a model to has minimum standard error and maximum adjusted R^2 values. However, the customer chose the first model for the roughness of the lathe.

Polynomial model has been built by applying least square estimator for model parameter. In this model, Standard Error of Estimation (SEE) is 0.373, R^2 is 87.9 % and adjusted R^2 is 87.1%. Therefore, it can be claimed that, this model can explain 87.9% of variation in Roughness. Moreover, according to the ANOVA table (Table 2) this model is statistically significant with 5% error rate. Durbin-Watson Statistic is equal to 0.432; it means autocorrelation does not exist. It can be claimed that this model is good enough to make prediction of faults.

As it is seen from the figure 30, the roughness at the beginning of the process approaches lower control limit. And at the and of the process it approaches upper control limit. Furthermore the halh of the process occurs under the mean and the rest half is on the mean, there fore trouble indicating patterns seem as a result of control chart. Nevertheless, the company’s definition of faults limited with specification limits.

The principle of the model-based fault detection approach consists of a comparison between the measured outputs of the system $y_{measured}$ (t) and the same estimated outputs $y_{estimated}$ (t), as shown in Figure 26. The same input value u (t) is applied to the system and to its model. Without fault, these two outputs are identical (to more or less the noise and the modeling errors) and r (t), signal called residual, tends towards zero. Any occurrence of fault results in a deviation of the measured output and r (t) is different from zero. A decision logic reduced to a thresholding makes it possible to determine the occurrence of a fault. The threshold is selected to take into account noise level and modeling uncertainties.

REFERENCES:

Abe, N., Zadrozny, B., and Langford, J. 2006. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, NY, USA, 504-509.

Abraham, B. and Box, G. E. P. 1979. Bayesian analysis of some outlier problems in time series. *Biometrika 66,* 2, 229-236.

Abraham, B. and Chuang, A. 1989. Outlier detection and time series modeling. *Technometrics 31,* 2, 241-248.

Addison, J., Wermter, S., and MacIntyre, J. 1999. Effectiveness of feature extraction in neural network architectures for novelty detection. In *Proceedings of the 9th International Conference on Artificial Neural Networks*. Vol. 2. 976-981.

Aeyels, D. 1991. On the dynamic behaviour of the novelty detector and the novelty filter. In *Analysis of Controlled Dynamical Systems- Progress in Systems and Control Theory*, B. Bonnard, B. Bride, J. Gauthier, and I. Kupka, Eds. Vol. 8. Springer, Berlin, 1-10.

Aggarwal, D. 2005. An empirical bayes approach for maintenance in manufacturing. *Advanced Manufacturing Systems.* Vol.2.IEEE Computer Society, Washington, DC, USA, 26-33.

Agarwal, D. 2006. Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems 11,* 1, 29-44.

Aggarwal, C. 2005. On abnormality detection in spuriously populated data streams. In *Proceedings of 5th SIAM Data Mining*. 80-91.

Aggarwal, C. and Yu, P. 2001. Outlier detection for high dimensional data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM Press, 37-46.

Aggarwal, C. C. and Yu, P. S. 2008. Outlier detection with uncertain data. In *SDM*. 483-493.

Agrawal, R. and Srikant, R. 1995. Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering*. IEEE Computer Society, Washington, DC, USA, 3-14.

Agyemang, M., Barker, K., and Alhajj, R. 2006. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis 10,* 6, 521-538.

Albrecht, S., Busch, J., Kloppenburg, M., Metze, F., and Tavan, P. 2000. Generalized radial basis function networks for classification and novelty detection: self-organization of optional bayesian decision. *Neural Networks 13,* 10, 1075-1093.

Aleskerov, E., Freisleben, B., and Rao, B. 1997. Cardwatch: A neural network based database mining system for credit card fraud detection. In *Proceedings of IEEE Computational Intelligence for Financial Engineering*. 220-226.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study. In *Proceedings of DARPA Broadcast News Transcription andUnderstanding Workshop*. 194-218.

Anderson, Lunt, Javitz, Tamaru, A., and Valdes, A. 1995. Detecting unusual program behav-
ior using the statistical components of NIDES. Tech. Rep. SRI{CSL{95{06, Computer Science Laboratory, SRI International. may.

Anderson, D., Frivold, T., Tamaru, A., and Valdes, A. 1994. Next-generation intrusion detection expert system (nides), software users manual, beta-update release. Tech. Rep. SRI{
CSL}95-107, Computer Science Laboratory, SRI International. may.

Ando, S. 2007. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. In *Proceedings of 7th International Conference on Data Mining*. 13-22.

Angiulli, F. and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag, 15-26.

Anscombe, F. J. and Guttman, I. 1960. Rejection of outliers. *Technometrics 2*, 2, 123-147.

Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In *Proceedings of 2nd International Conference of Knowledge Discovery and*
Data Mining. 164-169.

Augusteijn, M. and Folkert, B. 2002. Neural network classification and novelty detection.
International Journal on Remote Sensing 23, 14, 2891-2902.

Bakar, Z., Mohemad, R., Ahmad, A., and Deris, M. 2006. A comparative study for outlier detection techniques in data mining. *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, 1-6.

Baker, D., Hofmann, T., McCallum, A., and Yang, Y. 1999. A hierarchical probabilistic model for novelty detection in text. In *Proceedings of International Conference on Machine Learning*.

Barbara, D., Couto, J., Jajodia, S., and Wu, N. 2001a. Adam: a testbed for exploring the use of data mining in intrusion detection. *SIGMOD Rec. 30,* 4, 15{24.

Barbara, D., Couto, J., Jajodia, S., and Wu, N. 2001b. Detecting novel network intrusions using bayes estimators. In *Proceedings of the First SIAM International Conference on Data Mining*.

Barbara, D., Li, Y., Couto, J., Lin, J.-L., and Jajodia, S. 2003. Bootstrapping a data mining intrusion detection system. In *Proceedings of the 2003 ACM symposium on Applied computing*. ACM Press, 421-425.

Chan, P. K. and Mahoney, M. V. 2005. Modeling multiple time series for anomaly detection. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 90-97.

Chawla, N. V., Japkowicz, N., and Kotcz, A. 2004. Editorial: special issue on learning from
imbalanced data sets. *SIGKDD Explorations 6,* 1, 1-6.

Chen, D., Shao, X., Hu, B., and Su, Q. 2005. Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra. *Analytical Sciences 21,* 2, 161-167.

Chiu, A. and chee Fu, A. W. 2003. Enhancements on local outlier detection. In *Proceedings of 7th International Database Engineering and Applications Symposium*. 298-307.

Dasgupta, D. and Nino, F. 2000. A comparison of negative and positive selection algorithms in novel pattern detection. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. Vol. 1. Nashville, TN, 125-130.

Donoho, S. 2004. Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, New York, NY, USA, 420-429.

Du, W., Fang, L., and Peng, N. 2006. Lad: localization anomaly detection for wireless sensor networks. *J. Parallel Distrib. Comput. 66*, 7, 874-886.

Duda, R. O., Hart, P. E., and Stork, D. G. 2000. *Pattern Classification (2nd Edition)*. Wiley Interscience.

Edgeworth, F. Y. 1887. On discordant observations. *Philosophical Magazine 23,* 5, 364-375.

Escalante, H. J. 2005. A comparison of outlier detection algorithms for machine learning. In
Proceedings of the International Conference on Communications in Computing.

Eskin, E. 2000. Anomaly detection over noisy data using learned probability distributions. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 255-262.

Fan, W., Miller, M., Stolfo, S. J., Lee, W., and Chan, P. K. 2001. Using artificial anomalies to detect unknown and known network intrusions. In *Proceedings of the 2001 IEEE International Conference on Data Mining*. IEEE Computer Society, 123-130.

Forrest, S., Hofmeyr, S. A., Somayaji, A., and Longstaff, T. A. 1996. A sense of self for
unix processes. In *Proceedinges of the ISRSP96*. 120-128.

Forrest, S., Perelson, A. S., Allen, L., and Cherukuri, R. 1994. Self-nonself discrimination
in a computer. In *Proceedings of the 1994 IEEE Symposium on Security and Privacy*. IEEE
Computer Society, Washington, DC, USA, 202.

Fox, A. J. 1972. Outliers in time series. *Journal of the Royal Statistical Society. Series B(Methodological) 34,* 3, 350-363.

Fu, A. W.-C., Leung, O. T.-W., Keogh, E. J., and Lin, J. 2006. Finding time series discords
based on haar transform. In *Proceeding of the 2nd International Conference on Advanced Data Mining and Applications*. Springer Verlag, 31-41.

Fujimaki, R., Yairi, T., and Machida, K. 2005. An approach to spacecraft anomaly detection problem using kernel feature space. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM Press, New York, NY, USA, 401-410.

Ghosh, S. and Reilly, D. L. 2005. Fault monitoring with a neural-network *System Science*. Vol. 3. Los Alamitos, CA.

Ghoting, A., Parthasarathy, S., and Otey, M. 2006. Fast mining of distance-based outliers in high dimensional datasets. In *Proceedings of the SIAM International Conference on Data Mining*.

Gonzalez, F. A. and Dasgupta, D. 2003. Anomaly detection using real-valued negative selection. *Genetic Programming and Evolvable Machines 4,* 4, 383-403.

Grubbs, F. 1969. Procedures for detecting outlying observations in samples. *Technometrics 11,* 1, 1-21.

Gwadera, R., Atallah, M. J., and Szpankowski, W. 2005b. Reliable detection of episodes in event sequences. *Knowledge and Information Systems 7,* 4, 415-437.

Harris, T. 1993. Neural network in machine health monitoring. *Professional Engineering*.

Hartigan, J. A. and Wong, M. A. 1979. A k-means clustering algorithm. *Applied Statistics 28,* 100-108.

Hasan, K. 2005 Multilevel k-way partitioning scheme for Latte. *Journal of Parallel and Distributed Computing 48,* 1, 96-129.

Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In *Proceedings of 17th International Conference on Pattern Recognition*. Vol. 3. IEEE Computer Society, Washington, DC, USA, 430-433.

Hawkins, D. 1980. Identification of outliers. *Monographs on Applied Probability and Statistics*.

Hawkins, D. M. 1974. The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association 69,* 346 (june), 340-344.

Hodge, V. and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review 22,* 2, 85-126.

Ihler, A., Hutchins, J., and Smyth, P. 2006. Adaptive event detection with time-varying poisson processes. *Journal of Manufacturing and Engineering*. IEEE Press, New York, NY, USA, 207-216.

Jakubek, S. and Strasser, T. 2002. Fault-diagnosis using neural networks with ellipsoidal basis functions. In *Proceedings of the American Control Conference*. Vol. 5. 3846-3851.

Janakiram, D., Reddy, V., and Kumar, A. 2006. Outlier detection in wireless sensor Networks using bayesian belief networks. In *First International Conference on Communication System*
Software and Middleware. 1-6.

Jarmine, M. F., Tseng, S. S., and Su, C. M. 2001. Two-phase clustering process for outliers din machining. *Pattern Recognition Letters 22*, 6-7, 691-700.

Jin, W., Tung, A. K. H., and Han, J. 2001. Mining top-n local outliers in large databases. In
Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery
and data mining. ACM Press, 293-298.

Kadota, K., Tominaga, D., Akiyama, Y., and Takahashi, K. 2003. Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. *Chem-Bio Informatics 3,* 1, 30-45.

Kou, Y., Lu, C.-T., and Chen, D. 2006. Spatial weighted outlier detection. In Proceedings of
SIAM Conference on Data Mining.

Kumar, V. 2005. Parallel and distributed computing for cybersecurity. *Distributed Systems Online, IEEE 6,* 10.

Labib, K. and Vemuri, R. 2002. Nsom: A real-time network-based intrusion detection using self-organizing maps. *Networks and Security*.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282-289.

Lee, W., Stolfo, S. J., and Mok, K. W. 2000. Adaptive fault detection: A data mining approach. *Artificial Intelligence Review 14,* 6, 533-567.

Lee, W. and Xiang, D. 2001. Information-theoretic measures for anomaly detection. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE Computer Society, 130.

Li, M. and Vitanyi, P. M. B. 1993. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, Berlin.

Lin, S. and Brown, D. E. 2003. An outlier-based data association method for linking criminal
incidents. In *Proceedings of 3rd SIAM Data Mining Conference*.

Liu, J. P. and Weng, C. S. 1991. Detection of outlying data in bioavailability/bioequivalence
studies. *Statistics Medicine 10,* 9, 1375-89.

Lu, C.-T., Chen, D., and Kou, Y. 2003. Algorithms for spatial outlier detection. In *Proceedings of 3rd International Conference on Data Mining*. 597-600.

Ma, J. and Perkins, S. 2003a. Online novelty detection on temporal sequences. In *Proceedings*
of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining.
ACM Press, New York, NY, USA, 613-618.

Ma, J. and Perkins, S. 2003b. Time-series novelty detection using one-class support vector
machines. In *Proceedings of the International Joint Conference on Neural Networks*. Vol. 3.
1741- 1745.

Mahoney, M. V. and Chan, P. K. 2002. Learning nonstationary models of normal network traffic for detecting novel attacks. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, 376-385.

Mahoney, M. V. and Chan, P. K. 2003. Learning rules for anomaly detection of hostile network traffic. In *Proceedings of the 3rd IEEE International Conference on Data Mining*. IEEE Computer Society, 601.

Manson, G., Pierce, G., and Worden, K. 2001. On the long-term stability of normal condition
for damage detection in a composite panel. *Journal of Damage Assessment of Structures*. Cardi , UK.

Markou, M. and Singh, S. 2003a. Fault detection: a review-part 1: statistical approaches.
Signal Processing 83, 12, 2481-2497.

Markou, M. and Singh, S. 2003b. Fault detection: a review-part 2: neural network based
approaches. *Signal Processing 83,* 12, 2499-2521.

Nairac, A., Corbett-Clark, T., Ripley, R., Townsend, N., and Tarassenko, L. 1997. Choosing an appropriate model for novelty detection. In *Proceedings of the 5th IEEE International Conference on Artificial Neural Networks*. 227-232.

Noble, C. C. and Cook, D. J. 2003. Graph-based anomaly detection. *Knowledge discovery and data mining*. ACM Press, 631-636.

Odin, T. and Addison, D. 2000. Novelty detection using neural network technology. In *Proceedings of the COMADEN Conference*. Houston, TX.

Otey, M. E., Ghoting, A., and Parthasarathy, S. 2006. Fast distributed outlier detection in
mixed-attribute data sets. *Data Mining and Knowledge Discovery 12,* 2-3, 203-228.

Palshikar, G. K. 2005. Distance-based outliers in sequences. *Lecture Notes in Computer Science 3816,* 547-552.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C. 2002. Loci: Fast outlier detection using the local correlation integral. Tech. Rep. IRP-TR-02-09, Intel Research Laboratory, Pittsburgh, PA. July.

Parzen, E. 1962. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics 33,* 1065-1076.

Qin, M. and Hwang, K. 2004. Frequent episode rules for internet anomaly detection. In *Proceedings of the 3rd IEEE International Symposium on Network Computing and Applications*. IEEE Computer Society.

Ratsch, G., Mika, S., Scholkopf, B., and Muller, K.-R. 2002. Constructing boosting algorithms from svms: An application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24,* 9, 1184-1199.

Roberts, S. 1999. Novelty detection using extreme value statistics. In *Proceedings of IEEE -*
Vision, Image and Signal processing. Vol. 146. 124-129.

Roberts, S. 2002. Extreme value statistics for novelty detection in biomedical signal processing. In *Proceedings of the 1st International Conference on Advances in Medical Signal and Information Processing*. 166-172.

Rousseeuw, P. J. and Leroy, A. M. 1987. *Robust regression and outlier detection*. John Wiley
& Sons, Inc., New York, NY, USA.

Roussopoulos, N., Kelley, S., and Vincent, F. 1995. Nearest neighbor queries. In *Proceedings*
of ACM-SIGMOD International Conference on Management of Data.

Salvador, S. and Chan, P. 2003. Learning states and rules for time-series anomaly detection. Tech. Rep. CS-2003-05, Department of Computer Science, Florida Institute of Technology
Melbourne FL 32901.

Sarawagi, S., Agrawal, R., and Megiddo, N. 1998. Discovery-driven exploration of olap data
cubes. In *Proceedings of the 6th International Conference on Extending Database Technology*.

Springer-Verlag, London, UK, 168-182.

Sargor, C. 1998. Statistical anomaly detection for link-state routing protocols. In *Proceedings of the Sixth International Conference on Network Protocols*. IEEE Computer Society,Washington, DC, USA, 62.

Shekhar, S., Lu, C.-T., and Zhang, P. 2001. Detecting graph-based spatial outliers: algorithms
and applications (a summary of results). In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press, New York, NY, USA, 371- 376.

Song, Q., Hu, W., and Xie, W. 2002. Robust support vector machine with bullet hole image classification. *IEEE Transactions on Systems, Man, and Cybernetics { Part C:Applications and Reviews 32,* 4.

Song, S., Shin, D., and Yoon, E. 2001. Analysis of novelty detection properties of auto-
associators. In *Proceedings of Condition Monitoring and Diagnostic Engineering Management*. 577-584.

Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional anomaly detection. *IEEE*
Transactions on Knowledge and Data Engineering 19, 5, 631-645.

Sun, J., Qu, H., Chakrabarti, D., and Faloutsos, C. 2005. Neighborhood formation and anomaly detection in bipartite graphs. In *Proceedings of the 5th IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 418-425.

Sun, P. and Chawla, S. 2004. On local spatial outliers. In *Proceedings of 4th IEEE International Conference on Data Mining*. 209-216.

Sun, P. and Chawla, S. 2006. Slom: a new measure for local spatial outliers. *Knowledge and Information Systems 9,* 4, 412-429.

Sun, P., Chawla, S., and Arunasalam, B. 2006. Mining for outliers in sequential databases. In
In SIAM International Conference on Data Mining.

Tan, P.-N., Steinbach, M., and Kumar, V. 2005. *Introduction to Data Mining*. Addison-Wesley.

Tandon, G. and Chan, P. 2007. Weighting versus pruning in rule validation for detecting network and host anomalies. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press.

Tang, J., Chen, Z., chee Fu, A. W., and W.Cheung, D. 2002. Enhancing effectiveness of outlier detections for low density patterns. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 535-548.

Torr, P. and Murray, D. 1993. Outlier detection and motion segmentation. In *Proceedings of SPIE, Sensor Fusion VI, Paul S. Schenker; Ed.* Vol. 2059. 432-443.

Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Wei, L., Qian, W., Zhou, A., and Jin, W. 2003. Hot: Hypergraph-based outlier test for categorical data. In *Proceedings of the 7th Pacific-Asia Conference on Knowledge and Data Discovery*. 399-410.

Weigend, A. S., Mangeas, M., and Srivastava, A. N. 1995. Nonlinear gated experts for time-series - discovering regimes and avoiding overfitting. *International Journal of Neural*
Systems 6, 4, 373-399.

Weiss, G. M. and Hirsh, H. 1998. Learning to predict rare events in event sequences. In *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, R. Agrawal,P. Stolorz, and G. Piatetsky-Shapiro, Eds. AAAI Press, Menlo Park, CA, New York, NY, 359-363.

Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning*. AAAI Press, Menlo Park, California, 808-815.

Worden, K. 2007. Structural fault detection using a statistical measure. *Journal of Sound Vibration 201,* 1, 85-101.

Wu, M. and Jermaine, C. 2009. Fault detection by sampling with accuracy guarantees. *Journal of Signal Processing in Machining.* ACM, New York, NY, USA, 767-772.

Wu, N. and Zhang, J. 2003. Factor analysis based anomaly detection. In *Proceedings of IEEE*
Workshop on Information Assurance. United States Military Academy, West Point, NY, USA.

Yairi, T., Kato, Y., and Hori, K. 2001. Fault detection by mining association rules from house-keeping data. In *In Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space*.

Yang, J. and Wang, W. 2003. Cluseq: Efficient and effective sequence clustering. In *Proceedings of International Conference on Data Engineering*. 101-112.

Yel, N. and Fırat, Ü. 2010. Durum bazlı bakım sistemlerine veri madenciliği yaklaşımı. In *Üretim Araştırmaları Sempozyumu.*

Yu, J. X., Qian, W., Lu, H., and Zhou, A. 2006. Finding centric local outliers in machining. *Knowledge and Information Systems 9,* 3, 309-338.

Zeevi, A. J., Meir, R., and Adler, R. 2007. Time series prediction using mixtures of experts.
In *Advances in Neural Information Processing*. Vol. 9. MIT Press.

Zhang, J. and Wang, H. 2006. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and Information Systems 10,* 3, 333-355.

RESUME

Name-Surname	**:** Necla YEL
Birth of Date&Place	**:** 16.01.1986/ Istanbul
Education	
High School	**:** Haci Sabanci Highschool, İstanbul, 2003
Undergraduate	**:** Yildiz Technical University, Statistics, 2008
Graduate	**:** Marmara University, Industrial Engineering, 2008-...
Address	**:** Cubuklu Mah. Engurubag Cd.. No:06/03 Beykoz/ İstanbul
Telephone	**:** 536 600 9283
E-mail	**:** neclayel@gmail.com