AN ANALYSIS OF PECULIARITY ORIENTED INTERESTINGNESS MEASURES ON MEDICAL DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

CEM NURİ ALDAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

SEPTEMBER 2008

Approval of the Graduate School of Informatics

Prof. Dr. Nazife BAYKAL
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Nazife BAYKAL
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Tuğba Taşkaya TEMİZEL
Supervisor

Examining Committee Members

Prof. Dr. Yasemin YARDIMCI (METU, IS)________________

Asst. Prof. Dr. Tuğba TAŞKAYA TEMİZEL (METU, IS)________________

Assoc. Prof. Dr. Erkan MUMCUOĞLU (METU, MIN)________________

Prof. Dr. Osman SAKA (AKDENİZ UNV.)________________

Prof. Dr. Neşe YALABIK (METU, CENG)________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Cem Nuri Aldaş

Signature : __________________

ABSTRACT

AN ANALYSIS OF PECULIARITY ORIENTED INTERESTINGNESS MEASURES ON MEDICAL DATA

Aldaş, Cem Nuri
M.S., Informatics Institute, Department of Medical Informatics
Supervisor: Yrd. Doç. Dr. Tuğba Taşkaya Temizel

September 2008, 71 pages

Peculiar data are regarded as patterns which are significantly distinguishable from other records, relatively few in number and they are accepted as to be one of the most striking aspects of the interestingness concept. In clinical domain, peculiar records are probably signals for malignancy or disorder to be intervened immediately. The investigation of the rules and mechanisms which lie behind these records will be a meaningful contribution for improved clinical decision support systems.

In order to discover the most interesting records and patterns, many peculiarity oriented interestingness measures, each fulfilling a specific requirement, have been developed. In this thesis well-known peculiarity oriented interestingness measures, Local Outlier Factor (LOF), Cluster Based Local Outlier Factor (CBLOF) and Record Peculiar Factor (RPF) are compared. The insights derived from the theoretical infrastructures of the algorithms were evaluated by using experiments on synthetic and real world medical data. The results are discussed based on the interestingness perspective and some departure points for building a more developed methodology for knowledge discovery in databases are proposed.

Keywords: Peculiarity; Outlier detection; Interestingness Measures, Interestingness Analysis; Knowledge Discovery in Databases

ÖZ

OLAĞANDIŞILIK KAYNAKLI İLGİNÇLİK ÖLÇÜTLERİNİN TIBBİ VERİ ÜZERİNDE ÇÖZÜMLENMESİ

Aldaş, Cem Nuri
Yüksek Lisans, Enformatik Enstitüsü
Tez Yöneticisi: Yrd. Doç. Dr. Tuğba Taşkaya Temizel

Eylül 2008, 71 sayfa

Diğer verilerden dikkat çekici şekilde ayrılan ve göreceli olarak az sayıda bulunan örüntüler olağandışı veri olarak nitelenmekte ve bu tip veriler ilginçlik ölçümü için en çarpıcı adaylar arasında yer almaktadır. Klinik çalışmalarda olağandışı veriler hemen müdahale edilmesi gereken bir tümör ya da hastalığa ilişkin sinyal niteliği taşımaktadır. Bu verilere ilişkin kuralların ve bu verileri oluşturan düzeneklerin keşfi, daha gelişmiş tıbbi karar destek sistemlerinin oluşması için anlamlı bir katkı olacaktır.

En ilginç veri ve örüntülerin bulunabilmesi için olağandışılık kavramını temel alan pek çok ilginçlik ölçütü ortaya atılmış, ancak bu ölçütler gereksinimin ancak kısıtlı bir bölümünü karşılamıştır. Bu çalışmada LOF("Local Outlier Factor"; Yerel Aykırılık Ölçütü), CBLOF("Cluster Based Local Outlier Factor"; Küme Temelli

Yerel Aykırılık Ölçütü) ve RPF (“Record Peculiarity Factor”; Kayıt Olağandışılık Ölçütü) ilginçlik ölçütleri karşılaştırılmış, söz konusu ölçütlerin teorik altyapılarından kaynaklı öngörüler sentetik veriler ve gerçek tıbbi veri kullanılarak deneysel olarak değerlendirilmiştir. Sonuçlar ilginçlik bağlamında tartışılmış ve veri tabanlarında bilgi keşfi için daha gelişmiş bir yöntemin altyapısını oluşturacak hareket noktaları ortaya konmuştur.

Anahtar Kelimeler: Olağandışılık, Aykırı değer algılama, İlginçlik ölçütleri, İlginçlik Çözümlemesi, Veri Tabanlarında Bilgi Keşfi

To My Daughter, Yankı Ada

ACKNOWLEDGMENTS

I express sincere appreciation to my supervisor Asst. Prof. Dr. Tuğba Taşkaya Temizel for her patience, guidance and insight throughout the research. I greatly appreciate her share in every step taken in the development of the thesis. Thanks go to my domain expert Dr. Sezen Özkök for her participation to my questionnaire about meningitis and evaluation of my results. I also thank Serdar Murat Öztaner for his technical help for the development of MATLAB programs.

Encouragement of my wife, Gülsev, for my application to M.S. program after 14 years and moral assistance during the all period is also gratefully acknowledged. I also appreciate to my work fellows for coping with my frequent absences during the last two years.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ ……………………………………….. vi

DEDICATION ……………………………………………………………………… viii

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES …………………………………………………………………. xii

LIST OF FIGURES ……………………………………………………………….... xiii

CHAPTER

I. INTRODUCTION 1

I.1 Problem Statement.. 3

I.2 Summary of Results and Contributions.. 5

II. LITERATURE SURVEY...……………………………………...................... 7

II.1 Interestingness Measurement.. 7

II.2 Peculiarity Oriented Interestingness and Outlier Detection 13

II.3 Detailed Description of the Compared Measures 19

II.3.1 Peculiarity Factor (PF)……………………………………....... 19

II.3.2 Local Outlier Factor (LOF)... 21

II.3.3 Cluster Based Local Outlier Factor(CBLOF)................................ 24

II.3.4 Theoretical Evaluation of RPF, LOF and CBLOF....................... 27

II.3.5 Enhancements and Combinations of RPF, LOF and CBLOF........ 28

II.4 Interestingness, Peculiarity and Medical Data…................................. 29

II.4.1 Data Mining and Uniqueness Of Medical Data ..…………….…..... 29

II.4.2 Interestingness Measurements On Medical Data and Patterns...... 31

II.4.3 Peculiarity and Outlier Oriented Approaches On Medical Data.... 35

III. COMPARISON OF METHODS 39

III.1 Experimental Environment and General Definitions............................ 39

III.2 Experiments with Synthetic Datasets…………................................. 40

III.2.1 Normally Distributed Random Numbers……………………......... 40

III.2.2 Two Spherical Clusters Each of Which Consists of Normally Distributed Random Numbers 43

III.2.3 Two Spherical Clusters Each of Which Consists of Normally Distributed Random Numbers with Noise 46

III.2.4 Non-convex clusters and Abundant Noise 50

III.3 Experiments with Real World Medical Datasets…............................. 52

III.3.1 Wisconsin Breast Cancer Dataset ……………………..…......... 52

III.3.2 Meningoencephalidis Diagnosis Dataset ……………..…......... 56

IV. CONCLUSIONS 62

REFERENCES 67

LIST OF TABLES

Table II.1 Contingency Table for Rule Φ→Ψ……………………………………. 14

Table II.2 Characterization of Different Types of Discovered Rules ….................. 15

Table II.3 Formulae of the Top-Five Measures According to Ohsaki et.al............. 35

Table III.1 Color Scheme Definitions for Interestingness Levels..……………......... 40

Table III.2 RPF Results of WBC Dataset with $\mu+\sigma$ Threshold.……..…….......... 53

Table III.3 RPF results of WBC Dataset with $\mu+2\sigma$ Threshold…..………………. 53

Table III.4 LOF results of WBC Dataset…………………...……………..……. 54

Table III.5 CBLOF results of WBC Dataset………………...……………............ 54

Table III.6 Comparison of CBLOF and RPF Results for WBC Dataset…............... 54

Table III.7 Summarized Results of the Experiment on Meningoencephalidis Diagnosis Dataset.. 59

Table III.8 Comparison of Subjective Interestingness Points……….................... 60

Table III.9 Distances of the Measures from Subjective Evaluation…..................... 60

LIST OF FIGURES

Figure II.1 Guideline for Processing by Using Knowledge Goals ..………………… 9

Figure II.2 The Framework for OBE……………………………..………………… 16

Figure II.3 Algorithm for Finding Peculiar Records by RPF..……………..………… 21

Figure II.4 Simple 2-d Dataset Representing the Local Outlier Concept…………… 22

Figure II.5 Reachability Distance Representation………………..………………….. 23

Figure II.6 2-D Dataset Representing Cluster-Based Local Outlier Concept………... 25

Figure II.7 The FindCBLOF Algorithm..…………………………………………… 26

Figure II.8 Investigative Method for Medical Data Mining…..……………………… 31

Figure II.9 Construction of a Metric for Strength of an Hypothesis………………… 32

Figure II.10 Metric Implementation…………………………….………………...... 33

Figure II.11 The Methodology to Investigate Human Multi-perception……………… 37

Figure III.1 General View of Sample Dataset 1………………………..……………… 40

Figure III.2 Representation of RPF Values for Sample Dataset 1…..……………….. 41

Figure III.3 Representation of LOF Values for Sample Dataset 1…….…………… 41

Figure III.4 Representation of CBLOF Values for Sample Dataset 1……………..… 42

Figure III.5 General View of Dataset 2…………………………….………………… 43

Figure III.6 Representation of RPF Values for Sample Dataset 2……………….……… 44

Figure III.7 Representation of LOF Values for Sample Dataset 2…………………… 44

Figure III.8 Representation of CBLOF Values for Sample Dataset 2 with β=5………. 45

Figure III.9 Representation of CBLOF Values for Sample Dataset 2 with β=3……… 45

Figure III.10 General View of Sample Dataset 3……………………….…………….. 47

Figure III.11 Representation of RPF Values for Sample Dataset 3……………….……... 48

Figure III.12 Representation of LOF Values for Sample Dataset 3…………………. 48

Figure III.13 Representation of CBLOF Values for Sample Dataset 3 with β=5…… 49

Figure III.14 Representation of CBLOF Values for Sample Dataset 3 with β=3…..... 49

Figure III.15 General View of Sample Dataset 4……………………………..…… 50

Figure III.16 Representation of RPF Values for Sample Dataset 4…………….…… 51

Figure III.17 Representation of LOF Values for Sample Dataset 4……………….... 51

Figure III.18 Representation of CBLOF Values for Sample Dataset 4…………….. 52

Figure III.19 ROC Curve for RPF Results on WBC Dataset…………..…………… 55

Figure III.20 ROC Curve for CBLOF Results on WBC Dataset ……..…………… 55

Figure III.21 R-CBLOF Values for Sample Dataset 2 with β=3….…..…………… 57

CHAPTER I

INTRODUCTION

The process of Knowledge Discovery in Databases (KDD) is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. Despite of the fact that, this definition combines the accuracy and generalization aspects of the data mining process with the notion of differentiation and variety, the vast majority of the studies in KDD domain focus on the discovery of the correct and valid patterns. However, in practice maximizing predictive accuracy usually results in an end product of generally accepted and consequently not surprising patterns. [2] Although the accuracy and reliability of the mined patterns are important, the users also expect to obtain new knowledge about the domain, which are often categorized as peculiar and even contradictory according to his former beliefs. By using this knowledge, the curiosity of the user is triggered and this fact forces the user to examine his/her prejudgments.

The word "Interesting" in its lexicographic explanation refers to the terms curiosity, absorption and fascination [3] that are related with above mentioned properties of the desired patterns and knowledge to be derived from the KDD process. Hence, Interestingness is selected as the descriptive term in data mining world covering peculiarity, surprisingness, novelty and diversity.

Rule interestingness has become an active area of study in the fields of data mining (DM) and knowledge discovery in databases (KDD) in the last years. Many studies have been conducted concerning the formalization of rule interestingness, and concerning human substitutive evaluation of rules using formalized interestingness measures.]

Since interestingness concept is tightly connected with the belief system that the user has, some studies [4,5,6,7] come up with a suggestion that interestingness is subjective due to its nature and suppose methodologies which refer to domain experts and their knowledge, beliefs and preferences in order to evaluate the discovered patterns. These user-driven methodologies use the domain experts' declarations as a feedback to the system and try to build machine learning frameworks.

On the contrary, some other studies [8,9,10] concentrate on a data-driven approach, use the statistical properties of the data instead of domain knowledge and try to build more generic models independent of the application domains and users. The interestingness measures are mostly used at the post-processing phase of the knowledge discovery. The mined patterns are evaluated and sorted according to their scores derived from the measure(s) and the most interesting patterns are selected according to their scores.

But in the recent years interestingness measures have been also used in the pre-processing and data mining phases as well as the post processing phase. [2]

In the pre-processing step for calculating interestingness scores, first, the deviation of the record from others is measured. Second, only the tuples having interestingness values over a certain threshold are selected. This selection can also be done by using outlier detection schemes. Interestingness notion is closely related with outliers, which are defined as observations whose values lie outside the set of values considered likely according to some hypothesis (usually one based on other observations) .[11] Both of the two mentioned methods are based on an assumption that interesting data lead to interesting patterns. To use the objective interestingness measures on the pre-processing phase decreases the number of patterns to be analysed during the knowledge discovery process and its claim is to improve the quality of the patterns. But there are some problems originating from the nature of these interestingness measurement and outlier detection algorithms.

A major problem to identify peculiar records during the pre-processing step is the tendency of interestingness measurement algorithms and outlier detection schemes to mark the noise in the datasets as interesting data. Peculiarity measures and most of the outlier detection schemes are designed to work on noise-free data sets. But in the real world, the databases have significant amount of noise due to different reasons and the tendency of selecting them as interesting tuples will possibly end up with faulty, invalid or even misleading patterns. The decomposition of noise from peculiar data can be done by using domain knowledge or meta knowledge of database which come along with some subjectivity added into our objective methodology. On the other hand, a method distinguishing the noise from peculiar data candidates will preserve our consistency to be objective and practically appropriate especially for the large databases.

Another problem is the record deviation factor. The traditional distance-based methods handle the dataset on a global scale by considering the whole database with its all dimensions. An example is record peculiarity factor [12]. They proposed a measure called Peculiarity Factor (PF), in order to label the peculiar data in tables on attribute level. They used a distance measure to calculate the Record Peculiarity Factor (RPF) values for each tuple, focused on the whole dataset and tried to investigate a relatively small number of records deviating from the others. .The method assumes that the data set comes from a normal distribution and marks the points outside $\mu + \sigma$ interval as peculiar data.

On the other side, interestingness can be measured on a local scale by considering the density of the data set. An example approach is the local outlier factor (LOF) approach proposed by Breunig et al [13]. The *local outlier factor (LOF)* for each record is calculated by taking the ratio of the average density of the neighborhoods of the neighbors of the record over the density of the neighborhood of the record. The LOF approach focuses on the disadvantages of the current outlier detection schemes. In LOF, the dataset is not perceived as a whole like others such as Peculiarity Factor. The measure mainly aims to find out the records which deviate significantly from its neighborhood.

Another method which takes into account the density of the data set is Cluster Based Local Outlier Factor (CBLOF) approach of He et.al. 2002. [14] They combined the clustering methods with the locality concept of Breunig et al.[13] and defined the outliers from the point of view of clusters and identified those data points that do not lie in any large clusters as outliers. Firstly, they classified clusters as small and large. For this classification, they introduced two numeric parameters α and β, whose values can be modified according to the problem. Afterwards, CBLOF value of a record was determined by the size of its cluster, and the distance between the record and its closest large cluster (if this record lies in small cluster) or the distance between the record and the cluster it belongs to (if this record belongs to large cluster), which emphasizes the local data behavior.

I.1. PROBLEM STATEMENT

The scope of my study is to investigate the usage of interestingness measures at the pre-processing phase and to make a comparison between these three approaches that are used in order to discover the peculiar data hidden in databases. The nature of LOF, CBLOF and RPF originating from their algorithmic structure are demonstrated on synthetic datasets that have different characteristics. After making a critical review, a modification on CBLOF algorithm is proposed and four methods are evaluated on two real world datasets from medical domain.

The real world datasets are selected from medical domain due to the originality of the clinical decision making process compared with other domains. The medical decision making process cannot be regarded as a deterministic model that can be easily converted into numbers. In spite of the fact of increasing usage of IT technology and the development of the concept of evidence-based medicine, the subjective aspect of the medical experts and their experience are playing the most comprehensive role in their decision making. Hence, the medicine becomes the most appropriate domain for this study which aims to narrow the distance between the subjectivity of the medical expert and the objectivity of the KDD process by using interestingness measures as a lever.

On the other side, the clinical databases can be used as the appropriate sources of knowledge discovery process. They have been used for many years,, contain valuable data and information waiting to be converted into knowledge. But it has to be mentioned that, a traditional KDD process implemented on these databases are likely to produce a huge quantity of association rules with most of which will only tell the medical expert the basic truths that he or she knows at the beginning of the operation.

In order to cope with such situations, this study focuses on extraction of peculiar and consequently interesting rules which are low in support and high in confidence. In order to construct an infrastructure for the derivation of such rules, we need some objective measures in order to capture the outlier records which are definitely distant from the rest. of the data.. In terms of medical domain, this interesting data may refer to malignant samples, exceptional diagnoses or unpredicted formations in medical images. Rather than handling the all medical database, concentrating on these certain tuples will lead us to understand the mechanisms behind their production and obtain more useful rules at the end of our KDD study.

This study aims to compare the usage of three peculiarity oriented measures at the pre-processing phase and construct some guidelines that aid the medical experts to find out candidates which have potential to contain “interesting” knowledge.

I.2. SUMMARY OF RESULTS AND CONTRIBUTIONS

The experiments in my study are covered in two sections. In the first group of experiments, the nature of the LOF, RPF and CBLOF algorithms are demonstrated in section 3.1. The characteristics and behaviors of these measurements are investigated and discussed by means of different datasets having different distribution and properties that is thought to be reflecting variant aspects of real world. Three of these synthetic datasets that are used in this section were constructed by using MATLAB programming language and its random number producing functions. The experimental results are compared, discussed and a simple but an important modification on CBLOF is proposed.

In the second section of my study two popular medical datasets were used. Firstly, the records in Wisconsin Breast Cancer dataset were classified by using the three mentioned algorithms and the improved CBLOF. The results were discussed by using confusion matrices that divide a set of instances into true positive (TP), false positive (FP), true negative (TN), and false negative (FN) compartments. The results of this experiment were encouraging for the proposed method. Consequently for the last experiment, I needed to add some subjectivity in to my study.

In my last experiment, the Meningitis data set is used and the most interesting patterns were obtained. A questionnaire is prepared by mixing the non-interesting records with the ones that are characterized as interesting by one or more measure. The subjects are selected among the medical specialists that are working on infection oriented diseases. In the first step, two different specialist doctors made a diagnosis by only using the data in the questionnaire and expressed their certainty about the diagnosis. In the second step, the contradicted records between the real data and the diagnosis of the experts are revealed and they are requested to rank the real result by using a scale from totally non-interesting to very interesting. From these results the novelty and surprisingness of the data with respect to the experts were measured.

As a result, I found that all discussed measures do not fully satisfy the needs of a user who is investigating different types of peculiarity or novelty in a dataset. Each measure defines the peculiarity by means of its specific perception and only covers a little portion of the huge concept. The proposed measurements cannot be thought to be a magic stick on handling interestingness and some improvements such as our R-CBLOF modification should be suggested and tested.

On the other side, by making a prior analysis on the whole data and determining the characteristics of the data that can be evaluated as interesting from the users' aspect may direct the KDD process on a more convenient path.

CHAPTER II

LITERATURE SURVEY

This chapter of my study contains a literature survey on related work, that has been held on various courses of the interestingness topic. In the section II.1 the main aspects of the interesting concept and their implementations are introduced. In section II.2 the similarities and dissimilarities between the peculiarity and the interestingness are investigated. The link between outlier detection and peculiarity is attempted to be clarified. The third section of this chapter mainly focuses on the specific properties of data in medical domain. The interestingness concept in KDD research in clinical domain is evaluated in particular by using the peculiarity oriented approaches. The last section is mainly about the description and introduction of the three peculiarity based interestingness measures which forms the fundamentals of this study.

II.1 INTERESTINGNESS MEASUREMENT

In spite of the fact that interestingness has become a popular issue in KDD studies for the last years, no agreement on the definition of interestingness has been reached so far. Based on the diversity of definitions presented up to-date, interestingness is perhaps best treated as a broad concept that emphasizes peculiarity, diversity, novelty, surprisingness, conciseness, coverage, reliability, utility, and actionability.

If we examine the lexicographic definitions of the first four terms, we can observe that these are mainly originated from the notion of differentiation and variety while the others mainly focus on the generalization and usage aspects of interestingness. This variety of terms also leads to the different kinds of studies about the measurement of the interestingness of the patterns.

Subjective studies focus on both the data and the user of this data. Researches that use or develop subjective interestingness measurements are usually based on the novelty and surprisingness aspects of the concept. Since novelty term refers to the level of newness of the pattern for a person according to his/her prior knowledge and regarding the impossibility to represent the user's knowledge, novelty cannot be measured explicitly with reference to the user's knowledge. Surprisingness is a similar term with a nuance. While the key point for the novelty concept is the newness, the descriptive expression for surprisingness is the contradiction. The contradiction between the existing knowledge of the user and the pattern raises the surprisingness and due to the same reasons mentioned in novelty topic, the studies targeting to maximize surprisingness need to contain user involvement during the KDD process.

The user examines the data or founded patterns by using his/her background domain knowledge. The involvement of this knowledge to the KDD process is another discussion point. Some studies obtain this knowledge by directly interacting with the user during the KDD process, while others interpret the domain knowledge by utilizing various procedures and frameworks. [4,5,6,7].

Liu et. al in 1997 distinguished two types of knowledge as general impressions and reasonably precise knowledge. [15] They are only interested in the general impressions and designed a methodology based on the interpretation of the general impressions of domain experts in the KDD system. Besides, they propose some algorithms in order to analyze the founded patterns against the general impressions.

Sahar uses a different methodology to obtain the interesting patterns.[6] Instead of imitating the human interest, he uses the domain expert in order to eliminate non-interesting patterns recursively. The user's classification is not only used for pruning the uninteresting patterns, but also starts the construction of the domain knowledge base.

On the other hand, other studies regarding the user of the data as a subject, handle the notion of interestingness conceptually and try to define this term by a cognitive approach. Ram defined the term "Knowledge Goal" as the goals of a reasoner to acquire some piece of knowledge required for a reasoning task, as focusing criteria for inference control. [16] He connects this term with human interest and presents a theory of interestingness departed from this connection point. In his theory, he classifies the knowledge goals according to the type of understanding task that they arise from as text, memory, explanation and relevance goals. He also defines a control structure to use knowledge goals as a guideline for processing as shown in Figure II.1.

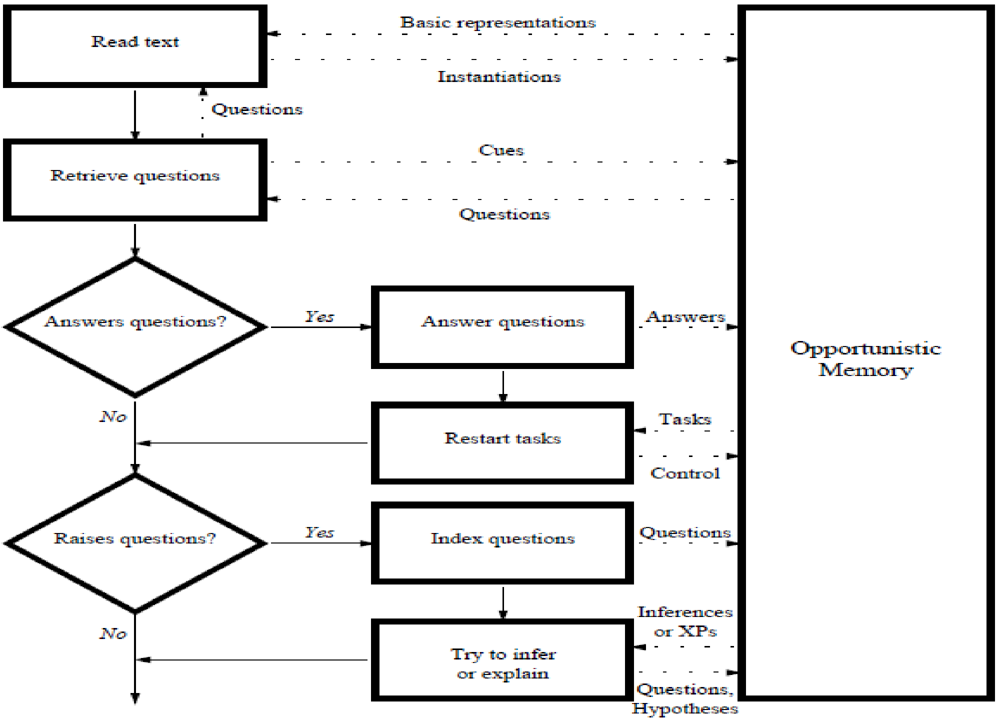


Figure II.1. Guideline for Processing by Using Knowledge Goals.[16]

The figure II.1 symbolizes that the interestingness notion is very related with questions asked. A fact is interesting if it satisfies a knowledge goal in memory, or if it gives rise to new knowledge goals. These correspond to the two diamonds in Figure II.1. The aim of these diamonds is to determine the facts in which the understander should focus on. The fact to be mentioned here is the unbalanced priority or weight of all questions and answers. The notion of interestingness is closely related with these weights and priorities. Hence, a set of heuristics in order to fulfill the requirement of determination of the priorities of the knowledge goals is also recommended by Ram. He called them interestingness heuristics, not measures, because he defines interestingness as a product of attention focusing: “Interestingness is a guess one thinks one might learn from paying attention to a fact or a question”. From this aspect, he does not bring up any objective interestingness measure; instead he recommends a question-driven subjective methodology to produce interesting patterns.

Silbershatz and Tuzhilin, also worked on the cognitive characteristics of interestingness and defined interestingness concept by means of two subjective arguments: unexpectedness and actionability. [4,5] In their point of view, a pattern is actionable and consequently interesting when the user can do something about it; in other words when the user can convert this knowledge to a reaction to maximize his or her advantage. On the other hand, they also admit the importance of unexpectedness while they contradict the users' expectations, which in fact depend on their settled system of beliefs. Based on these two concepts, they classify the patterns as unexpected/actionable, expected/actionable, expected/nonactionable and unexpected/nonactionable. Even they accept that some unexpected/nonactionable patterns might be interesting, they do not avoid arguing that unexpectedness is a good approximation for actionability and vice versa. They prefer to address actionability through unexpectedness and proposed a method to define interestingness of a pattern as a measure of its unexpectedness.

Silbershatz and Tuzhilin, emphasizes on the relation of the beliefs and the unexpectedness of a pattern. They classify the beliefs into hard and soft beliefs according to their subjectivity and beliefs, suggest a methodology for both types of beliefs, hard and soft in order to measure the level of unexpectedness.

Silbershatz and Tuzhilin, in their work refer to the early studies of Piatetsky-Shapiro and Matheus [17], in which they used a discovery system KEFIR in order to study subjective measures of interestingness of the healthcare insurance claims. Their aim was to uncover key findings which were the statements about the most important deviations from the norms for different attributes. KEFIR defines the interestingness in terms of the estimated benefits of taking corrective actions that are done in order to restore the deviation to its norm. The corrective actions in this context were specified by a domain expert. In this study, the scope is deliberately narrowed for the use of a specific domain by pre-classifying the patterns into a finite set which are to be assigned a corrective action. Silbershatz and Tuzhilin generalized the practical domain-specific solution of Piatetsky-Shapiro and Matheus, interchanged their term of the estimated benefits of taking corrective actions to accountability and made a mathematical definition by exposing the three main propositions.

PROPOSITION 1. *An interestingness of a pattern relative to a belief system B does not change if any belief(s) in B are replaced by its (their) complements.*

PROPOSITION 2. *Let* α *be a belief, such that* $0.5 < d(\alpha|\varepsilon < 1$. *Let p be a pattern confirming belief* α ($\alpha|$=p) *such that* $0.5 < d(\alpha| !p,\varepsilon)$ *and* $d(\alpha| p,\varepsilon) < 1$ *Then,* $I(p, \alpha,\varepsilon) < I(! p, \alpha, \varepsilon)$

PROPOSITION 3. *Let D be the old (historical) data stored in a database, and* ΔD *be the new data that was just added to D. Let B be a belief system about the data stored in the database. If there is a belief* α *in B such that* $d(\alpha| \Delta D, D) \neq d(\alpha | D)$, *then there exists a pattern p in*

ΔD such that $I(p,B,D) \neq 0$.

By using these propositions, especially the last, they form a theoretical justification for a belief-driven discovery scheme, and they applied this scheme in [18].

Conceptual studies such as above usually concentrate on the semantics and explanations of the derived rules, and they try to draw the borders of interestingness from their aspect. The common characteristics of the above mentioned papers are to handle the interestingness issue by means of utility and actionability. The point of difference is the fact that Ram offers a Question-Driven approach complying with his concept Knowledge Goal, while Silbershatz and Tuzhilin suggests a Belief-Driven approach originating from their emphasis on unexpectedness as a source of actionability.

These conceptual approaches reflect the similar properties such as involving domain knowledge from user and are usually considered in the group of subjective studies. But in the classical subjective studies the domain knowledge is only about the data itself and even represented in a similar format with the discovered pattern. In the semantic based approach, the user's goals and expectations reflected to the scene by using some utility function. The domain expert in this approach, is not only the person to view and evaluate the data and patterns, besides he or she is regarded to be in a strategic position which is responsible for the utilization and application of the results derived from KDD process. This kind of studies are called "semantic" measures of interestingness in the literature.

The last but not the least type of interestingness measures is produced by the "objective" studies which are held by using the data and/or the constraints as the only source. These studies use the theories of statistics, probability and information theory. Ignoring some exceptions, that can be said that they concentrate on the conciseness, generality, reliability, peculiarity and diversity aspects on which all depend only on the data and patterns.

The objective studies about the interestingness firstly focused on the generality concept which measures the comprehensiveness of a pattern. The main assumption motivating these kind of studies was the acceptance of the fact that if a pattern characterizes more information in a dataset, then it is regarded as interesting. Agrawal and Srikant developed the Apriori algorithm [19] which have been used for finding frequent itemsets by featuring "support" and confidence as two main measures which are still broadly used in KDD world.

Studies that target to maximize conciseness of the patterns used some concepts such as monotonicity and confidence invariance in order to end up with patterns or pattern set containing less attributes. Padmanabhan and Tuzhilin combined the subjective measure of unexpectedness with conciseness even by not addressing the latter term directly and proposed new methods for discovering a minimal set of unexpected patterns that discover orders of magnitude fewer patterns and yet retain most of the truly interesting ones. [9] In their method they used the monotonicity property and suggest an algorithm MinZoominUR in order to decrease the number of association rules by keeping the interestingness threshold. Padmanabhan and Tuzhilin work on decreasing the number of the resultant rules, whereas Bastide et. al. focus on decreasing the complexity of the derived rules. [8] Their main objective is to lead to the most informative rules which have a minimal antecedent (left-hand side) and a maximal consequent (right-hand side). In order to realize this objective, they use the semantic for association rules based on the Galois connection and they recommend an algorithm called “Close” in order to extract frequent closed itemsets and their generators.

Conciseness and generality often coincide with each other . Concise patterns tend to have more coverage. Webb and Brain, presented two hypotheses about the relationship between conciseness and generality. [20] Their first hypothesis was that the accuracy on unseen data of the more general rule would be more likely to be closer to the accuracy on unseen data of a default rule for the class than would the accuracy on unseen data of the more specific rule. The second hypothesis was that the accuracy on previously unseen data of the more specific rule would be more likely to be closer to the accuracy of the rules on the training data than would the accuracy of the more general rule on unseen data.

Reliability can be defined as the quantity and percentage of the applicable cases in which the relationship declared by the pattern occurs. The studies on the reliability concept usually prefer the ranking and categorizing the existing measures more than implementing a new measurement. For example Tan et. al in 2002 made a comparative study of the twenty-one different interestingness measures from the aspect of reliability. [21] Ohsaki et.al. in 2004 conducted a similar study for medical data set and compared the results of the interestingness measures by real human interest. [22]

Diversity and peculiarity oriented studies differ from the above mentioned studies with one important assumption. The conciseness and generality, and partially reliability concepts usually focus on the generalization and usage aspects of interestingness. The researches on these concepts usually originate from the idea that frequent patterns lead to interestingness and they concentrate on the majority of the data during the KDD process. On the other side, peculiarity and diversity are based on the notion of differentiation and variety and the studies referring to these terms usually focus ona small portion of data setthat are patterns which convey clues about the marginal behaviors. The diversity is usually defined as a common factor for measuring the interestingness of summaries and the summaries are usually examined with respect to their probability distributions . Anomalies from the normal or uniform distribution are supposed to contain more interestingness.

The peculiarity concept, which is usually regarded as the distant data and patterns from the other data and patterns based on any distance measure is the main research area of this study. The related work on peculiarity oriented interestingness approaches are explained in the following section.

II.2 PECULIARITY ORIENTED INTERESTINGNESS AND OUTLIER DETECTION

Aforementioned, peculiarity is usually defined as the distant data and patterns relatively far from the other data and patterns for any distance measure. For an objective approach, peculiarity can be regarded as the most appropriate measure, because the idea of peculiarity is based on some distance measure between the data or discovered patterns. Oxford dictionary defines peculiarity as distinguished in nature, character, or attributes from others; unlike others [11]. In contrast with the studies based on generality and conciseness, the studies regarding peculiarity come up with an assumption that the high frequent patterns are probably well-known facts (common knowledge) and they concentrate on low frequency patterns. Hence determining the peculiar data, which are relatively few in number and significantly different from the rest of the data become more important in order to find peculiar patterns. [23,24]. This is also an intersection point of peculiarity and outlier concepts.

The outlier data is defined by Hawkins, as an observation that deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism. [25] On the contrary of the spirit in this definition, in most data mining and knowledge discovery applications the outliers are usually classified as noise and the only generation mechanism of their appearance was reduced to measurement or implementation errors. Many data preprocessing techniques [26,27,28] have been developed in order to minimize the effect of outliers on the data indicators or to eliminate them. But in the view of peculiarity, the outliers are regarded as valuable and important patterns in a data set . Some studies on outlier detection used this theoretical principle and laid a foundation for some important applications such as credit card fraud detection, discovering criminal behaviors in e-commerce, discovering computer intrusion, weather prediction and customer segmentation.[29,30] In fact, the peculiar data which are in form of records are called as outliers [6]. Mostly, it is used interchangeably with outliers [7].

Zhong et. al. introduced the peculiarity rules as a new class of rules, which can be discovered from a relatively low number of peculiar data by searching the relevance among them. [31] In their study, they differentiated peculiar rules from association and exception rules. In their point of view, the characteristics of a rule $\Phi \to \Psi$ can be summarized by the contingency table given as Table II.1.

Table II.1. Contingency Table for Rule $\Phi \to \Psi$

	Ψ	$\neg \Psi$	Totals
Φ	a	b	**a + b**
$\neg \Phi$	c	d	**c + d**
Totals	**a + c**	**b + d**	**a+b+c+d=n**

The generality of Φ is defined by $G(\Phi) = (a+b) / n$. In other words if $G(\Phi) = \alpha$, then the $100\alpha\%$ of objects in universal set U satisfy Φ. The absolute support of Ψ provided by Φ, which shows the degree to which Φ implies Ψ is $AS(\Phi \to \Psi) = a / (a+b)$. The change of support (CS) of Ψ provided by Φ is the difference between the $AS(\Phi \to \Psi)$ and $G(\Phi)$. From these equations that can be argued that the high support and confidence values of the will not always lead to a significant change in support(CS) , and may not reflect the true association between Φ and Ψ. Hence the algorithms that were developed to find association rules with the support and confidence parameters may end up with invalid or trivial rules and also may fail in identifying interesting patterns.

An exception rule, which is defined as a deviational pattern to a common sense, exhibits unexpectedness and is sometimes extremely useful. [32] If an association rule is defined by the $\Phi \to \psi$ expression and if the $\Phi \Lambda \Phi' \to \neg\psi$ is associated with the main rule as an extension, then Φ' is called as the condition of exception for rule $\Phi \to \psi$.

Zhong et.al. introduced the term peculiar rules and suggested a methodology to discover them. According to them, in order to derive peculiar rules, attention should be paid to records that have attributes quite different from the other records. After completing this phase, which can be called as peculiar data identification, the peculiar rules which have low support and high confidence is searched. According to the approach of Zhong et. al, although peculiar rules have some common properties with exception rules such as being expressed in confidence and support, there is an important semantic difference. Peculiar rules are not defined as complements of any association rules, on the other hand the definition of exception rules lies beyond a contradiction with an a-priori association rule.

These three kinds of rules also emphasize the difference between the three aspects of interestingness. Association rules with their high support and confidence point to general facts and represents the coverage aspect of interestingness. As mentioned before, surprisingness is strongly coupled with contradiction with the users' current beliefs. Exception rules find out the patterns in a dataset which contradict with association rules which can be interpreted as general beliefs. Hence exception rules can be regarded as models representing surprisingness. On the other hand, peculiar rules do not have any claim about contradiction; instead they focus on the small portions of the datasets and try to discover similar behaviors of such records.

Based on the above discussion, Table II.2 figures out the difference between association, exception and peculiar rules in a more systematic representation.

Table II.2. Characterization of Different Types of Discovered Rules [31]

Rule	**Form**	**G (support)**	**AS (confidence)**	**CS**	**Semantics**
Association rule	$\Phi \rightarrow \psi$	High	High	Unknown	Common-sense
Exception rule	$\Phi \rightarrow \psi$ $\Phi \Lambda \Phi' \rightarrow \neg \psi$	High Low	High High	Unknown High	Exception
Peculiar rule	$\Phi \rightarrow \psi$	Low	High	High	Common-sense

In order to realize principles of peculiar rule mining as defined above, several peculiar data or outlier detection methodologies have been suggested.

Some studies are concentrated on learning the model represented by the dataset using statistical models [33], neural networks [34] and support vector machines.[35] Such studies focus on the revelation of the basic characteristics of the model and aim to discover the outliers that do not match with these characteristics.Yamanishi et.al developed an outlier detection engine SmartSifter which identifies outliers in an online process through the on-line supervised learning of a probabilistic model of the information source.[36]

In contrast with previous examples, Zhu et.al. [37] developed the Outlier-by-Example(OBE) system which allows users to give some examples of outliers and tries to discover the outlierness model on the given examples. OBE performs outlier detection in three stages: In the feature extraction step, all objects are mapped into the MDEF-based [38] feature space, where the MDEF plots of objects capturing the degree of "outlier-ness," as well as the scales at which the "outlier-ness" represented by vectors appears, . After this phase, the examples provided by the users are augmented by adding artificial data, whose feature values (i.e., MDEF values) are greater than those of the given outlier examples. In the third and the last step of OBE, Zhu et.al. uses support vector machines (SVM) in order to learn the interestingness with respect to user. Their framework for outlier detection can be explained by Figure II.2.

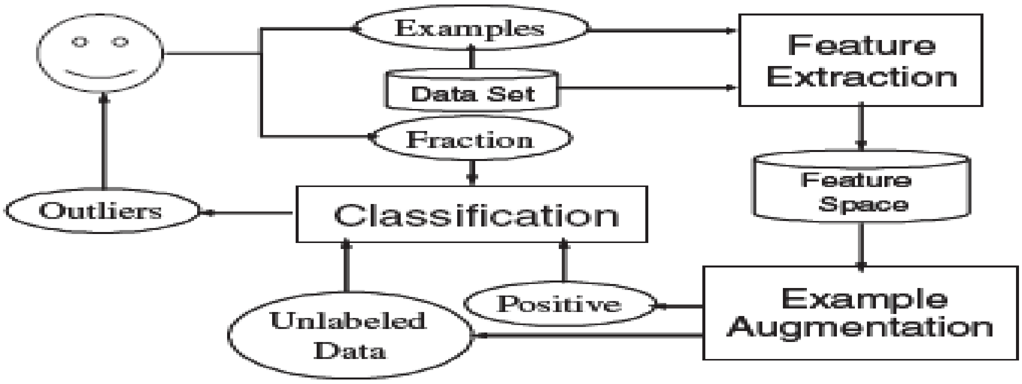


Figure II.2. The Framework for OBE [37]

Some studies use statistical approaches and use discordance tests depending on data distribution, distribution parameters and number of expected outliers. One of the earliest studies about outlier detection is the Grubb's Test. [39] which assumes that the univariate data is and comes from the normal distribution and detects one outlier at a time by using the his test statistic $G = \frac{|x-\mu|}{\sigma}$ where x is the value of the attribute and μ and σ are the mean and the standard deviation respectively.

A different approach is based on the assumption that the dataset consists of samples from a mixture model containing M and A, where M refers the majority and A refers to anomalous distributions. By using this model and accepting the probability of any element to be originated from A as λ, Eskin [40] defined the generative distribution(GD) of the above mentioned mixture model as $GD = (1 - \lambda)$ M+ λ A. Detecting anomalies in the method is equivalent to which determining the source of a given element as A or M. At the beginning all samples are supposed to be in M distribution and the set of anomalies is an empty set. He defines the log likelihood function for the whole distribution as given below.

$$LL_t(GD) = |M_t| \log(1-\lambda) + \sum_{xi \in M_t} \log(P_{M_t}(X_i)) + |A_t| \log(\lambda) + \sum_{Xi \in At} \log(P_{A_t}(X_i))$$

To measure the likeliness of an element x_i, he compares the difference in the log likelihood of the distribution if the element is removed from the majority distribution M_{t-1} and included in the anomalous distribution A_{t-1}. If this difference (LL_t – LL_{t-1}) is over a threshold c, then the element is declared to be an anomaly or outlier and permanently move it from M to A. Otherwise it remains in the majority distribution. As a result of a repetition of this process for every element a partition of the dataset into a set of normal and outlier subsets.

The main problem with the methods that use statistical approaches is that they assume that the underlying data distribution is known a prior. However, for many applications, it is an impractical assumption and the cost for fitting data with standard distribution is significantly considerable. In addition ,most of the distributions used are univariate and this restriction makes them unsuitable for multidimensional datasets.

A depth-based solution is recommended by Ruts and Rousseeuw for outlier mining. [41] Data objects are organized in convex hull layers in data space according to depth, and outliers are expected to be found within data objects with low depth values. This method avoids the necessity of distribution fitting, and conceptually allows multidimensional data objects to be processed. In theory, depth-based methods could work in high dimensional data space. However, due to relying on the computation of k–d convex hulls, these techniques have a lower bound complexity of $\Omega(N^{k/2})$, where N is number of data objects and k is the dimensionality of the dataset. This makes these techniques infeasible for high dimensional large datasets.

Unlike the model-based methods mentioned above, the distance-based method does not focus on the model itself. Instead a distance function, such as Euclidean, Mahalanobis or Hamming is determined, the distance measure is used in order to define neighborhoods by using a fixed radius threshold or a fixed number of neighbors and finally distance-based outliers are discovered with respect to these neighborhoods.

Knorr and Ng [23] made a formal definition of the DB-outliers : an object O in a dataset D is a DB(p;dmin) outlier if at least fraction *p* of the objects in T lies greater than distance dmin from O. Afterwards they implemented two algorithms (index-based and nested loop) both having a complexity of $O(k\ N^2)$ and a cell-based algorithm which is linear with respect to N but exponential with respect to k where *k* is the dimensionality and *N* is the number of objects in the dataset. They showed that the cell-based algorithm outperforms the other algorithms where k ≤ 4. In addition to this, they also suggested an improvement on this cell-based algorithm that guarantees at most three passes over a dataset for disk resident datasets. According to Knorr and Ng, the usage of the nested-loop algorithm is more convenient than others when k > 4.

Ramaswamy et al. extended the distance-based outlier concept to be based on the distance of a point to its k nearest neighbors. [42] After ranking the points with respect to its k nearest neighbors, the top-k points were identified as outliers. Moreover , they developed a highly efficient *partition-based* algorithm for mining outliers. This algorithm first partitions the input data set into disjoint subsets, and then prunes entire partitions as soon as it is determined that they cannot contain outliers.

In a more recent study of Bay and Schwabacher modification was suggested on the nested-loop algorithm which has almost linear time performance on many large datasets by implementing a simple pruning rule. [43] Their main contribution is keeping track of the found closest neighbors and assigning a score value origination from any monotonically decreasing function of the nearest neighbor distances. When an example's closest neighbors achieve a score lower than the cutoff, they removed the example because it has already lost the chance to become an outlier. While processing more examples, the algorithm finds more extreme outliers and the cutoff increases along with pruning efficiency.

In this section, the peculiarity oriented interestingness are listed and the similarities and main differences in various approaches were discussed. Roughly the peculiar data or pattern detection methodologies can be classified in two groups. First group focuses on the model which produces the normal or outlier data, in some cases both. After understanding the mechanism by using different methodologies, such as machine learning on a limited amount of data and tries to apply this knowledge to the rest of the dataset. The second group concentrates on the obtained data and somewhat ignores the mechanism that produces them. The peculiarities in the data are investigated through some distance, density or clustering based measure and usually these methods come up with a score metric like [43]. This score metric not only overruns our objective of discovering outliers, it also permits the user to make a comparison between the outliers that have already been discovered. All three methods of peculiar data detection, namely Record Based Peculiarity Factor(RPF), Local Outlier Factor (LOF) and Cluster Based Local Outlier Factor (CBLOF) that are in the scope of this study should be considered as different aspects of this approach and they will be explained in detail in the next section.

II.3 DETAILED DESCRIPTION OF THE COMPARED MEASURES

Although a general scope is outlined for the interestingness measurement studies and the relation between peculiarity and outlier detection is expressed until this point, the main scope of this study is to analyze the following three peculiar data or outlier detection schemes. In this section their theoretical background, main properties and the similarities and dissimilarities with each other will be explained.

II.3.1. Peculiarity Factor (PF)

Peculiarity measures often do have a score factor which indicates how much a record deviates from others. An example of this approach is the implementation of the score function by Bay and Schwabacher. [43] Since their usage objective is not to rank the data according to the score factor, but only to make an early pruning, they did not focus on the details of the score function. They confined themselves by only mentioning its significant property of monotonicity. On the other hand, some other studies which will be mentioned below use score functions in order to discover and rank peculiar data.

Zhong et. al. proposed a measure called Peculiarity Factor (PF), in order to label the peculiar data in tables on attribute level. [31] They used a distance measure to calculate the Peculiarity Factor (PF) values as described below, focused on the whole dataset and combined them with density based approaches.

Let a relation A be with attributes $a_1, a_2, \dots , a_k$ and x_{ij} be the value of a_j of the i^{th} record and N be the number of records. The peculiarity of x_{ij} is calculated as :

$$PF\ (x_{ij}) = \sum_{r=1}^{N} distance\ \ (x_{ij}, x_{rj})^{\beta}$$

where *distance* denotes the conceptual distance, β is a parameter which can be adjusted by user, and $\beta = 0.5$ is used as default. This formula evaluates whether x_{ij} occurs in relatively low number and is very different from other data x_{rj} by calculating the sum of square root of the conceptual distance between x_{ij} and x_{rj}.The two significant characteristics of this formulization are its capability of being used independent of data type (continuous or categorical) and its elasticity to the supported background knowledge. The latter property can be implemented by adjusting the conceptual distance values.

In order to lead to a record based formulation from PF, they came up with the definition of Record Peculiarity Factor (RPF). [Zhong et al 2004]

Let X denote a record set in a relation A (with attributes $a_1, a_2, \ldots, a_k$,) that is X = {X_1, X_2, ….., X_N}. A record X_i is represented by {X_{I1}, X_{I2},…..,X_{IJ}, ……, X_{IN}} where x_{ij} denotes the value of the X_i on attribute a_j. The peculiarity of X_i can be evaluated by :

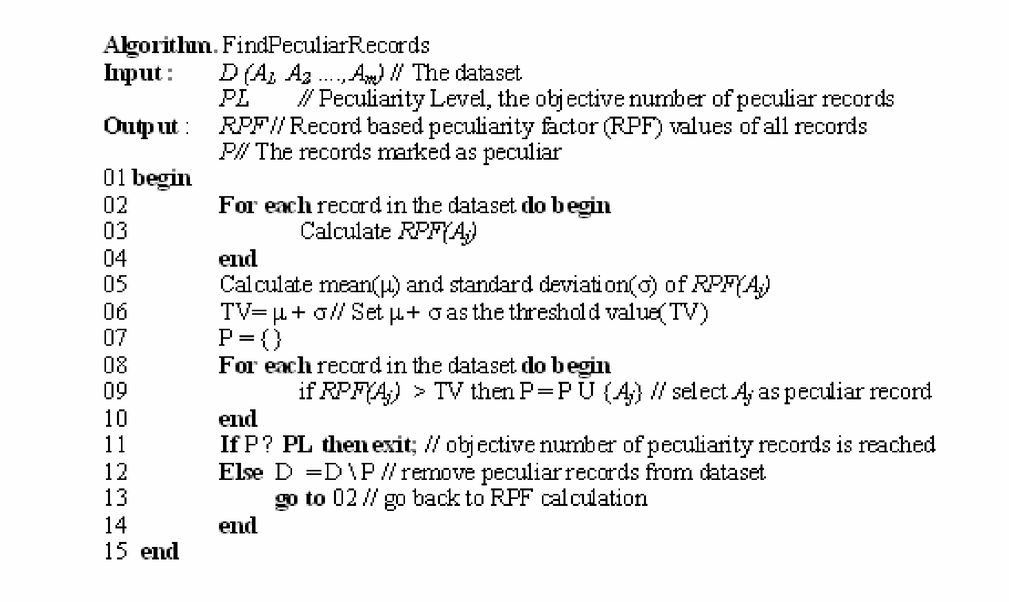
$$\mathrm{RPF(Xi)} = \sum_{m=1}^{N} \sqrt{\sum_{j=1}^{k} \alpha_j (\mathrm{PF}(x_{ij}) - \mathrm{PF}(x_{mj}))^2}$$

where α_j is the weight of an attribute which depends on the knowledge provided by a user. $\alpha_j = 1$ is used as default. By using the α_j factor users can modify the peculiarity calculation according to their specific needs. As a general guideline the key attributes which do not carry any information and just used for the identification of the record must be excluded from above formulation by assigning α_j value to 0.

Originating from the nature of the distance-based metrics, standardization must be done in order to prevent the effects of numeric differentiation between the attributes. This standardization will be easily applied to continuous variables, but it may cause some problems for other kinds of variables. Instead of standardizing the variable values, they suggest to normalize the Peculiarity Factor as follows.

$PF\ (X_{ij}) = PF(X_{ij}) / max(PF(X_{rj}))$ while r represents all records from 1 to N.

To realize the RPF calculation in order to derive peculiar records, Zhong et. al. exhibit an algorithm given in Figure II.3. The algorithm is based on the RPF values derived from normalized PF values and uses a threshold which is a linear combination of the mean and standard deviation of the Record based Peculiary Factors. This can be formulized as TV= $\mu + (\alpha * \sigma)$ where μ is the mean, σ is the standard deviation values and α is the adjustment coefficient determined by the user in order to control the number of peculiar data.



```
Algorithm. FindPeculiarRecords
Input :    $D(A_1, A_2 ...., A_m)$ // The dataset
           PL      // Peculiarity Level, the objective number of peculiar records
Output :   RPF // Record based peculiarity factor (RPF) values of all records
           P// The records marked as peculiar
01 begin
02         For each record in the dataset do begin
03                 Calculate $RPF(A_j)$
04         end
05         Calculate mean(μ) and standard deviation(σ) of $RPF(A_j)$
06         TV= μ + σ // Set μ + σ as the threshold value(TV)
07         P = { }
08         For each record in the dataset do begin
09                 if $RPF(A_j)$ > TV then P = P U {$A_j$} // select $A_j$ as peculiar record
10         end
11         If P ? PL then exit; // objective number of peculiarity records is reached
12         Else D =D \ P // remove peculiar records from dataset
13               go to 02 // go back to RPF calculation
14         end
15 end
```

Figure II.3. Algorithm for Finding Peculiar Records by RPF [24]

II.3.2. Local Outlier Factor (LOF)

As mentioned in section II.2 Knorr and Ng formalized the outlier notion of Hawkins and constructed a distance-based outlier detection methodology. Breunig et.al. criticized this formulization as resulting in similar type of outliers. [13] According to Breunig et. al, Knorr's method [23] takes a global view of the dataset and the outliers which are discovered by this method can be viewed as "global" outliers.

To illustrate this situation, they use a simple two-dimensional dataset shown in Figure II.4, which has 502 objects. In this dataset there are 400 objects in the first cluster *C*1, 100 objects in the cluster *C*2, and two additional objects *o*1 and *o*2. In this example the framework of distance-based outliers can only manage to point out *o*1 as an outlier. If for every object q in *C*1, the distance between *q* and its nearest neighbor is greater than the distance between *o*2 and *C*2 (i.e., *distance*(*o*2,*C*2)), that can be showed that there is no appropriate value of *pct* and *dmin* such that *o*2 is a DB(*pct*,*dmin*)-outlier but the objects in *C*1 are not.

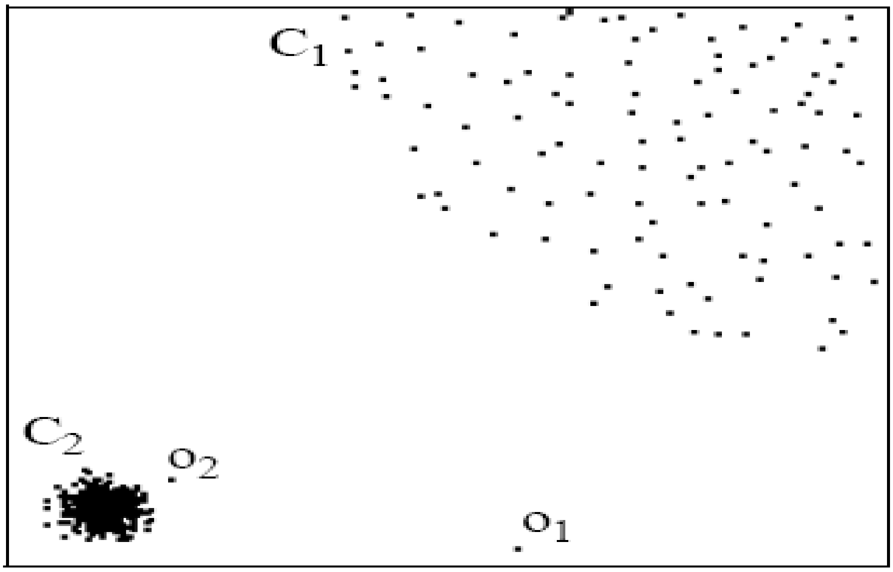


Figure II.4. Simple 2-d Dataset Representing the Local Outlier Concept [13]

However due to the spirit of the definition of Hawkins and *C*2 being denser than *C*1 both o1 and o2 can be called outliers, whereas objects in *C1* and *C2* should not be. The above example shows that the global view taken by the DB(pct, dmin)-outliers end up with adequate results only under certain conditions. The approach of Knorr and Ng has serious problems when clusters of different densities exist. Breunig et. al. represented the "local outliers" concept in order to overcome this type of problems.

They firstly developed a formal description of local outliers depending on four preliminary definitions most of which are borrowed from density based clustering approach.

1. k-distance of an object p: For any positive integer *k* and dataset D, the k-distance of object *p*, denoted as *k-distance(p),* is defined as the distance *distance(p,o)* between *p* and an object *o* ε D such that:
(i) for at least *k* objects *o* ε D \ {p} it holds that *distance(p,o') <= distance(p,o),* and
(ii) for at most *k-1* objects *o'* ε D \ {p} it holds that *distance(p,o') < distance(p,o).*

2. *k*-distance neighborhood of an object *p*) : Given the k-distance of *p*, the k-distance neighborhood of *p* contains every object whose distance from *p* is not greater than the k-distance, i.e. $N_{k\text{-}distance(p)}(p)$ *= { q* ε *D\{p} | distance(p, q) <= k-distance(p) }.* Objects *q* are called the k-nearest neighbors of *p.*

3. reachability distance of an object *p* w.r.t. object *o*: Let *k* be a natural number. The

reachability distance of object *p* with respect to object *o* is defined as: *reach-dist$_k$(p, o) = max {k-distance(o), distance(p, o) }*

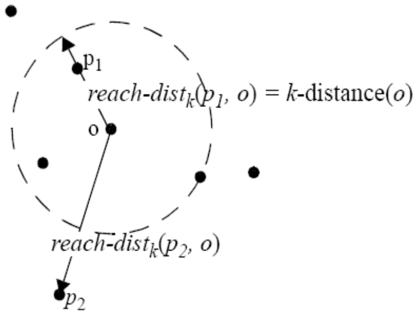


Figure II.5. Reachability Distance Representation [13]

Figure II.5 illustrates the idea of reachability distance with $k = 4$. That can be seen from the figure that object *p2* is far away from *o*, then the reachability distance between the *p2* and *o* is simply their actual distance. On the other side, *p*1 and o are "sufficiently" close, so the actual distance will be replaced by the *k*-distance of *o*. Hence the statistical fluctuations of distance(p,o) for the closest objects will be reduced and this reduction can be controlled by the parameter k.

4. local reachability density of an object p : The *local reachability density(lrd)* of *p* is defined as ;

$$lrd_{MinPts}(p) = 1 / \left(\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p, o)}{|N_{MinPts}(p)|} \right)$$

According to Breunig et. al., the local reachability definition plays an important role in connecting the "local outlier" notion to density based clustering. Typical density based clustering techniques use two parameters. One of them is the minimum number of points to be located in one cluster and the other is a parameter specifying the volume. During clustering, objects and regions are connected if their neighborhood densities exceed one of the thresholds. For the sake of the determination of the pseudo-clusters, Breunig et.al. kept the MinPts as the only parameters and used the values reach-dist$_{MinPts}$(*p, o*), for *o* ε $N_{MinPts}(p)$, as a measure of the volume and reached the local reachability density definition.

Basing on above mentioned concepts and definitions, they define their local outlier factor

(LOF) score as below. From this formula LOF can be defined as the average of the ratio of the local reachability density of p and those of of p's MinPts-neighborhoods.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

After making a formulization, they performed a detailed analysis on the properties of LOF and conducted some experiments on synthetic and real datasets to demonstrate the capabilities of local outliers.

II.3.3. Cluster-Based Local Outlier Factor (CBLOF)

Most clustering algorithms, especially those developed in the context of KDD (e.g. CLARANS [44], DBSCAN [45], BIRCH [46], OPTICS[47]) have been used for discovering outliers or peculiar data. Usually the objects that are not located in the clusters formed by the algorithms are regarded as outliers. This approach has two significant problems. First of all, because of the fact that the main objective of the a clustering algorithm is to find clusters, they are developed to optimize clustering, and not to optimize outlier detection, the outliers are usually regarded as noise and typically tolerated or ignored in order to increase the efficiency of the clustering operation. In addition to this point, they only sign the objects in a duality of being an outlier and not being an outlier. There is no quantification as to how outlying an object is.

Due to our knowledge base, Jiang et.al [48] firstly addressed the concept of clustering-based outlier detection and came up with an approach to regard small clusters as outliers. But they did not suggest a measure for identifying the degree of each object as to be an outlier. For most of the data points in a dataset that are not outliers, it is important to identify only top n outliers. Hence, the method proposed by Jiang et al. failed to fulfill this task effectively. Furthermore, how to distinguish small clusters form the rest was not addressed in their method.

He et al., 2002 contributed Jiang's approach by proposing a cluster-based local outlier definition. [14] Their basic idea is to combine the clustering view of Jiang et al., with the locality approach of Breunig et. al., 2000. [13] Their approach was illustrated in Figure II.6.

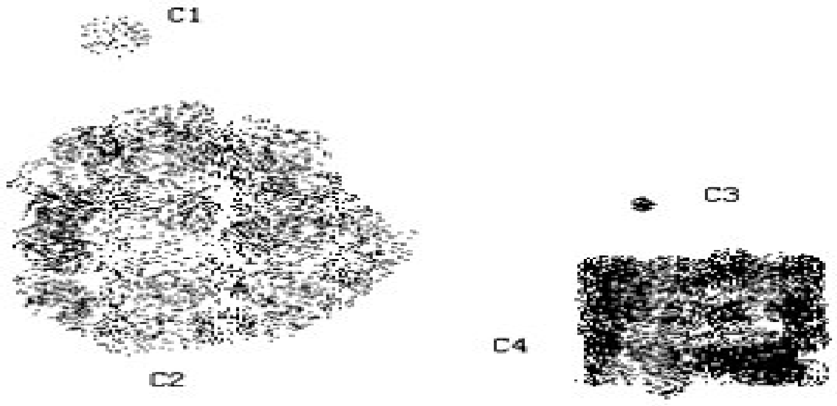


Figure II.6 2-D Dataset Representing the Cluster-Based Local Outlier Concept [14]

In Figure II.6 , there are four clusters C1, C2, C3 and C4. Obviously, the data points in both C1 and C3 should be regarded as outliers and captured by proposed definitions. Intuitively, we call data points in C1 and C3 outliers because they do not belong to the cluster C2 and C4. Thus, it is reasonable to define the outliers from the point of view of clusters and identify those data points that do not lie in any large clusters as outliers. Here, the number of data points in C2 and C4 are dominant in the data set. Furthermore, to capture the spirit of ''local'' proposed by Breunig et al. [13], the cluster-based outliers should satisfy that they are local to specified clusters. For example, data points in C1 are local to C2.

Firstly, they fulfilled the requirement for the classification of the clusters as small and large. For this classification, they introduced two numeric parameters α and β, whose values can be modified according to needs.

C = {C_1, C_2, …, C_k} represents set of clusters formed by implementing a clustering algorithm on whole dataset D. |C| is the number of data points in cluster C and the C set is in the sequence that $|C_1| >= |C_2| >= |C_3| \ldots |C_{k-1}| >= |C_k|$. Given two parameters α and β, b is defined as the boundary of large and small clusters if one of the following formulas holds.

$$|C_1| + |C_2| + \ldots + |C_b| >= |D|\,\alpha \qquad (1)$$

$$|C_b| / |C_{b+1}| >= \beta \qquad (2)$$

From (1) and (2) ;

The large clusters are defined as; LC = { C_i, | i<=b) ,

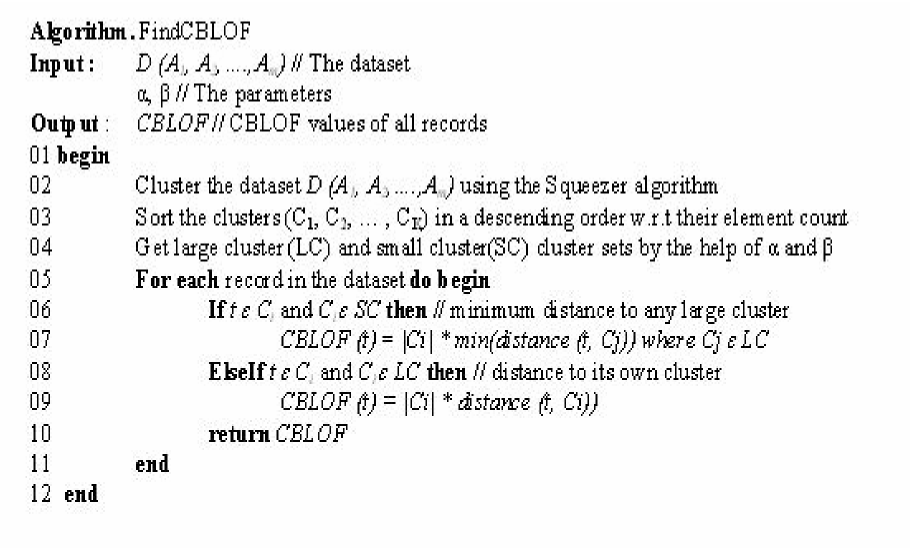
The small clusters are defined as; SC = { C_j, | j>b)

The first formula considers the fact that most data points in the data set are not outliers. α represents the percentage of the data points with respect to the total number of data points in the dataset D. For example, if α is selected as 90%, the clusters containing 90% of the records are regarded as *large clusters*. The second focuses on the fact that, the small and large clusters should have significant differences in size. If β is set to 5, the smallest cluster in LC is at least five times greater than the largest cluster in SC.

Originating from the basic idea stated above and the definitions of large and small clusters, He et.al, 2002, [14] formulated the cluster-based local outlier factor (CBLOF) by means of the size of its cluster, and the distance between the record and its closest cluster (if this record lies in small cluster) or the distance between the record and the cluster it belongs to (if this record belongs to large cluster), which emphasizes the local data behavior.

In the same study, He et. al. also came up with the findCBLOF algorithm, which uses Squeezer algorithm for clustering the dataset, and calculates the CBLOF values for each data point. They explained their choice of selecting Squeezer from the point of producing good clustering results and good scalability.

The findCBLOF algorithm is shown in the Figure II.7. But the details of the Squeezer algorithm were not covered here. There are two inputs of the algorithm: the first one is the dataset itself and the second one is α and β parameters declared in Equations 1 and 2. The output is the CBLOF values of all records.



```
Algorithm. FindCBLOF
Input:   $D(A_1, A_2, \ldots, A_m)$ // The dataset
         α, β // The parameters
Output:  CBLOF // CBLOF values of all records
begin
    Cluster the dataset $D(A_1, A_2, \ldots, A_m)$ using the Squeezer algorithm
    Sort the clusters ($C_1, C_2, \ldots, C_K$) in a descending order w.r.t their element count
    Get large cluster(LC) and small cluster(SC) cluster sets by the help of α and β
    For each record in the dataset do begin
        If $t \varepsilon C_i$ and $C_i \varepsilon SC$ then // minimum distance to any large cluster
            $CBLOF(t) = |Ci| * min(distance(t, Cj))$ where $Cj \varepsilon LC$
        ElseIf $t \varepsilon C_i$ and $C_i \varepsilon LC$ then // distance to its own cluster
            $CBLOF(t) = |Ci| * distance(t, Ci))$
        return CBLOF
    end
end
```

Figure II.7. The FindCBLOF algorithm [14]

II.3.4. Theoretical Evaluation of RPF, LOF and CBLOF

In this section, the theoretical backgrounds of the algorithms will be compared in various aspects.

The three algorithms have common properties such as constructing their interestingness model upon peculiarity, handling the peculiarity as a product of distance and density and proposing definite scores for outlier determination. On the other hand, they are differentiate from each other in terms of their perceptions of outlierness. In fact, this is the reason for me to select these three peculiarity oriented interestingness measures among a great number of candidates for my study.

The first difference among all is the assumption of each method about the way a peculiar record appears in a data set although each method argues that they are capable of finding all peculiar records.

For example RPF algorithm regards the dataset as a whole, and focuses on the points which are located in the peripheries of the dataset. Unlike RPF, the LOF algorithm comes up with the subjective hypothesis that the outliers are usually located near the denser pseudo-clusters. Consequently, LOF mainly focuses on locality originating from this point. CBLOF may be regarded as a hybrid approach on this issue, which takes into account both locality and globality in a data set for finding interesting records. In contrast with RPF and LOF, the CBLOF algorithm is concentrated on the investigation of the clusters that exhibit common characteristics and lead to find the peculiar data which forms of small clusters.

The second difference is the manner of handling noise, which is an important problem in peculiar data mining. All peculiarity measures are designed to work on noise-free datasets. Tendency of perceiving noise as peculiar data may cause wrong decisions, which may lead to inaccurate pattern constructions. On the other hand if we remember the idiom “One person’s noise is another person’s signal.” that can be argued that being more sensitive about the noise probability can cause the user to miss peculiar data. The decomposition of noise from peculiar data can be carried out by using domain knowledge or meta knowledge of database. On the other hand, a method distinguishing the noise from peculiar data candidates will be more appropriate especially for the large databases.

Examining the formulations from this perspective, RPF algorithm seems to be useful in limited circumstances in particular when the users know that the data set does not have any noise or they can easily distinguish the noise from the dataset by using domain experts or data quality techniques. This situation is related with the similarity between the noise and the algorithms' perception of peculiarity. On the other hand, the LOF and CBLOF algorithms mainly focus on local differentiations and this property can help them to differentiate between noise and outliers.

The third difference between the algorithms is the way of handling different types of data. Firstly, the original CBLOF algorithm's distance measures are based on support values and this measure can only be used for categorical attributes. Binning or using different techniques in order to calculate the distance may be thought as a solution to this problem. The other algorithms do not have any problems with the usage of different types of data.

II.3.5. Enhancements and Combinations of RPF, LOF and CBLOF

As a result of the theoretical and experimental studies on RPF, LOF and CBLOF, some researchers suggested some improvements.

Aggmeyang and Ezeife criticized LOF algorithm because of its huge repetitive computation and comparison need for every object before outliers are detected. [49] They emphasized on high computational expense of computing the reachability distance. To cope with this expense, they proposed the LSC-Mine algorithm based on the distance of an object and those of its k nearest neighbors. LSC-Mine improves upon the response time of LOF by avoiding the computation of reachability distances and local reachability densities. In addition, data objects that are not likely outlier candidates are pruned as soon as they are identified.

Another study by using LOF as a representative of density based outlier detection schemes held by Tang et. al [50] presented a unified model for several existing outlier detection schemes, and proposed a compatibility theory, which establishes a framework for describing the capabilities for various outlier formulation schemes in terms of matching users' intuitions. In their study they made a comparison of the density-based scheme and the distance-based scheme on various datasets and introduced a connectivity-based scheme that improves the effectiveness of the density-based scheme when a pattern itself is of similar density as an outlier. Finally, connectivity-based and density-based schemes are comparatively evaluated on both real-life and synthetic datasets.

Duan et. al. [51] proposed a new clustering algorithm LDBSCAN relying on a local-density-based notion of clusters. In this technique, they simplified the selection of appropriate parameters and took the advantage of the LOF to detect the outliers comparing with other density-based clustering algorithms. Their basic departure point is to improve the accuracy of both outlier detection and clustering. In their study, they combined the LOF and CBLOF and proposed a new measure CBOF which uses the cluster sizes, distances and local reachability densities as inputs.

Like Duan et. al., Sheresta et. Al [52] also constructed a composite z score which calculates overal peculiarity by combining values of other measures, namely Knorr's distance-based outlier detection method, Zhong's peculiarity factor and Breunig's LOF.

Apart from this measure, they defined the record, attribute, frequency, interval, sequence, and sequence of changes views of data. They introduced a framework which analyzes a data set based on these perspectives and detects peculiar data according to each by using Knorr's distance-based outlier detection method, Zhong's peculiarity factor and Breunig's LOF measures including composite z-score.

II.4 INTERESTINGNESS, PECULIARITY AND MEDICAL DATA

In this section, I will discuss the applications of peculiarity oriented interestingness measurements in the medical domain. Firstly, the challenges about the KDD process on medical data are briefly expressed. Secondly, the studies about the interestingness measurement of medical data is explained. Thirdly,the implication of peculiarity and outlierness in medical domain is investigated and the key implementation areas of peculiarity oriented interestingness measures, especially LOF, CBLOF and RPF on medical data are investigated.

II.4.1. Data Mining and Uniqueness Of Medical Data

The early studies on KDD focused on the transactional, retail based data sets, but implementing the know-how obtained from these experiences to find previously unknown knowledge from data collected in other domains, including medicine, have become an area of growing interest and specialization in the recent years.

However, some researches sharing their experience with science society agrees on the complex and problematic character of implementing a KDD process on medical data originating from the nature of it. [53, 54]

Cios and Moore classified the medical data as unique and pointed out the main headings that form this uniqueness. [53] In their opinion, the medical data is coming from different sources and in addition to the relational data that is used to be meet in KDD studies, it contains of various types of images, sounds and graphs taken from mechanical and electronical devices. Furthermore, the medical data is not limited with the electronic means. Unstructured and usually non-readable interview notes and comments written by health care providers complement the picture of heterogeneity of the medical data. They also stated that the general problems in KDD processes, such as redundant, inconsistent, imprecise and missing data plus noise are dramatically in question for medical data originating from its character to be a primarily a patient-care activity than an area of research.

Besides, medicine is a very human oriented domain and the usage of the medical data is restricted with social, ethical and legal limitations, such as data ownership, privacy and security. They also emphasized on the discordance between statistical methods and medical data. The ideal test conditions such as the minimal sample size may not be provided for medical data originating from technical or social problems.

Another important point they mentioned in [53], is their concern about incorporating medical domain knowledge into the mechanisms of clustering. They called attention to the risk of clustering problems to be computationally infeasible or to end up with results that do not make sense without partial human supervision.

While Cios and Moore concentrated on problems, Roddick et.al. merely focused on solutions in particular for exploratory data mining. [54] First of all they recommended the miners to use an investigative method for medical data mining as illustrated in Figure II.8. [55]

The investigative model uses an initial hypothesis as a starting point, from which other hypotheses are generated and tested. The generated hypotheses are validated initially by known constraints, then by the revised theory. When the data mining results support the hypothesis, the confidence of the conceptual model is increased, otherwise an indication of a conceptual model modification is produced.

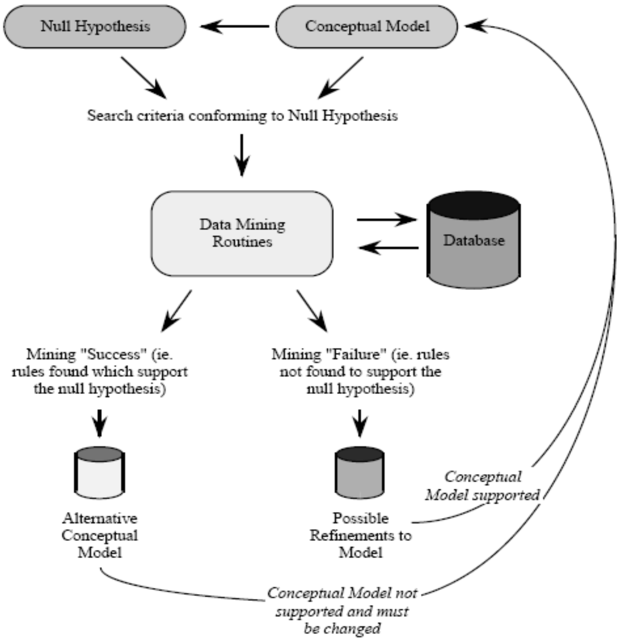

Null Hypothesis
Conceptual Model
Search criteria conforming to Null Hypothesis
Data Mining Routines
Database
Mining "Success" (ie. rules found which support the null hypothesis)
Mining "Failure" (ie. rules not found to support the null hypothesis)
Alternative Conceptual Model
Possible Refinements to Model
Conceptual Model supported
Conceptual Model not supported and must be changed

Figure II.8. Investigative Method for Medical Data Mining [54]

They also emphasized the significance of the episodic data in order to complete an efficient KDD process on medical data. The diagnosis and treatment process of any disease is closely linked with the orders, intervals and locations of events. Consequently, the KDD process should be enriched by adding space and time dimensions. From this aspect, Roddick et.al. suggested two techniques related directly with our topic. One of them is called difference detection performed by the investigation of statistically significant differences between rules derived from different but comparable datasets. The other is the anomaly detection, which refers to the discovery of the changes in data values or rule strengths that differ significantly from the expected patterns or values across that time period. Both terms, especially anomaly detection is strongly coupled with the peculiarity, which is the main concept of this study.

II.4.2. Interestingness Measurements On Medical Data and Patterns

The interestingness measurements have particular problems while working with different domains. Each domain has its specific characteristics and requirements originating from the technical, cultural and social nature of the area of interest. The ideal solution is being flexible to allow for cross domain applications, but also being specific enough to provide functionality which caters for the nuances of each domain. [56]

Motivated from the realization of the above declared statement, Shillabeer and Roddick targeted to reconceptualize the determination of interestingness for the medical domain. [56]

For example, they used the terms “hypothesis” and “strength” instead of “rule” and “interestingness” respectively. According to them the main reason of this change is the need of compliance with the terms commonly used in medical domain. In addition to this, they expressed that the usage of hypothesis and strength terms are more likely to define the real work carried out during KDD process. They kept the idea that the generic data mining algorithms were intended to produce inappropriate sets of hypotheses, [54] and underlined the requirement for understanding the factors that form the value of interestingness for different users.

Their definition of rule interestingness (“hypothesis strength”) depends on six criteria: novelty, provability, understandability, validity, applicability and representiveness and each is symbolized by using their first letter as the sub-index of the general strength S. In order to prioritize the criterion according to their importance for medicine W_x weight factors are used and the following formula is obtained :

$$S = [S_a*W_a] + [S_n*W_n] + [S_p*W_p] + [S_r*W_r] + [S_u*W_u] + [S_v*W_v]$$

By accepting the fact that each criteria is constructed from one or more metrics, the construction of a general metric can be illustrated as Figure II.9.

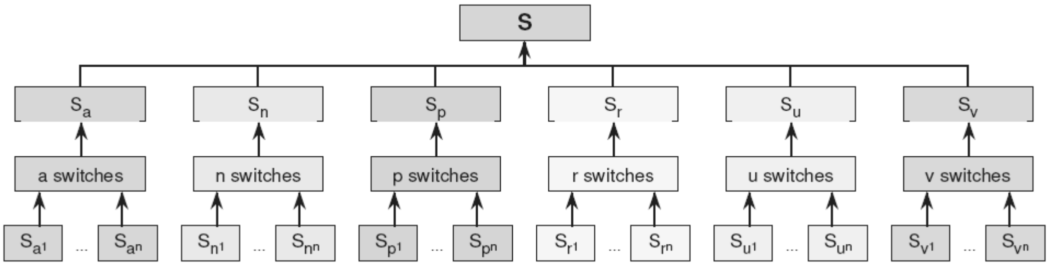


Figure II.9. Construction of a Metric for Strength of an Hypothesis [56]

In order to realize the construction mechanism, they also recommended a flow chart given in Figure II.10. The metrics that are defined to be in the scope are automatically propagated to next level. The main role of the switches in Figure II.10 is to filter the hypotheses as they attempt to propagate upwards through three levels. Any metric achieving within the user defined scope would automatically propagate to the next level. A metric which does not achieve a level within scope would be filtered through the switch. If the metric is critical (needed) to determine strength then the switch will be turned on and that hypothesis is discarded. If the metric is not critical then the switch will be turned off. As a result, the metric value and hypothesis are included in the next level if no other critical metrics fail for that hypothesis.

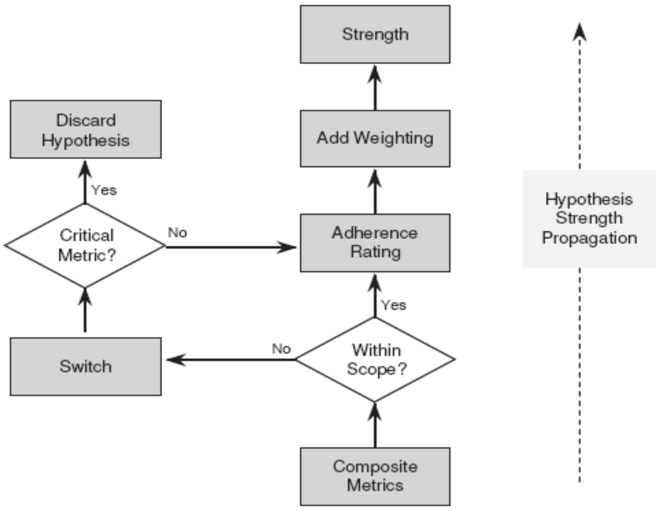


Figure II.10. Metric Implementation [56]

The research of Shillabeer and Roddick suggests a hypothesis selection engine as an annex to traditional data mining process which is fed by both domain knowledge and objective metrics. The proposed infrastructure may lead to contributive results, but in my opinion it has to be evaluated by using appropriate methods and data.

In contrast to the theoretical approach of Shillabeer and Roddick, Ohsaki et.al. examined the interestingness concept in medical data mining by using practical methods and experiments [57]. In this study, they used the output (association rules) of two former case studies [58,59] as an input for their experiments. The methodology of their work is similar with [60], where the different objective measures and subjective ideas of domain experts are compared on 8 datasets.

The medical expert evaluates the rules obtained from a clinical dataset as Not-Interesting (NI), Interesting (I) and Not-Understandable (NU), while the same evaluation on the same rules is done by using various objective interestingness measures. After the evaluations, the agreement degree of these two sets of evaluations is examined by researchers. Similar with the S measure of Shillabeer and Roddick; they also proposed a hybrid metric CMC defining the combined interestingness. They estimated the lower boundaries of meta-criteria assuming a random allocation of quality labels to rules, and evaluate the relative usefulness of objective measures to the lower boundaries.

They tested their method and measure on two medical datasets: Hepatitis and meningitis data sets. By using the results of two experiments, they labeled the Accuracy, x2-M1, RR, UncNeg, and Peculiarity measures as top five. Their CMC values are 1.7 times greater than the lower boundary, hence these objective measures are at least 1.7 times more capable of correctly identifying interesting rules than a random allocation.

When the formulized infrastructures of these top-five measures are examined, it can be easily seen that PF is the only measure representing the uniqueness of a rule, where others concentrate on generality. This shows that the medical experts placed importance on not only correctness, but also uniqueness. On the other hand, there is an other uniqueness based measure which is used in experiments is the GBI (Gago and Bento's Interestingness) measure. While PF focuses on the differences in the attribute values, GBI is usually interested in the differences between the attributes of rules. In the experiments, Ohsaki et. al. demonstrated that the performance of PF is superior than GBI. Hence the medical experts favored the differences in the attribute values of rules over the differences in the attributes of rules. The formulae of top-five measures are illustrated in Table II.3. Let A→C is the defined rule and TP, TN, FP and FN implies True Positive, True Negative, False Positive and False Negative in binary classification respectively.

Table II.3. Formulae of the Top-Five measures According to Ohsaki et. al. [57]

Measure	**Formula**	**CMC**
Accuracy	P(A Λ C) + P(⌐A Λ ⌐C) or in binary classification (TP+TN) / (TP+TN+FP+FN)	0.5685
χ^2-M1	T_{event} and O_{event} are theoretical and observational values of the number of instances contained in A→C respectively, $\sum_{event} \frac{(T_{event} - O_{event})^2}{T_{event}}$	0.5540
Relative Risk (RR)	P(C|A) / P(C|⌐A)	0.5536
Uncovered Negative (UncNeg)	P(⌐A Λ ⌐C) or in binary classification (TN) / (TP+TN+FP+FN)	0.5531
Peculiarity Factor (PF)	$\mathrm{RPF}(X_j) = \sum_{m=1}^{N} \sqrt{\sum_{j=1}^{k} \alpha_j (\mathrm{PF}(x_{ij}) - \mathrm{PF}(x_{mj}))^2}$ $PF\ (x_{ij}) = \sum_{r=1}^{N} distance\ \ (x_{ij}, x_{rj})^{\beta}$	0.5412

II..4.3. Peculiarity and Outlier Oriented Approaches On Medical Data

The results derived from the study of Ohsaki et.al. approved that the peculiarity oriented knowledge discovery approaches have a potential to meet medical experts' needs. They also concluded that the peculiarity of the attribute values end up with more interesting rules. In this section the applications of peculiar data and outlier detection methods in medical domain will be reviewed.

Like other domains, the outliers are detected by using standard statistical methods. Then, Laurikkala et. al. used the box plot method in order to detect outliers.[61] They implemented their study on vertigo and female urinary incontinence data. The box plot is a well-known simple display of the five-number summary (lower extreme, lower quartile, median, upper quartile, upper extreme. The lower threshold is defined as lower quartile - step and upper threshold is defined as upper quartile + step. Step is 1.5 times the interquartile range (upper quartile - lower quartile) which contains 50% of the data. If a value is greater than upper threshold or smaller than lower threshold then it is identified as an outlier.

They had two objectives in outlier detection, firstly they wanted to observe the effect of the removal of outliers to the classification methods by considering them as noise. Their study showed that the discarding outliers balances the class distribution and leads to an easier realization of nearest neighbor and discriminant analysis classification methods. Secondly instead of treating the outliers as noise and removing from the dataset, they prefer to concentrate on them. By this method they aimed to determine the interestingness level of the outlier records. In their experiments, some cases where the records are identified as abnormal by domain experts resulted with meaningful and valuable results especially in heterogeneous diagnostic groups.

The study of Laurikkala et.al.used the box plot thresholds as outlier detection measures. On the other hand, Podgorelec et. al. defined an outlierness score directly called confusion score metric that is based on the classification results of a set of classifiers. [62] The classification is made by using an evolutionary decision tree induction algorithm.

Podgorelec et.al. based their work on exceptional behavior of a single data record which is classified differently by different classifiers and produces contradictory information. This approach is closely coupled with the concept of the exception rules which is also based on capturing contradictions. Podgorelec et. al. restricted their study by only detecting the outliers before the generation of training sets and improving the generality and efficiency of learning mechanism by removing these outliers. They implemented their method on two cardio-vascular data sets and achieved considerable improvements in smaller and less complex datasets in particular. But it has to be noted that this improvement was not stable when the classification method was changed.

Another study based on the interestingness of exception rules belongs to Suzuki and Tsumoto [63]. They adopted a hypothesis-driven approach and used the method of Suzuki [64], which was introducing the problem as finding a pair of complementary rules with one of which represents a common sense rule where the other can be regarded as an exception rule. They carried out their experiments on the Meningitis data set, which is also used in this study. The domain expert evaluated each rule pair based on the validness, novelty, unexpectedness and usefulness aspects. Results of the study demonstrated that attributes with higher scores of novelty and unexpectedness are almost ignored during the classical data mining process. They also pointed out that some interesting discoveries of unknown mechanisms can be revealed by using these attributes.

A more recent study of Zhong et. al. [65] used the PF which was suggested as an interestingness measure by his group. [31] In this study, they came up with interesting results on modeling, transforming and mining multiple human brain data derived from visual and auditory psychological experiments by using the fMRI. In figure II.11. the global view of their methodology can be seen.

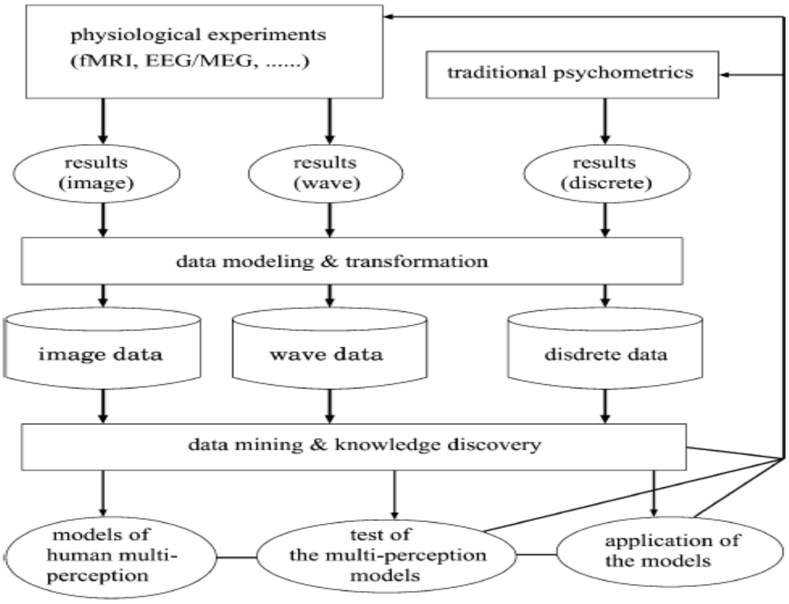

physiological experiments
(fMRI, EEG/MEG, ...)
traditional psychometrics
results (image)
results (wave)
results (discrete)
data modeling & transformation
image data
wave data
disdrete data
data mining & knowledge discovery
models of human multi-perception
test of the multi-perception models
application of the models

Figure II.11. The methodology to investigate human multi-perception [65]

Physiological experiments, such as fMRI, EEG, and traditional psychometrics are synchronously used in order to obtain the raw data for modeling the real human multi-perception mechanism. In order to formalize, clean and conceptualize the multimedia data, MED(x)[66] software package was used. The package transformed the image data from fMRI into a relational data model which could be used in the knowledge discovery process. The mental arithmetic problems of addition which were also used in traditional psychological experiments are auditory and visual stimuli to measure cortical areas of human calculation processing.

The knowledge discovery process was implemented by applying two different methods, one of which contained the prior knowledge of Brodmann areas, where the other did not. PF was utilized to find the peculiar records obtained from fMRI-oriented relational data.

By comparing the results of peculiarity oriented mining in auditory and visual calculation data, they confirmed the hypothesis which suggests that the auditory information might be transferred into visual information, in some cases of advanced information processing such as calculation.

All of the researches reviewed in this section exhibit various usages of the peculiarity and outlierness oriented methodologies in medical data which has unique features as described in Section 3.1. The all results derived from different studies commonly demonstrated the success of peculiarity oriented approaches in medical domain. This study follows the main ideas that examine the relation between peculiarity and interestingness in medical data sets. Moreover it focuses on the comparison and analysis of the diffent interestingness measures in order to reach to a more descriptive model which defines the relation between the peculiarity concept and the real human interest concentrated on medical data.

CHAPTER III

COMPARISON OF METHODS

In this chapter, the design and results of the experiments done by using synthetic and real world medical data will be discussed. In the first section the experimental environment is introduced and general definitions are given. In the second section the experiments which are carried out by using synthetic datasets were examined with an objective approach. In the third and latest section, the objective results obtained from real world medical datasets will be discussed and evaluated by domain specific knowledge.

III.1. Experimental Environment and General Definitions

The first three synthetic datasets were constructed by programs which are developed by using the randomization functions of MATLAB[67]. The mean and standard deviation values of the produced clusters are selected deliberately in order to construct a demonstrative view of samples. The latest synthetic dataset is obtained from the MS thesis study of Cosku Erdem. [68]

The implementations of RPF and CBLOF are realized by MATLAB [66] programs developed specifically for this study. In order to handle numerical data DBSCAN [45] is used for the clustering phase of CBLOF instead of the Squeezer algorithm. For LOF measure, the maxlof function presented in dprep package developed for R software [69] is used.

In all diagrams representing any LOF, RPF or CBLOF values, the peculiarity of the records with respect to mean (*μ*) and standard deviation (*σ*) values are illustrated by the color palette defined in Table III.1. The values are classified as very ordinary, ordinary, normal, interesting and very interesting according to their lower and upper bounds to aid comprehension. In the figures related with CBLOF, the points belonging to large clusters are symbolized with "*", and the outliers and the small cluster elements are represented with "+" and "o" respectively.

Table III.1. Color Scheme Definitions of Interestingness Levels

Class	Lower Bound	Upper Bound	Color	Class Number
Very Ordinary	*N/A*	μ-2σ	Gray	0
Ordinary	μ-2σ	μ-σ	Silver	1
Normal	μ-σ	μ	Cyan	2
Normal	μ	$\mu+\sigma$	Blue	3
Interesting	$\mu+\sigma$	$\mu+2\sigma$	Orange	4
Very Interesting	$\mu+2\sigma$	*N/A*	Red	5

III.2. Experiments with Synthetic Datasets

In this section the LOF, RPF and CBLOF measurements are used in order to find the peculiar data in four synthetic datasets which have different statistical properties.

III.2.1. Normally Distributed Random Numbers

According to the central limit theorem the sum of a large number of independent random variables each with finite mean and variance will be approximately normally distributed. Departing from this point, it will be appropriate to start our investigation process from normal distribution. In this experiment the behavior of the three measures in normal distribution will be observed and the results of the different outlierness definitions will be discussed.

Our dataset seen in Figure III.1 consists of a Gaussian distribution of 10.000 data points with two dimensions. The mean and standard deviation of the first dimension is 5.00 and 1.00, where the values for the second dimension are 2.00 and 1.00 respectively.

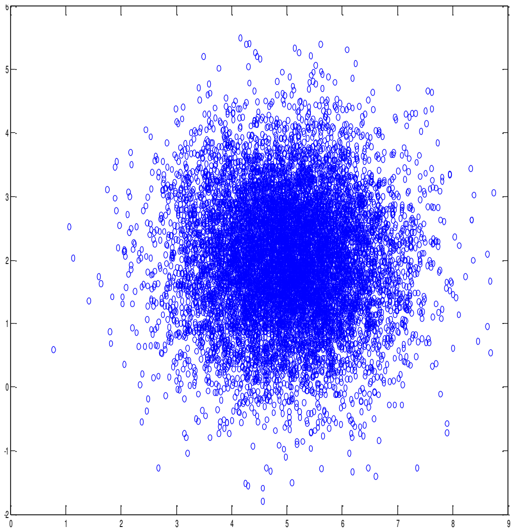


Figure III.1. General View of Sample data set 1

The mean and standard deviation of the RPF values of the cluster are 24.00 and 6.13. By using the color scheme defined in Table III.1., if $\mu+\sigma$ is selected as the threshold for being recognized as peculiar record, the orange and red points in Figure III.2, are pointed as peculiar records. If the criteria of being peculiar is changed to $\mu+2\sigma$, then only the red points will be regarded as peculiar.

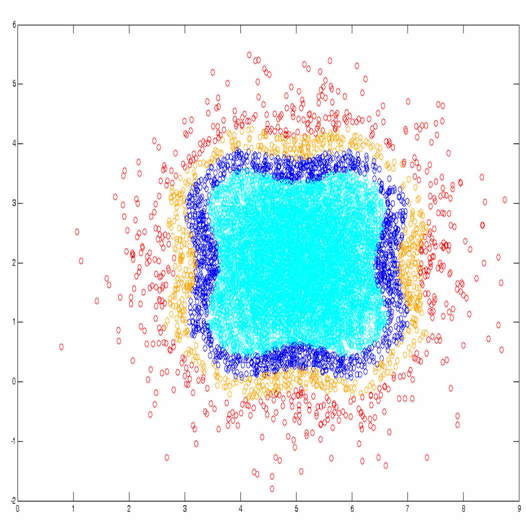


Figure III.2. Representation of RPF Values for Sample Dataset 1

If the LOF values are calculated, the mean of the LOF values of the dataset is 1.02 and the standard deviation is 0.07. The results of these experiments are given in Figure III.3.

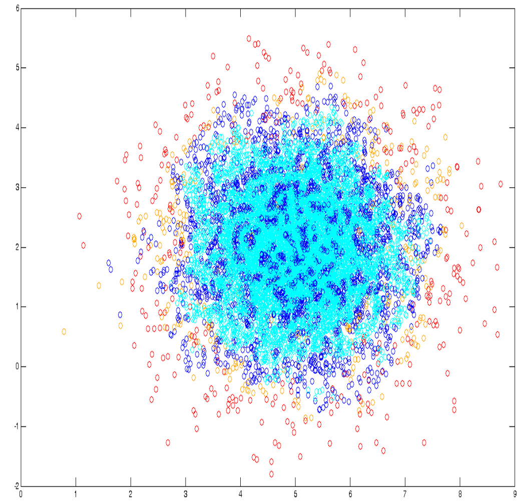


Figure III.3. Representation of LOF Values for Sample Dataset 1

When CBLOF is used for outlier detection, the algorithm firstly divides the dataset into 11 clusters. The densest cluster consists of 9328 elements, and it is the largest cluster among 11 clusters. The remaining 558 points which do not exist in any of the clusters were identified as outliers. The CBLOF results for this dataset is illustrated in Figure III.4.

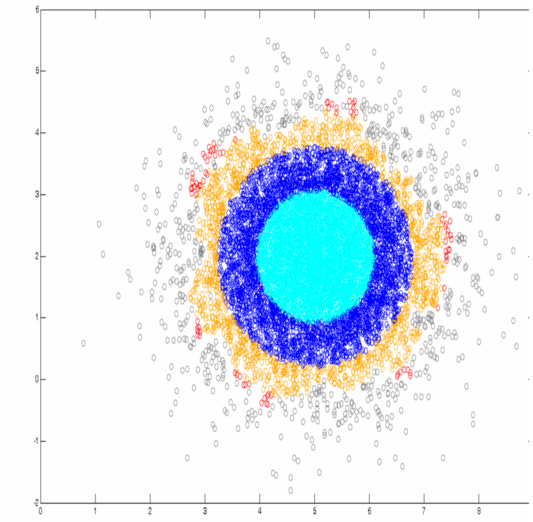


Figure III.4. Representation of CBLOF Values for the Sample Dataset 1

Originating from the distance based approach of RPF, the highest peculiarity values are found at the periphery of the data set. The lower the peculiarity value the closer it is to the center of the data set. The results derived from LOF do not show this characteristic so definitely, but the tendency of being outlier is increased at the peripheries of the dataset.

If the records identified as local outlier by LOF algorithm and peculiar by RPF handled together, some interesting results is reached. 81.81% of the local outliers are also detected as peculiar data by RPF algorithm. This percentage increases to 82.50%, when the threshold is set as $\mu+2\sigma$. On the other side, only the 44.77% of peculiar records are also identified as local outlier. Moreover, if the threshold is increased to $\mu+2\sigma$, then this percentage is decreased to 40.86%. This analysis demonstrates that the RPF algorithm is more selective than LOF algorithm when normal distribution conditions are set.

The results of CBLOF algorithm is similar to RPF results, but with one important difference. The peripheral points that are shown in gray color belong to the outliers which have the most minimum CBLOF values. The number of the points that are regarded as a member of the found cluster is 9.328. As mentioned by He et. al. the CBLOF value has a coefficient $|C_i|$ which represents the number of elements in the cluster. In those points this coefficient increases the CBLOF value and by the multiplication of the relatively high inner cluster distances, they become the points that have the highest CBLOF values.

III.2.2. Two Spherical Clusters Each of Which Consists of Normally Distributed Random Numbers

Some studies such as [40] mentioned in Section II.2 are based on the assumption that the dataset consists of samples from a mixture model containing M and A, where M refers the majority and A refers to anomalous distributions. In order to evaluate this assumption the following dataset is produced and tested by three measures. Number of points that belong to clusters were chosen in order to demonstrate the effect of β parameter in CBLOF algorithm.

In this experiment, two non-overlapping clusters are created which comprise two dimensional tuples, with the values in each dimension following a normal distribution. The denser cluster to be called as C1, consists of 8000 objects with $\mu_1=2, \sigma_1=1, \mu_2=1, \sigma_2=1$ and the other cluster C2 contains 2000 points with $\mu_1=8, \sigma_1=1, \mu_2=3, \sigma_2=1.$ C2 can be regarded as *small cluster* according to the definition proposed by He et.al in 2002 [14], while β value is selected as 3. The combination of two clusters constitutes our second dataset. This 2 dimensional dataset is illustrated in Figure III.5.

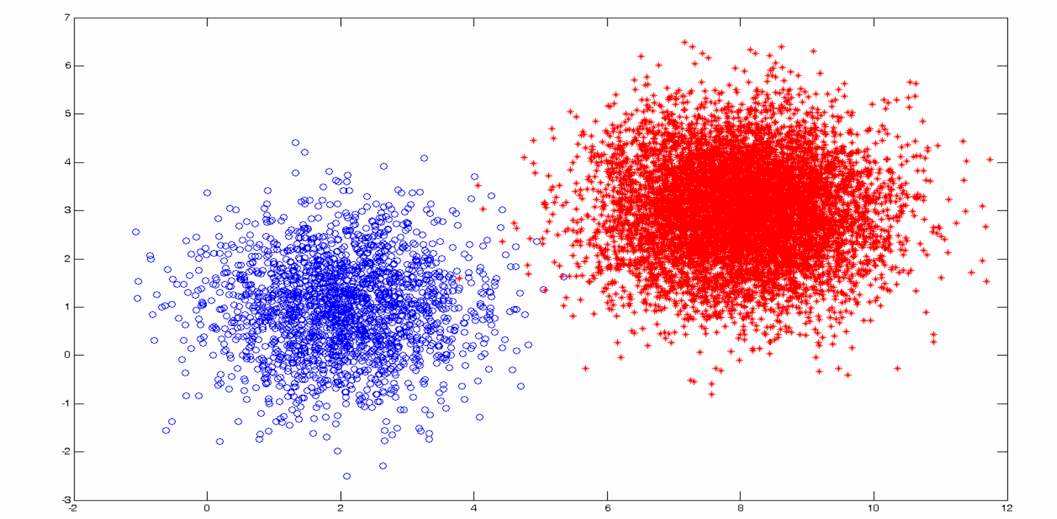


Figure III.5. General View of Sample Dataset 2

The RPF values of the tuples are shown in Figure III.6. In C2 μ and σ of the RPF values of the tuples are found as 44.1 and 9.21 respectively. The same values for C1 are found as 27.38 and 3.13 respectively.

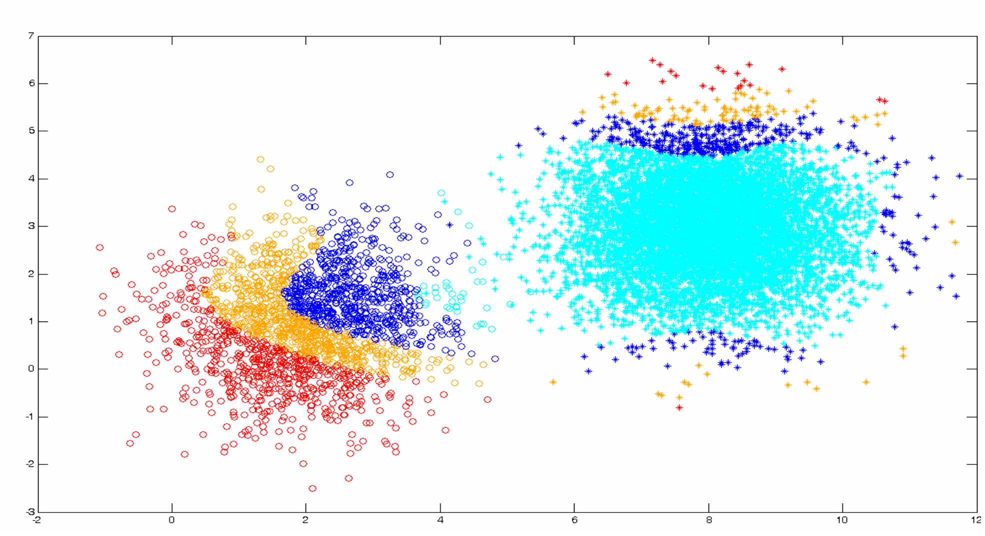


Figure III.6. Representation of RPF for Sample Dataset 2

The LOF values of the tuples are shown in Figure III.7. In C2, μ and σ values of the LOF values of the tuples are found as 1.04 and 0.10 respectively. In C1, these values are found as 1.03 and 0.08. Out of 623 points, 223 points (35.79%) that are marked as local outlier belong to the C2 and the remaining 400 belong to the other one. It is observed that local outliers tend to appear more in less dense clusters but not as much as RPF. The percentage of local outliers which come from the denser cluster decreases from 64.21% to 59.57% when the threshold is set as $\mu+2\sigma$.

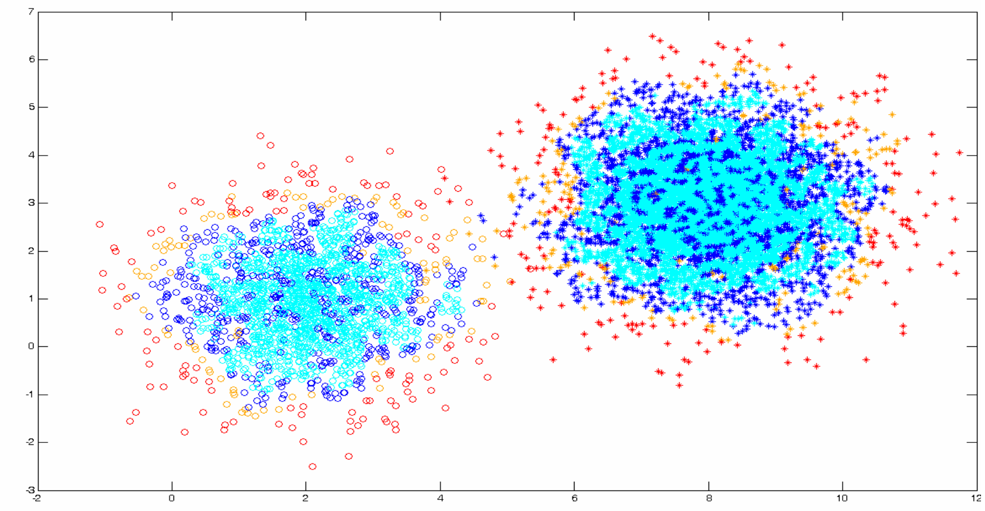


Figure III.7. Representation of LOF Values for Sample Dataset 2

Figure III.8 and III.9 show the results of CBLOF algorithm. The β parameter is adjusted in order to demonstrate the difference in the results of the experiment while it directly affects the small and large cluster identification. The DBSCAN algorithm [45] that is the initial step of the CBLOF algorithm, divides the dataset into 12 clusters. The first cluster consists of 7657 and the second cluster comprises 1678 points. 555 points were identified as outliers (DBSCAN puts them into one cluster) and the remaining points are distributed among 9 clusters, in which the number of elements in each cluster varies from 6 to 23.

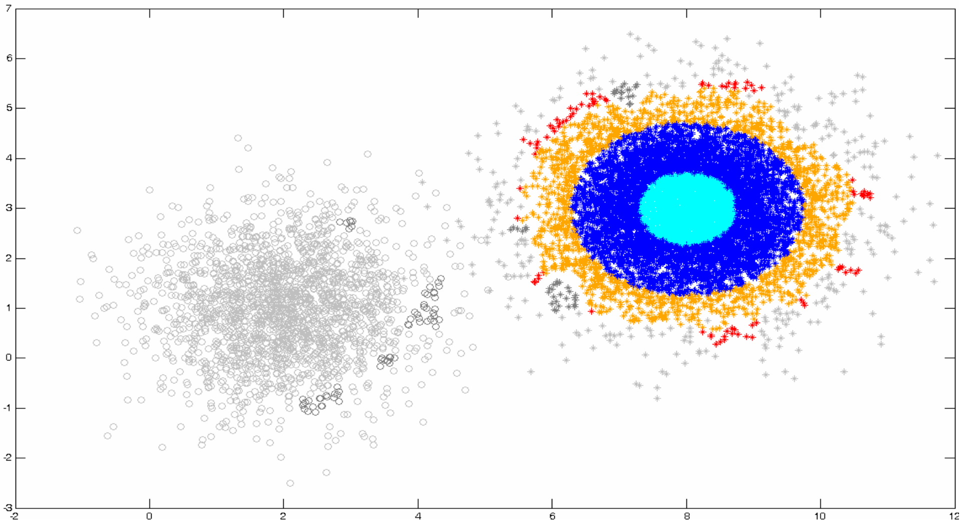


Figure III.8. Representation of CBLOF Values for Sample Dataset 2 with β=5

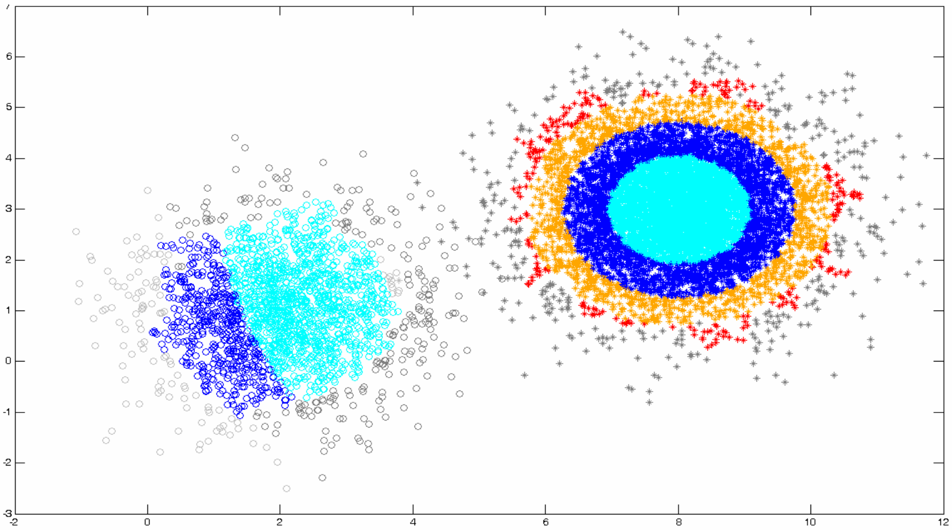


Figure III.9. Representation of CBLOF values for Sample Dataset 2 with β=3

This dataset is a combination of two Gaussian distributions and shows the difference between the spirits of two methods. In LOF method, the resultant records are placed in the periphery of each cluster, whereas RPF method handles the dataset as a whole and determines a de-facto centre. It means that if the attribute values come from mixture of Gaussians, it will take into account the center of Gaussians not the mean of each Gaussian. The measure is sensitive to the density of attributes in a data set. If the peculiar data are few and lies between the two clusters which are equally distant from the peculiar data, the method may not favor the expected values as peculiar but only data resided in the peripheries. For example, as can be shown in Figure III.6, the points between the clusters are not as interesting as the points in the peripheries of the clusters according to RPF.

Another issue is that the method seems to exhibit similar characteristics with that of distance-based clustering [24]. They stated that the calculation used in their method is based on the distances between individual attribute values rather than all. But this also leads to a problem. If two attributes do not appear as outliers individually but together do, the method will ignore this tuple. For example, let attribute A_1={male, female} and A_2={housewife, engineer}. If there exists a record having {male, housewife}, the algorithm will not label it as outlier. To handle this, anomaly scores proposed by [70] can be used. Here, supports for all item sets are calculated and if they are under a threshold and so its subsets, they reciprocated the support of each item set and summed up to obtain an anomaly score.

The CBLOF values of the dataset are sensitive to α and β thresholds that are set in order to distinguish small and large clusters. In addition, the values of the attributes of the dataset, which are used in distance calculation between any record t and large clusters affect the CBLOF value. The minimum of all distances calculated between any record t and b number of large clusters can be significantly greater than the average distance between any record t and large cluster (see Section II.3). For example, when β is selected as 3 in dataset 1, the large cluster will contain 7657 elements and all the calculated distances can be less than one. Hence, as the number of points in a cluster $|C_i|$ can dominate the distance factor in the calculation, the result will affect the coefficient in the CBLOF formula and favor the records near the large cluster at most.

III.2.3. Two Spherical Clusters Each of Which Consists of Normally Distributed Random Numbers with Noise

It is a generally agreed point that real world databases are handicapped with noise. Errors originated from data entry, conversion and processing processes lead to a specific amount of noise in datasets. It is a desired property for an interestingness measure to have ability to distinguish peculiar data and noise. This experiment aims to compare the interestingness measures on noisy database conditions.

In this experiment two non-overlapping clusters are created which comprise two dimensional tuples, with the values in each dimension following a normal distribution. But in this experiment 25% of the points in each cluster are omitted and instead of these points uniformly distributed noise was added to the data set. The distribution interval of the noisy data is between -2 and 12 in the first dimension and -3 and 7 in the second dimension. The C1 cluster in this experiment consists of 6000 objects with with $\mu_1=2, \sigma_1=1, \mu_2=1, \sigma_2=1$ where the C2 cluster contains 1500 points with $\mu_1=8, \sigma_1=1, \mu_2=3, \sigma_2=1.$ The points, which are represented with + symbols, refer to the points that are distributed uniformly to the plane. The cluster C2 can be again regarded as small cluster according to the definition proposed by He et.al in 2002,[24] while the β value is selected as 3. The dataset is illustrated in Figure III.10.

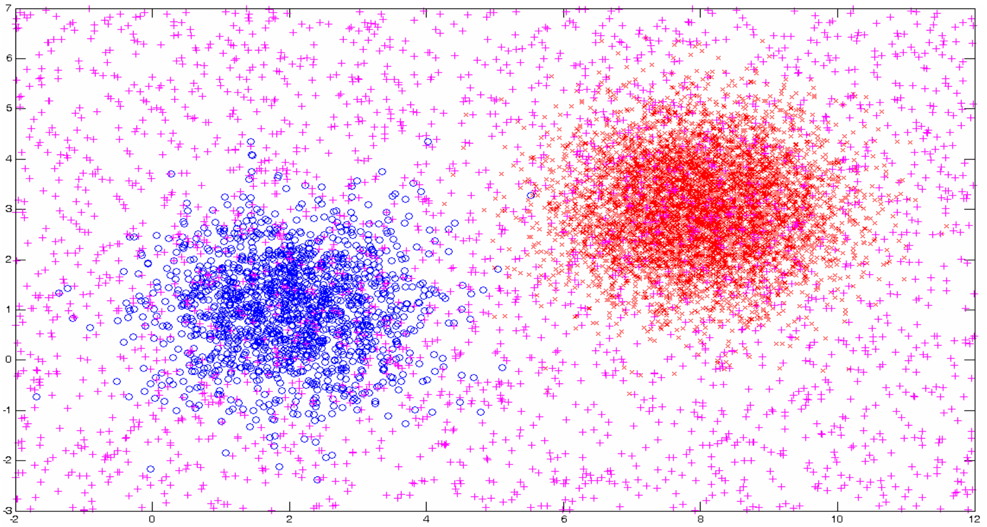


Figure III.10. General View of Sample Dataset 3

The RPF values of the tuples are shown in Figure III.11. For C2, μ and σ of the RPF values of the tuples are found as 44.11 and 9.21 respectively. For C1, the same values are calculated as 27.38 and 3.13 respectively. In spite of the fact that some data points located in the denser cluster C1 are marked as peculiar, most of the peculiar records are originating from C2 , which is less dense than the other one.

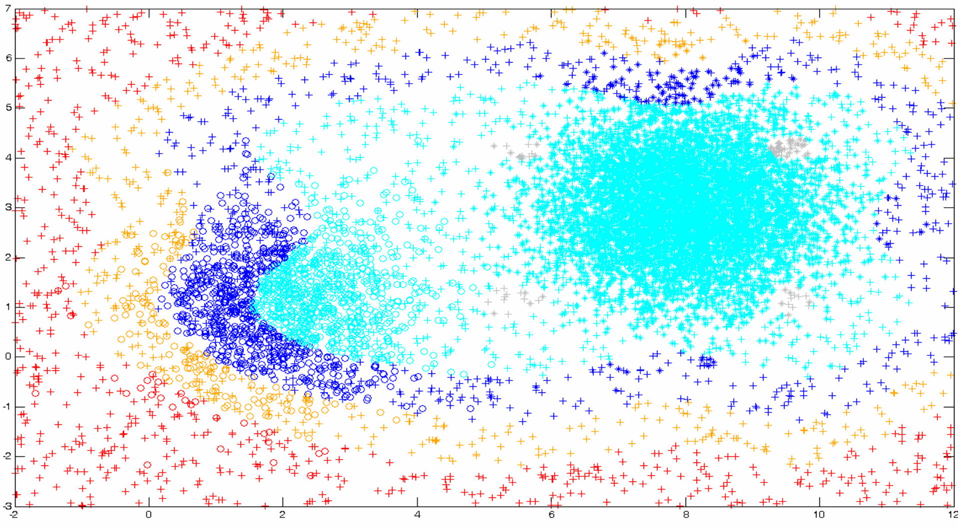


Figure III.11. Representation of RPF Values for Sample Dataset 3

The distribution of the LOF values with respect to μ and σ are shown in Figure III.12. In C2, $\mu_{.}$ of the LOF values of the records is found as 1.02. For C1, μ of the LOF values of the records is found as 1.02. For the remaining records, μ of the LOF values of the records is found as 1.03. No significant difference is observed among the standard deviation values.

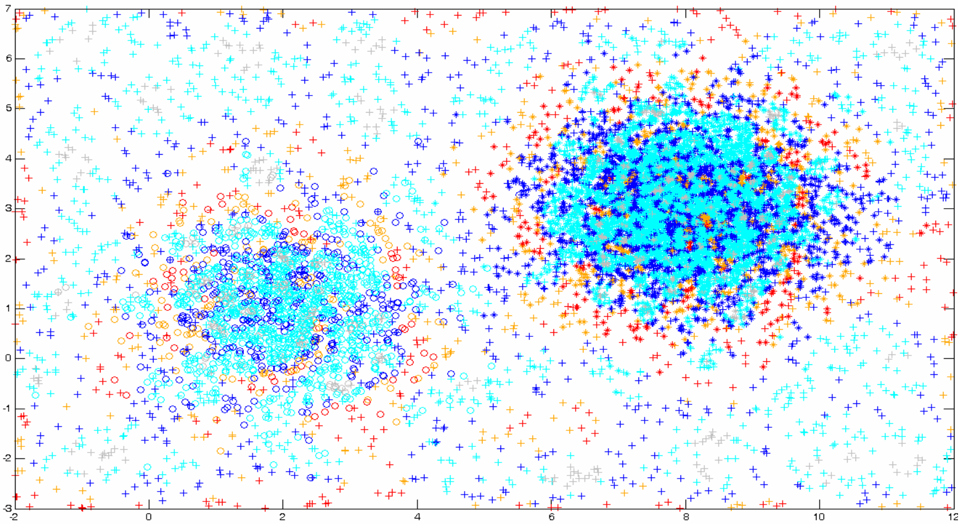


Figure III.12. Representation of LOF Values for Sample Dataset 3

The distribution of the CBLOF values with respect to μ and σ are shown in Figures III.13 and III.14 when β is set as 5 and 3 respectively.

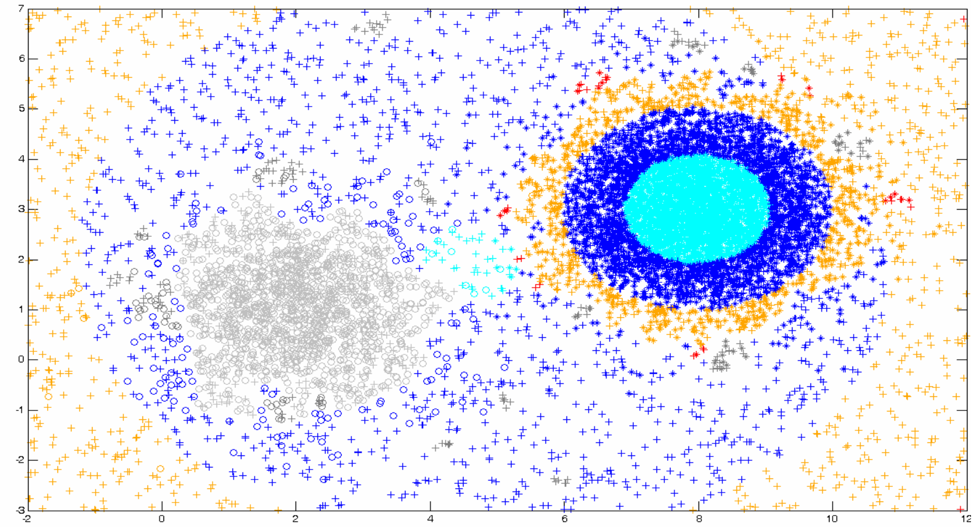


Figure III.13. Representation of CBLOF Values for Sample Dataset 3 with β=5

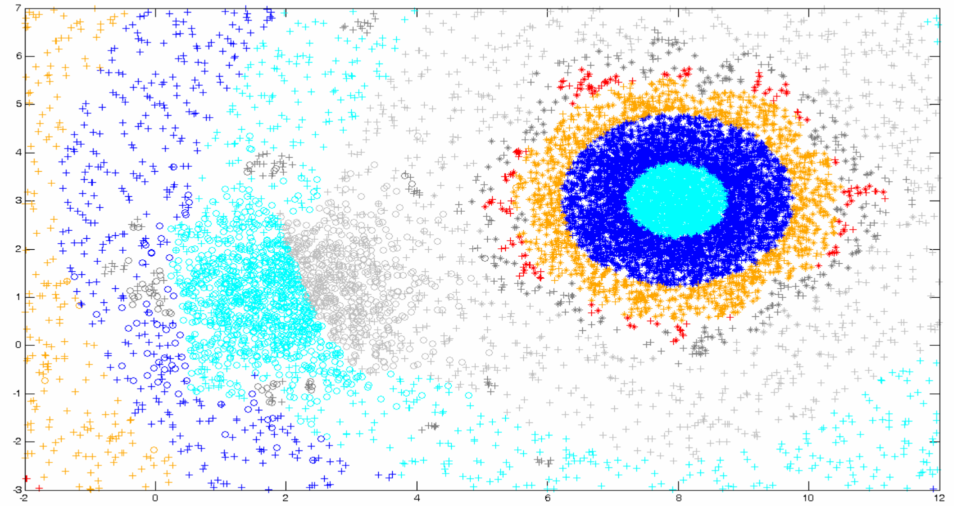


Figure III.14. Representation of CBLOF values for Sample Dataset 3 with β=3

Despite the fact that density plays a role in RPF's decision, the distance measure becomes significant enough to be selected. On the other side, the LOF method favors the points very close to clusters most with some exceptions. These points originate from relatively dense regions or low density of clusters including noisy data.

On the other hand, CBLOF produces different results in different β values. When the β value is selected as 5, only one cluster is identified as large cluster and the points having the highest CBLOF values are agglomerated along the peripheries of this cluster. If 3 is chosen as β, then the highest CBLOF values appear at the leftmost part of our dataset plane.

III.2.4. Non-convex clusters and Abundant Noise

While the previous synthetic examples represent some aspects of real world datasets, it has to be mentioned that real world datasets are not limited with the ones consisting spherical clusters. In most cases, the records in a dataset form non-convex clusters and different techniques are proposed in order to find out such structures. This experiment is accomplished to represent such formations that are likely to be seen in medical datasets.

This set is produced at the stage of prototype implementation of DBCM [68]. It consists of 2,783 data points which form two visually differentiable non convex clusters and abundant amount of noise in the background.. This dataset is illustrated in Figure III.15.

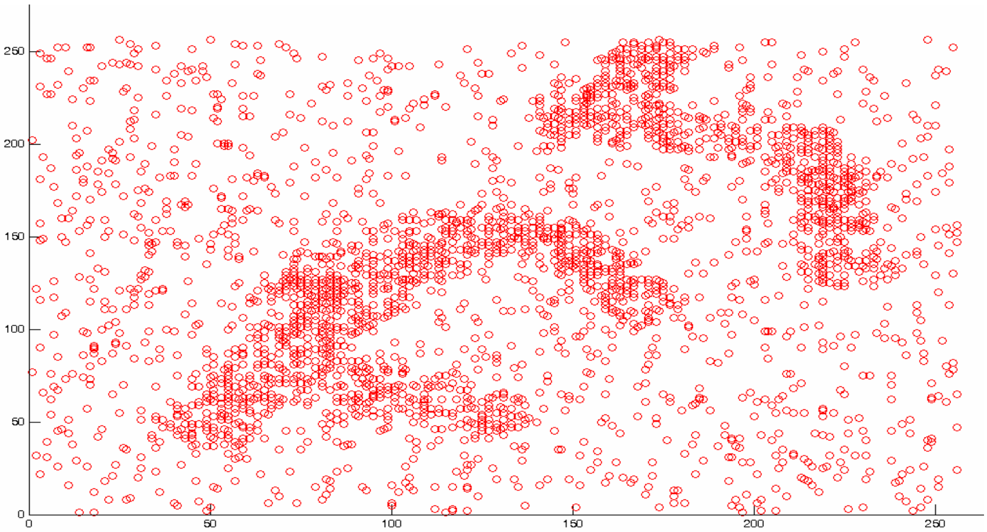


Figure III.15. General View of Sample Dataset 4

Concentrating on the orange and red points in Figure III.16 that can be said that, even there is an accumulation near the S-shaped cluster when confidence level is selected as 84.1%, ($\mu+\sigma$) , these points lose their peculiarity characteristics when confidence level is set as 97.7% ($\mu+2\sigma$).

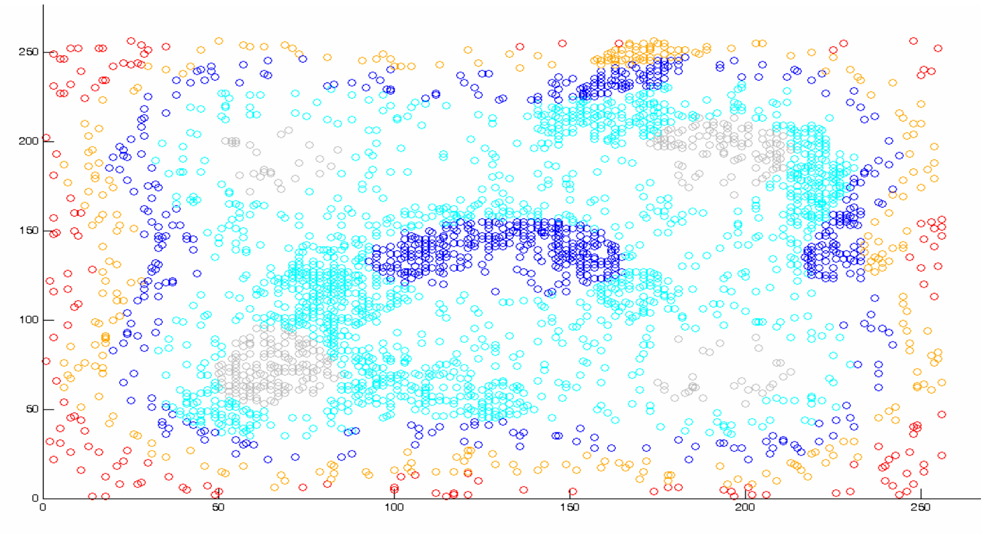


Figure III.16. Representation of RPF Values for Sample Dataset 4

The local outliers definitely exhibit their main characteristics and they are located adjacent to the F-shaped and S-shaped pseudo clusters as illustrated in Figure III.17.

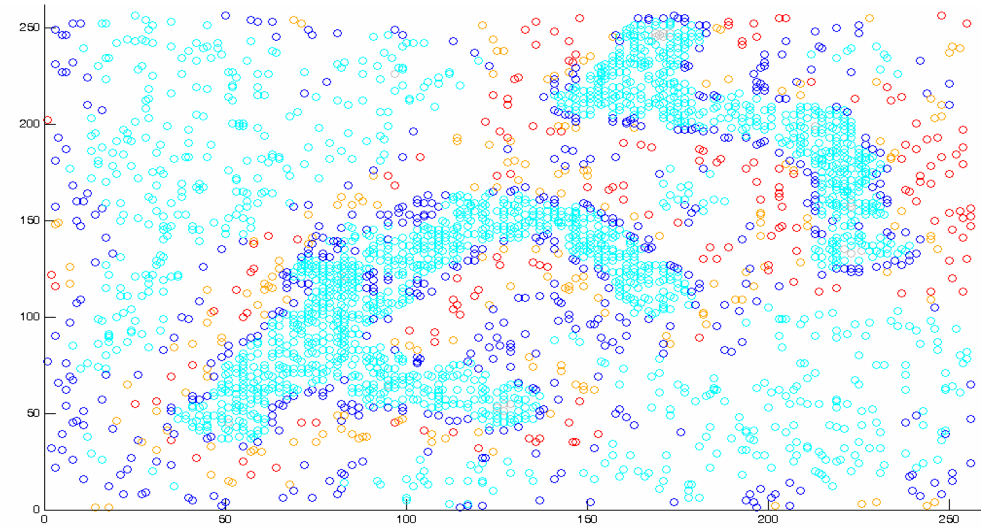


Figure III.17. Representation of LOF Values for Sample Dataset 4

Figure III.18 shows the CBLOF result of data set 4. The adjustment on the β parameter on this dataset is not illustrated, because this adjustment does not lead to a change in status of clusters being small or large.

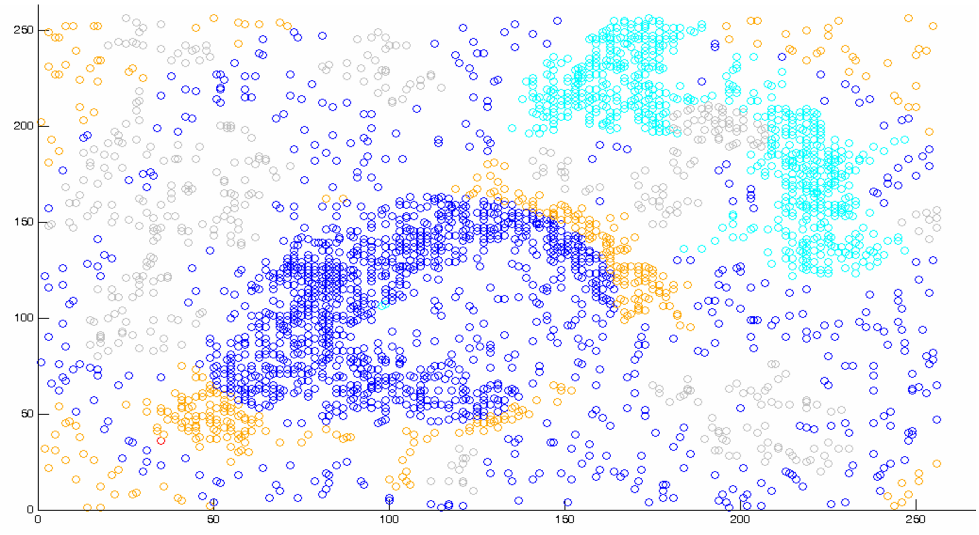


Figure III.18. Representation of CBLOF Values for Sample Dataset 4

In the experiment RPF again focused on the furthest points according to distance and density measures, while LOF concentrated on cluster peripheries. This example well demonstrates the differences between the two methods in terms of their objective.

Before CBLOF values are calculated, the dataset is divided into 21 clusters. The two clusters S and F are pointed as large clusters and the other 18 are chosen as small clusters. The remaning points are marked as outliers and put into one cluster by the algorithm regardless of their dissimilarity.The points having greater CBLOF values are mainly originating from large clusters as a result of cluster size coefficient. CBLOF results are heavily affected by the results of the initial clustering algorithm.

III.3. Experiments with Real World Medical Datasets

Experiments carried out with synthetic datasets demonstrated the general structure of the compared measures. In this section, the same measures are compared due to their performance on real medical data.

III.3.1. Wisconsin Breast Cancer Dataset

The first real world dataset to be used in our study is the Wisconsin Breast Cancer (WBC) dataset from the UCI Machine Learning Repository. The original Wisconsin Breast Cancer Dataset consists of 699 records with 9 attributes. The tenth dimension contains the class information which takes either “benign” or “malignant” value. As all the attributes are defined in 1 to 10 scale, no normalization was required. Sixteen instances with missing values were deleted from the original dataset. In addition, some malignant records were removed from the system and the reduced final dataset included of 483 records, 39 of which comprises malignant and 444 of which consists of benign records. This process was carried out inline with [14]. Unlike the synthetic datasets, Wisconsin Breast Cancer dataset has 9 dimensions, and is not suitable for visual representations. Instead, the interestingness measures of all records were calculated and evaluated on their performance in finding malignant records by using contingency tables. The results obtained from the RPF measure is given at Table III.2 and Table III.3. In Table III.2. $\mu+\sigma$ is selected as threshold, where $\mu+2\sigma$ is preferred in Table III.3.

Table III.2. RPF Results of WBC Dataset with $\mu+\sigma$ Threshold

	Real World Results		
Test Results	**Benign**	**Malignant**	**Total**
Benign	433	0	433
Malignant	11	39	50
Total	444	39	483

Table III.3. RPF Results of WBC Dataset with $\mu+2\sigma$ Threshold

	Real World Results		
Test Results	**Benign**	**Malignant**	**Total**
Benign	442	10	452
Malignant	2	29	31
Total	444	39	483

When the same experiment was carried out with LOF algorithm by using the $\mu+\sigma$ threshold, the following results at Table III.4 were achieved. 36 of the 435 records that have been identified as local outlier by LOF algorithm are malignant. On the other side, 45 out of 48 records that have been declared as local outliers which refer to “malignant” records are in fact benign. The results show that the LOF algorithm and the local outlier concept are not appropriate for this dataset.

While the results lead to a clear failure of LOF in identifying malignant records, even when the threshold is selected as $\mu+\sigma$, the test was not repeated by using a higher threshold such as $\mu+2\sigma$.

Table III.4. LOF Results of WBC Dataset

	Real World Results		
Test Results	**Benign**	**Malignant**	**Total**
Benign	399	36	435
Malignant	45	3	48
Total	444	39	483

The clustering phase of the CBLOF algorithm classifies the whole dataset into two groups. It forms one cluster representing the ordinary records and the outliers. When the ordinary records are regarded as benign samples and the outliers as malignant ones, the contingency table III.5 is obtained.

Table III.5. CBLOF results of WBC dataset

	Real World Results		
Test Results	**Benign**	**Malignant**	**Total**
Benign	429	0	435
Malignant	15	39	54
Total	444	39	483

He et. al. used the same dataset on the study [14] in which they proposed the cluster based outlier concept and findCBLOF algorithm. The criterion they chose to evaluate the performance of their algorithm is the ratio of detected malignant records over the records which have the top CBLOF value. Their results are combined with the results derived RPF experiment and illustrated in Table III.6.

Table III.6. Comparison of CBLOF and RPF Results for WBC Dataset

Top %	**RPF**	**CBLOF**
1%	4 (10.26%)	4 (10.26%)
2%	9 (23.08%)	7 (17.95%)
4%	19 (48.72%)	14 (35.90%)
6%	27 (69.23%)	21 (53.85%)
8%	35 (89.74%)	27 (69.23%)
10%	38 (97.44%)	32 (82.05%)
12%	39 (100.00%)	35 (89.74%)
14%	39 (100.00%)	38 (97.44%)
16%	39 (100.00%)	39 (100.00%)

If a ROC curve is drawn upon this knowledge, the following results illustrated in Figure III.19 and Figure III.20 are achieved. The scales for the figures are adjusted in order to obtain reasonable views. The area under ROC curve for RPF is found as 0.997, where the same value for CBLOF is calculated 0.979. The results demonstrate the efficiency of both algorithms in order to select malignant records from the WBC dataset.

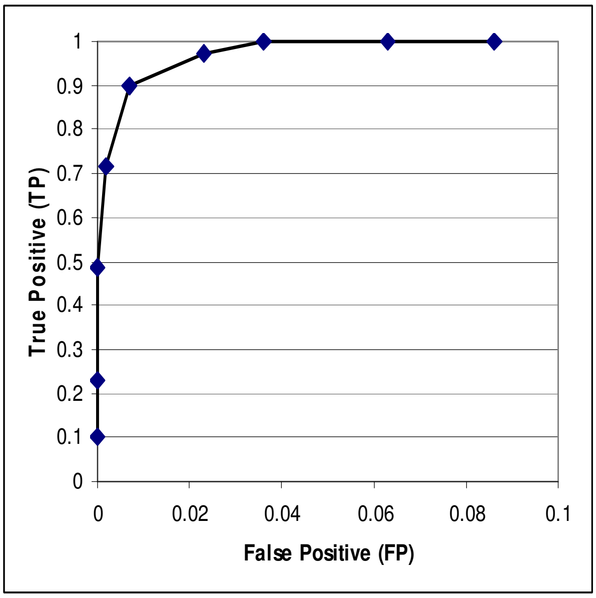


Figure III.19. ROC Curve for RPF Results on WBC Dataset

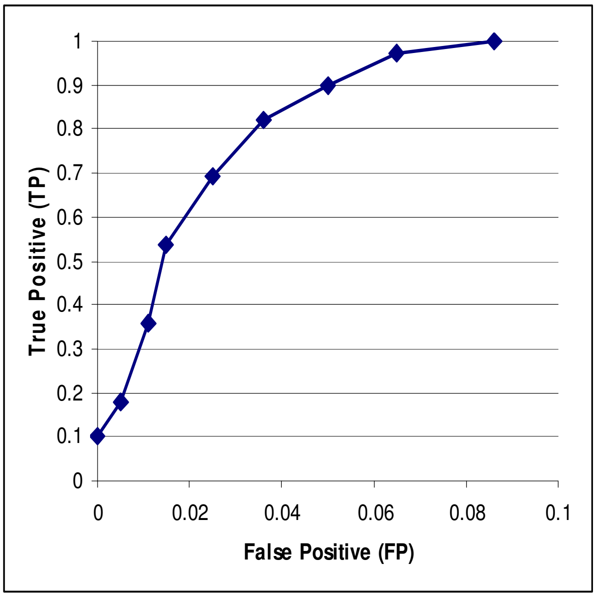


Figure III.20. ROC Curve for CBLOF Results on WBC Dataset

The failure of LOF algorithm in this experiment can be explained as the infeasibility of using the locality concept for detecting malignant samples in the data set. If a dataset having two main classes such as benign and malignant in its solution space can be solved by a decision plane and a discriminant function, the usage of LOF is not appropriate. In such datasets the algorithms having global aspects will be more successful as to be observed from the RPF results.

RPF algorithm leads to more meaningful results. While this algorithm depends on the distance measures, it can easily find the tuples that have different characteristics with respect to the others. But the selection of a threshold value plays a key role in evaluating RPF results, because smaller threshold may possibly trigger false positives, while a higher threshold may end up with false negatives. The threshold should be selected by taking the characteristics of the research area (domain information) into consideration.

The outlier objects found by using CBLOF algorithm also corresponds to malignant samples. But if we order the records according to their RPF values, the top 12% of the records cover all the malignant samples. The same value for CBLOF algorithm is 16%.Hence this analysis demonstrates that the RPF algorithm distinguishes the malignant samples more accurate than the CBLOF algorithm for the Wisconsin Breast Cancer dataset.

III.3.2. Meningoencephalidis Diagnosis Dataset

In this experiment, the Meningoencephalidis Diagnosis dataset is used in order to evaluate the interestingness of the patterns obtained from RPF, LOF and CBLOF algorithms. The dataset consists of 140 tuples with 38 attributes, but only 27 attributes that are used in diagnosis process is selected for our study. The eliminated attributes refer to the results achieved from the diagnosis and treatments that are provided to the patient. The numeric attributes are normalized to a 0 to 1 scale and the binary attributes are converted to 0 and 1 values.

All records in the dataset end up with a disease diagnosis, hence the contingency table approach used in III.3.1. is not appropriate for this experiment. Instead, the results derived from the interestingness measurements are evaluated by a domain expert. A similar evaluation strategy was also carried out in [57] by using the same dataset.

Another difference of this experiment is the addition of a revised version of CBLOF as a new measure into the experiments. As mentioned in the discussions in sections III.2.1 and III.2.2. the number of points in a cluster $|C_i|$ dominate the distance factor in the calculation, the result is affected by the coefficient in the CBLOF formula and favor the records near the large cluster at most. In order to minimize this effect, we suppose a revision on the CBLOF formula by replacing the |C$_i$| coefficient by |C$_i$|/β. At the rest of the document, this algorithm will be referred as R-CBLOF, where R represents the word "Revised". A preliminary trial of R-CBLOF is held on dataset 2. The results which exhibit the differentiation of the CBLOF values by this modification are illustrated in Figure III.21.

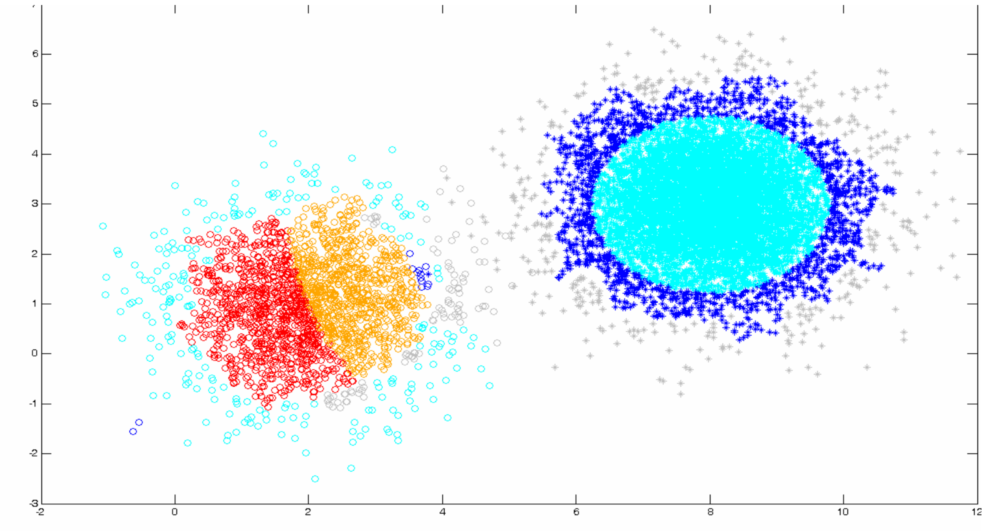


Figure III.21. R-CBLOF Values for Sample Dataset 2 with β=3

The results of this trial as illustrated in Figure III.21 exhibit the differentiation of the CBLOF values by this modification. When these results are compared with the results of the standard CBLOF algorithm given in Figure III.9, it can be observed that the CBLOF values of the denser cluster were decreased and no points from this cluster were regarded as interesting. On the other side the CBLOF values of the points that belong to the small cluster in terms of He's definitions increased. Since the philosophy of CBLOF is based on the identification of small clusters rather than the large ones, the R-CBLOF can be a good alternative to reduce the effect of |C$_i$| coefficient. After this encouraging result, we decided to add R-CBLOF to the experiments in order to test its effect on a real world medical data.

In the first phase of this experiment, the interestingness of all records is measured by RPF, LOF, CBLOF and R-CBLOF algorithms. For each record, the interestingness values obtained from the algorithms are classified in a scale of 0-5 according to Table III.1, where 4 and 5 points out the records that are identified as interesting.

Our objective in the evaluation phase is to find out the similar points between the results of objective measures and the subjective scoring of the domain expert. In order to realize this objective, a questionnaire was prepared by using 30 records. 12 out of the records were selected randomly from the tuples which were not marked as interesting by any algorithm. The other records were chosen by taking the interestingness model of the whole dataset into account.

The domain expert, who is a research assistant of infectious diseases in Baskent University, firstly analyzed the values of the attributes and she declared her subjective diagnosis for each record. In this stage, she is blinded to the real diagnosis. After finishing the subjective evaluation stage, the real diagnosis results were shared with the expert, and she was asked to grade the interestingness of the results in a 1-5 scale, where 1 refers to the least interesting, 5 refers to the most interesting patterns. The results of this experiment are illustrated at Table III.7.

The REC# column refers to the record number in the dataset. The RPF, LOF, CBLOF and R-CBLOF columns contain the interestingness class of the record according to the algorithm. The agreement column gives information about the coincidence of the objective and subjective diagnosis results, where "+" symbolizes the agreement and "-" symbolizes the disagreement. The interestingness column is the subjective interestingness value given in 1-5 scale by the expert. The expert did not graded the interestingness of 2 records, namely 2nd and 53rd ones, and they are regarded to be equal to the mean interestingness value (2.43) of the all questionnaire dataset.

Table III.7. Summarized Results of the Experiment on Meningoencephalidis Diagnosis Dataset

REC#	RPF	LOF	CBLOF	R-CBLOF	AGREEMENT	INTERESTINGNESS
1	3	5	1	4	-	5
2	2	3	4	2	+	
3	2	2	4	2	-	3
4	3	2	5	2	+	1
5	4	3	2	4	-	5
6	4	2	2	4	+	1
7	4	2	2	4	-	2
13	3	3	3	2	+	1
15	3	3	3	2	-	4
16	1	2	2	2	+	1
17	2	3	3	2	-	4
18	2	2	2	2	-	2
20	2	3	3	2	+	1
23	2	2	3	2	+	1
26	4	4	2	4	+	1
27	3	2	2	4	-	3
28	3	4	1	4	+	1
29	3	2	1	4	-	5
32	4	5	1	4	-	5
33	3	3	2	4	-	3
34	2	3	3	2	-	2
37	2	3	3	2	+	1
38	2	4	3	2	+	1
39	2	3	3	2	-	3
40	1	2	2	2	-	2
41	2	5	1	3	-	2
42	1	2	2	2	+	2
43	2	4	3	2	-	3
53	3	2	4	2	-	
64	3	4	4	2	-	3

Firstly the data is handled in 5 groups according to the measure(s) which identify it as an outlier. The Group-NI is reserved for the non-interesting records which are not declared as interesting by any measure.

Interestingness comparison between subjective evaluation and objective measurements can be done by different methods. Ohsaki et. al. in their study [22] used an hybrid measure namely CMC based on Pearson Correlation Coefficient and Precision Recall curves. In our study this mission is realized by using interestingness ratio and Manhattan distances.

The total interestingness is the maximum interestingness point that can be obtained by each group. While 5 is the maximum grade for an interesting pattern, total interestingness (TI) can be formulized as TI = 5 * Number of Records. Obtained Subjective Interestingness (OI) is the sum of the interestingness grades that are given to tuples in a specific group. An interestingness ratio is assigned to each group by dividing the obtained interestingness point to the maximum interesting point that is equal to the value. (IR=OI/TI) The results are illustrated in Table III.8.

Table III.8. Comparison of Subjective Interestingness Points

Group	Number of Records (NR)	Total Interestingness (TI=5*NR)	Obtained Subjective Interestingness (OI)	Interestingness Ratio (IP)
Group-NI	12	60	24.00	40.00%
RPF	5	25	14.00	56.00%
LOF	8	40	21.00	52.50%
CBLOF	5	25	11.86	47.44%
R-CBLOF	10	50	31.00	52.00%

These results demonstrate that the peculiar records identified by using RPF lead to the most interesting results according to the evaluations of domain expert. The proposed revision lead to an increase of nearly 10% w.r.t. CBLOF.

Another analysis was made by using the total distance between the objective interestingness measure and subjective evaluation. The same scale was used for both types of interestingness evaluation, hence no normalization was needed. The total Manhattan distances between each objective measure and subjective evaluation are given at Table III.9. Ohsaki et.al. only regarded the agreements and disagreements between the subjective and objective interestingness evaluations. In this study a more precise distance measurement between two evaluations are done by using a 1 to 5 scale in both methods.

Table III.9. Distances of the Measures from Subjective Evaluation

Interestingness Measure	Total Distance from Subjective Evaluation
RPF	33.00
LOF	35.00
CBLOF	43.14
R-CBLOF	33.86

These results demonstrate the power of RPF and LOF over CBLOF on this dataset. The R-CBLOF has a distance value which is 27.50% smaller than CBLOF.

Another area of interest on this study may be the relation between the agreement and subjective interestingness points. This investigation may bring up some clues that will be useful to discover the concealed logic behind the grading behavior of the expert. The mean of the interestingness values of the records which are not truly identified by the domain expert is 3.18. On the other side 10 records out of 12 are graded as 1 in the subjective evaluation. One record is graded as 2 and the last record's interestingness was not declared. This result may demonstrate that the expert's perception of interestingness is coupled with surprisingness concept which concentrates on the contradiction with the person's existing knowledge.

The standard deviation of the grades given to unidentified tuples is 1.16, which is respectively high w.r.t. mean which ends up with a coefficient of variation (CV) of 36.47%. While the identified tuples are usually characterized as "Ordinary", the source of the interestingness on this experiment can be reduced to this differentiation.

Since the subjective evaluation was implemented by using only one expert due to time and human resource limitations, the sufficiency and reliability of this experiment can be criticized. The reorganization of the subjective evaluation mechanism will be realized and cross-checking between the evaluations of two or more experts will be provided in future studies.

CHAPTER IV

CONCLUSION

In this study, the relation between the peculiarity and interestingness concepts has been investigated and some widely used algorithms, namely Local Outlier Factor (LOF), Cluster Based Local Outlier Factor (CBLOF) and Record Peculiarity Factor(RPF) were compared.

The results have shown that no method in this study's scope fully satisfies the needs of a user who is investigating interesting records with different characteristics in a dataset. Despite this fact, the experiments in synthetic datasets prove that the peculiarity oriented approaches are capable of detecting outliers defined from various aspects. The experiments carried out with real world medical data demonstrates the power of peculiarity oriented approaches in clinical domain. The algorithms used in this study do not only manage to capture the records which have outlier characteristics, but also they usually come up with interesting patterns that contradict with the user's current beliefs.

We suggest the following criteria to select an appropriate interestingness measure for a given dataset

- Statistical distribution of the data set
- The density differentiations in the dataset
- Existence and characteristics of noise
- The types and values of the attributes
- The objective(s) of the study held on the dataset

The statistical distribution of the data directly affects the results. The RPF algorithm regards the dataset as a whole, and focuses on the points which are located in the peripheries of the dataset. When the dataset comes from a normal distribution, this method produces proper results. But when the dataset exhibits a bimodal distribution or a Gaussian-Mixture, then the statistical approach and the lack of locality concept in RPF algorithm leads to the selection of peculiar records from the global peripheries. Even this selection is affected by the density of the different Gaussians, some points that may have some peculiarity in different perspectives can be missed. This situation can easily be seen in synthetic datasets 2 and 4.

Unlike RPF, the LOF algorithm mainly focuses on locality and it is very capable of the inspection of local differentiations in bi-modal distributions and clusters as can be seen in experiments 2 and 4. But in some real problems such as the Wisconsin Breast Cancer dataset the ignorance of the global perspective leads to inaccurate results.

The CBLOF algorithm is a hybrid approach, which takes into account both locality and globality in a data set for finding interesting records. The CBLOF algorithm is concentrated on the investigation of the small clusters that exhibit common characteristics symbolized by their distance to larger clusters. By using proper α and β values the effect of the distribution of data is minimized.

The existence and eliminability of noise is important in the process of interestingness measure selection. It is certainly known that the real world datasets are noisy in general. The important problem in peculiar data mining is to distinguish the peculiar data from the noise. Peculiar data and noise have very similar characteristics, especially if the definition of outlierness is laid on distance concept. The results of the third and fourth synthetic datasets demonstrate the lack of ability of the RPF algorithm in noisy datasets. It may only be useful in limited circumstances in particular when the users know that the data set does not have any noise or they can easily distinguish the noise from the dataset by using domain experts or data quality techniques.

Even in noisy conditions the LOF algorithm marks the peripheries of the clusters as interesting and very interesting, which can be observed in Figure III.12. If experiments 2 and 3 are examined together, that can be seen that the outliers captured in experiment 2 also were marked as outlier in experiment 3. The LOF values of noisy tuples added in Experiment 3 is respectively low as illustrated in Figure III.12. A few noisy tuples on the peripheries of the data set are found as interesting and very interesting according to LOF, but it is due to the high percentage (25%) of noise in the dataset. The noisy points also form dense regions by chance and some local outlier records are observed in the peripheries of such structures. The performance of LOF in separating noise and outlier is demonstrated also on the fourth dataset which is a combination of non-convex clusters and noise. (Figure III.17)

The CBLOF algorithm is not capable of coping with noisy datasets compared to LOF algorithm. The high majority of the interesting and very interesting points belong to noisy tuples. But unlike RPF, it marks some tuples from the peripheries of the densest cluster as interesting and very interesting, so it can be said that its ability of separating noise from peculiar data is higher than RPF. Besides the outliers are removed at the first stage of the algorithm, so the users can carry out their further analyses by ignoring these outliers.

The types and values of the dataset play an important role in selection of interestingness measures. This situation can definitely be observed on CBLOF data set. Firstly, the CBLOF algorithm's distance measures are based on support values [14]. As it is mentioned in [14], this measure can only be used for categorical attributes. Since the datasets contain numeric data, we have replaced this measure with Euclidean distance. In our experiments, this situation did not cause a major effect on the results of the algorithm, but this point may be an area of interest for further studies.

The $|C_i|$ coefficient in the CBLOF formula leads to a significant difference in the member points of the large and small clusters. If we recall the results of Experiment 3 with $\beta=3$, all interesting points except the noisy points belong to the densest dataset. The number of members in this cluster is 6250 and the other cluster comprises 1573 records. Let C_1 the densest cluster and $x_1 \in C_1$. Let C_2 is the second densest cluster and $x_2 \in C_2$. In this situation we have to make a comparison between the distance (x_1,C_1) and the distance (x_2,C_1), in which C1 is regarded as the only large cluster. In CBLOF calculation, we can easily observe that x_2 may have a higher CBLOF value than x_1 if and only if distance (x_2,C_1) is at least 4 times greater than distance(x_1,C_1). The value 4 in this proposal is the evidence of the fact that the CBLOF algorithm is affected from the real values of the attributes, especially when the continuous numeric data is used. In order to decrease the effect of cluster size in CBLOF calculation, we proposed a change in the formula by replacing the $|C_i|$ coefficient with $|C_i/\beta|$.

R-CBLOF, is mainly tested on the Meningoencephalidis Diagnosis dataset as represented in section III.3.2. The results show that R-CBLOF decreased the coefficient effect $|Ci|$ in the calculation and increased the accuracy of outlier detection in Meningoencephalidis Diagnosis dataset. The R-CBLOF or similar suggestions will be analyzed and improved on further studies.

For the last but not the least point, the selection of interestingness criteria, mostly depends on the user's subjective idea about the definition of interesting record in this context. If the characteristics of the interesting data, which is investigated in a given data set is known in advance, the selection of an appropriate method will be simpler. Breunig et. al. [13], used the the soccer dataset, which includes the information about German Football League players. Unlike distance based algorithms, LOF is concentrated on the records that deviate from its cluster mates. The inspection of the scoring goal-keeper Hans-Jorg Butt and Michael Schjönberg, who played only 15 matches and scored 6 goals despite the fact of being a defender demonstrates the power of LOF in such situations. But if the the user's perception about interestingness is somewhat different from Breunig et. al., the derived results from LOF may not be regarded as interesting in subjective evaluations.

The subjectivity issue is also valid for medical datasets. For example the results of the experiments held in this study on medical datasets usually point out the superior performance of the distance based measure RPF w.r.t. LOF and CBLOF. But this result cannot be generalized for all clinical datasets. As mentioned in Section II.4.1., medical data has unique characteristics which lead to some difficulties in selection of interestingness measures. In addition to this, "clinical data" term comprises various forms of data originating from different sources and having dissimilar characteristics. Hence, rather than making a generalization for the selection of interestingness measure, a methodology that accepts subjective noise and peculiar data definitions as an input and applies machine learning approaches to construct outlierness model, such as [36,37], may be more appropriate for the outlier detection process in medical data.

In addition to this, the medical domain is directly related with the human life and it is not appropriate to handle the medical data isolated from this reality. The social aspect of medical domain will drive the user to look for some other alternatives other than using objective approaches without any inquiry. For example in the experiment carried out with Wisconsin Breast Cancer Dataset, the objective of the study is very clear: identifying the malignant records in the sample. As a consequence, to make a selection between CBLOF and RPF, we have to take into account the motivations behind the study. By using a numeric approach, the usage of RPF with $\mu+2\sigma$ threshold seems to be the best solution for this dataset, but in the real world the user should be aware that the risk of identifying a malignant record as benign cannot be compensated. Originating from this approach, we have to select CBLOF or RPF with $\mu+2\sigma$ despite their poor performance which produces false positives.

Future Work

In this thesis, the comparison of LOF, CBLOF and RPF was made and their behaviors are demonstrated by using various synthetic and real world medical data sets. Their advantages and deficiencies were demonstrated in different statistical, experimental and social conditions. In addition to this, a revision of CBLOF algorithm, namely R-CBLOF was proposed in order to cope with the handicaps of CBLOF in handling numerical data. As a future work we have two directions: In order to improve cluster based peculiarity approach the relation between the clustering algorithms and CBLOF results can be inspected. As mentioned in III.1, we used the DBSCAN algorithm instead of Squeezer due to the existence of numerical data. Although He et. al. declared that the selection of the clustering algorithm can be made freely [14], our preliminary studies based on different clustering algorithms has resulted with some doubts on this issue. R-CBLOF presented in this study lead to an increase in the interestingness of identified outliers w.r.t. CBLOF. But its mathematical infrastructure must be analyzed in detail and the experiments should be repeated with more complex and representative datasets. The other direction may be to combine these three algorithms, which have different motivations and perceptions of peculiarity by using a multi-objective selection strategy. The composite z-score measure suggested by Shrestha et. al. [52] is an example of a similar approach, but the straightforward style of their calculation and the lack of user participation are the disadvantages of this study. Further studies focused on the treatment of these deficiencies will end up with more functional frameworks.

REFERENCES

[1].Fayyad, U.M, Piatetsky Shapiro, G., and Smyth,P., “Advances in Knowledge Discovery and Data Mining”, Chapter 1, Pages l-34. 1996.

[2].Freitas, A.A,“Are we really discovering the interesting knowledge from data?”, University of Kent, UK, 2005

[3].“The Grolier International Dictionary”,Grolier Incorporated, 1981

[4].Silberschatz, A and Tuzhilin, A, 1995, “On subjective measures of interestingness in knowledge discovery. ”, *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, pp. 275–281.

[5].Silberschatz, A and Tuzhilin, A, 1996b, “What makes patterns interesting in knowledge discovery systems”. *IEEE Transactions on Knowledge and Data Engineering* 8(6), 970–974.

[6].Sahar, S, 1999, “Interestingness via what is not interesting”, *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, CA*, pp. 332–336.

[7].Chen, X and Wu, Y.F., “Personalized Knowledge Discovery: Mining Novel Association Rules from Text”, from http//www.siam.org/meetings/sdm06/proceedings/067chenx.pdf

[8].Bastide, Y., Pasquıer, N., Taouıl, R., Stumme, G., AND Lakhal, L. 2000. “Mining minimal nonredundant association rules using frequent closed itemsets.”, *Proceedings of the Ist International Conference on Computational Logic*. London, UK. 972–986.

[9].Padmanabhan, B. and Tuzhilin, A. 2000. Small is beautiful: “Discovering the minimal set of unexpected patterns.”, *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*(*KDD 2000*). Boston, MA. 54–63.

[10].Freitas A.A., “On Objective Measures of Rule Surprisingness”, *Proc. Second European Symp. Principles of Data Mining and KnowledgeDiscovery (PKDD ’98)*, pp. 1-9, 1998.

[11].“Oxford Dictionary”,Claderon Press, Danbury, Connecticut, 1993

[12].Zhong, N., Yao, Y. Y., and Ohshima, M. “Peculiarity oriented multidatabase mining.”, *IEEE Transactions on Knowledge and Data Engineering* ,Vol.15, 4, 952–960.,2003

[13].Breunig, M. M., Kriegel H.-P., Ng, R. T. and Sander, J. , “LOF: Identifying Density-Based Local Outliers”, *Proceedings of the ACM SIGMOD Conference*, 2000

[14].He, Z., Xu, X. and Deng, S. “Discovering Cluster Based Local Outliers.”, *Pattern Recognition Letters*, 2003, 24 (9-10): 1651-1660.

[15].Liu, B., Hsu, W. and Chen, S, “Using General Impressions to Analyze Discovered Classification Rules.” *Proceedings of the Third International Conference on KnowledgeDiscovery and Data Mining (KDD 97)”,* pp. 31-36,1997

[16].Ram, A., “A Knowledge goals: A Theory of Interestingness.”, *Proceedings of the 12th annual conference of the cognitive science society*, pages 206–214, Cambridge, 1990

[17].Piatetsky-Shapiro, G. and Matheus, C.J., "The Interestingness of Deviations", *Proceedings AAAI* '94 *Workshop Knowledge Dzscovery in Databases,* pp 25-36,1994

[18].Tuzhilin, A. and Silberschatz, “A Belief-Driven Discovery Framework Based on Data Monitoring and Triggering”, 1996

[19].Agrawal, R. and Srikant, R., “Fast algorithms for mining association rules.”, *Proceedings of the 20th International Conference on Very Large Databases.* Santiago, Chile. 487–499, 1994

[20].Webb, G. I. and Brain, D., “Generality is predictive of prediction accuracy.”, *Proceedings of the 2002Pacific Rim Knowledge Acquisition Workshop* (*PKAW 2002*). Tokyo. 117–130, 2002

[21].Tan, P., Kumar, V., and Srivastava, J., “Selecting the right interestingness measure for association patterns.”, *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining* (*KDD 2002*). Edmonton, Canada. 32–41, 2002

[22].Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., And Yamaguchi, T., “Evaluation of rule interestingness measures with a clinical dataset on hepatitis.”,*Proceedings of the 8th European Conference on Principles of Data Mining and Knowledge Discovery* (*PKDD 2004*)., Pisa, Italy. 362–373, 2004

[23].Knorr, E.M. and Ng, R.T., “Algorithms for mining distance based outliers in large datasets”, *Proceedings of VLDB_98*, New York, USA, pp. 392–403,1998

[24].Zhong, N., Liu C., Yao, Y.Y., Ohshima, M., Huang, M. and Huang, J. , “Relational Peculiarity Oriented Data Mining”, *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM’04)*, Brighton, UK, 2004

[25].Hawkins, D.: “*Identification of Outliers*”, Chapman and Hall, London, 1980.

[26].Kubica, J. And Moore A., “Probabilistic noise identification and data cleaning”,*Proceedings of ICDM*, FL, USA., 2003

[27]. Zhu, X., Wu, X. and Chen Q. , “Eliminating class noise in large datasets”, *Proceedings of 20th ICML*, Washington D.C., USA, 2003

[28].Gamberger, D., Lavrac, N. and Dzeroski, S., “Noise Detection and Elimination in Data Preprocessing: experiments in medical domains.”, *Applied Artificial Intelligence 14*, 205-223, 2000

[29].Bolton, R. J. and Hand, D. J., “Statistical fraud detection: A review (with discussion).”,

Statistical Science, 17(3):235-255, 2002.

[30].Zhang, J. and Zulkernine, M., "Network Intrusion Detection Using Random Forests", *Proceedings of the Third Annual Conference on Privacy, Security and Trust*, St. Andrews, New Brunswick, Canada, October 2005.

[31].Zhong, N., Yao, Y.Y.,Ohshima M. and Ohsuga, S., "Interestingness, Peculiarity, and Multi-Database Mining", *Proceedings. IEEE International Conference. Data Mining (ICDM '01)* ,566-573 2001.

[32].Suzuki, E. and Kodratoff Y., "Discovery of Surprising Exception Rules based on Intensity of Implication", *Principles of Data Mining and Knowledge Discovery (PKDD'98), LNAI1510,* Springer, Berlin, 10-18, 1998

[33].Barnett, V., Lewis, T., "Outliers in Statistical Data",John Wiley and Sons, New York, 1994.

[34]. Harkins, S., He, H., Willams, G.J., Baster, R.A., "Outlier detection using replicator neural networks.", *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, Aix-en-Provence, France, 170–180, 2002

[35].Lazarevic, A., Ertoz, L., Ozgur, A., Srivastava, J. and Kumar, V., "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings. Of SIAM Conf. Data Mining*, 2003.

[36].Yamanishi, K., Takeuchi, J., Williams,G.J., and Milne, P.,"On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms.", *Data Mining and Knowledge Discovery*, 8(3):275-300, 2004.

[37].Zhu, C., Kitagawa, H., Papadimitriou, S. and Falautsos, C., "OBE:Outlier by Example", *Proceedings PAKDD*, 222-234, 2004

[38].Papadimitriou, S., Kitagawa, H., Gibbons, P.B. and Faloutsos, C., "LOCI: Fast Outlier Detection Using the Local Correlation Integral. ", *Proceedings ICDE,* 315-326, 2003.

[39].Grubbs, F.,E.,"Procedures for Detecting Outlying Observations in Samples.", *Technometrics*,11: 1–21, 1969

[40].Eskin, E., "Anomaly detection over noisy data using learned probability distributions. *Proceedings of the International Conference on Machine Learning*, 2000.

[41].Ruts, R., Rousseeuw, P., "Computing depth contours of bivariate point clouds.", *Journal of Computational Statistics and Data Analysis*,23,153–168, 1996

[42].Ramaswamy, S., Rastogi, R., Kyuseok, S., "Efficient algorithms for mining outliers from large data sets.", *Proceedings of SIGMOD_00*, Dallas, Texas, pp. 93–104, 2000

[43].Bay, S. D. and Schwabacher, M., "Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, August 24-27, 2003

[44]. Ng, R. T. and Han, J., “Efficient and effective clustering methods for spatial data mining.”,*Proceedings of 20th International Conference on Very Large Data Bases*, 144–155, Santiago de Chile, Chile, 1994.

[45].Ester M., Kriegel, H., Sander, J. and Xu, X. “A density-based algorithm for discovering clusters in large spatial databases with noise.”, *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231, Portland, Oregon, 1996.

[46]. Zhang, T., Ramakrishnan R. and M. Livny. “BIRCH: an efficient data clustering method for very large databases.”, *Proceedings of 1996 ACM SIGMOD International Conference on Management of Data*, 103–114,Montreal, Quebec, Canada, 1996.

[47].Ankerst, M., Breunig, M. M., Kriegel, H.P.and Sander, J., “OPTICS: Ordering points to identify the clustering structure.”, *Proceedings of ACM SIGMOD International Conference on Management of Data*, 49–60, Philadephia, Pennsylvania, U.S.A., 1999.

[48].Jiang, M.F., Tseng, S.S., Su, C.M.,“Two-phase clustering process for outliers detection.” *Pattern Recognition Letters 22* (6/7), 691–700, 2001.

[49]. Agyemang, M. and Ezeife, C.I., “LSC-Mine: Algorithm for Mining Local Outliers.” *Proceedings of the 15th Information Resource Management Association (IRMA) International Conference*, New Orleans, 5-8, 2004

[50]. Tang, J., Chen, Z., Fu, A.W., Cheung D.W., “Capabilities of outlier detection schemes in large datasets, framework and methodologies”, *Knowledge and Information Systems* 11(1),45–84, 2006

[51].Duan L., Xu L., Guo, F.,Lee J. and Yan. B., “A local-density based spatial clustering algorithm with noise”, *Information Systems* ,32 ,978–986, 2007

[52].Shrestha, M., Hamilton, H.J., Yao, Y.Y., Konkel, K., and Zhong, N., “The PDD Framework for Detecting Categories of Peculiar Data”, *Proceedings of the Sixth International Conference on Data Mining*, 562-571, 2006.

[53].Cios, K. J. and Moore G. W., “Uniqueness of medical data mining.”,*Artificial Intelligence in Medicine*, vol.26, no.1–2, 1–24, 2002

[54]. Roddick, J. F., Fule, P. and Graco, W. J., “Exploratory medical knowledge discovery : Experiences and issues”, *SigKDD Explorations* 5(1), 2003

[55]. Roddick J.F . and Lees, B.G ., “ Paradigms for spatial and spatio-temporal data mining.”, *Geographic Data Minting and Knowledge Discovery*, *Research Monographs in Geographic Information Systems.* Taylor and Francis, 2001 .

[56]. Shillabeer, A. and Roddick, J.F., “Reconceptualising interestingness metrics for medical data mining.”, *Australian Workshop on Health Data mining, http://acrc.unisa.edu.au/groups/health/hdw2005/Shillabeer.pdf* , 2005

[57].Ohsaki, M., Abe, H., Tsumoto, S., Yokoi H., Yamaguchi,T.,“Evaluation of rule interestingness measures in medical knowledge discovery in databases”, *Artificial Intelligence in Medicine* ,41, 177—196, 2007

[58]. Ohsaki, M., Kıtaguchı, S., Okamoto, K., Yokoi, H., And Yamaguchi, T., “Evaluation of rule interestingness measures with a clinical dataset on hepatitis.” *Proceedings of the 8th EuropeanConference on Principles of Data Mining and Knowledge Discovery* (*PKDD 2004*). Pisa, Italy. 362–373., 2004

[59]. Hatazawa H, Abe H, Komori M, Tachibana Y, Yamaguchi T., “Knowledge discovery support from a meningoencephalitis dataset using an automatic composition tool for inductive applications.”, *Post-Proceedings of the joint JSAI-2001 workshop on new frontiers in artificial intelligence.Lecture notes in artificial intelligence,* vol. 2253., Berlin: Springer;. 500-507, 2002

[60]. Carvalho DR, Freitas AA, Ebecken N., “Evaluating the correlation between objective rule interestingness measures and real human interest.”, *Proceedings of the 16th European conference on machine learning and the 9th European conference on principles and practice of knowledge discovery in databases ECML/PKDD-2005. Lecture notes in artificial intelligence,* vol. 3731. Berlin: Springer; 453-461, 2005

[61]. Laurikkala, J., Juhola, M. & Kentala, E., “Informal Identification of Outliers in Medical Data.”,*Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000* , Berlin, Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000, 2000

[62].Podgorelec, V., Herizko, M., and Rozman, I., “Improving Mining of Medical Data by Outliers Prediction”, *Proceedings of 18th IEEE International Symposium on Computer-Based Medical Systems*, Trinity College Dublin, Ireland, June 23-24, 91-96, 2005

[63].Suzuki, E., and Tsumoto, S.,”Evaluating hypothesis-driven exception-rule discovery with medical data sets.” *Knowledge discovery and data mining.Lecture Notes in Artificial Intelligence* 1805. Berlin: Springer; 86–97,2000.

[64]. Suzuki, E.,"Scheduled Discovery of Exception Rules", *Discovery Science (DS'99), LNAI nZl,* Springer, Berlin, 184-195,1999

[65]. Tsumoto S. "Comparison of Data Mining Methods using Common Medical Datasets", *ISM Symposium: Data Mining and Knowledge Discovery in Data Science,* The Inst, of Statistical Math., Tokyo, 63-72, 1999.

[66]. MEDx 3.4 : http://medx.sensor.com/products/medx/overview.html

[67]. MATLAB 6.5: http://www.mathworks.com/, December 2006

[68]. Erdem, C., “Density Based Clustering Using Mathemetical Morphology” Unpublished *M.S. Thesis,* Middle East Technical University The Graduate School of Informatics,2006

[69]. The R Project for Statistical Computing: http://www.r-project.org/

[70]. M.E.Otey, A. Ghoting, and S. Parthasarathy., “Fast Distributed Outlier Detection in Mixed-Attribute Data Sets *Data Mining and Knowledge Discovery Vol.12,* Springer, 2005, pp.203-228