COMPARISON OF DIMENSIONALITY REDUCTION TECHNIQUES APPLIED TO DRIVING EVENT DATASET

(SÜRÜŞ OLAYLARI VERİ SETİNDE BOYUT İNDİRGEME TEKNİKLERİNİN KARŞILAŞTIRILMASI)

by

Can ÇETİN, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

October 2016

ACKNOWLEDGEMENTS

Firstly, I would first like to thank my thesis advisor Prof. Dr. Tankut ACARMAN for the continuous support of my Master study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

Also I thank to my colleagues from VASUP Bilişim supporting me and my Master education. And I would like to thank to Infotech Information and Communication Tecnologies Ltd. For providing me the dataset obtained from more than 3.000 vehicle's navigation unit.

Last but not the least, I would like to thank to my family for supporting me spiritually throughout writing this thesis and my life in general.

October 2016
Can Çetin

TABLE OF CONTENTS

LIST OF SYMBOLS

API	**:** Application Programming Interface
MSE	**:** Mean Squared Error
PCA	**:** Principal Component Analysis
JVM	**:** Java Virtual Machine
KDP	**:** Knowledge Discovery Process
DM	**:** Data Mining
FSS	**:** Feature Subset Selection
FE	**:** Feature Extraction
WEKA	**:** Waikato Environment for Knowledge Analysis
GUI	**:** Graphical User Interface

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

When investigating on problematical and indefinite areas with data exploring tools such as machine learning or data mining algorithms, weight of data attributes effecting classification result is generally unknown issue. Using entire feature set might cause the low classification success. Dependency existence among features, (near) zero variance features, outlier and missing data on feature may harm classification accuracy. To increase classification success and learn feature effect on classification, dimension reduction techniques such as feature subset selection and feature extraction are used.

To estimate driver habits leading to car accidents and make robust machine learning algorithms, dimension reduction methods are applied to event driver dataset. Driving event dataset contain sudden harsh breaking and acceleration, harsh left and right hand cornering, link speed exceeding etc.., Global Position System (GPS) measurement, speed and time information. Events are thowed on each second or according to event occurance. Driving event dataset are enriched with Traffic Information Centers' (TRAMER) accident data that include accident time, accident cost, crash type, replaced auto parts etc.,. Even if dataset is discrete, for 600 drivers, there are more than 400 million record per month. Because of huge data per driver and high dimensional spaces, feature subset selection and feature extraction methods are applied to find more robust, high accurate results and most affecting habits that cause to car accident.

Feature subset selection and feature extraction are the two different applied methods for reducing the dimension set. While feature subset selection methods is focusing to find the most important features that affect the classification result, feature extraction methods are dealing with the creating of new attributes as a linear or non linear combination of initial feature set. Both methods are used on the investigation of classification, clustering and regression problems. Feature extraction methods sacrifice

the explanation of the problems, when they combine the existent features to create new ones. On the other hand, features subset selection methods help to pick the most important features by ordering attributes according to their ranking methods. If classification researches are not satisfying or contribution of the attributes that affecting the classification is not known deeply, both methods can be used to understanding the importance of the attributes, and increase classification accuracy.

Principal Component Analysis (PCA) is used to implement feature extraction on driving event dataset. PCA helps to reduce initial feature dimension size by the creation of new uncorrelated features that are linear combination of the initial feature set. Extracted features are not depending on the machine learning algorithms and ordered according to their contribution. In our research, PCA generally helps to increase on the classification accuracy and generally produce robust attributes.

For feature subset selection, filter and wrapper methods are applied to driving event dataset. Information gain (IG) and Support Vector Machines for attribute evalution (SVM attr. eval.) are used for filter methods. Filter methods considers the relation between attributes, covariance, variance etc.. and ranks the attributes according to contribution to the methods. At each step, one attribute which is ranked as minimum importance according to filter methods, is removed from the initial feature dataset and evaluate the classification results. Main aim is to achieve maximum accuracy and minimum feature set. If two different feature sets give the same accuracy, feature set with less features is selected. Driving event dataset produce more increasing accuracy or less feature set than the initial feature set.

As a wrapper technique, genetic algorithm (GA) is implemented to solve the feature subset selection problem. Each feature is represented as a bit string and one gene. 1 indicates that feature is included and 0 indicates feature is excluded. Decision of the feature selection based on evolutionary process such as mutation probability, mutation strategy, population count, evolution probability, initial population, fitness funtion etc... Although the main problem of wrapper techniques are the huge computational costs, the best results are gathered from this approach.

Driving event data set is transformed into penalty scores of the drivers to apply machine learning algorithms. Although the results after applying the dimension reduction techniques results were acceptable in accuracy but far from the robust system. After changing the penalty scoring system to find the drivers usage routines such as link speed exceeding habits, usage of too much break event etc. system accuracy are increased to over %80.

ÖZET

Veri madenciliği veya makina öğrenmesi gibi veri inceleme yöntemleriyle kesin çözüm getirilemeyen belirsiz alanlarda çalışmalar yapılırken, veri özelliklerinin sınıflandırma üzerindeki etkileri genellikle bilinmez. Verinin araştırılan tüm özellik kümesini kullanmak ise genellikle düşük sınıflandırma başarımına yol açabilir. Özellikler arasındaki bağımlılık, özelliğin sıfır veya düşük varyansa sahip olması, özelliğin üzerindeki aykırı uç veya kayıp veriler yüzünden sınıflandırma başarıları düşebilir. Bu gibi durumlarda sınıflandırma başarımını arttırmak için veri indirgeme teknikleri kullanılmaktadır.

Kazalara neden olan sürücü alışkanlıklarını bulmak ve sağlıklı sınıflandırma tahminlemesi yapmak için veri indirgeme teknikleri kullanılmıştır. Bu bağıntı için sürücü alışkanlıklarını barındıran, üzerinde sefer bilgisi, ani yavaşlama, ani hızlanma, sola savrulma, sağa savrulma, tümseğe hızlı girme, hız ve konum bilgilerini bulunduran olay tabanlı kaza olay veri seti kullanılmaktadır. Bu bilgiler olay oluşunca veya saniyede bir yollanmaktadır. Veri seti üzerindeki sürücünün kullanım alışkanlıkları ve ceza bilgisi, Trafik Bilgi Merkezi (TRAMER) üzerindeki kaza bilgisi ile birleştirilip, veriler anlamlandırılmaya çalışılmıştır. Veri seti ayrık yapılandırılmış olsa da, yaklaşık 600 sürücü için kullanım sürelerine bağlı olarak aylık 400 milyonun üzerinde veri üretmektedir. Çalışma boyunca veri sayisi ve özellik kümesi fazlalığından, sınıflandırma başarımı çeşitli nedenlerde düşmektedir. Kazaya etki eden dinamiklerin bulunması ve sınıflandırma başarımını arttırmak için, veri madenciliği öncesi veriyi işlemeden önce, özellik boyut indirgemesi yöntemlerinden, öznitelik arama veya özellik alt küme seçimi metodları kullanılmaktadır.

Özellik alt küme seçimi ve boyut indirgemesi, özellik uzayındaki özellik kümesini indirgemek için kullanılan iki farklı yöntemdir. Özellik alt küme seçiminde, daha etkin

öznitelikler bulunurken, boyut indirgemesinde orjinal özellik kümesinden lineer veya lineer olmayan kombinasyonlar yardımıyla yeni öznitelikler üretilir. İki yöntemde sınıflandırma, gruplama ve regresyon problemlerini incelemede kullanılmaktadır. Boyut türetme yöntemi, hali hazırdaki özelliklerden yeni öznitelikler oluşturarak problemin tanımlanabilirliğinden fedakarlık gösterir. Özellik alt küme seçimi yöntemi ise, en iyi özellik kümesi seçmek için, daha az önemli özellikleri eleme yöntemini seçer. İki yöntemde sınıflandırma başarımını en fazla etkileyen özellikler ya aynen kullanarak yada onlardan yeni özellikler oluşturarak, yeni azaltılmış özelllik kümesi yaratmayı amaçlar. Bu yöntemler istenen sınıflandırma başarısı için yeterli değilse, sınıflandırma başarısını arttırmak için veya veri uzayında sınıflandırma için önemli parametreleri bulup onları daha fazla anlamlandırmak için kullanılmalıdır.

Öznitelik arama yöntemi olarak ise temel birleşenler analizi yöntemi kullanılmıştır. Temel birleşenler analizi, eldeki veri seti üzerinden, özellik kümesinin lineer yöntemler yoluyla birbirleri arasındaki korelasyonu sıfırlayarak, özellik kümesinin sayısı azaltmak ve verinin daha az özellik kümesiyle gösterilmesini sağlamaktadır. Kaza olay veri setinde, temel bileşelenler analizi ile bazı durumlarda sınıflandırma başarımı önemli oranda artsa da, bazı durumlarda bu başarım azalmıştır.

Özellik alt küme seçimi için filtre yöntemi ve kapsayıcı yöntemleri kullanılmıştır. Filtre yöntemi olarak temel birleşenler analizi ve bilgi kazanımı yöntemleri, eldeki veri seti üzerinden özellikler arasındaki birbirlerine bağımlılık ve sonuca etki etme derecesi gibi etkenlere göre verileri önem sırasına göre düzenlemektedir. Olgunlaştırılmamış araştırmalarda, özelliklerin eldeki veriler üzerinden çıkarılan önem sıralaması, en önemsiz özelliğin sonuca etkisinin az olduğunu belirtmektedir. Biz bu sıralandırmayla, yani en önemsiz özellikten en önemli özelliğe doğru her bir özelliği çıkarıp, sınıflandırma başarımını ölçerek veri setini azaltma ve daha etkin veri kümesi bulmayı amaçladık. Verileri sırayla çıkarırken elde ettiğimiz sonlanma fonksiyonu, daha iyi sınıflandırma başarımı elde etmeme durumudur. Örneğin 12 elemanlı ve 15 elemanlı veri setinde de aynı başarımı elde ettiysek, daha az olan 12 veri setini seçtik. Bizim çalışmalarımızda veri seti üzerindeki çalışmalarımızda, sınıflandırma başarımında artış sağladık.

Kapsayıcı yöntemler adı altında ise genetik algoritma yöntemiyle özellik alt küme seçimi yöntemi uygulanmıştır. Kapsayici yöntemler, özellik kümesinden alt küme seçerken, makina öğrenmesi algoritmalarını kullanırlar. Özellik alt küme seçimi probleminin evrimsel bir yapı kazandırmak için her bir geni bir özellik ile gösterip o genin dahil edilmesini 1 edilmemesini ise 0 ile gösterdik. Her bir özellik populasyona dahili edilip edilmeme kararı mutasyon olasılığı, mutasyon stratejisi, algoritmanın populasyona birey seçimi stratejisi, evrim olasılığı, iterasyon sayısı, başlangıç kümesi, seçilen uyumluluk fonksiyonu gibi parametrelere göre belirlenir. Genetik algoritma sonunda çıkan sonuç o populasyonun en iyi grubu olabilir ama her çalıştırmada farklı sonuçlar vermektedir. Kapsayıcı yöntemler kaza olay veri setinde, genellikle olduğu gibi, filtre yöntemlerinden daha yüksek başarımı sağlamaktadır, ama işlem zamanı çok daha fazladır.

Olay kaza veri seti üzerindeki çalışmalar sırasında öncelikle ceza tabanlı özellik kümesi üzerinden çalışmalar yapılmıştır. Veri indirgeme yöntemleri ile başarım sağlansa da bunu arttırmak için veri seti sürücü kullanım saatlerine göre, hız limitini geçme sıklıklarına vb. sürüş alışkanlıklarına göre yeniden biçimlendirilmiştir. Bu ikinci özellik kümesinde ise öncekine oranla sürücü bazında daha tutarlı sonuçlar alınmış ve %80 üzerinde sistem başarımı elde edilmiştir.

1. INTRODUCTION

Knowledge Discovery Process (KDP) extracts information from any raw data. Anil (2011) explained the KDP as a searching data patterns automatically, and creating useful knowledge from data. When exploring the unsolved and indefinite classification problems, KDP services can help people to make informed decisions based on patterns observed in collected data.

There are various steps that are involved in KDP. Soundararajan et al., (2005) explained the KDP according to following steps:

- *Problem Domain:* Prior knowledge before dealing the problem
- *Target Data Set:* Data set selection on which KDP is applied
- *Data Processing:* Missing values and outliers handling strategy is defined.
- *Data Reduction and Projection*: Finding the most useful features by reducing existing feature sets or creating new features from the original features.
- *Selecting the Data Mining (DM) Function*(s): According to features of original data, techniques such as classification, clustering, regression are selected.
- *Selecting the DM Algorithm(s):* Selecting the appropriate methods related to data. Methods such as decision tree, regression analysis, statistical algorithms, time series analysis can be applied on the target data.
- *Applying DM :* Finding the data patterns.
- *Visualization and Interpretation:* After DM process, visualization and interpretation steps are needed to investigate the created patterns. Redundant patterns are removed and according to success criteria, KDP process can be restarted.
- *Using the Knowledge:* If success criteria is met, created new information is used by taking actions.

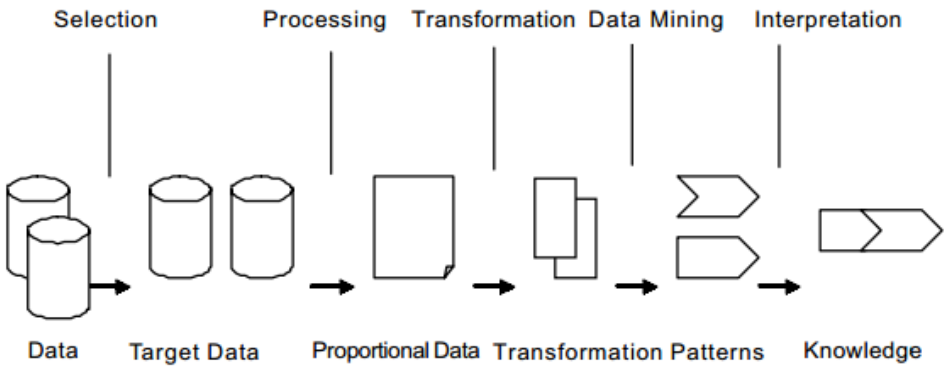

Figure 1.1: Knowledge Discovery Process (Anil 2011).

Thanks to advance in technology, data scientists can find huge amount of features on any spesific classification problem. When investigating new problems, due of lack of domain spesific knowledge, people tends to increase feature set on any classification task. Contrary to general opinion, increasing number of features on the data exploring problems might cause to poor classification performance, Besides from that, waste of computational resources are the another result of huge amount of features.

The studies in the literature showed that when feature dimensions enlarge linearly, sample size for training phase must be enlarged exponentially. (Dash et al., 2008). Figure 1.2 shows an example of the curse of dimensionality. Dash et al., (2008) also explained that as the number of dimension increases, the mean square error also increases.

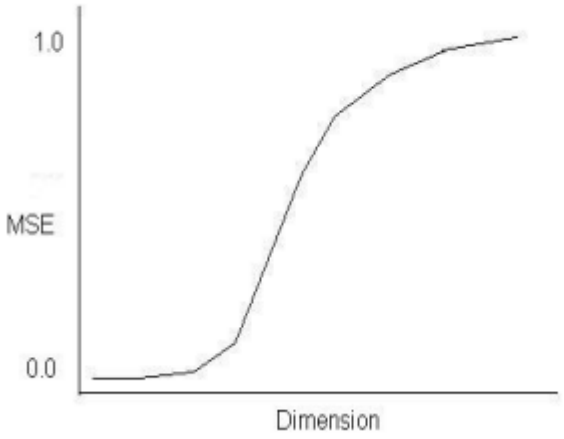

Figure 1.2: Curse of dimensionality (Dash et al., 2008).

Huge amount of dimensions generally include irrelevant features that might cause to curse of dimensionality phenomenon. When classification problem does not fit in the input attribute set, Dimensionality Reduction (DR) process becomes mandatory task. Manoranjan et al., (2008) explained that the existance of redundant, irrevelant and zero variance or near zero variance features might cause the data over-fitting problem, increase computational costs. He also added that, DR is the effective solution of curse of dimensionality phenomenon.

Other important disadvantage of the huge amount of dimensions are the waste of computational resources. Generally, computation time is tightly related to dimension space when dealing with KDP. Rosaria et al., (2014) explained that, the computational time of Backward Elimination Method with 231 attributes cause the unserviceable execution time.

Dunja explained the main reasons of applying the DR techniques as improving prediction success and increasing learning efficiency, providing faster responses by using less features from the initial feature set. DR techniques are not only needed to decrease waste of computational resources or curse of dimensionality problem causing to over-fitting issue but also help to improving prediction performance and learning efficiency.

Attribute subset selection and attribute extraction are the two different approachs of DR techniques that reduce the attribute space of feature set. While in feature subset selection (FSS) method, a subset of the original attributes is selected, attribute extraction method generates linear or non-linear combinations of the original attribute set.

Driver's driving dynamics that cause to accident event is a perfect example of totally unsolved problem. The work undertaken in this thesis is to implement different DR techniques to find the relation between driver and accident events on indefinite area. Captured real time accident events and drivers with the historical accident records (TRAMER) are the used for classification criteria. Driver's all driving events are collected from boards, that are entegrated over 3000 cars. Boards are programmed to throw events on any spesific driving events such as harsh acceleration, harsh break, ignition on, ignition off and speed limit exceeded warning. Over 40 features that are based on drivers' driving characteristic are carefully selected and examined for detecting driving patterns that lead to car accident.

It is obvious that, on some issues driver is not directly responsible of car accident event. Measuring the car accident according to costs and finding relation with driving dynamics probably may not give us the desired relation on some occasions. Other obvious thing is the drivers' driving patterns that lead to crash event or events which are converging to some indefinite patterns.

Our aim is to apply the DR methods and evaluate the relation between driver's accident probability and driving dynamics. Classification accuracy is selected as a evoluation method. 10 fold cross validation technique is used to evaluation of the accuracy.

2. LITERATURE REVIEW

Although there are significant amount of research implemented on DR methods, the main difference in the DR methods are the different categorization of issue. Manoranjan et al. [2008] briefly explained the different categorisation techniques as feature selection or feature extraction (FE), linear or nonlinear, supervised or unsupervised, and local or global. The first method is the same as this thesis's approach. In linear methods, original features are combined to create new ones by linear mapping. Like linear methods, non-linear methods use non-linear mapping to create new features. While supervised methods consider class information on feature selection phases (classification), unsupervised methods do not have this class information (clustering). Local methodologies try to find class features for each category separately. Global methodologies select class features by taking into consideration of all categories. All of these terms are frequently encountered on DR methodologies.

In the literature, creation of filter methods algorithms which is the sub brach of FSS methodologies, are detaily discussed in different perspectives. Avrim et all [6], explained the idea of creating filter methods as a suitable example of solving heuristic search methods and also approaches to the all FSS steps as the heuristic search methods. Four steps are required to be fulfill the heuristic search algorithms to apply to any FS problems. First item is the starting point of the domain space which affects the search operators and direction. One approach is initiated with no attributes and add attributes according to creation of successful states. Another approach is starting with adding all attributes to search space and remove them according to successful criteria of the methodologies. First method is known as forward selection and the second method is named as backward elimination.

A second step to formulize the filter selection techniques is the selection of search techniques. Because of huge computational costs, search spaces are generally not proper for brute force approaches. Greedy algorithms which try to find locally optimum strategy are suitable brach of the search techniques. As an instance of greedy algorithms, hill climbing algorithm that add or remove attribures on stepwise selection or elimination phases and select the best subset according to success criteria perfectly fits the search algorithm conditions. Not only greedy algorithms fulfill the needs, but also best first search algorithms which have more computational costs than greedy algorithms are useful for search techniques.

Finding the proper strategy to evaluate the all alternative feature subsets is the third step. Although many of them is criterion based on training data set, but others are only focusing on measuring of accuracy on training set and evaluation.

Finally, halting criteria must be made for the search algorithms. For example, attribute adding or removing are unneccesary when none of the options improve the success criteria. Halting condition is activated when search space is completed or none of the options improve the success criteria or more of options degrade the accuracy. Table 2.1 gives the example classification of well known DR methods.

Table 2.1: The Process of Feature Subset Selection (Avrim et al. 1997)

	Starting Point	**Search Control**	**Halting Criteria**	**Indıction Algorithm**
Almuallim (FOCUS)	None	Breadth first	Consistency	Dec. tree
Cardie	None	Greedy	Consistency	Near. neigh.
Koller and Sahami	All	Greedy	Threshold	Tree./Bayes
Kira and Rendell(RELI	-	Ordering	Threshold	No

EF)				
Kubat et all.	None	Greedy	Consistency	Naive Bayes
Schlimmer	None	Systematic	Consistency	Yes
Singh and Provan	None	Greedy	No info. gain	Bayes Net

The general thought is about wrapper method that it costs too much computational resources because it use induction algorithms at each search steps and compare the results. Although computational resources are limited, it is not true that wrapper methods are not suitable for real time applications. Zamalloa et al (2012), compared the Genetic Algorithm (GA) with SVM as a wrapper method with PCA and linear dicriminant analysis (LDA) as FE methods and stated that, optimization is a batch job and recognition is an online job. Recognition with proposed the GA method take less time than FE methods' result.

Feature exploring with GA has been researched widely in the literature. Haleh Vafaie eet al (1994), compared greedy like search algorithms with GA and retrieved better results from GA based feature selection. D. Asir Antony Gnana Singh et al (2016), compared GA with different mutation techniques like one point, and operator, xor operator, or operator and evaluate the medical diognasis accuracy by using IB1, Naive Bayes and KNN ML algorithms.

3. DIMENSIONALITY REDUCTION

Dimensionality Reduction (DR) is dealing with elimination of irrevelant variables and reducing feature set. It helps to increase classification accuracy and also helps to learn the attributes have the most contribution on the classification result.

Classification, clustering, regression like problems are suitable for all DR methods. In this thesis, driving event dataset as a classification problem is handled to investigate the effects of DR.

DR methods can be divided into FSS and FE methods. Both methods's effects are examined on the driving habit dataset.

3.1. FEATURE SUBSET SELECTION

Feature Subset Selection is the popular branch of DR methods. For FSS methods, features are ordered according to their weighting algorithm. Number of initial input parameters are shown by *d*. FSS methods choose the most valuable *k* parameters to use in induction algorithm by increasing or descreasing order where the $k<d$ condition is satisfied. Therefore *d-k* parameters are not included to be used on machine learning algorithms anymore.

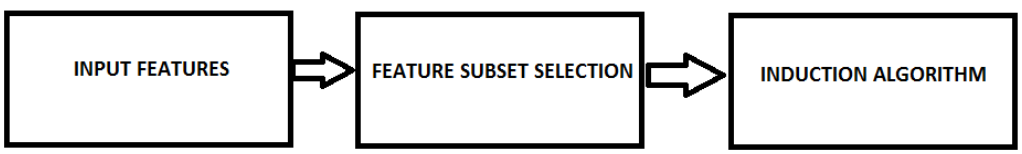

Figure 3.1 The Feature Subset Selection (Kohavi et al. 1997)

The FSS methods help to reduce the computational time complexity and provide more simpler explanations of the problem.

Janecek et al. (2008) categorized the FSS methods as Embeded methods, Filter methods and Wrapper methods. Filter-based feature selection methods does not depend on any machine learning algorithms. Their main aim is to analyze the intrinsic properties of data instead of using the induction algorithm.

The other approach of attribute subset selection methods are wrapper methods that are heavily depending on machine learning algorithms.

3.1.1 Filter-Based Feature Selection

Machine learning algorithms play a vital role while determining classifier accuracy. Kohavi et al, indicated that main disadvantages of the filter-based feature selection methods are the not using of any machine learning algorithm on feature selection step. Because of the classifier ignorance of this approach, eliminated feature subsets are tend to include redundant features that might have decreasing effect on classifier performance. Besides from that, filter-based feature selection methods are more suitable to be used on huge feature sets and generally tend to produces much more quick responses than any other wrapper methods.

SVM attribute evaluation and infogain methods are implemented as an example of the filter-based feature selection methods on the thesis.

3.1.1.1 SVM Attribute Evaluation

Gayatri et al (2012), explained that SVM based attribute evaluation method uses weights as an importance indicator. By calculating the support vectors, weights are calculated and features are ordered according to square of the weights. At FSS for multiclass problems, features are eliminated with considering one-vs-all principle.

Mladenić et al. (2004), gives the weight calculation equations of trained linear SVMs,

$$w = \sum_{i} a_i \, x_i \tag{1}$$

Equation 1 is shows that each x attributes have weight multipliers correspond to the feature. We will get the w1.x1 + w2.x2 + ... + wn.xn equation. When absolute value of weights are ordered descendly, small valued weights which do not have huge influence for the classification and corresponding attributes assumed as less important.

3.1.1.2 INFORMATION GAIN

To explain information gain (IG) term, entropy issue must be examined. IG is the amount of information contained in the class by knowing the presence and absence of the attribute.

$$I(S) := -\sum_{i=1}^{C} p_i \log_2 p_i \tag{5}$$

where the total number of classes denoted by C, and p_i the probability of instances that belong to class i. Equation 5 is the definition of the entropy value ranges from 0 to 1 according to purity, impurity of the variable. Information gain calculates the information retrieved from the attribute according to resulting criteria. Information gain is calculated for each feature A of S according to equation 6.

$$IG(S,A) = Entropy(S) - Entropy(S \mid A) = I(S) - \sum_{v \in A} \frac{|S_{A,v}|}{|S|} I(S_{A,v}) \tag{6}$$

$I(S_{A,v})$ is the entropy of the S that is only depends to A feature according to changing v values (Janecek et al. 2008).

When using IG on building tree, weighted average of entropy of selected children is substracted from entropy of parent, value of information gain is found. IG is tighty related to importance of the value that effects the result. When IG of the attribute x is higher than other values, it signifies that, x attribute is more relevant to the information.

3.1.2 WRAPPER METHOD

Janecek et al. (2008), explained the wrapper methods as searching the all feature space set and calculating the accuracy of induction algorithm according to add or remove features from the desired feature subset. ZHUO Li et al. (2008) wrapper methods can produce better evaluation performance, although they consume high computational power.

Zena et al (2015), divide the wrapper types into two different areas as deterministic wrappers and randomised wrappers. Deterministic wrappers are more prone to stuck with local optima. Randomised wrapper can handle the local optima problem but can take intensive computational time.

3.1.2.1 GENETIC ALGORITHM

The Genetic Algorithm (GA) is a type of search and ML algorithm which is inspired from the processes of evolution. Evaluation is totally related to limited resources. Resources are generally limited, individuals race to get the limited resources against each others. Strongest of the population have more chance to survive and spread their genes more than others.

GA is belonged to evolutionary algorithms, that are the subsection of heuristic methods of the optimization problems. Different application areas can use GA as a novel method, when brute force approaches do not provide feasible solutions. Evalution steps of GA are inheritance, crosover, selection, mutation. Those operators help to induction

algorithms reaching optimum solutions when it is possible but optimum solution is not guaranteed.

Evalution starts with randomly produced individuals belonged to population. Each individual in the population is reproducted and evaluated according to fitness function and the best fitting individuals are choosed and used to create better population for the next generations. Best ranked and selected individuals of the population are generally not mutated and directly selected according to elitism rules. Since the elites are the best individuals, they are valuable candidates to organise the new individuals using crossover operator. They do not have to be used in crossover operation, but also can be copied (remain unchanged) into the new population. The new population will go through the same steps and it might probably have the mutated genes. Mutation helps to discover and increase the finding probability of the potentially global optimum candidates. Rather than sticking with the local optimum solution, finding a more fitting solution will help to succeed the halting criteria more faster.

Reproduction operators are handled in three different ways such as pure reproduction, crosover and mutation. Directly copying the individual of the population to the next generation is the pure reproduction. Genes of two selected individuals are crossed at some point is a common example of crosover techniques. Mutation feature is the changing of one gene (bit) in selected individuals. Commonly, mutation rarely used operator in reproduction.

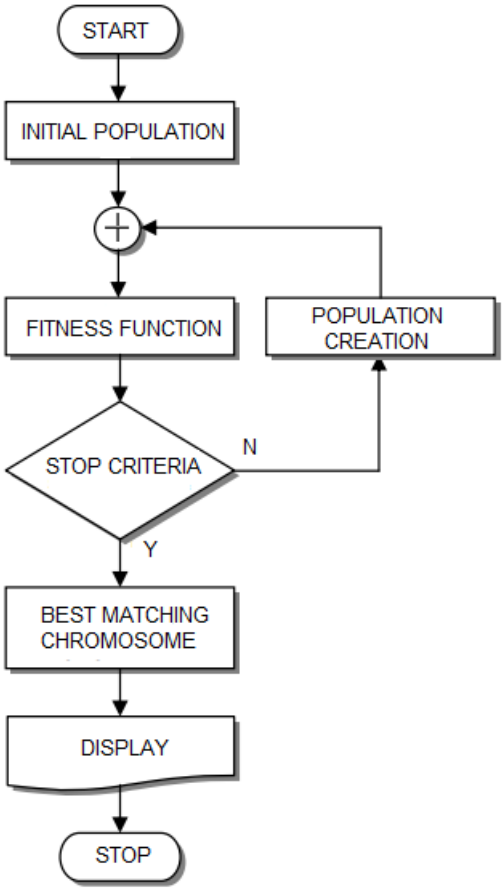

Fig. 3.2. GA Flow for Wrapper Based Attribute Selection Method.

Generally GA is halted when fitness rules are satisfied or maximum number of generations are reached. Generation count, mutation probability, crossover types and probability parameters are carefully tuned to find the better solution sets. Genaral flow chart of the GA is depicted in figure 3.2.

Figure 3.3 shows that transformation of input patterns that is a pre-process task to evaluate the fitness function according to selected classifier. To evaluate best feature set in the population, input matrix that is multiplied by GA matrix is classified and then new GA matrix is evaluated according to GA production operators.

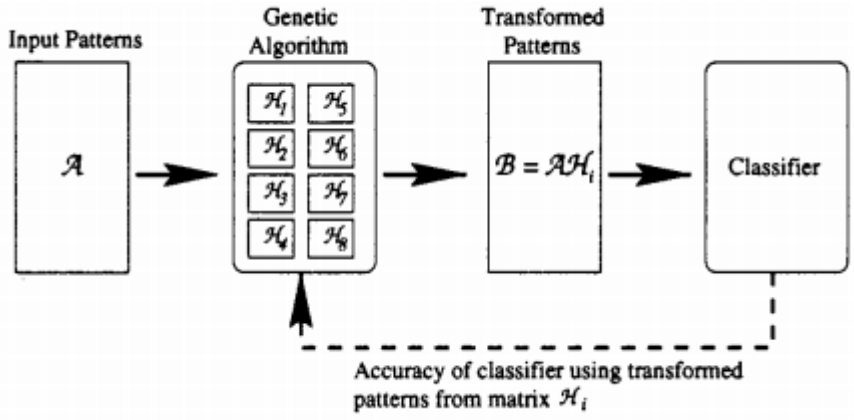

Fig. 3.3. Feature Extractor using Classification Accuracy as Fitness Function (Raymer et al. 2008).

Raymer et al. 2008 explained that, transformation matrix filters the input patterns and creates new transformed patterns. According to classifier accuracy, GA operators are used to create new transformed patterns to meet desired accuracy.

3.2. FEATURE EXTRACTION

Feture extraction is the another branch of DR process. On FE process, original attributes are combined linearly or non-linearly to create new attributes. New attributes are described by using original attributes. Janecek 2005 stated that on FE methods, original attribute set is represented as a minimum number of combined attributes.

Figure 3.4 shows the FE process. LDA, PCA are widely used implementations of the FE methods.

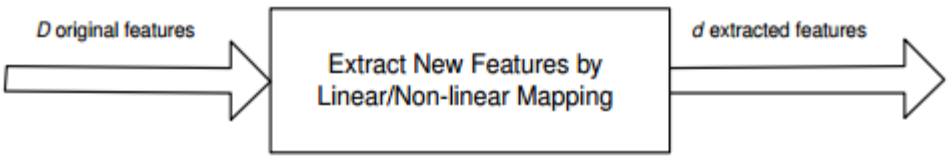

Fig. 3.4. Feature Extraction (Dash et al. 2008).

3.2.1 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is widely used DR technique in the literature. Jolliffe (1986) defined the aim of applying PCA to create new features that can be accomplished by completing the following rules. Created new attributes should be linear combination of first feature sets. Secondly, attributes are created according to maximizing the variance of new attribute set and orthogonal to each other. When creating new attributes, desired data variance can be captured, that helps lowering the noise, also achieves DR by small number of PC's.

PCA is an example of multivariate statistical analysis. Firstly, mean & scatter matrices or covariance matrix will be calculated. Scatter or covariance matrices are used to creation of eigenvectors. After calculation of the eigenvalues and corresponding m eigenvectors are selected to the largest m eigenvalues.

Tan et al. (2005) explained the four main criteria of PCA:

(i) Covariance should be zero between two extracted features

(ii) Features are ordered according to their containing information

(iii) The first feature contains the most variance of the information according to other features

(iv) Each feature contains as much as information possible.

4. DRIVING EVENT DATASET

To find driver habits that cause to car accident and make robust machine learning algorithms, DR methods are applied to event driver dataset. Driving event dataset contain sudden harsh breaking and acceleration, harsh left and right hand cornering, link speed exceeding, gps, speed and time information. Events are thowed on each second or according to event occurance. For example, if car suddenly accelerate, harsh accelerate event is thrown directly from board in real time.

Driving event dataset are also enriched with TRAMER accident data that include accident time, accident cost, crash type, replaced auto parts etc... TRAMER information give us usage of supervised learning methods with different parameters like accident cost segmentation, crash type categorisation, crash count segmentantion etc... Crash time and crash related events are applied to find the causes of the crash.

Moreover, without changing anything on the board, new events that are related to existing features can be created. For instance, according to consequent left and right (or right, left) gyro events, wobble event can be detected. Besides from that, it can be detected the drivers usage habit such as weekend, weekday, or give penalty who prone to use break event more frequently etc... Board initiated and created events based on penalty are depicted at Table 4.1.

Table 4.1 Penalty based v1 Features List

Event	Event Description
drive_duration_point	drive_duration_point
hl_total	harsh_left_penalty
hr_total	harsh_right_penalty
hb_total	harsh_break_penalty
ha_total	harsh_acceleration_penalty
hj_total	harsh_jump_penalty
left_right	(harsh_left_proportion / harsh_right_proportion)
acc_break	(harsh_acceleration_proportion/harsh_break_proportion)
ha_soft	soft_harsh_acceleration_usage
ha_average	average_harsh_acceleration_usage
ha_hard	hard_harsh_acceleration_usage
hb_soft	soft_harsh_break_usage
hb_average	average_harsh_break_usage

hb_hard	hard_harsh_break_usage
hl_soft	soft_harsh_left_usage
hl_average	average_harsh_left_usage
hl_hard	hard_harsh_left_usage
hr_average	average_harsh_right_usage
hr_hard	hard_harsh_right_usage
hj_soft	soft_harsh_jump_usage
hj_average	average_harsh_jump_usage
hj_hard	hard_harsh_jump_usage
harsh_right_criteria	harsh_right (event over 100 km) and gyro> 400
harsh_break_criteria	harsh_break (event over 100 km) and gyro> 400
harsh_left_criteria	harsh_left (event over 100 km) and gyro> 400
harsh_jump_criteria	harsh_jump (event over 100 km) and gyro> 400
speed_exceeded_criteria	speed_exceeded (event over 100 km) and gyro> 400

harsh_acceleration_criteria	harsh_acceleration (event over 100 km) and gyro> 400
usage_timing_habit	night/day usage
usage_location_habit	in city or highway
usual_path_proportion	usual path proportion
driving_habit	for work, weekdays, weekends
detected_car_wobble_props	wobble per months
detected_car_wobble_features	combination of speed, accelerometer and gyro
detected_event_patterns_before_crash_event1	Pattern that occured right before crash event
detected_event_patterns_before_crash_event2	Pattern that occured right before crash event
detected_event_patterns_before_crash_event3	Pattern that occured right before crash event
detected_event_patterns_before_crash_event4	Pattern that occured right before crash event
detected_event_patterns_before_crash_event5	Pattern that occured right before crash event
detected_event_patterns_before_crash_event6	Pattern that occured right before crash event
detected_event_patterns_before_crash_event7	Pattern that occured right before crash event

detected_event_patterns_before_crash_event8	Pattern that occured right before crash event

To meet success criteria in the system, various data sets are applied to retrieve the best accuracy. Although penalty based v1 feature list satisfied the accuracy criteria of the system, to improve the resulting accuracy and accomplish second aim of the system that is finding the most related habits causes the crash events, habit based v2 feature list is created. Penalty-based v1 feature list depend on punishment of predefined driving acts and it is more related to flag of undesired actions that are mostly detected right before the crash and depend on magnitude of the actions. Usage habits based created features focused on frequency of the selected undesired actions. Table 4.3 indicates the habit-based feature list.

Table 4.2 Usage Habit based Feature List v2

Event	**Event Description**
hl_30	harsh_left_30km
hl_3050	harsh_left_3050
hl_5070	harsh_left_5070km
hl_70U	harsh_left_70km
hr_30_1	harsh_right_30km
hr_3050_1	harsh_right_3050km
hr_5070_1	harsh_right_5070km

hr_70U_1	harsh_right_70km
ha_30_2	harsh_acceleration_30km
ha_3050_2	harsh_acceleration_3050km
ha_5070_2	harsh_acceleration_5070km
ha_70U_2	harsh_acceleration_70km
hb_30_3	harsh_break_30km
hb_3050_3	harsh_break_3050km
hb_5070_3	harsh_break_5070km
hb_70U_3	harsh_break_70km
ls_30_4	linkspeed_exceeded_30km
ls_3050_4	linkspeed_exceeded_3050km
ls_5070_4	linkspeed_exceeded_5070km
ls_70U_4	linkspeed_exceeded_70km
works	Work_hours_activity
daily	Daily hours_activity

night	Night hours_activity
overnight	Overnight hours_activity
avg. speed	Average Speed
std. deviation	Standart Deviation

Even if dataset is discrete, for 600 drivers, more than 400 million events are throwed per month. Because of huge amount of data per driver and high dimensional spaces, FSS and FE methods are applied to find more robust, high accuracy and most affecting habits that cause to car accident.

5. IMPLEMENTATION DETAILS

In this thesis, we have applied the DR methods to improve the classification accuracy and investigate the important events that have much more impact on car accident event. Our aim is to explore the effects of DR methods on the solution of uncertain areas to eliminate the unnecessary events and to find out the more compact form of solution. When dealing with problemetical and totally indefinite areas, DR methods are used not only helping the understanding of the problems' important features, but also they can increase the accuracy of the estimations. FSS and FE methods are compared according to classification accuracy and number of reduced attribute set.

To process 400 million data per month, Oracle 12c is choosed as database management system.All records are uploaded to and processed via Oracle 12c. For our data exploration tests, WEKA is selected as a classifier tool thanks to its maturity. There are other options like knime, R, or RapidMiner but WEKA has lots of online documentations. Although GUI of WEKA is easy to use, to automate the experiments, API option of WEKA in Java 8 programming language. Otherwise, it would be nearly impossible task to find the best subset with all DR methods. Besides with filter based FSS methods, FSS with GA as a wrapper approach is also implemented in Java 8 programming language.

After tool and data selection steps, we have applied the standart data cleaning and transformation steps. Outliers and missing data are cleaned. For instance, drivers with less than total 3 hours driving time in one month are not included. At the transformation step, min-max data normalization technique is employed. At min-max data normalization technique, data is normalized and scaled into pre-defined interval.

When completing the pre-process steps of DM, different machine learning algorithms such as Naive Bayes, Multilayer Perception and J48 that is open source implementation of c4.5 algorithm are picked as referance induction algorithm are implemented to process event driving dataset. Naive Bayes and J48 algorithms are categorized as statistical classifier. Multilayer Perception model is based on artificial neural network. These algorithms are implemented at Java 8 to automate our tests.

Before applying the DM algorithms, data should be separated to the traning and testing sets. On the separation process, k fold cross validation technique are used. At k fold cross validation technique, input set n is divided into k equal parts, and at each part, k elements are used for testing set and n-k elements are used for training until all k elements are used on testing process.

Infogain and SVM attribute classifier like filter methods rank the attributes descendenly. Our main aim is to find the best combination of the attributes while applying the DR methods and find the most important features acording to DR methods. To accomplish these tasks, the least significant attributes indicated by the Infogain and SVM attribute classifier methods are eliminated one by one. At each step, classification accuracy are evaluated and best accuracy with least attribute set are picked as the best feature subset according to corresponding DM algorithms. This approach help us not only the removing irrevelant features accurding to filter methods but also increasing classification accuracy.

PCA initially choose the number of attributes according to its combination result. We used the default linear kernel option of the PCA that creates 9 attributes at first feature set and 16 attribute at second feature list. To test the result of pca, we eliminate one attribute and applied induction algorithms to see the behaviour of the classification performance. At all scenarios, classification performance is degraded. PCA give us the best option on all induction algorithms. That means all features of PCA, contribute to the result positively.

```
public boolean GetChrosomeFitness(Chromosome chromosome){

    //Transform CurrentGene To Feature Driver Set
    for each gene in the chromosome
        if chromosome(gene) is selected then
            attributeSelectionArr[gene]=1
            writeToFile(feature[attribute]);
        else
            attributeSelectionArr[gene]=0
        end if
    end for

    //create featureSet
    for each driver in the featureSet
        for each feature in the featureSet{
            if attributeSelectionArr[feature] is selected   then
                this.gene.featureSet[driver][feature] = InitialFeatureSet[driver][feature];
            else
                //do nothing
            end if
        end for
        writeToCurrentChromosomeFile(this.gene.featureSet[driver]);
    end for

    featureMatrix=readFromCurrentChromosomeFile();

    curAccuracy=Apply_DM_Algorithm(featureMatrix);// NAIVEBAYES|MULTILAYERPERCEPTION|J48

    if(curAccuracy<accuracy) // mutation decision
        return 0;
    else then
        this=chromosome;    // add to the population of current generation
        accuracy=curAccuracy;
        return 1;
    end if;
}
```

Figure 5.1. Psoude Code of Fitness Function

GA is implemented to handle dimention reduction problems. Each feature in the attribute set is mapped into genes. Each chromosome has one bit representation and each bit represents one feature. If bit is valued 1 means that corresponding feature included to the population, 0 means not included. Initial population populated randomly and size of population equals to size of selected features.

Fitness function named as getChromosomeFitness is depicted at Figure 5.2. Fitness function takes initial, random, crossovered or mutated genes and transforms it to feature matrix. After creation of feature matrix, new feature subset are created and evaluated according to corresponding induction algorithm. Results are evaluated according to accepted threshold limit or producing more accuracy than previous chromosome. If resulting accuracy are more with newly created chromosome, generation add this chromosome to create new generations accurding to evalutionary rules and new threshold is updated.

Crossover function is accomplished by single point crossover function. Figure 5.2 represents the one point crosover function. Two parents with 8 genes copy their genes on the same index to the childs to create new generations. Mutation operator is handled

by randomly altering value of the one gene. As a fitness function, DM algorithms accuracy measure are selected. Aim is to maximize the accuracy and eliminates the features. If the population can not qualify for the fitness function, according to parameters decision of crossover and mutation functions are handled.

P_1:	1	0	1	0	0	1	1	1
P_2:	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**0**
C_1:	1	0	1	0	**0**	**0**	**1**	**0**
C_2:	**1**	**1**	**1**	**1**	0	1	1	1

Figure 5.2: Single Point Crossover (D. Asir Antony Gnana Singh 2016).

After implementing the GA, according to our problem, domain spesific parameters such as mutation probability, max mutated chromosome count in generation, choosing the selection strategy (less, more), generation count, permitting mutation at inherited chromosomes flag according to elism rule, directly inherited chromosomes count should be tuned. Attributes are selected according to evolution process.

6. RESULTS & CONCLUSION

We investigate the problem as a result of driver's driving dynamics and happening sequence of labelled dangerous occuring events that is based on penalty scoring system and drivers usage habits. On these two approaches, we have created two different feature sets. As a evaluation parameter, TRAMER information of the driver is used for classification criteria.

V1 Feature Set's classification results without DR methods are far from satisfying and classification accuracy are oscillating between %67-72. We have applied the DR methods to learn the most important attributes and also picked the attributes according to best classification accuracy result. At table 6.1, classification results with and without DR methods are given.

Table 6.1 Classification Accuracy of v1 Feature Set

	J48	Multilayer Perception	Naive Bayes	Average Accuracy
SVM attribute classifier	%72.97 with 27 features	%76.67 with 19 features	%70.27 with 24 features	%73.30
Infogain	%76.67 with 34 features	%72.97 with 17 features	%70.27 with 14 features	%73.30
PCA	%72.97 with 9 features	%76.67 with 9 features	%72.27 with 9 features	%74.2
Genetic Algorithm	**%78.37 with 8 features**	**%86.22 with 7 features**	**%81.08 with 7 features**	**%81.89**
Without Reduction	%67.74 with 43 features	%72.97 with 43 features	%67. 74 with 43 features	%69.48

We have compared that SVM attribute classifier, Infogain, PCA and GA results according to the resulting classification performance based on driving event data set. Created v2 attribute list was more robust on v1 attribute list according to the drivers accident prediction. V2 Feature Set's classification results gives more better accuracy

than v1 feature set's result. Classification accuracy without DR methods are over %74. DR methods again increased the classification result. At table 6.2, classification results with and without DR methods are given.

Table 6.2 Classification Accuracy of v2 Attribute Set.

	J48	Multilayer Perception	Naive Bayes	Average Accuracy
SVM attribute classifier	%78.03 with 11 features	%78.65 with 16 features	81.34% with 11 features	%79.34
Infogain	%78.03 with 11 features	%79.21 with 22 features	**82.51% with 16 features**	%79.25
PCA	%74.11 with 16 features	**% 81.17 with 16 features**	78.65% with 16 features	%77.97
Genetic Algorithm	**%85.75 with 17 features**	%79.64 with 15 features	80.93 % with 15 features	**%82.10**
Without Reduction	%74.62 with 26 features	%78.65 with 26 features	78.82 % with 26 features	%77.38

Table 6.3 depicts the best ten attributes according to SVM attribute classifier method. Table 6.4 is organized according to Infogain method and GA's selected best attribute set are showed at Table 6.5. DR methods have produced three common attributes that are represented at Table 6.6.

Table 6.3 Best 10 attributes according to SVM attribute classifier method

Event	**Event Description**
hl_30_2	harsh_acceleration_30km_alt
Avg. SPEED	Average Speed
ha_50_2	harsh_acceleration_3050km_arasi

hb_50_3	harsh_break_3050km_arasi
ls_30_4	linkspeed_exceeded_30km_alt
ls_70_4	linkspeed_exceeded_5070km_arasi
ls_50_4	linkspeed_exceeded_3050km_arasi
ha_70U_2	harsh_acceleration_70km_ustu
hr_70_1	harsh_right_5070km_arasi
hb_70_3	harsh_break_5070km_arasi

Table 6.4 Best 10 attributes according to Information Gain method

Event	**Event Description**
ls_70U_4	linkspeed_exceeded_70km_ustu
hr_50_1	harsh_right_3050km_arasi
hb_70_3	harsh_break_5070km_arasi
hb_50_3	harsh_break_3050km_arasi
ha_70U_2	harsh_acceleration_70km_ustu
ha_30_2	harsh_acceleration_30km_alti

hb_70U_3	harsh_break_70km_ustu
hl_70U	harsh_left_70km_ustu
hr_70_1	harsh_right_5070km_arasi
ha_50_2	harsh_acceleration_5070km_arasi

Table 6.5 Retrieved best 10 attributes according to GA attribute selection methods

Event	**Event Description**
hl_50	harsh_left_3050km_arasi
hl_70U	harsh_left_70km_ustu
hr_30_1	harsh_right_30km_alti
hr_70_1	harsh_right_5070km_arasi
ha_70_2	harsh_acceleration_5070km_arasi
ha_70U_2	harsh_acceleration_70km_ustu
hb_30_3	harsh_break_30km_alti
hb_70_3	harsh_break_5070km_arasi
ls_30_4	linkspeed_exceeded_30km_alti

ls_70U_4	linkspeed_exceeded_70km_ustu

Table 6.6 Infogain, SVM and GA FS methods' common attributes

Event	Event Description
hr_70_1	harsh_right_5070km_arasi
ha_70U_2	harsh_acceleration_70km_ustu
hb_70_3	harsh_break_5070km_arasi

We compared the DR methods and the resulting classification performance with DR methods based on event-driver data sets meet success requirement. We have created completely two different datasets grouped as habit based and penalty scores.

- Frequency based dataset, performed more robust and accurate than penalty based dataset.
- GA give us the best average on two dataset. But on all induction methods give us the different attribute set.
- All DR methods (except pca on the second dataset j48 algo) provide better feature set than initial feature set. It is the result of existance of irrevelant features.
- PCA is a candidate for the second best performer according to average accuracy at first dataset, but in the second dataset performed badly.
- Infogain method give the best result on j48 algorithm, with 34 attributes that is relatively high attribute set at first table.
- Filter-based DR methods are generally performed not quite well, although their average accuracy is higher than default dataset's accuracy at first dataset.
- Filter-based DR methods reduced the initial attribute not consistently for different induction algorithms at first dataset.

- At first dataset GA and the second dataset SVM attribute eval. give us the minimum attribute set for all induction methods.
- GA that is the best performer give us the more accurate results on labeled drivers with accident info than the drivers with no accident label.
- From driving event dataset new information can be produced to find a relation with accident probability and driving habits.
- From driving event dataset new feature list is produced to find a relation with accident probability and driving habits. We have produced more robust results with the new attribute set.
- harsh_right_5070km_arasi, harsh_acceleration_70km_ustu and harsh_break_5070km_arasi features are the common features among three DR method

REFERENCES

Andreas Janecek, Wilfried Gansterer, Michael Demel, Gerhard Ecker (2008). On the Relationship Between Feature Selection and Classification Accuracy, *JMLR Workshop Conference Proceedings 4*. pp. 90–105.

Anil Kumar Dhiman. (2011). Knowledge Discovery in Databases and Libraries. DESIDOC Journal *of Library & Information Technology*, Vol. 31, pp. 446-451

D. Asir Antony Gnana Singh, E. Jebamalar Leavline, R. Priyanka, P. Padma Priya, *"Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis"*, International Journal of Intelligent Systems and Applications(IJISA), IJISA, Vol. 8, No. 1, January 2016, ISSN: 2074-904X (Print), ISSN: 2074- 9058 (Online), DOI: 10.5815/ijisa

Dunja Mladenic, (2005). Feature Selection for Dimensionality Reduction. *SLSFS*, pp. 84-102.

Dunja Mladenić, Janez Brank, Marko Grobelnik, and Natasa Milic-Frayling. (2004). *Feature selection using linear classifier weights: interaction with classification models.* In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04). ACM, New York, NY, USA, pp. 234-241.

Gayatri, N. and Nickolas, S. and Reddy, A. V. (2012). *ANOVA Discriminant Analysis for Features Selected through Decision Tree Induction Method*. Global Trends in Computing and Communication Systems: 4th International Conference, ObCom 2011, Vellore, TN, India, December 9-11, 2011. Proceedings, Part I, pp. 61-70

Haleh Vafaie and Ibrahim F. Imam, *"Feature selection methods: genetic algorithms vs. greedy-like search,"* International Conference on Fuzzy and Intelligent Control Systems 1994.

Jolliffe, I. (1986). Principal Component Analysis. Springer Verlag.

Kohavi, Ron and John, George H., (1997), Wrappers for Feature Subset Selection, Artif. Intell., pp. 273–324.

Manoranjan Dash, Huan Liu (2008). Dimensionality Reduction. *Wiley Encyclopedia of Computer Science and Engineering*

M. Zamalloa, L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel and J. P. Uribe. (2008). "*Feature dimensionality reduction through Genetic Algorithms for faster speaker recognition*," Signal Processing Conference, 16th European, Lausanne, pp. 1-5.

Rosaria Silipo, Iris Adae, Aaron Hart and Michael Berthold (2014). Seven Techniques for Dimensionality Reduction Missing Values, Low Variance Filter, High Correlation Filter, PCA, Random Forests, Backward Feature Elimination, and Forward Feature Construction.

Soundararajan E., Joseph J.V.M., Jayakumar C. and Somasekharan M. (2005). Knowledge Discovery Tools and Techniques, pp. 141-145

Tan, P.-N., Steinbach, M.,, Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley. ISBN: 0321321367

Zena M. Hira and Duncan F. Gillies (2015), A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, *Advances in Bioinformatics*, vol. 2015, Article ID 198363, 13 pages.

Zhou Li, Zheng Jing, Wang Fang, LI Xia, Ai Bin, Qian Jun-ping, (2008*) A genetic algorithm based wrapper feature selection method for classification of hyper spectral data using support vector maching*, GEOGRAPHICAL RESEARCH. pp. 493-501.

BIOGRAPHICAL SKETCH

Can Çetin was born in Jane 15, 1986 in Izmir. He received his high school education in Manisa Turgutlu Halil Kale Science High School. Furthermore, he received his Bachelor of Science in Computer Engineering from Istanbul Technical University in 2010.

PUBLICATION