Handling sparse data sets by applying contrast set mining in feature selection

Dijana Oreski and Bozidar Klicek

***Abstract*—**A data set is sparse if the number of samples in a data set is not sufficient to model the data accurately. Recent research emphasized interest in applying data mining and feature selection techniques to real world problems, many of which are characterized as sparse data sets. The purpose of this research is to define new techniques for feature selection in order to improve classification accuracy and reduce the time required for feature selection on sparse data sets. The extensive comparison with benchmarking feature selection techniques conducted on 128 data sets was conducted. Results of the 1792 analysis showed that in the more than 80% of the 128 analyzed data sets contrast set mining techniques are superior to benchmarking feature selection techniques. This paper provides a study on the new methodologies that have tried to handle the sparse datasets and showed superiority in handling data sparsity.

***Keywords*—**Data characteristics, Contrast set mining, Feature selection, Neural network classification.

I. Introduction

IN the last two decades there is a huge increase in the amount of data that is stored in digital format. According to The Economist, 150 exabytes (EB) of data were created in 2005; in 2010 that number was expected to rise in 1200 EB. Owing to today's technology data collection ceases to be a problem and in the focus of interest is their analysis and obtaining valuable information from the data (knowledge). Central for this issue is the process of knowledge discovery in data. The process consists of several steps: data understanding, data preparation, modelling, evaluation and implementation [18].

Data preparation, which includes data cleaning and feature selection take away from 60% to 95% time of the whole process. Main idea of feature selection, clearly the most important stage of this step, is to choose a subset of features by eliminating those with little predictive information. Benefits of feature selection include reducing dimensionality, removing irrelevant and redundant features, facilitating data understanding, reducing the amount of data for learning, improving predictive accuracy of algorithms, and increasing interpretability of models [21; 4; 36; 8]. Feature selection plays an important role in many areas and has found extensive application. In this study, a contrast set mining based feature selection techniques are proposed. The objective is to investigate potential of contrast set mining techniques for improving feature selection.

Contrast set mining is subfield of data mining and was first proposed in 1999 as a way to identify those features that significantly differentiate between various groups (or classes). Contrast set mining is being applied in many diverse fields to identify features that provide greatest contrast between various classes. It has been successfully applied in e.g. market based analysis [46] and medicine [29]. This approach has the advantage that the complexity and size of the data is reduced while most of the information contained in the original raw data is being preserved, which is the main idea behind feature selection. Thus, this paper recognizes potential of contrast set mining techniques for application in feature selection.

Previous research in data mining field recognized that optimal choice of mining algorithm depends on the characteristics of the data set employed [10;11;43]. We can not select an algorithm and claim its superiority over competing algorithms without regard to data characteristics as well as the suitability of the algorithm to such data. This paper focuses on data sparsity characteristic since enforcing sparsity can greatly improve on methods [12;49]. The subject of this research is an application and evaluation of contrast set mining techniques as techniques for feature selection and validation of such techniques for feature selection on sparse data sets. The extensive empirical research is conducted in order to determine do contrast set mining techniques outperform classical feature selection techniques, and obtaining general answer can we use contrast set mining techniques as a superior feature selection techniques, and whether they can eliminate the bottleneck of the entire process of knowledge discovery in data. Comparison of contrast set mining techniques in feature selection with benchmarking feature selection techniques is performed on 128 data sets.

The process of knowledge discovery in data can be performed in order to achieve one of the following tasks: classification, clustering, visualization, summarization, deviation detection or estimation [18]. Classification is considered to be the basic task [18] and, thus, is in the focus of interest in this paper. Evaluation of contrast set mining techniques was carried out in order to perform classification on the datasets with two classes of dependent feature.

The paper is organized as follows. In section 2 we briefly describe some of the feature selection techniques comparison`s reported in the literature. Data sparsity as an important data set characteristics is presented in section 3. Section 4 describes contrast set mining techniques, STUCCO and Magnum Opus, because we used them to define our approach. In section 5 we present experimental framework and research hypothesis. In section 6 the proposed approach is described. Evaluation experiments are shown and discussed in section 7 and conclusions are drawn in section 8.

II. Feature selection

Feature selection is an active field in computer science [32;44]. The problem of feature selection can be seen as a search problem on the power set of the set of available features [25]. The goal is to find a subset of features that allows us to improve, in some aspect, a learning activity. It has been a fertile field of research and development and has produced many new feature selection techniques. Here we will not focus on many studies that concentrate on explaining details of particular techniques. Rather, we will take a look at studies comparing feature selection techniques with widely differing capabilities focusing on the research scope (number of techniques in comparison, number of data sets used and used criteria). Overview is given in a chronological order.

John et. al. described a technique for feature subset selection using cross-validation that is applicable to any induction algorithm, and discuss experiments conducted with ID3 and C4.5 on artificial and real datasets [25]. Kohavi and Sommerfield compared forward and backward selection on 18 data sets [28]. Koller and Sahami introduced information theory based feature selection technique. They have tested new technique on 5 data sets [26]. Kohavi and John introduced wrapper approach and compare it to Relief, a filter approach to feature selection. Significant improvement in accuracy is achieved for some datasets for the two families of induction algorithms used: decision trees and Naive-Bayes [27].

Dash and Liu gave comprehensive overview of many existing techniques from the 1970's to the 1997 and categorize the different existing techniques in terms of generation procedures and evaluation function [14]. Furthermore, they chose representative techniques from each category. Their comparative analysis was performed on 3 data sets. Weston et. al. introduced a feature selection technique for Support Vector Machines [47]. The method was superior to some standard feature selection algorithms on the 4 data sets tested. Liu et al. tested Relief algorithm on 16 data sets [32]. Geng et. al. introduced new feature selection technique based on the similarity between two features. New approach was tested on 2 data sets [20].

Alibeigi et. al. suggested new filter feature selection technique and compared it with 3 techniques on 3 data sets [3]. Janecek compared feature selection techniques on 3 data sets from two fields. Drugan and Wiering proposed feature selection technique for Bayes classifier and tested it on 15 data sets [17]. Cehovin and Bosnic compared 5 feature selection techniques: ReliefF, random forest feature selector, sequential forward selection, sequential backward selection and Gini index by means of classification accuracy of 6 classifiers including decision tree, neural network and Naive Bayes classifier [9]. Lavanya and Usha Rani investigated performance of feature selection techniques on 3 data sets regarding breast cancer issue. Results didn`t indicate superiority of one technique on all data sets. They used classification accuracy and time required for feature selection as comparison criteria [31]. Novakovic et al. compared 6 feature selection techniques on 2 data sets and used classification accuracy as a criterion [35]. Haury et.al compared 8 feature selection techniques on 4 data sets [23]. Silva et. al compared 4 existing feature selection techniques (information gain, gain ratio, chi square, correlation) on 1 data set from the domain of agriculture [39].

Survey of previous research pointed out classification accuracy and time required for performing feature selection (elapsed time) as most important criteria for feature selection techniques performance. However, methodology for evaluation has not been standardized so far and differs from one research to another. Thus, it is difficult to draw out conclusion or make comparisons of feature selection techniques.

Furthermore, analysis of previous research indicated lack of comprehensiveness as main disadvantage, to be more specific:

- narrow choice of feature selection techniques,
- use of a single classifier which makes it impossible to establish connection between performances of classifiers and feature selection techniques,
- small and simulated data sets which does not represent real world problems,
- number of data sets in the analysis was very small,
- only one criterion was used in the comparison.

These paper efforts are largely motivated by aforementioned problems and research presented here has following properties:

- feature selection techniques comparison is conducted on 128 data sets,
- 7 feature selection techniques are compared,
- 2 different classifiers are used in the learning process.

III. Data sparsity

In this section we investigate the relationship between the dimensionality of data and the number of samples required to model the data accurately. This relationship is not trivial and Van der Walt [43] defined measure that captures this relevant factor.

A. *Theoretical background*

In this section we explain in detail measure to quantify whether the number of samples in a data set is sufficient to model the data accurately. The measure measures how sparse data is by taking the dimensionality, number of classes and number of samples in data set into account. Thus, data sparsity is defined through relationship of dimensionality and number of instances sufficient to model the data accurately. Relationship between dimensionality (d) and the number of samples (N) can be linear, quadratic or exponential. Van der

Walt uses theoretical properties of classifiers to describe each of the three types of relationship. To test if a linear relationship holds between d and N we will employ the normality test and examine correlations between features [43]. The number of parameters that must thus be estimated is 2dC + C. To test if this quadratic relationship between d and N holds we will measure the homogeneity of class covariance matrices as well as the normality of the class data. Thus, the total number of parameters that must be estimated is:

$$d^2 + DC + C.$$

How do we decide which of the three relationships between N and d is most appropriate? A linear relationship can be tested by employing tests for multivariate normality and correlation. Quadratic relationships can be tested by testing for multivariate normality and the homogeneity of class covariance matrices. If the linear and quadratic relationships don't hold, an exponential relationship between N and d is possible. When we have determined the relationship between d and N we need to quantify whether there are enough samples in the training set to model the structure of the data accurately. For each of the four relationships mentioned above, we define a measure (Nmin), which sets the scale for the minimum number of samples that is required to model the data accurately.

If the data are normally distributed and uncorrelated, a linear relationship between d and N will exist and the minimum number of samples that are required will be in the order of

$$N_{l\,(\min)} = 2dC + C$$

If the data are normally distributed, correlated and the classes have homogeneous covariance matrices, then a quadratic relationship will exist between d and N and the minimum number of samples that are required will be proportional to

$$N_{q1(\min)} = 2d^2 + dC + C$$

If the data are normally distributed, correlated and the classes have non-homogeneous covariance matrices, then a quadratic relationship will exist between d and N and the minimum number of samples that are required will be on the order of

$$N_{q2(\min)} = Cd^2 + dC + C$$

If the data are not normally distributed, an exponential relationship between d and N will be assumed and the number of samples that are required may be as plentiful as

$$N_{e\,(\min)} = D^{d}_{steps}$$

where Dsteps is the discrete number of steps per feature. We will now quantify if the number of samples are sufficient to model the data accurately by defining a ratio between the actual number of samples and the minimum number of samples that are required. We define a measure of data sparsity as follows:

$$DSR = \frac{N}{N_{min}}$$

where Nmin is the appropriate minimum number of samples measure and N the actual number of samples in the data set. We also define a measure to indicate if the number of samples are sufficient by inverting equation as follows:

$$DS = \sqrt[d]{N}$$

where N is the number of samples in the data set and d the dimensionality of the data set.

B. *Measuring data sparsity*

Hereinafter, we define two categories of sparsity for dana sets: low and high. Data set has small sparsity if real number of samples is higher or equal to number of samples required to model the data accurately. Data set sparsity is high if real number of samples is smaller or equal to number of samples required to model the data accurately. On the case of one data set used in our research, data set vote, [42], we explain how did we measure data sparsity. Standard measures for vote data set are following:

Standard measures	Measure for *vote*
Dimensionality	17
Number of instances	435

By applying Kolmogorov Smirnov test we have tested normality of the distribution. In vote data set no feature has normal distribution. Thus, exponential relationship between features in data set exsists. Thus, required number of samples for accurate modelling is calculated as:

$$2^{17} = 131072$$

Since minimal number of instances required for accurate modelling (131 072) is higher than actual number of instances (435), we can conclude there is not enough instances for precise modelling and data sparsity of vote data set is HIGH.

IV. BACKGROUND ON CONTRAST SET MINING

Data mining is one of the most exciting information science technologies in twenty-first century. It has become an important mechanism that is able to interpret the information hidden in data to human-understandable knowledge [33]. Involvement in a wide range of practices led to the development of specialized sub-areas within the data mining. One of the newest is contrast set mining field. While data mining has traditionally concentrated on the analysis of a static world, in which data instances are collected, stored, and analyzed to derive models that describe the present, there is growing consensus that revealing how a domain changes is equally important as producing highly accurate models [7]. Nowadays, developing methods for analyzing and understanding change is seen as one of the primary research issues when dealing with evolving data [7]. Led by this practical need, subfield of data mining for analyzing changes was developed called contrast set mining. Contrast set mining has started to develop in 1999, and today is one of the most challenging and vital techniques in data mining research [33].

The objective of contrast set mining is to quantify and describe the difference between two data sets using concept of contrast set. Contrast set is defined as „conjunctions of attributes and values that differ meaningfully in their distribution across groups" [5]. To differ meaningfully an item set's support difference must exceed a user-defined threshold. Description of STUCCO and Magnum Opus is provided in following two sections.

A. STUCCO algorithm

Concept of contrast sets was first proposed by Bay and Pazzani, to describe the difference between two data sets by contrast sets which they defined as „conjunctions of attributes and values that differ meaningfully in their distribution across groups“ [5]. To discover contrast sets Bay and Pazzani proposed the STUCCO (Search and Testing for Understandable Consistent Contrast) algorithm [5]. STUCCO algorithm performs a breadthfirst search in the item set lattice. It starts with testing the smallest item sets, then tests all next-larger ones, and so on. To overcome complexity problems, the algorithm prunes the search space by not visiting an item set's supersets if it is determinable that they will not meet the conditions for contrast sets or if their support values are too small for a valid chi-square test [7]. Formally defined, it looks like this.

The data is a set of groups $G_1, G_2 \dots G_l$. Each group is a collection of objects $O_1 \dots O_u$. Each object O_i is a set of k feature-value pairs, one for each of the features`$A_1 \dots A_k$. Feature A_j has values drawn from the set $V_{j1} \dots V_{jm}$. A contrast set is a set of feature-value pairs with no attribute A_i occuring more than once. This is equivalent to an itemset in association-rule discovery when applied to attribute-value data. Similar to an itemset, we measure the support of a contrast set. However, support is defined with respect to each group. The support of a contrast set *cset* with respect to a group G_i is the proportion of the objects $o \in G_i$ and is denoted as $supp(cset, G_i)$. Contrast set discovery seeks to find all contrast sets whose support differs meaningfully across groups. This is defined as seeking all contrasts sets *cset* that satisfy following:

$\exists ij P(cset|G_i) \neq P(cset|G_j)$ (Eq 1) and

$\max(i, j)\, |support(cset, G_i) - support(cset, G_j)| \geq \delta$ (Eq 2

where δ is a user-defined threshold called the *minimum support-difference*. Contrast sets for which Eq. 1 is statistically supported are called *significant* and those for which Eq. 2 is satisfied are called *large*. When both equations are satisfied, the contrast set is called the *deviation*. Eq. 1 provides the basis of a statistical test of `meaningful,' while Eq. 2 provides a quantitative test. The statistical significance of Eq. 1 is assessed using a chi-square test to assess the null hypothesis that contrast set support is independent of group membership.

B. Magnum Opus

Magnum Opus is a commercial implementation of the OPUS AR rule-discovery algorithm. OPUS stands for **O**ptimized **P**runing for **U**nordered **S**earch. It provides association-rule-like functionality, but does not use the frequent-itemset strategy and hence does not require the specification of a minimum-support constraint.

At the heart of Magnum Opus is the use of *k*-optimal (also known as top-*k*) association discovery techniques. Most association discovery techniques find frequent patterns. Many of these will not be interesting for many applications. In contrast k-optimal techniques allow the user to specify what makes an association interesting and how many (*k*) rules they wish to find. It then finds the *k* most interested associations according to the criteria the user selects.

Under this approach the user specifies a rule value measure and the number of rules to be discovered, *k*. This extends previous techniques that have sought the single rule that optimizes a value measure for a pre-specified consequent [45;2]. Rule value measures are central to the enterprise of *k* optimal rule discovery. We explain five such measures. The available criteria for measuring interest include *lift*, *leverage*, *strength* (also known as *confidence*), *support* and coverage. These measures are defined in more detail, by using *following* notation: D = data set, X = LHS (Left Hand Side) and Y = RHS (Right Hand Side). The *coverage* of the rule is the number of cases that contain the LHS. The *support* of the rule is the number of cases that contain both the LHS and the RHS. The *strength* is the support divided by the coverage. This represents the proportion of the cases that contain the LHS that also contain the RHS. It can be thought of as an estimate of the probability that the RHS will occur in a case if the LHS occurs. The *lift* is the strength divided by the strength that would be expected if there were no relationship between the LHS and the RHS. A value of 1.0 suggests that there is no relationship between the two. Higher values suggest stronger positive relationships. Lower values suggest stronger negative relationships (the presence of the LHS reduces the likelihood of the RHS). The *leverage* is the support minus the support that would be expected if the LHS and RHS were unrelated to one another. A positive value suggests a positive relationship and a negative value suggests a negative relationship. Value *p* is the result of a statistical evaluation of the significance of the rule. Lower *p* value means the less likely that this rule is spurious, either because the LHS and RHS are unrelated to one another, or because one or more of the values in the LHS do not contribute to the association with the RHS.

C. Contrast set mining techniques discussion

Another approach used to distinguish two or more groups is to use a decision tree. This has the advantage of being fast in generating understandable models. However they have major disadvantages: (1) Decision trees are not complete because they achieve speed by using heuristics to prune large portions of the search space and thus they may miss alternative ways of distinguishing one group from another, (2) decision trees focus on discrimination ability and will miss group differences that are not good discriminators but are still important. (3) Rules obtained by decision tree are usually interpreted in a fixed order where a rule is only applicable if all previous rules were not satisfied. This makes the interpretation of individual rules difficult since they are meant to be interpreted in context. Finally, (4) it is difficult to specify useful criterion such as minimum support.

Area closely related to contrast sets is association rule mining [2]. Association rules express relations between variables of the form X->Y . In market basket data X or Y are items such as beer or salad. In categorical data X and Y are attribute-value pairs such as occupation = professor. Both, association rules and contrast sets require search through a space of conjunctions of items or attribute-value pairs. In association rule, we look for sets that have support greater than a certain cutoff (these sets are then used to form the rules) and for contrast sets we seek those sets which represent substantial differences in the underlying probability distributions.

Since both techniques have a search element there are many commonalities. Actually, in order to enhance contrast set algorithms we build on some of the search work developed for association rule mining is applied. Though, contrast sets approach differs substantially from association rules because contrast set work with multiple groups and have different search criteria. Idea to apply association rule mining algorithms to find contrast sets wouldn`t work effectively. For example, one approach would be to mine the large item sets for each group separately, and then, compare them. Separately mining of the groups would lead to the poor pruning opportunities which can greatly deteriorate efficiency.
Alternatively, we could encode the group as a variable and run an association rule learner on this representation. But, this will not return group differences, and the results will be difficult to interpret, since it is difficult to tell what is different between the two groups. First, there are too many rules to compare, and second, the results are difficult to interpret because the rule learner does not use the same attributes to separate the groups [15]. But, even with matched rules, we need a statistical test for comparison to see if differences are significant. In contrast sets that is clearly specified and that is their advantage.

V. RESEARCH METHODOLOGY

Research follows steps of knowledge discovery in data and consists of: (1) feature selection, (2) classification and evaluation, (3) comparison of the results. First, data sets of different characteristics are collected. Sources of data sets are public repositories containing referent data sets with accompanying documentation for each set. In order to extract the features with maximum information for classification, feature selection is performed on each data set. Comparisons of contrast set mining techniques with benchmarking feature selection techniques are performed. For the first time contrast set mining techniques are applied here as feature selection techniques. Classification is performed on selected features by applying classifiers that represent different approaches to classification: a statistical approach (discriminant analysis) and neural computing approach (neural networks). The classification is performed by applying each classifier on each data set that meets the requirements of algorithm. Feature selection techniques` performance relates to: (1) elapsed time (time of processor required to perform feature selection) and (2) accuracy of classifier. Accuracy of classification algorithms is the ability of the algorithm to accurately classify a large number of samples from the data set. To do performance comparison, we conduct statistical testing for assessing the statistical significance of differences between individual techniques in time and accuracy. The purpose of the test is to determine whether the differences of the estimated mean values of classification accuracy and elapsed time are significant. Thus, we want to gather evidence about the degree to which the results are representative for the generalization about the behavior of the feature selection techniques [24]. By performing analysis we want to determine do contrast set mining techniques outperform benchmarking feature selection techniques in terms of speed and classification accuracy.

A. Research hypothesis

Following research hypothesis are set up:

H1: Contrast set mining techniques will faster conduct feature selection than benchmarking feature selection techniques.

H2: Application of contrast set mining techniques in feature selection will provide more accurate classification than use of benchmarking feature selection techniques.

H3: Contrast set mining techniques will perform feature selection efficiently on sparse data sets than benchmarking feature selection techniques.

H3.a. Contrast set mining techniques will perform faster feature selection on sparse data sets than benchmarking feature selection techniques.

H3.b. Contrast set mining techniques will perform more accurate feature selection on sparse data sets than benchmarking feature selection techniques.

We will accept hypothesis H1 if contrast set mining techniques will faster select features than benchmarking feature selection techniques in more than 50% of analyzed data sets. Comparison is performed on 128 data sets.

We will accept hypothesis H2 if application of contrast set mining techniques in feature selection will result with more accurate classification than use of benchmarking feature selection techniques in more than 50% of analyzed data sets. Comparison is performed on 128 data sets in case of neural networks as classifier and 64 data sets in case of discriminant analysis as classifier.

We will accept hypothesis H3a if contrast set mining techniques will faster select features on sparse data sets than benchmarking feature selection techniques in more than 50% of analyzed data sets. Comparison is performed on 64 sparse data sets. We will accept hypothesis H3b if application of contrast set mining techniques in feature selection on sparse data sets will result with more accurate classification than use of benchmarking feature selection techniques in more than 50% of analyzed data sets. Comparison is performed on 64 sparse data sets in case of neural networks as classifier and 16 sparse data sets in case of discriminant analysis as classifier.

Literature review pointed out following feature selection techniques as benchmarking: Relief, Gain ratio, information gain, linear forward selection and voting technique [22]. Those techniques were used in hypothesis testing.

VI. CONTRAST SET MINING FOR FEATURE SELECTION

This paper proposes feature selection techniques that are created by combination of:

- feature evaluation measure to assign individual preference values to each feature,
- cutting criterion to choose the number of features selected.

Arauzo – Azofra et. al. suggested five measures for feature evaluation. The description of the measures follows [4]:

- **Mutual information**, also known as information gain, measures the quantity of information that a feature gives about the class. It is defined as the difference between the entropy of the class and the entropy of the class conditioned to knowing the evaluated feature.
- **Gain ratio** is defined as the ratio between information gain and the entropy of the feature. In this way, this measure avoids favoring features with more values, which is the natural behavior of previous measure.
- **Gini index** represents probability of two instances randomly chosen having a different class.
- **Relief-F** is an extension of the original Relief developed by Kononenko [21]. It can handle discrete and continuous attributes. Despite evaluating individual features, Relief takes into account relation among features. This makes Relief-F to perform very well.
- **Relevance** is a measure that discriminates between attributes on the basis of their potential value in the formation of decision rules [16].

Arauzo – Azofra et. al. described six general cutting criteria [4]:

- **Fixed number** (n) simply selects a given number of features. The selected features are the ones with greater evaluation.
- **Fraction** (p) selects a fraction, given as a percentage, of the total number of available features.
- **Threshold** (t) selects the features whose evaluation is over a user given threshold.
- **Threshold given as a fraction** (pm) selects the features whose evaluation is over a threshold, where this threshold is given as a fraction of the range of evaluation function.
- **Difference** (d) selects features, starting from the one with greater evaluation and following the sorted list of features, until evaluation difference is over a threshold.
- **Slope** (s), on the sorted list of features, selects best features until the slope to the next feature is over a threshold.

In this section, we explain in detail proposed techniques called:

SfFS (**S**tucco **f**or **F**eature **S**election) and

MOFS (**M**agnum **O**pus **F**eature **S**election).

Proposed methodology utilizies feature independence assumption. In literature we can find variety advantages of this assumption: simplicity, scalability and effectiveness in dealing with large data sets [48]. It was used by: Kudo & Sklansky, 1998 [30]; Blum & Langley, 1997 [6]; Guyon & Elisseeff, 2003 [21] and Abe, Kudo, Toyama, & Shimbo, 2006 [1]. Feature independence assumption implies use of an evaluation function which assigns evaluation measure to each attribute. After feature evaluation, those with the highest values are selected. To complete the selection process, cutting criterion is applied that determines where the selection stops.

Arauzo-Azofra, Aznarte and Benitez argue that one can`t generally recommend one evaluation measure and one cutting criterion [4]. Therefore, we analyzed papers that cite Arauzo-Azofra, Aznarte and Benitez in the database Scopus to see whether it is in one of the later studies made an evaluation. Of the eight papers that cite Arauzo-Azofra, Aznarte and Benitez in the database Scopus, one proves that the most effective cutting criterion is threshold [38]. Guided by their results, contrast set mining techniques in feature selection are using threshold as cutting criterion. As an evaluation measure, relevance is used. It is defined as a measure which discriminates between features on the basis of their potential in forming rules [16]. The reason for this lies in the fact that the contrast set mining techniques, STUCCO and Magnum Opus, are essentially defined in such a way to give as the result rules and measures of the quality of rules (measure that differes features with respect to their potential in defining rules). Measures are: deviation in case of SfFS and leverage in case of MOFS. Deviation is only measure STUCCO provides, whereas leverage is the best in case of STUCCO according to Piatetsky-Shapiro. He argues that many measures of rule value are based on the difference between the observed joint frequency of the antecedent and consequent, support(X!Y), and the frequency that would be expected if the two were independent, cover(X) × cover(Y) [37]. He asserts that the simplest such measure is leverage. Leverage is of interest because it measures the number of additional records that an interaction involves above and beyond those that should be expected if one assumes independence [37]. This directly represents the volume of an effect and hence will often directly relate to the ultimate measure of interest to the user such as the magnitude of the profit associated with the interaction between the antecedent and consequent.

The techniques considered in this paper utilize evaluation functions that assign an evaluation value to each feature. Once features have been evaluated, techniques based on individual evaluation always select those features with best evaluation. However, this is not all. To complete feature selection, they need to determine how many features are selected and how many are discarded. Contrast set mining techniques in feature selection apply relevance as evaluation measure and threshold defined by user as cutting criterion. The procedures of the proposed methodology for both algorithms are described below.

First, MOFS pseudocode is in Fig 1.

```
Input: Data set $S = \{A_1, A_2, \dots A_n\}$

$Leverage_{min}$ // minimum value of leverage defined by user

m// number of generated contast sets

p// statistical significance of the rule

i, feature //auxiliary variables

Output: Selected subset of features P

    1) $P_0 = generate\_contrast\_set$ (S) //generate contrast sets in form: LHS->RHS

    2) FOR (i=1; i<= m; i++){
```

Figure 1. MOFS pseudocode

MOFS algorithm calculates leverage value and statistical significance of the rule (p value). All features on the left side of statistically significant rules (rules with p<0.05) with leverage value higher than user defined value are selected in subset.

STUCCO algorithm finds contrasting sets that are deviations. Deviation is contrast set that is significant and large. Contrast set for which at least two groups differ in their support is significant. To determine the significance chi-square test is performed with the null hypothesis that the support of contrast set is equal between groups. In calculating, chi square test checks the value of the distribution. The value must be less than the defined threshold of statistical significance (p=0.05). Contrast set for which the maximum difference between the support is greater than the value mindev (minimum deviation) is large. In SfFS selected are those features which are on the left side of the contrast set that is significant and large. SfFS pseudocode is in Fig 2.

```
Input: Data set $S = \{A_1, A_2, \dots A_n\}$

    $min\_dev$ //minimum value of deviation defined by user

    m// number of generated contast sets

    p// statistical significance of the rule

    i, feature //auxiliary variables

Output: Selected subset of features P

    1)$P_0 = generate\_contrast\_set$ (S) //generate contrast sets in form: LHS->RHS

    2) FOR (i=1; i<= m; i++){

            IF

            ($p < 0.05$ $AND$ $deviation \geq \min\_dev$ )
```

Figure 2. SfFS pseudocode

VII. RESEARCH DESCRIPTION

The goal of this work is to compare feature selection techniques taking into account all factors, so a complete experimental setup has been used. In this setup, the number of independent experiments is the number of the possible combinations of the three factors: number of data sets (128), number of feature selection techniques (7) and number of classifiers (2). We designed and conducted an extensive and rigorous empirical study, out of which meaningful conclusions may be drawn. In this section, we provided a detailed description of the experimental setup. The main measures considered to evaluate the feature selection techniques are: classification accuracy and elapsed time.

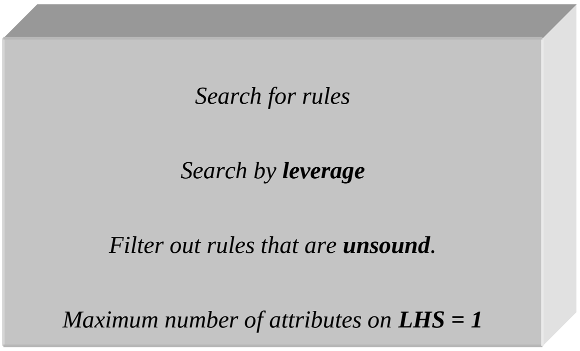
In order to get reliable estimates for classification accuracy, every experiment has been performed using 10-fold cross-validation. Any result shown is always the average of the 10-folds. The significance of results is assessed using statistical test, Friedman test. The Friedman test is a non-parametric equivalent of the repeated-measures ANOVA. It was used here since all ANOVA`s assumptions were not met. Friedman test ranks techniques for each data set separately, the best performing algorithm getting the rank of 1, the second best rank 2, and so on [13]. To get more details on Friedman test see Demsar, 2006 [16].

In order to include a wide range of classification problems, the following publicly available repositories have been explored seeking for representative problems with diverse data set characteristics (different number of features and instances, data distribution, level of noise, correlation,..): UCI Machine Learning Repository [42], StatLib - Carnegie Mellon University [41], Sociology Data Set Server of Saint Joseph`s University in Philadelphia [40], Feature selection datasets at Arizona State University [19]. Finally, 128 data sets were chosen. In order to estimate the quality of feature selection performed by each technique, the selected features are tested in a complete learning scenario of classification problems. The following well known learning methods are considered: neural networks and discriminant analysis. These methods have been chosen to cover the categories of methods most used.

This section provides empirical comparison of benchmarking feature selection techniques with contrast set mining techniques, for the first time used in feature selection. Techniques are demonstrated on the example of one data set, vote from University of California repository.

A. Feature selection with MOFS

MOFS is applied as described in section 5. The feature selection techniques considered have some parameters that must be set before running the algorithms. MOFS parameters are in figure 3.



Search for rules

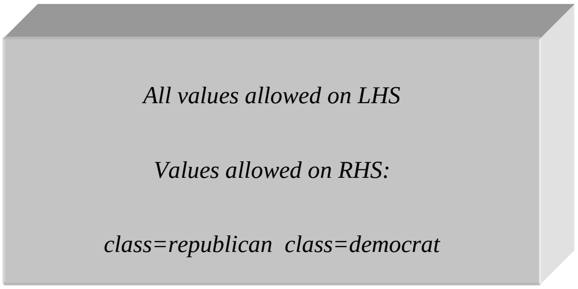
Search by ***leverage***

Filter out rules that are ***unsound****.*

Maximum number of attributes on ***LHS = 1***

Figure 3. MOFS settings

Defined settings determine the following aspects. Measure is impact of the quality of the rule, and features are ranked with respect to the value of the measure. As the filter, unsound option is used. Unsound filter extracts only statistically significant rules that have the value of p <0.05. Furthermore, only one feature is allowed on the left side of the rule. When applying Magnum Opus in feature selection this setting is extremely important because it is not taking into account multiple features on the left side and interaction of the features is avoided. In the feature selection with Magnum Opus, through the rules, we want to see the impact of single feature on the class attribute, but not the impact of group features to the class attribute. Hence on the right side is just one feature, that is class feature (has two values: republican and democrat).



All values allowed on LHS

Values allowed on RHS:

class=republican class=democrat

Figure 4. Allowed values in the rules

As a result of the execution of Magnum Opus, 10 statistically significant rules were produced. Below is one rule, which serves as an example.

physician-fee-freeze=y -> class=republican

[Coverage=0.407 (177); Support=0.375 (163); Strength=0.921; Lift=2.38; ***Leverage=0.2176*** *(94.6); p=4.55E-095]*

The first line of the rule gives contrast set. The values in parentheses are measures of the quality of the rule (from coverage to leverage), followed by p value - statistical significance of the rule. Leverage is bold because based on this measure features are evaluated in the process of feature selection.

As the result of feature selection process, features which are found on the left side of the rule with every value, are selected. For vote data set, they are the following four features:

- *physician-fee-freeze*
- *adoption-of-the-budget-resolution*
- *el-salvador-aid*
- *education-spendin*

B. Feature selection with SfFS

SfFS is performed under the following settings: *minDev* = *0.1, alpha = 0.05, surprisingThreshold = 0.*2. Thus, SfFS seeks for the statisticaly significant sets on the level of *p<0,05* with 0.1 as minimal difference in support.

These are also default values from successful Webb`s research [46]. Four significant and large contrast sets are gained. One of them is below.

```
==== Node: SUPERFUND_RIGHT_TO_SUE = y;
Contingency table:
republican      democrat
T:              4           4
F:              0           1
P:         1,00000      0,800000
```

Four selected features are:

SUPERFUND_RIGHT_TO_SUE
EDUCATION_SPENDING
CRIME
WATER_PROJECT_COST_SHARING

The selected features are used in the further steps of knowledge discovery in data.

VIII. EXPERIMENTAL RESULTS

The experiments described generated a large amount of resulting data. An appropriate summarizing analysis is necessary to interpret them and achieve conclusions. The results are described in four parts. First, a comparison of the feature selection techniques is provided in case of neural network classifier accuracy. Second, a comparison of techniques regarding discriminant analysis accuracy and, than, the comparison of the elapsed time of feature selection. These three parts of results relates to whole group of data sets used, that is, 128 data sets elapsed time and neural network accuracy measuring and 64 data sets for discriminant analysis accuracy measuring. In the fourth part we present results on sparse data sets. From the total of 128 data sets, 64 of them were sparse data sets.

For every classifier, all feature selection techniques have been compared. In this way, we can compare the effect of feature selection on each classification algorithm

A. Classification accuracy

Results of neural network classification revealed following. Of the 128 data set in the 82,03% cases (105 data sets) contrast set mining techniques in feature selection yielded statistically significantly more accurate classification compared to other feature selection techniques. In 17,97% of cases (23 data sets) yielded poorer (lower classification accuracy) results than others or not statistically significantly better than others.

- On the 23 data sets contrast set mining techniques obtained worse accuracy:

o For 12 data sets Relief obtained better accuracy,

o For 4 data sets InfoGain obtained better accuracy,

o For 2 data sets Linear forward selection obtained better accuracy,

o For 2 data sets contrast set mining techniques were better, but the difference in accuracy between them and other techniques was not statistically significant.

Figure 5 shows the comparison of feature selection techniques in terms of average neural network accuracy on all 128 data sets. The two best techniques are SfFS and MOFS. It is important to notice that these results show how two techniques based on contrast set mining have good classification performance.

With an intension to find out whether the same feature selection technique may lead to best results for various datasets on various classification algorithms, experiments are conducted with two different classification algorithms. Classifiers that possess different nature and biases may have a different effect on feature selection. For example, classifiers with one type of bias may be more (or less) suited to selecting relevant features from a dataset than classifiers with another type of bias.

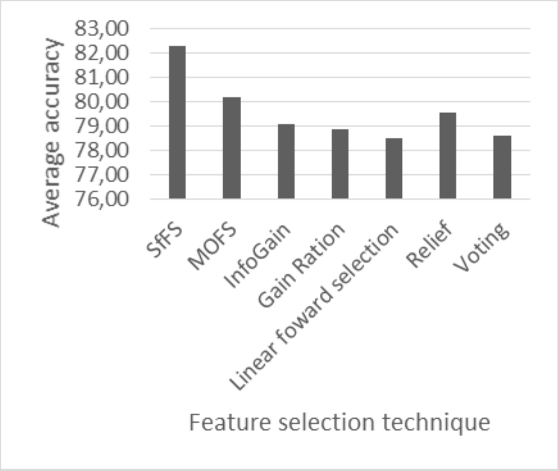


Figure 5. Neural network accuracy results

With This may be due to the fact that the biases made by one of the classifiers match (or do not match) the underlying biases and characteristics of the dataset used [13]. The next classification algorithm used in the evaluation of feature selection techniques' performance is discriminant analysis. The verification of the performance is conducted in the same manner as in case of the neural network classifier.

Discriminant analysis was performed on 32 data sets that have met the requirements of discriminant analysis. In 78,12% of cases (25 data sets) contrast set mining techniques in feature selection resulted with more accurate classification.

On the 7 data sets contrast set mining techniques obtained worse accuracy:

o For 3 data sets Relief obtained better accuracy,

o For 1 data set InfoGain obtained better accuracy,

o For 1 data set Linear forward selection obtained better accuracy,

o For 2 data sets contrast set mining techniques were better, but the difference in accuracy between them and other techniques was not statistically significant.

Figure 6 shows graphically discriminant analysis accuracy results. Feature subset selected by SfFS provides highest classification accuracy, followed by MOFS.

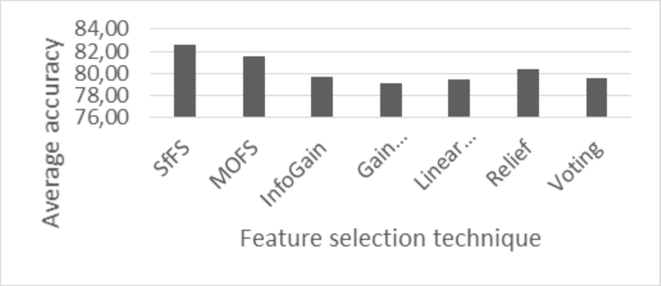


Figure 6. Discriminant analysis accuracy results

We can observe that the average accuracy results of neural network seem to be similar to the results obtained with discriminant analysis. In both cases, the same four feature selection techniques had the best behavior.

B. *Elapsed time*

Big data analysis intends to be performed in real time and elapsed time is also important to be measured. Thus, in this experiment, the effectiveness of the proposed techniques was evaluated in a two-stage scheme. Hereinafter are results of feature selection techniques comparison regarding elapsed time. By elapsed time we mean CPU time required for the implementation of the feature selection. Elapsed time was measured in seconds.

Of 128 data sets, for 60,94% of them contrast set mining techniques executed feature selection quicker than benchmarking feature selection techniqus.

In 39,06% of cases contrast set mining techniques achieved worse resutls or there were not significant differences between the results obtained by different techniques:

- For 20 data sets InfoGain yielded beter results,
- For 9 data sets Gain Ratio yielded beter results,
- For 3 data sets Relief yielded beter results,
- For 18 data sets difference between MOFS and other techniques were not statistically significant.

Figure 7 shows the comparison of feature selection techniques in terms of average time of selection on all of 128 data sets.

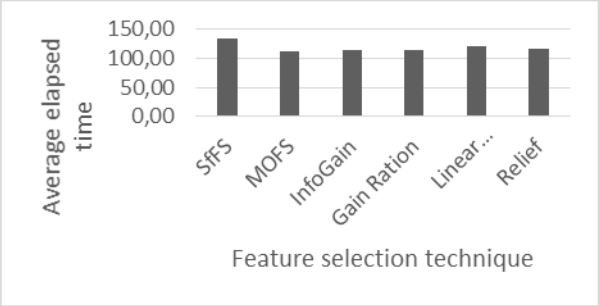


Figure 7. Average elapsed time results

As shown on figure, average elapsed time is the lowest for MOFS, followed by Info Gain and Gain Ratio. SfFS has maximum elapsed time. The reason is because SfFS is implemented as interpreter, whereas the other techniques are compilers.

C. *Results on sparse data sets*

Results of neural network classification on 64 data sets with high sparsity revealed following. On the 57 data sets contrast set mining techniques in feature selection yielded statistically significantly more accurate classification compared to other feature selection techniques. On the 7 data sets contrast set mining techniques yielded poorer (lower classification accuracy) results than others or not statistically significantly better than others:

- For 5 data sets Relief obtained better accuracy,
- For 2 data sets InfoGain obtained better accuracy.

Results of discriminant analysis classification on 16 data sets with high sparsity revealed following. On the 14 data sets contrast set mining techniques in feature selection yielded statistically significantly more accurate classification compared to other feature selection techniques. On the 2 data sets contrast set mining techniques yielded poorer (lower classification accuracy) results than others or not statistically significantly better than others:

- For 1 data set Relief obtained better accuracy,
- For 1 data set Linear forward selection obtained better accuracy.

Figure 8 demonstrates comparison of the results obtained by neural network and discriminant analysis classification. Results are presented as percentage of data sets.

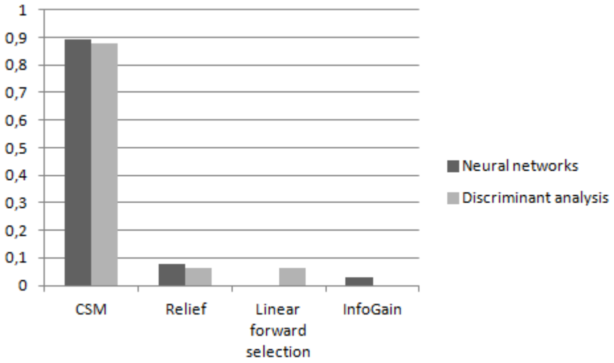


Figure 8. Comparison of classification results

We conduct additional experiments to observe the speed capability of contrast set mining on sparse data sets. The feature selection techniques based on contrast set mining outperform all other techniques in account, when applying on sparse data sets. The speed of the two contrast set mining methodologies is better in almost 90% analysed data sets. Out of the 64 data sets, contrast set mining faster conducted feature selection on 56 data sets. Figure 9 gives full statistics of the performance metrics.

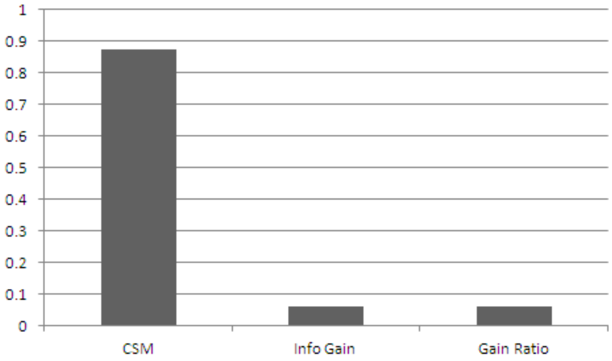


Figure 9. Techniques speed on sparse data sets

Contrast set mining techniques are superior to other feature selection techniques at the significant level of 5% or above.

IX. CONCLUSION

The central claim of this paper is that feature selection for classification most effectively can be accomplished on the basis of contrast set mining approach. A feature selection algorithms were implemented and empirically tested to support this claim.

In the field of binary classification problems, an extensive empirical study (128 data sets was analyzed) on feature selection techniques based on contrast set mining has been conducted and presented in this paper. These techniques, created by combination of the relevance as evaluation measure and threshold as cutting criterion, are explored and compared with benchmarking feature selection techniques. The results indicate that the optimal feature subset selected by the proposed techniques has a good classification performance and it is performed quickly. Furthermore, these techniques shown to be particularly effective on sparse data sets. Thus, all three our research hypotheses are accepted.

The research contributions for feature selection and data mining are: (1) inovative feature selection techniques based on contrast set mining (called SfFS and MOFS). Research results indicated their superiority in terms of accuracy and speed. (2) Research imposes new challenges in terms of evaluation in data mining field: in-depth comparison was done regarding number of data sets used in comparison (128), number of feature selection techniques (7) and number of classifiers. Since machine learning research has traditionally concentrated on small number of data sets and has routinely used small number of techniques in evaluation, this research represents step forward. (3) points out the need to investigate data sets characteristics prior of applying feature selection.

Nevertheless, there are some limitations that should be considered when interpreting the results of this research: (1) contrast set mining techniques in feature selection are defined with the assumption of feature independence. Although it has numerous advantages, this is limitation when some features interact. (2) Techniques are evaluated only on datasets with two classes. In future research it can be extended performing the evaluation on data sets with multiple classes.

For the future work, we intend to investigate whether data set some other characteristics (e.g. number of features, number of instances, noise, class imbalance [34]) affect feature selection techniques´ performance. Based on the results, we could develop recommender system which is able to suggest feature selection technique for data set of particular characteristics.

References

[1] N. Abe, M. Kudo, J.Toyama, H. Shimbo, « Classifier-independent feature selection on the basis of divergence criterion », Pattern Anal Applic 9 (2006), pp. 127–137.

[2] R. Agrawal, T. Imielinski, A. Swami, « Mining association rules between sets of items in large databases », In Proceedings, ACM SIG-MOD Conference on Management of Data, Washington D.C., 1993, pp. 207-216.

[3] M. Alibeigi, S. Hashemi, A. Hamzeh, « Unsupervised feature selection based on the distribution of features attributed to imbalanced data sets », International Journal of Artificial Intelligence and Expert Systems 2 (1) (2011), pp.14-22.

[4] A. Arauzo – Azofra, J.L. Aznarte, J.M. Benitez, « Empirical study of feature selection methods based on individual feature evaluation for classification problems », Expert Systems with Applications 38 (7) (2011), pp. 8170-8177.

[5] S.D. Bay, M.J. Pazzani, « Detecting group differences: Mining contrast sets », Data Mining and Knowledge Discovery 5 (3) (2001), pp. 213-246.

[6] A.L.Blum, P. Langley, « Selection of relevant features and examples in machine learning », Artificial Intelligence 97 (1-2) (1997), pp.245-271.

[7] M. Boettcher, « Contrast and change mining », Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (3) (2011), pp.215–230.

[8] J.M. Cadenas, C.M. Garrido, R. Martinez, « Feature subset selection: Filter-Wrapper based on low quality data », Expert systems with applications 40 (2013), pp. 6241-6252.

[9] L. Cehovin, Z. Bosnic, « Empirical evaluation of feature selection methods in classification », Intelligent data analysis 14 (2010), pp.265-281.

[10] F.F.Chang, « Characteristics Analysis for Small Data Set Learning and the Comparison of Classification Methods », 7th WSEAS Int. Conf. on ARTIFICIAL INTELLIGENCE, KNOWLEDGE ENGINEERING and DATA BASES (AIKED'08), University of Cambridge, UK, Feb 20-22, 2008, pp. 122-127.

[11] F.M.Chang, « The Characteristics of Learning in Limited Data and the Comparative Assessment of Learning Methods »,WSEAS Transaction on Information Science & Applications, 5 (5), 2008, pp.407-414

[12] A., Chimienhti, P., Dalmasso, R., Nerino, G., Pettiti, M., Spertino, « Surface reconstruction from sparse data by a multiscale volumetric approach », Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Malta, September 15-17, 2005, pp.35-40.

[13] K.A. Chrysostomou, « The Role of Classifiers in Feature Selection: Number vs Nature ». Doctoral thesis, School of Information Systems, Computing and Mathematics Brune University (2008).

[14] M. Dash, H. Liu, « Feature Selection for Classification », An International Journal of Intelligent Data Analysis 1 (3,1) (1997), pp. 131-156.

[15] J. Davies, D. Bilman, « Hierarchical categorization and the effects of contrast inconsistency in an unsupervised learning task », In Garrison W. Cottrell (ed.), Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society, 750 (1996).

[16] J. Demsar, « Statistical Comparisons of Classifiers over Multiple Data Sets », Journal of Machine Learning Research 7 (2006), pp.1–30.

[17] M.D. Drugan, M.D., M.A.Wiering, « Feature selection for Bayesian network classifiers using the MDL-FS score », Internation journal of approximate reasoning 51 (2010) pp. 695-717.

[18] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, « From data mining to knowledge discovery in databases », AI magazine 17 (3) (1996), pp.37-54.

[19] Feature selection datasets at Arizona State University, available at: http://featureselection.asu.edu/datasets.php, last accessed: 20.01.2013.

[20] X. Geng, T.Y. Liu, T. Qin, T., H. Li, « Feature selection for ranking », SIGIR: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (2007), pp.407- 414.

[21] I. Guyon, A. Elisseeff, « An introduction to variable and feature selection », Journal of Machine Learning Research 3 (2003), pp. 1157–1182.

[22] M.A. Hall, G. Holmes, « Benchmarking attribute selection techniques for discrete class data mining », IEEE transactions on knowledge and data engineering 15 (3) (2003), pp. 1-16.

[23] A.C. Haury, P. Gestraud, J.P. Vert, « The Influence of Feature Selection Methods on Accuracy », Stability and Interpretability of Molecular Signatures, PLoS ONE 6(12): e28210. doi:10.1371/journal.pone.0028210 (2011)1-16.

[24] N. Japkowicz, M. Shah, « Evaluating learning algorithms: A classification perspective ». (1st ed.) Cambridge University Press, New York, 2011.

[25] G.H.John, R. Kohavi, K. Pfleger, « Irrelevant features and the subset selection problem ». Machine Learning: Proceedings of the Eleventh International Conference, edited by William W. Cohen and Haym Hirsh, (1994), pp.121-129.

[26] D. Koller, M. Sahami, « Toward optimal feature selection ». Proceedings of the Thirteenth International Conference on Machine Learning (1996), pp. 284-292.

[27] R. Kohavi, G.H. John, « Wrappers for feature subset selection », Artificial Intelligence 97(1-2) (1997), pp.273–324.

[28] R. Kohavi, D. Sommerfield, « Feature subset selection using the wrapper method: overfitting and dynamic search space topology », Proceedings of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (1995), pp.192-197.

[29] P. Kralj Novak, N. Lavrac, D. Gamberger, A. Krstacic, « CSM-SD: Methodology for contrast set mining through subgroup discovery », Journal of Biomedical Informatics 42(1) (2009), pp.113–122.

[30] M. Kudo, J. Sklansky, « Classifier-independent feature selection for two-stage feature selection », In: Amin A, Dori D, Pudil P, Freeman H (eds) Proceedings of the Joint IAPR International Workshops on SSPR'98 and SPR'98, (1998), pp. 548–555.

[31] D. Lavanya, K. Usha Rani, « Analysis of feature selection with classification: breast cancer datasets », Indian Journal of Computer Science and Engineering (IJCSE) 2 (5) (2011), pp.756-763.

[32] H. Liu, H. Motoda, L. Yu, « Feature selection with selective sampling », Proceeding ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning (2002), pp. 395-402.

[33] X. Lu, « Applying pattern discovery methods to a healthcare data ». PhD thesis, UniSA School of Computer and Information Science, 2009.

[34] S. Maldonado, R. Weber, F. Famili, « Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines », Information Sciences (2014),DOI: 10.1016/j.ins.2014.07.015

[35] J. Novakovic, P. Strbac, D. Bulatovic, « Toward optimal feature selection using ranking methods and classification algorithms », Yugoslav Journal of Operations Research 21 (1) (2011), pp. 119-135.

[36] S. Oreski, D. Oreski, G. Oreski, « Hybrid System with Genetic Algorithm and Artificial Neural Networks and its Application to Retail Credit Risk Assessment », Expert systems with applications 39 (16) (2012), pp. 12605–12617.

[37] G. Piatetsky-Shapiro, « Discovery, analysis, and presentation of strong rules », In: Piatetsky-Shapiro, G., Frawley,W. (eds.): Knowledge Discovery in Databases, Menlo Park, CA: AAAI Press, (1991), pp. 229-248.

[38] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz, M. Garcia-Torres, « Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches », Expert Systems with Applications 39 (2012), pp.11094–11102.

[39] L.O.L.A. Silva, M.L.Koga, C.E.Cugnasca, A.H.R. Costa, « Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings », Compters and electronics in agriculture 97 (2013), pp.47-55.

[40] Sociology Data Set Server of Saint Joseph`s University in Philadelphia, available at: http://sociology-data.sju.edu/, last accessed: 14.12.2012.

[41] StatLib - Carnegie Mellon University, available at: http://lib.stat.cmu.edu/, last accessed: 10.12.2012.

[42] UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ml/datasets.html, last accessed: 29.10.2013.

[43] C.M.,Van der Walt, « Data measures that characterise classification problems », Master's Dissertation, http://upetd.up.ac.za/thesis/available/etd-08292008-162648/, downloaded: 21.09.2014.

[44] J.L.,Vasquez, J. Vasquez, J.C., Briceño, E. Castillo, C.M. Travieso, « Feature selection of RAPD haplotypes for identifying peach palm (Bactris gasipaes) landraces using SVM, WSEAS Transactions on Computers archive, 9 (3), 2010, pp. 205-214.

[45] G. I. Webb, « Efficient Search for Association Rules », In R. Ramakrishnan and S. Stolfo (Eds.), Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000) Boston, MA. New York: The Association for Computing Machinery, (2000), pp. 99-107.

[46] G. I.Webb, S. Butler, D. Newlands, « On detecting differences between groups ». KDD 2003., Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM., (2003), pp. 739 – 745.

[47] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, « Feature selection for SVMs », Advances in Neural Information Processing Systems 13 (2001), pp. 668 –674.

[48] L. Yu, H. Liu, « Efficient Feature Selection via Analysis of Relevance and Redundancy », Journal of Machine Learning Research 5 (2004) pp. 1205–1224.

[49] S.Zhang, X. Zhao, B. Lei, « Facial Expression Recognition Using Sparse Representation », WSEAS transactions on systems,8 (11), 2012, pp. 440-452.