

# Handling Sparse Data Sets by Applying Contrast Set Mining in Feature Selection

Dijana Oreški, Mario Konecki\*

Faculty of Organization and Informatics, Pavlinska 2, 42000 Varaždin, Croatia.

\* Corresponding author. Email: mario.konecki@foi.hr

Manuscript submitted September 15, 2015; accepted December 1, 2015.

doi: 10.17706/jsw.11.2.148-161

---

**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 on 64 sparse data sets was conducted. Results have shown superiority of contrast set mining techniques in more than 80% of the analysis on sparse data sets. This paper provides a study on the new methodologies and detected superiority in handling data sparsity.

**Key words:** Classification, contrast set mining, data characteristics, data sparsity, feature selection.

---

## 1. Introduction

Modern organizations deal with the constant growth of available electronic data. Due to modern technologies, collecting data has become easy and it ceased to be a problem. Instead, the data analysis and understanding of the data reduction in a usable way has come into focus. The main part of this problem is the process of knowledge discovery in data. According to Fayyad and Piatetsky-Shapiro this process consists of the following steps: developing an understanding of the application domain, creating a target data set, data cleaning and preprocessing, data reduction and transformation, data mining, evaluation, constructing discovered knowledge [1]. Data preparation consists of data cleaning and data reduction and takes from 60% to 95% of the whole process time. Data reduction decreases the complexity of the data set by reducing the number of features and the size of the instance space. There are two groups of dimensionality reduction: selection of a smaller subset of inputs (feature selection) and feature reduction. Feature selection is in the focus of this paper. Main idea of feature selection 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 [2]-[5]. 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 on sparse data sets.

Contrast set mining is a subfield of data mining and it 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 [6], [7]. The advantage of this approach is 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 [8]. It is not possible to select an algorithm and claim its superiority over competing algorithms without taking into consideration the data characteristics as well as the suitability of the algorithm for such data. This paper focuses on data sparsity characteristic since enforcing sparsity can greatly improve on performance of methods [9]. 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 whether contrast set mining techniques outperform classical feature selection techniques when dealing with data sparsity. Comparison of contrast set mining techniques in feature selection with benchmarking feature selection techniques is performed on 64 sparse data sets.

The paper is organized as follows. In Section 2 feature selection techniques comparisons reported in the literature are briefly presented. Data sparsity and important data set characteristic are presented in Section 3. Section 4 describes contrast set mining techniques, STUCCO and Magnum Opus as a basis for the defining of new approach, proposed in this paper. In Section 5 experimental framework and research hypothesis are presented. In Section 6 the proposed contrast mining approach is described. Research is described in Section 7 and results are discussed in Section 8.

## 2. Feature Selection

Feature selection is an active field in computer science [10]. Feature selection is defined as a search problem on the power set of the set of available features [11]. Aim of feature selection is to find a subset of features that, in some aspect, enables improvement of learning activity. Feature selection is a rich field of research and development and many new feature selection techniques are emerging. In this paper details of particular techniques will not be explained. Instead, the paper will focus on studies which compared feature selection techniques to identify their research scope (number of techniques in comparison, number of data sets and criteria). John *et al.* describe a technique for feature 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 [11]. Kohavi and Sommerfield compared forward and backward selection on 18 data sets [12]. Koller and Sahami introduced information theory based feature selection technique. They have tested new technique on 5 data sets [13]. Kohavi and John introduced wrapper approach and compared it to Relief, filter approach to feature selection. Significant improvement in accuracy is achieved for datasets in the two families of induction algorithms used: decision trees and Naive-Bayes [14].

Dash and Liu gave comprehensive overview of many existing techniques from the 1970's to the 1997 and categorized the different existing techniques in terms of generation procedures and evaluation function [15]. Furthermore, they have chosen representative techniques from each category. Their comparative analysis was performed on 3 data sets. The method was superior to some standard feature selection algorithms on 4 data sets tested. Liu *et al.* tested Relief algorithm on 16 data sets [10]. Geng *et al.* introduced new feature selection technique based on the similarity between two features. New approach was tested on 2 data sets [16]. Alibeigi *et al.* suggested new filter feature selection technique and compared it with 3 techniques on 3 data sets [17]. 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 [18]. 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 [19]. Lavanya and Usha Rani investigated performance of feature selection techniques on 3 data sets. Results did not indicate superiority of one technique on all data sets. They used classification accuracy and time required for feature selection as comparison criteria [20]. Novakovic *et al.* compared 6 feature selection techniques on 2 data sets and they used classification accuracy as a criterion [21]. Haury *et al.* compared 8 feature selection techniques on 4 data sets [22]. 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 [23].

Survey of previous research pointed out classification accuracy and elapsed time of feature selection as 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, several limitations of previous comparative analysis have been identified:

- 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 do not represent real-world problems
- number of data sets in the analysis was very small
- only one criterion was used in the comparison
- Conducted research was aimed at overcoming these limitations as follows:
- feature selection techniques comparison has been conducted on 64 data sets
- 7 feature selection techniques have been compared
- 2 different classifiers have been used in the learning process
- data sparsity has been taken into account

### 3. Data Sparsity

In this section the relationship between the dimensionality of data and the number of samples required to model the data accurately is investigated and discussed. This relationship is not trivial and Van der Walt [8] defined measure that captures this relevant factor.

#### 3.1. Theoretical Background

In this section the measure used to quantify whether the number of samples in a data set is sufficient to model the data accurately is explained in detail. This 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 [8]. To test if a linear relationship holds between  $d$  and  $N$  the normality test can be employed and correlations between features [40] can be examined. Thus, the number of parameters that must be estimated is  $2dC + C$ . To test if this quadratic relationship between  $d$  and  $N$  holds the homogeneity of class covariance matrices can be measured as well as the normality of the class data. Thus, the total number of parameters that must be estimated is:

How it can be decided 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 do not hold, an exponential relationship between  $N$  and  $d$  is possible. When the relationship between  $d$  and  $N$  is determined it can be quantified 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, a measure ( $N_{min}$ ) is defined, which sets the scale for the minimum number of samples that is required to model the data accurately.

$$d^2 + DC + C. \quad (1)$$

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:

$$N_{1(min)} = 2dC + C \quad (2)$$

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:

$$N_{q1(min)} = 2d^2 + dC + C \quad (3)$$

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 in the order of:

$$N_{q2(min)} = 2d^2 + dC + C \quad (4)$$

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:

$$N_{e(min)} = D^d \text{ steps} \quad (5)$$

where  $D^d$  steps is the discrete number of steps per feature. The next step is to quantify if the number of samples is 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. Thus, a measure of data sparsity is defined as follows:

$$DSR = N/N_{min} \quad (6)$$

where  $N_{min}$  is the appropriate minimum number of samples measure and  $N$  is the actual number of samples in the data set. A measure to indicate if the number of samples is sufficient by inverting equation is defined as follows:

$$DS = \sqrt[d]{N} \quad (7)$$

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

### 3.2. Measuring Data Sparsity

Hereinafter, two categories of sparsity for data set are defined: low and high. Data set has low sparsity if real number of samples is higher or equal to the number of samples required to model the data accurately. Data set sparsity is high if number of samples is smaller or equal to the number of samples required to model the data accurately. Data set vote, [24], [25] serves as an example to explain data sparsity measurement. Standard measures for vote data set are the following: standard measures: vote, dimensionality: 17, number of instances: 435.

By applying Kolmogorov Smirnov test the normality of the distribution has been tested. In vote data set no feature has normal distribution. Thus, exponential relationship between features in data set exists. Based on this, required number of samples for accurate modeling is calculated as:  $2^{17} = 131072$ .

Since minimal number of instances required for accurate modeling (131 072) is higher than actual number of instances (435), it can be concluded that there is not enough instances for precise modeling and that data sparsity of vote data set is HIGH.

## 4. Background on Contrast Set Mining

One of the newest trends in data mining field 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 a growing consensus that revealing how a domain changes is equally important as producing highly accurate models [26]. Nowadays, developing methods for analyzing and understanding of changes are seen as one of the primary research issues when dealing with evolving data [26]. Led by this practical need, subfield of data mining for analyzing changes was developed and named 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.

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 [27]. To differ meaningfully an item set's support difference must exceed a user-defined threshold. Description of STUCCO and Magnum Opus is provided in the following two sections.

### 4.1. 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 [27]. To discover contrast sets Bay and Pazzani proposed the STUCCO (Search and Testing for Understandable Consistent Contrast) algorithm [27]. 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 [26].

### 4.2. Magnum Opus

Magnum Opus is a commercial implementation of the OPUS AR rule-discovery algorithm. OPUS stands for Optimized Pruning for Unordered Search. 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 center 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 that 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 [28]. Rule value measures are central to the enterprise of k optimal rule discovery. Five such measures are stated. The available criteria for measuring interest include lift, leverage, strength (also known as confidence), support and coverage.

### **4.3. Contrast Set Mining Techniques Discussion**

Another approach used to distinguish two or more groups is to use a decision tree. This approach has the advantage of being fast in generating understandable models. However it also has some 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  $X \rightarrow Y$ . In market basket data  $X$  or  $Y$  are items such as beer or fruit. 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, sets that have support greater than a certain cutoff (these sets are then used to form the rules) are observed as well as contrast sets, where those sets which represent substantial differences in the underlying probability distributions are sought.

Since both techniques have a search element, there are many commonalities. Actually, in order to enhance contrast set algorithms some of the search work developed for association rule mining is applied. Although, 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 would not work effectively. For example, one approach would be to mine the large item sets for each group separately, and then, compare them. Separate mining of the groups would lead to the poor pruning opportunities which can greatly deteriorate efficiency.

Alternatively, the group as a variable could be encoded and ran an association rule learner on this representation. But, this would not return group differences, and the results would 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 [29]. But, even with matched rules, a statistical test for comparison to conclude whether differences are significant is needed. In contrast sets this is clearly specified and that is their advantage.

## **5. 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, a statistical testing for assessing the statistical significance of differences between individual techniques in speed and accuracy has been conducted. The purpose of the Friedman test is to determine whether the differences of the estimated mean values of classification accuracy and elapsed time are significant. Thus, it is necessary to gather evidence about the degree to which the results are representative for the generalization about the behavior of the feature selection techniques [30]. By performing analysis it is possible to determine whether contrast set mining techniques outperform benchmarking feature selection techniques in the terms of speed and classification accuracy.

### **5.1. Research Hypothesis**

The following research hypotheses are given:

H1: Contrast set mining techniques will perform feature selection on sparse data sets better than benchmarking feature selection techniques.

H1.a. Contrast set mining techniques will perform faster feature selection on sparse data sets than benchmarking feature selection techniques.

H1.b. Contrast set mining techniques will perform more accurate feature selection on sparse data sets than benchmarking feature selection techniques.

The hypothesis H1a will be accepted if contrast set mining techniques will select features on sparse data sets faster than benchmarking feature selection techniques in more than 50% of analyzed data sets. Comparison is performed on 64 sparse data sets. The hypothesis H1.b. will be accepted if application of contrast set mining techniques in feature selection on sparse data sets will result with more accurate classification than in the case of using 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 on 16 sparse data sets in case of discriminant analysis as classifier.

Literature review pointed out the following feature selection techniques as benchmarking: Relief, Gain ratio, information gain, linear forward selection and voting technique [31]. Those techniques were used in hypothesis testing.

## **6. 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 and cutting criterion to choose the number of features selected. In this section, proposed techniques are explained in detail and these techniques include: SfFS (Stucco for Feature Selection) and MOFS (Magnum Opus Feature Selection).

Proposed methodology utilizes feature independence assumption. In literature a variety advantages of this assumption can be found: simplicity, scalability and effectiveness in dealing with large data sets [32]. It was used by: Kudo & Sklansky, 1998 [33]; Blum & Langley, 1997 [34]; Guyon & Elisseeff, 2003 [2] and Abe,

Kudo, Toyama, & Shimbo, 2006 [35]. Feature independence assumption implies the 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.

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 [36]. 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 the rules and measures of the quality of rules (measure that differs features with respect to their potential in defining rules) as the result. 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, 1991 [37]. 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,  $\text{cover}(X) \times \text{cover}(Y)$  [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 the best evaluation. However, this is not all. To complete feature selection, there is a 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.

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  $\text{mindev}$  (minimum deviation) is large. In SfFS selected features are on the left side of the contrast set that is significant and large.

## **7. Research Description**

The goal of this work is to compare feature selection techniques taking into account all factors. Therefore, a complete experimental setup has been used. In this setup, the number of independent experiments is the number of the possible combinations of three factors: number of data sets (64), number of feature selection techniques (7) and number of classifiers (2). An extensive and rigorous empirical study, out of which meaningful conclusions can be drawn, has been designed and conducted. In this section, a detailed description of the experimental setup is provided. 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 always represents 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 1, the second best rank 2, and so on [38].

In order to include a wide range of classification problems, the following publicly available repositories have been explored, seeking for representative problems with sparsity data set characteristic: UCI Machine Learning Repository [24], StatLib - Carnegie Mellon University [25], Sociology Data Set Server of Saint Joseph's University in Philadelphia [39], Feature selection datasets at Arizona State University [40]. Finally, 128 data sets have been chosen. In order to estimate the quality of feature selection performed by each technique, the selected features have been tested in a complete learning scenario of classification problems. Following learning methods have been considered: neural networks and discriminant analysis. This section provides empirical comparison of benchmarking feature selection techniques with contrast set mining techniques. Techniques are demonstrated on the example of one data set, vote, from University of California repository.

### 7.1. Feature Selection with MOFS

MOFS is applied as described. The feature selection techniques considered have some parameters that must be set before running the algorithms. Those are as follows: search for rules, search by leverage, filter out rules that are unsound, maximum number of attributes on LHS = 1.

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, the impact of single feature on the class attribute tries to be determined, but not the impact of group features to the class attribute. Hence on the right side is just one feature that is a class feature (has two values: republican and democrat).

As a result of the execution of Magnum Opus, 10 statistically significant rules have been produced. One of these rules is given below.

*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 represent 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.

## 8. Experimental Results

The experiments described generated a large amount of resulting data. An appropriate summarizing analysis is necessary to interpret the data and achieve conclusions. The results are described in three 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 is given and finally, the comparison of the elapsed time of feature selection is provided. These three parts of results relate to the whole group of data sets used, that is, 64 data sets elapsed time and neural network accuracy measuring and 16 data sets for discriminant analysis accuracy measuring.

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

### 8.1. Classification Accuracy

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

- 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 (that have met the assumptions of analysis) revealed the following. On 14 data sets, contrast set mining techniques in feature selection yielded statistically significantly more accurate classification compared to other feature selection techniques. On 2 data sets, contrast set mining techniques yielded poorer (lower classification accuracy) results than other techniques or not statistically significantly better results than other techniques:

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

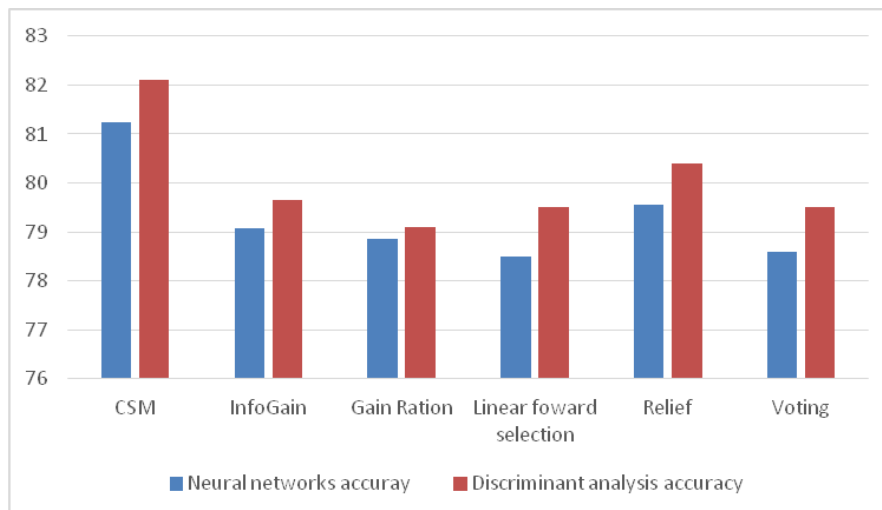


Fig. 1. Comparison of classification results.

Fig. 1 demonstrates comparison of the results obtained by neural network and discriminant analysis classification. Results are presented as percentage of data sets. Additional experiments to observe the speed capability of contrast set mining on sparse data sets have been conducted. The feature selection techniques based on contrast set mining outperformed all other considered techniques, when applied on sparse data sets. The speed of the two contrast set mining methodologies was better in majority of analyzed data sets. Contrast set mining techniques were superior to other feature selection techniques at the significance level of 5% or above. Results are presented in the next subsection.

### 8.2. Elapsed Time

Big data analysis intends to be performed in real time and elapsed time is also important and needs 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. Elapsed time denotes the CPU time required for the implementation of the feature selection. Elapsed time was measured in seconds.

Of 64 data sets, for 65,63% of them contrast set mining techniques executed feature selection quicker

than benchmarking feature selection techniques.

In 34,37% cases contrast set mining techniques achieved worse results or there were not significant differences between the results obtained by different techniques:

- For 8 data sets InfoGain yielded better results
- For 7 data sets Gain Ratio yielded better results
- For 1 data sets Relief yielded better results
- For 6 data sets difference between MOFS and other techniques was not statistically significant

Fig. 2 shows the comparison of feature selection techniques in terms of average time of selection on all of 64 data sets.

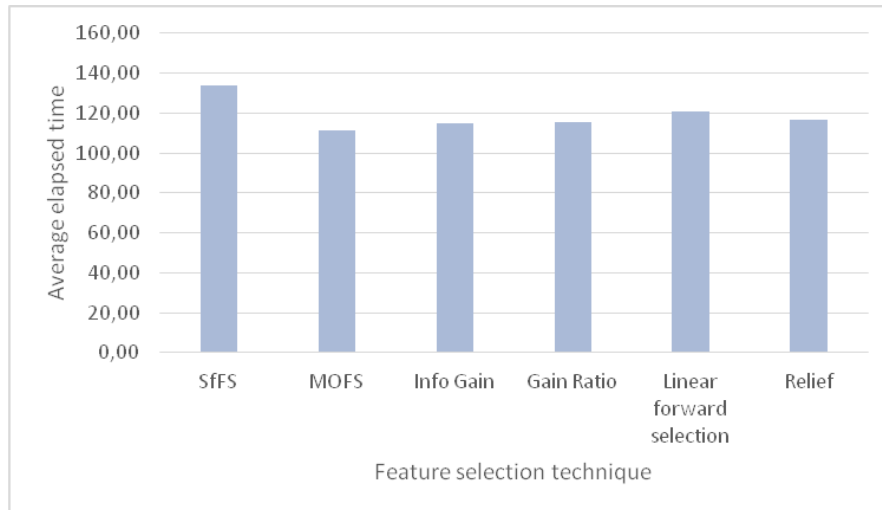


Fig. 2. Average elapsed time results.

As shown in Fig. 2, average elapsed time is the lowest for MOFS, followed by Info Gain and Gain Ratio. SfFS has maximum elapsed time. The reason for this is the fact that SfFS is implemented as interpreter, whereas the other techniques are compilers. That is one of the limitations of this work.

## 9. Conclusion

The hypothesis of this paper is that feature selection for classification on sparse data sets can be accomplished most effectively on the basis of contrast set mining approach. The feature selection algorithms were implemented and empirically tested to support this claim. In the field of binary classification problems, an extensive empirical study on feature selection techniques based on contrast set mining has been conducted and presented in this paper. These techniques have been 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 that it is performed quickly. Thus, the research hypotheses are accepted.

This research contributed towards advancement of the field of data mining in following aspects:

- 1) 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 (64), 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.
- 2) Research points out the need to investigate data sets characteristics prior to applying feature selection.

Main contribution of the research is in the feature selection domain. A novel methods have been proposed, namely, STUCCO for Feature Selection and Magnum Opus for Feature Selection, in order to train high performance models to generate significant features. A novel combination of metrics: threshold and relevance is proposed to evaluate the redundancy of targeting features. Experimental evaluations reveal that proposed feature mining and selection approaches outperform state-of-the-art techniques when dealing with data sparsity.

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 this has numerous advantages, there is a limitation when some features interact. (2) Techniques are evaluated only on datasets with two classes. In future research this can be extended to performing the evaluation on data sets with multiple classes. (3) Only one data set characteristic, data sparsity was examined.

In the future research, the possibilities of contrast set mining solutions for solving the following three problems will be explored: (1) feature noise, (2) low intrinsic dimensionality and (3) class imbalance. Since contrast set mining techniques have shown promising results, there are reasons to believe they could handle other data set characteristics. Especially ubiquitous is the class imbalance data [37], in which the classes are not equally represented and the minority of classes includes a much smaller number of examples than other classes. The conventional classifiers tend to be overwhelmed by the large classes while ignoring the smaller classes. In the future research this challenging task to overcome limitations of the prior techniques will be considered and addressed.

## References

- [1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-54.
- [2] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- [3] Arauzo - Azofra, A., Aznarte, J. L., & Benitez, J. M. (2011). Empirical study of feature selection methods based on individual feature evaluation for classification problems. *Expert Systems with Applications*, 38(7), 8170-8177.
- [4] Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, 39(16), 12605-12617.
- [5] Cadenas, J. M., Garrido, C. M., & Martinez, R. (2013). Feature subset selection: Filter-wraer based on low quality data. *Expert Systems with Applications*, 40(16), 6241-6252.
- [6] Novak, P. K., Lavrač, N., Gamberger, D., & Krstačić, A. (2009). CSM-SD: Methodology for contrast set mining through subgroup discovery. *Journal of Biomedical Informatics*, 42(1), 113-122.
- [7] Webb, G. I., Butler, S., & Newlands, D. (2003). On detecting differences between groups, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 739-745). New York: ACM.
- [8] Van der Walt, C. M. (2008). Data measures that characterize classification problems. Master's Dissertation. Retrieved September 21, 2014, from <http://upetd.up.ac.za/thesis/available/etd-08292008-162648/>.
- [9] Chimienhti, A., Dalmasso, P., Nerino, R., Pettiti, G., & Spertino, M. (2005). Surface reconstruction from sparse data by a multiscale volumetric approach, *Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision* (pp. 35-40). Malta, September 15-17.

- [10] Liu, H., Motoda, H., & Yu, L. (2002). Feature selection with selective sampling. *Proceeding ICML '02 Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 395-402).
- [11] John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 121-129).
- [12] Kohavi, R., & Sommerfield, D. (1995). 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* (pp. 192-197).
- [13] Koller, D., & Sahami, M. (1996). Toward optimal feature selection. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 284-292).
- [14] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2), 273-324.
- [15] Dash, M., & Liu, H. (1997). Feature selection for classification. *An International Journal of Intelligent Data Analysis*, 1(1), 131-156.
- [16] Geng, X., Liu, T. Y., Qin, T., & Li, H. (2007). Feature selection for ranking. *Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 407-414).
- [17] Alibeigi, M. H. & S. Hamzeh, A. (2011). Unsupervised feature selection based on the distribution of features attributed to imbalanced data sets. *International Journal of Artificial Intelligence and Expert Systems*, 2(1), 14-22.
- [18] Drugan, M. D., & Wiering, M. A. (2010). Feature selection for Bayesian network classifiers using the MDL-FS score. *International Journal of Approximate Reasoning*, 51(6), 695-717.
- [19] Cehovin, L., & Bosnic, Z. (2010). Empirical evaluation of feature selection methods in classification. *Intelligent data analysis*, 14(3), 265-281.
- [20] Lavanya, D., & Usha Rani, K. (2011). Analysis of feature selection with classification: Breast cancer datasets. *Indian Journal of Computer Science and Engineering (IJCSE)*, 2(5), 756-763.
- [21] Novakovic, J., Strbac, P., & Bulatovic, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 21(1), 119-135.
- [22] Haury, A. C., Gestraud, P., & Vert, J. P. (2011). The influence of feature selection methods on accuracy. *Stability and Interpretability of Molecular Signatures*.
- [23] Silva, L. O. L. A., Koga, M. L., Cugnasca, C. E., & Costa, A. H. R. (2013). Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings. *Computers and electronics in agriculture*, 97, 47-55.
- [24] UCI Machine Learning Repository. Retrieved October 29, 2014, from <http://archive.ics.uci.edu/ml/datasets.html>.
- [25] StatLib - carnegie mellon universit. Retrieved December 10, 2012, from <http://lib.stat.cmu.edu/>.
- [26] Boettcher, M. (2011). Contrast and change mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 215-230.
- [27] Bay, S. D., & Pazzani, M. J. (2001). Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3), 213-246.
- [28] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIG-MOD Conference on Management of Data* (pp. 207-216). Washington D.C.
- [29] Davies, J., & Bilman, D. (1996). Hierarchical categorization and the effects of contrast inconsistency in an unsupervised learning task. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 750-755).

- [30] Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. New York, Cambridge University Press.
- [31] Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(3), 1-16.
- [32] Yu, L. L., & Liu, L. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205-1209.
- [33] Kudo, M., & Sklansky, J. (1998). Classifier-independent feature selection for two-stage feature selection. *Proceedings of the Joint IAPR International Workshops on SSPR'98 and SPR'98* (pp. 548-555).
- [34] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2), 245-271.
- [35] Abe, N., Kudo, M., Toyama, J., & Shimbo, H. (2006). Classifier-independent feature selection on the basis of divergence criterion. *Pattern Analysis and Applications*, 9(2-3), 127-137.
- [36] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data Sets. *Journal of Machine Learning Research*, 7, 1-30.
- [37] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*. 229-248, Menlo Park, CA: AAAI Press,.
- [38] Chrysostomou, K. A. (2008). The role of classifiers in feature selection: Number VS nature. Doctoral Thesis. School of Information Systems, Computing and Mathematics, Brunel University.
- [39] Sociology Data Set Server of Saint Joseph's University in Philadelphia. Retrieved December 14, 2012, from <http://sociology-data.sju.edu/>.
- [40] Feature selection datasets at Arizona State University. Retrieved January 20, 2013, from <http://featureselection.asu.edu/datasets.php>.



**Dijana Oreški** received her master's degree in 2008 at the Faculty of Organization and Informatics (FOI), University of Zagreb in the field of information system development. Since April 2009 she works as a research and teaching assistant at FOI. She teaches the following courses: knowledge discovery in data, intelligent systems, knowledge based systems, development of information systems and informatics. She has worked on several national and international projects. She defended her PhD thesis entitled "Evaluation of contrast mining techniques for feature selection in classification" in February 2014. Her research interest is in development and application of data mining techniques in social sciences.



**Mario Konecki** is a senior research and teaching assistant at the Faculty of Organization and Informatics in Varaždin, Croatia. During his former scientific work he has published over 40 scientific papers and has been actively involved in 8 scientific projects. He is also active in professional work in the field of programming, design and education. His main scientific interests are: intelligent systems, development of programming languages, education in the area of programming, design of user interfaces and web technologies.

His main research projects are: "Determining the possibility of including visually impaired in the activities of graphical user interfaces design" and "New methods of teaching programming with an emphasis on teaching visually impaired students".