**RESEARCH PAPERS**
**FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA**
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2013 Special Number

# IMPACT OF DATA CHARACTERISTICS ON FEATURE SELECTION TECHNIQUES PERFORMANCE

Dijana ORESKI

## ABSTRACT

*Feature selection is a step in knowledge discovery in databases which takes away most of the time of the entire process. Therefore, the effective implementation of feature selection significantly improves the overall process. This paper suggests examining data characteristics before applying feature selection and hypothesize that data characteristics significantly affect feature selection techniques performance. Our experimental comparison of five previously used feature selection techniques reveals significant difference in feature selection techniques performance when dealing with data sets of different characteristics.*

## KEY WORDS

*feature selection, data characteristics, classification accuracy*

## INTRODUCTION

Very often data sets contain a large number of features, which can influence the performance of the entire learning in classification. Large dimensionality of the database can be reduced by using appropriate techniques. These techniques fall into two groups: those that transform the fundamental meaning of the features (feature extraction techniques) and those that preserve the semantics. Feature selection techniques belong to second group (which selects a small set of features on the basis of the evaluation function) and they are in the focus of this paper. Feature selection is a very active and fruitful area of research in machine learning, statistics and data mining (Ramaswami and Bhaskaran, 2009). In the process of knowledge discovery in databases preparation of data takes away 60% - 95% of the time (De Veaux, 2005). Feature selection, the most important part of this step, refers to the problem of selection of features that give the highest predictive information with respect to the output. The main objective of carrying out the feature selection is to select a subset of input features in order to remove features that are not relevant and do not provide predictive information, and finally, achieving high classification accuracy (Ramaswami and Bhaskaran, 2009). Feature selection, in theory and in practice, proved to be effective in increasing the efficiency of learning, forecasting accuracy and reducing complexity of the results (Koller and Saham, 1996).

Dijana ORESKI - University of Zagreb, Faculty of Organization and Informatics, Varazdin, Pavlinska 2, dijana.oreski@foi.hr

On different data sets different techniques respond differently and differ in the accuracy of classification. This paper explores the relation of the data set characteristics and performance of the feature selection techniques. For this purpose, research presented here compares five feature selection techniques on two data sets of different characteristics. The paper is organized as follows. The second section defines feature selection whereas third section provides a theoretical overview of the feature selection techniques that will be used in empirical research and discusses the characteristics of the data important for classification task. Fourth section describes the research and discusses the obtained results. The last chapter concludes the work.

## RELATED WORK

Feature selection can be defined as follows. „Suppose $F$ is the given set of original features with cardinality $n$ (where $n$ symbolizes the number of features in set $F$), and $\overline{F}$ is the selected feature subset with cardinality $\overline{n}$ (where $\overline{n}$ symbolizes the number of features in set $F$ ), then $\overline{F} \subseteq F$ . Also, let $J((\overline{F})$ ) be the selection criterion for selecting feature set $\overline{F}$ . We assume that a higher value of $J$ indicates a better feature subset. Thus, the goal is to maximise $J()$. The problem of feature selection is to find a subset of features $\overline{F} \subseteq F$ such that „(Chrysostomou, 2009)

$$J(\overline{F}) = \max_{Z \subseteq F, |Z| = n} J(Z) .$$

Dash and Liu define four steps of feature selection process: generating subset, subset evaluation, stopping criterion and validation (Dash and Liu, 1997). A review of the literature showed that there are many approaches to the problem of feature selection. However, essentially all approaches include two key components: search strategy that explores the set of all subsets of features in a purposeful way and evaluation criterion. The most common classification of feature selection techniques is: filter and wrapper techniques. The main difference between those two approaches is in the evaluation of the subset. Wrapper approach evaluates subset within learning algorithm, whereas when applying filter approach feature selection and classification are separated. Filter approach very often uses heuristic in which evaluation function is not directly related to the effectiveness of a particular classifier. Instead, the result depends on the internal characteristics of the data. Features are evaluated according to criteria such as the distance measure, the Pearson correlation coefficient, entropy or other measures of information (e.g. Devijver and Kittler, 1982. or Guyon and Elisseeff, 2003). We can say that the filter techniques present a general approach to feature selection providing a solution suitable for a wide set of classifiers. The filter techniques are very fast, and as such are useful for high-dimensional problems where other methods are not competitive with respect to computational complexity. However, the selected optimal features do not necessarily guarantee the best performance of the classifier.

Wrapper techniques evaluate subset of features by estimating the accuracy of the learning algorithm. Search strategy use prediction accuracy as a function of leading the search for the best subset and is looking for those features that maximize accuracy. Of course, features are optimized for the previously selected algorithm, and very likely not optimal for another learning algorithm. Wrapper techniques require a lot of computation, because a large number of models of classification must be done during the process of searching for the best subset of

the attributes. Speed could be achieved by using efficient search strategy. But that search becomes almost impossible with increasing dimensionality, especially when working with a computer - intensive methods of learning. Wrappers often result with overtraining. Despite this, some authors (e.g. John et. al., 1994. or Kohavi and John, 1997) showed that the accuracy is better than in the case of filter techniques. Traditionally, feature selection techniques are evaluated based on the time that they need for the performance and quality of the selected subsets of attributes (Jain and Zongker, 1997). The methodology for the evaluation of the results is not standardized and varies from article to article. It is therefore very difficult to draw conclusions or make comparisons between feature selection techniques. As criteria for comparison of techniques classification performance and algorithm execution time mainly have been used previously. Jain and Zongker argue that the algorithm execution time is less important criterion than the final classification performance (Jain and Zongker, 1997). The most commonly used criterion is the error rate (or classification accuracy) of the selected learning algorithm, which was implemented using a set of features derived by feature selection technique. Thus, this is criterion used in research presented here.

Wrapper methods often produce more accurate results than the filter methods, but the execution time is much larger. Therefore, when dealing with problems consisting of several thousand features wrapper methods are not applicable. While some argue that the biggest disadvantage of filter methods is ignoring the impact on the accuracy of the selected subset on the learning algorithm (Guyon and Elisseeff, 2003). Other authors (e.g. Abe et al. 2006) independence of the feature selection techniques emphasize as an advantage because it is best to select a subset of features that gives good results for each classifier (Abe et. al. 2006). Research presented in this paper uses both, filter and wrapper techniques in evaluation.

## METHODOLOGY DESCRIPTION

This section describes two main methodological aspects of research: feature selection techniques and characteristics of data set which was recognized in the previous research as important for task of classification

Five feature selection techniques are used in this research: *Relief*, *linear forward selection*, *information gain*, *gain ratio* and *voting* technique. Kira and Rendell introduce an algorithm *Relief* which assigns a relevance weight to each feature. Feature's weight represents ability of the feature to distinguish between class values. Features are ranked by weight and those that are higher than certain threshold are selected to form the final subset. A *linear forward selection* is a search where new features are added to a set one feature at a time. At each stage, the chosen feature is one that, when added to the current set, maximizes the objective. The algorithm terminates when the best remaining feature worsens the objective, or when the desired number of features is reached. *„Information gain* is a feature selection technique which provides a ranking for each feature describing the given training tuples. The feature with the highest information gain minimizes the information needed to classify the tuples in the resulting partitions and reflects the lowest degree of randomness or ''impurity'' in these partitions" (Oreski, Oreski and Oreski, 2012). The information gain technique is biased toward tests with many outcomes. An extension to information gain is known as *gain ratio*, which attempts to overcome this bias. „It applies a kind of normalization to information gain using a ''split information'' value" (Oreski, Oreski and Oreski, 2012).

Previous empirical studies have shown that the choice of optimal classifier in the process of knowledge discovery in databases depends on the data set employed (Michie, Spiegelhalter and Taylor, 1994). Van der Walt (Van der Walt, 2008) investigated the properties of data that influence classification performance and developed data measures that are specifically

defined to measure such data properties. These measures provide them to define the relationship between data characteristics and classifier performance. Measures were grouped into the following categories: standard measures, data sparseness measures, statistical measures, information theoretic measures, decision boundary measures, topology measures and noise measures. This research examines characteristics of two data sets included in the classification and performs feature selection in order to identify are there any differences in feature selection techniques performance when dealing with data sets of different characteristics.

## RESEARCH DESCRIPTION

The comparisons were carried out in two datasets coming from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Table 1 shows a summary of the characteristics of these datasets (*credit* and *spectf* data set) used in this paper to assess the performance of the five feature selection techniques: info gain, gain ratio, relief, linear forward selection and voting technique. In order to evaluate the performance of a feature selection techniques, the accuracy of the classifier (neural network) trained on those features selected by the aforementioned techniques will be compared.

DATA SETS CHARACTERISTICS Table 1

| Data characteristic | Data set 1 (credit) | Data set 2 (spectf) |
|---|---|---|
| *Standard measures* | Number of features: 15<br>Number of instances: 690 | Number of features: 45<br>Number of instances: 80 |
| *Data sparseness measures* | Linear relationship | Exponential relationship |
| *Statistical measures* | Correlation:0,117<br>Normality: yes<br>Homogeneity of covariance matrices: no | Correlation:0,223<br>Normality: no<br>Homogeneity of covariance matrices: no |
| *Information theory measures* | Intrisnic dimensionality: 0,733 | Intrisnic dimensionality: 0,355 |
| *Noise measures* | Feature noise:0,267 | Feature noise:0,644 |

As seen from table 1, presented data sets difeer significantly in number of instances and the level of data sparsity. Whereas first data set has low level of data sparsity and assumes linear relationship between features, second data set has high level of sparsity and assumes exponential relationship. Furthermore, data sets differ in the distribution (first data sets has normal distribution, whereas second does not have) and feature noise (second data set has significantly higher feature noise).

Five feature selection techniques were applied on described datasets. All techniques selected 9 features from first data set and 16 features from second data set. In the statistical evaluation of the feature selection techniques performance, we compare the achieved scores from classifier, neural networks, as follows.

Using the Friedman test we tested the null hypothesis stating that all feature selection techniques perform equally. The results of the tests showed that the Friedman statistic for equality of feature selection techniques performances has the p-value of 0.0021 for first data set and p-value of 0.0181 for second data. These results reject null hypothesis and indicates that the difference exist in the performance of feature selection techniques for both data sets. To analyze the performance differences, post-hoc Nemenyi test was performed in order to identify which technique provided better results by giving the p-values of the performance

comparisons between pairs of feature selection techniques. Results indicate that linear forward selection achieved significantly better compared to the other techniques on first data set**,** whereas Relief feature selection technique achieved significantly better accuracy on second data set.

## CONCLUSION

This paper presented empirical evaluation of five feature selection techniques on two data sets. The emphasis was on the different characteristics of the used data sets, which has not been empirically evaluated in the previous research. Hypothesis of the research involved measuring of how feature selection techniques react on data sets which differ in their characteristics. The testing results have shown that linear forward selection performed best in terms of low number of features and high number of instances, low correlation and low feature noise. Relief feature selection technique achieved highest accuracy in the situation of higher number of features and lower number of instances, higher correlation and higher feature noise. This conclusion speaks in favor of hypothesis that feature selection techniques performance deeply depends on data characteristics and thus examining of data characteristics is necessary prior of applying of feature selection.

## REFERENCES


1. ABE, N., KUDO, M., TOYAMA, J., SHIMBO, M. 2006. Classifier-independent feature selection on the basis of divergence criterion. *Pattern Analysis & Applications*, **9**(2), 2006, pp. 127.-137.
2. CHRYSOSTOMOU, K. 2009. Wrapper Feature Selection. In *J. Wang (Ed.), Encyclopedia of Data Warehousing and Mining*, Second Edition, pp. 2103-2108. Hershey, PA: Information Science Reference. doi:10.4018/978-1-60566-010-3.ch322
3. DEVIJVER, P. A., KITTLER, J. 1982. Pattern Recognition: A Statistical Approach. Prentice Hall.
4. De VEAUX, R., Predictive Analytics: Modeling the World, OR/MS Seminar, 2005. available at: *student.som.umass.edu/informs/slides/Predictive.pdf*
5. GUYON, I., ELISSEEFF, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, pp. 1157–1182.
6. JAIN, A., ZONGKER, D. 1997. Feature selection: Evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(2), pp. 153–158, 1997.
7. JOHN, G. H., KOHAVI, R., PFLEGER, K. 1994. Irrelevant features and the subset selection problem. In *W. W. Cohen and H. Hirsh, editors, Proceedings of the 11th International Conference on Machine Learning*. San Francisco, CA, Morgan Kaufmann Publishers, pp. 121–129,
8. KOLLER, D., SAHAMI, M. 1996. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 284–292.
9. MICHIE, D., SPIEGELHALTER, D.J., TAYLOR, C.C. 1994. *Machine learning, neural and statistical classification*. Ellis Horwood Limited, Hemel Hempstead.
10. 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, pp. 12605–12617.
11. RAMASWAMI, M., BHASKARAN, R. 2009. A Study on Feature Selection Techniques in Educational Data Mining. *Journal of computing*, **1**(1), pp. 7- 11.

12. UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ml/
13. VAN DER WALT, C. 2008. *Data measures that characterise classification problems, Master Thesis*. University of Pretoria.