Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets

Mina Alibeigi alibeigi@cse.shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Sattar Hashemi s_hashemi@shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Ali Hamzeh hamzeh@shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Abstract

Since dealing with high dimensional data is computationally complex and sometimes even intractable, recently several feature reduction methods have been developed to reduce the dimensionality of the data in order to simplify the calculation analysis in various applications such as text categorization, signal processing, image retrieval and gene expressions among many others. Among feature reduction techniques, feature selection is one of the most popular methods due to the preservation of the original meaning of features. However, most of the current feature selection methods do not have a good performance when fed on imbalanced data sets which are pervasive in real world applications.

In this paper, we propose a new unsupervised feature selection method attributed to imbalanced data sets, which will remove redundant features from the original feature space based on the distribution of features. To show the effectiveness of the proposed method, popular feature selection methods have been implemented and compared. Experimental results on the several imbalanced data sets, derived from UCI repository database, illustrate the effectiveness of the proposed method in comparison with other rival methods in terms of both AUC and F1 performance measures of 1-Nearest Neighbor and Naïve Bayes classifiers and the percent of the selected features.

Keywords: Feature, Feature Selection, Filter Approach, Imbalanced Data Sets.

1. INTRODUCTION

Since data mining is capable of finding new useful information from data sets, it has been widely applied in various domains such as pattern recognition, decision support systems, signal processing, financial forecasts and etc [1]. However by the appearance of the internet, data sets are getting larger and larger which may lead to traditional data mining and machine learning algorithms to do slowly and not efficiently. One of the key solutions to solve this problem is to reduce the amount of data by sampling methods [2], [3]. But in many applications, the number of instances in the data set is not too large, whereas the number of features in these data sets is more than one thousands or even more. In this case, sampling is not a good choice. Theoretically, having more features, the discrimination power will be higher in classification. However, this theory is not always true in reality since some features may be unimportant to predict the class labels or even be irrelevant [4], [5]. Since many factors such as the quality of the

data, are responsible in the success of a learning algorithm, in order to extract information more efficiently, the data set should not contains irrelevant, noisy or redundant features [6]. Furthermore, high dimensionality of data set may cause the “curse of dimensionality” problem [7]. Feature reduction (dimensionality reduction) methods are one of the key solutions to all these problems.

Feature reduction refers to the problem of reducing the dimension by which the data set is described [8]. The general purpose of these methods is to represent data set with fewer features to reduce the computational complexity whereas preserving or even improving the discriminative capability [8]. Since feature reduction can brings a lot of advantages to learning algorithms, such as avoiding over-fitting and robustness in the presence of noise as well as higher accuracy, it has attracted a lot of attention in the three last decades. Therefore, vast variety of feature reduction methods suggested which are totally divided into two major categories including feature extraction and feature subset selection. Feature extraction techniques project data into a new reduced subspace in which the initial meaning of the features are not kept any more. Some of the well-known state-of-the-art feature extraction methods are principal component analysis (PCA) [5], non-linear PCA [13] and linear discriminant analysis (LDA) [13]. In comparison, feature selection methods preserve the primary information and meaning of features in the selected subset. The purpose of these schemes is to remove noisy and redundant features from the original feature subspace [13]. Therefore, due to preserving the initial meaning of features, feature selection approaches are in more of interest [8], [9].

Feature selection methods can be broadly divided into two categories: filter and wrapper approaches [9]. Filter approaches choose features from the original feature space according to pre-specified evaluation criterions, which are independent of specified learning algorithms.
Conversely, wrapper approaches select features with higher prediction performances estimated according to specified learning algorithms. Thus wrappers can achieve better performance than filters. However, wrapper approaches are less common than filter ones because they need higher computational resources and often intractable for large scale problems [9]. Due to their computational efficiency and independency to any specified learning algorithm, filter approaches are more popular and common for high dimensional data sets [9].

As was stated above, feature selection has been studied intensively [4], [5], [6], [8], [9] but its importance to resolving the class imbalance problem was recently mentioned by researchers [10]. The class imbalance problem refers to the issue that occurs when one or more classes of a data set have significantly more number of instances (majority class) than other classes of that data set (minority class) [10].In this type of data sets, the minority class has higher importance than the majority class. Since, nowadays, imbalanced data sets are pervasive in real world applications such as biological data analysis, text classification, web categorization, risk management, image classification, fraud detection and many other applications, it is important to propose a new feature selection method which is appropriate for imbalanced data sets.

Therefore, in this study, we present a new filter unsupervised feature selection algorithm which has the benefits of filter approaches and is designed to have a high performance on imbalanced data sets. The proposed approach chooses more informative features considering the importance of the minority class, according to relation between the distributions of features which are approximated by probability density function (PDF). The main idea of the proposed scheme is firstly approximating the PDF of each feature independently in an unsupervised manner and then removing those features for which their PDFs have higher covering areas with the PDFs of other features which are known as redundant features.

The rest of this paper is organized as follow. Section 2 discusses the related researches for unsupervised feature selection. Section 3 explains the proposed method for unsupervised feature selection applications. Our experimental results are given in section 4 and section 5 concludes the paper by a conclusion part.

2. RELATED WORK

Conventional feature selection methods evaluate various subsets of features and select the best subset among all with the best evaluation according to an effective criterion related to the application. These methods often suffer from high computational complexity through their searching process when applied to large data sets. The complexity of an exhaustive search is exponential in terms of the number of features of the data set. To overcome these shortcomings, several heuristic schemas have been proposed such as Branch and Bound (B&B) method which guarantees to find the optimal subset of features with computational time expectedly less than the exponential under the monotonicity assumption [12]. B&B starts from the full set of features and removes features by a depth first search strategy until the removing of one feature can improve the evaluation of the remaining subset of features [12]. Another popular approach is Sequential Forward Selection (SFS) which searches to find the best subset of features in an iterative manner starting from the empty set of features. In each step, SFS adds that feature to the current subset of selected features which yields to maximize the evaluation criterion for the new selected feature subset [13]. However, heuristic approaches are simple and fast with quadratic complexity, but they often suffer from lack of backtracking and thus act poorly for nonmonotonic criterions. In [24], another heuristic method called Sequential Floating Forward Selection (SFFS) was proposed which performs sequential forward selection with the backtracking capability at the cost of higher computational complexity.

The former methods can be applied in both supervised and unsupervised schemas according to their evaluation criteria. Since the interest of this paper is developing an unsupervised feature selection method, here, we investigate only the unsupervised methods. These methods can be generally divided into two divisions: filter and wrapper approaches [4], [8], [13]. The principle of wrapper approaches is to select subset of features regarding a specified clustering algorithm. These methods find a subset of features that using them for training a specified clustering; the highest performance can be achieved. Some examples of these approaches are [14], [15], [16]. Conversely, filter methods select features according to an evaluation criterion independent of specified clustering algorithm. The goal of these methods is to find irrelevant and redundant features and remove them from the original feature space. In order to find irrelevant and redundant features, various dependency measures have been suggested such as correlation coefficient [6], linear dependency [18] and consistency measures [19].
In this paper, we propose a feature subset selection based on the distribution of features which is able to handle the nonlinearity dependency between features in an unsupervised framework with a high performance for imbalanced data sets because of considering higher importance of the minority class which is the most important class in an imbalanced data set. The following section explains the proposed method in details.

3. THE PROPOSED UNSUPERVISED FEATURE SELECTION METHOD ATTRIBUTED TO IMBALANCED DATA SETS

The proposed unsupervised feature selection which is a filter approach attributed to imbalanced data sets, includes four steps. In the first step features are scaled in the range [0, 1]. Then, the probability density function (PDF) of each feature is estimated which gives a good overview about the distribution of instances for a specific feature. The third step is computing the number of times that the PDF of one feature is similar to PDF of other remaining features. At last, features with higher counter of being similar to other features are removed. Each step is described in details as follows.

The proposed method finds the relation between each two features as if they are similar or not according to their PDFs and removes those features which are more similar to other features as redundant features because all or most of their information is repeated in other features.
As was explained before, the first step in the proposed feature selection approach is scaling feature values in the range [0, 1]. Afterwards, PDF is estimated for each feature. The methods for estimating probability density functions can be totally categorized into parametric and non-parametric approaches [21]. The parametric methods assume a particular form for the density,

such as Gaussian, so that only the parameters (mean and variance) need to be estimated. In comparison, non-parametric methods do not assume any knowledge about the density of the data and computes the density directly from the instances and because of this reason they are in more of interest. The general form of non-parametric probability density estimation methods is according to the following formula:

$$p(x) \cong \frac{k}{N * V} \qquad (1)$$

where, *p(x)* is the value of the estimated probability density function for instance *x*, *V* is the volume surrounding *x*, *N* is the total number of instances and *k* is the number of instances inside *V*. Two basic approaches can be adapted to practical non-parametric density estimation methods based on the status of *k* and *V*. Fixing the value of *k* and determining the corresponding volume *V* that contains exactly k instances inside, leads to methods commonly referred to as *K Nearest Neighbor* (*KNN*) methods. On the other hand, when the volume *V* is chosen to be fixed and *k* is determined, the non-parametric estimation method is called *Kernel Density Estimation* (*KDE*). Generally, the probability densities that estimated via *KNN* approaches are not very satisfactory because of some drawbacks. Because, *KNN PDF estimation* methods are prone to local noise. Moreover, the resulting PDF via KNN method is not a true probability density since its integral over all the instance space diverges [25]. In spite of these reasons, in this study, we estimate probability density functions through the *KDE* method with Gaussian kernel. It is noted that our proposed feature selection algorithm is not sensitive to any particular estimation method. However, using more accurate estimation methods cause the algorithm to perform more efficiently.

In order to compare PDFs of different features, all feature values are scaled into the [0, 1] interval because the range of various features may be different. Afterwards, the probability density functions for each of the features are computed according to *KDE* methods.
Having estimated the probability density function for each feature, the similarity between each of the two features is calculated. Two features are considered as similar features if the Mean Square Error (MSE) of their PDFs be less than a user specified threshold. Similar features contain nearly the same information because their PDFs are sufficiently similar. Thus, one of the similar features can be removed without a considerable loss of information. Among similar features, features which are similar to more other features of the whole feature space are removed. By removing the feature which has higher frequency of being similar with other features, the loss of information is minimized. Also, as the instances of all classes contribute equally for estimating the PDF of each feature, then instances of the minority classes are given higher importance in the PDF estimation process. Thus, features which are more informative according to minority classes are given higher chance to be selected. Algorithm 1 represents the steps of the proposed feature selection approach.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The comparisons were carried out in three data sets coming from the UCI Machine Learning Repository including Ecoli, Ionosphere and Sonar which are all imbalanced. Table I shows a summary of the characteristics of the data sets used in this paper to assess the performance of the proposed method. The first column of Table I shows the name of the data set. Number of features and number of classes are showed in the second and third columns, respectively. The last column in each row is the number of instances per each class.
In order to evaluate the performance of a feature selection method, the performance of classifiers trained on the features selected by the mentioned feature selection method, is compared to the performance of classifiers trained on the full set of features named as baseline performance. There are many classifiers in machine learning domains with different biases. The most well-known classifiers for evaluating a feature selection method are *Naive Bayes* (*NB*) [11] classifier and *K- Nearest Neighbor* (*KNN*) classifier [13]. Naïve Bayes is a simple probabilistic classifier based on the assumption of class conditional independence of features [25]. K-Nearest Neighbor

is a lazy learning algorithm which classifies each new test instance based on its *K* nearest training instances [25].
For imbalanced data sets, classifiers have difficulties to classify instances from the minority class because they simply classify instances as the majority class achieving a high accuracy. So, in these data sets, accuracy is not a good performance measure. There are a number of other statistics such as AUC (Area Under receiver operating characteristic Curve) and F-measures [26]. AUC and F1-measure are two of the statistics which are commonly used to evaluate classifiers focusing on the importance of the minority class. In this paper, we evaluate the performance of different feature selection methods based on AUC and F1-measure evaluation statistics.

Algorithm 1: The steps of the proposed unsupervised feature selection method.

Unsupervised Feature Selection Based on the Distribution of Features

Input: $D=\{d_1, d_2, \ldots, d_N\}$ // N is the number of instances
Input: $F=\{f_1, f_2, \ldots, f_n\}$ // n is the number of features
Output: $F^{(S)}$ // The selected subset of features, at the beginning, $F^{(S)} = F$

Begin

Step 1. Scale each feature in range [0,1]
Step 2. Estimate the probability density function (PDF) for each feature
Step 3. For i=1 to n-1
    For j=i+1 to n
        Calculate MES(density of feature i, density of feature j)
    If MSE <= epsilon
        Consider features i and j to be similar
        Increment the similarity counter of both features i and j
Step 4. Between each similar features, remove feature with higher similarity counter
Step 5. Return list of remaining features as the list of selected features ($F^{(S)}$)

End

Name	# Features	# Class	# Instances Per Class
Sonar	60	2	97, 111
Ionosphere	34	2	126, 225
Ecoli	7	8	143, 77, 52, 35, 20, 5, 2, 2

TABLE 1: Characteristics of data sets used in this study for experimental evaluations.

Comparisons are done in Weka framework [22]. To show the effectiveness of the proposed method, we compared our method with two of the commonly used supervised approaches proposed by Hall et al. [19] and Lie et al. [4] named as Correlation-based Feature subset Evaluation and Consistency-based feature Subset Evaluation, which are abbreviated in results as CfsSubsetEval and ConsistencySubsetEval, respectively. We also compared the proposed method with an unsupervised Sequential Forward Selection (SFS) scheme for which Entropy is used as the evaluation criterion. This method is illustrated as SFS with entropy in experiments. The entropy criterion for this method is defined according to formula (2).

$$Entropy = -\sum_{p=1}^{l}\sum_{q=1}^{l}(sim(p,q)*\log(sim(p,q)) + (1-sim(p,q))*\log(1-sim(p,q))$$

$$Sim(p,q) = e^{-\alpha D_{pq}} \quad (2)$$

$$D_{pq} = [\sum_{j=1}^{M}(\frac{x_{p,j} - x_{q,j}}{\max_j - \min_j})^2]^{1/2}$$

where D_{pq} is the distance between two instances p and q and $x_{p,j}$ denotes the jth feature value for *instance p*. max_j and min_j are respectively the maximum and minimum values for the jth feature and M denotes the number of features. In (2), α is a positive constant which is set as $\alpha = \frac{-\ln 0.5}{\overline{D}}$ where $\overline{D}$ is the average distance between all instances.
Tables 2-4, separately illustrate the experimental results on each of the introduced data sets. The fist column of these tables, is the name of the feature selection method. The second column of each table, is the number of selected features by the corresponding feature selection method. AUC and F1 performance of Naïve Bayes (NB) classifier are shown in third and forth columns, respectively. Also, AUC and F1 performance of K-Nearest Neighbor (KNN) classifier are shown in the last two columns of each table.

As the results show in Tables 2-4, the AUC and F1 performance of the proposed method is fairly comparable to the performance of the Baseline method while the proposed method removes some redundant features (about half of the original features, see Figure 1) which lead to less computational complexity. This illustration acknowledges that feature selection is a key solution for classifiers on high dimensional imbalanced data sets.

Also, the proposed feature selection method has higher 1-NN classifier performance than CfsSubsetEval and ConsistencySubsetEval feature selection schemes in terms of both AUC and F1 evaluation measures and is comparable to both mentioned feature selection methods in terms of AUC and F1 performance of NB classifier. However, it is noticeable that the proposed method is an unsupervised approach which has access to less information in comparison with CfsSubsetEval and ConsistencySubsetEval feature selection methods which are supervised methods and have access to the class labels. Furthermore, our method has higher performance in comparison with other rival unsupervised feature selection scheme named as SFS with Entropy in experiments in terms of both AUC and F1 performances of 1-NN and NB classifiers. In general, it can be concluded that the proposed feature selection approach is more efficient than the other rival unsupervised feature selection and is comparable to the commonly used supervised feature selection schemas considered in experiments.

Figure 1 shows the comparison among rival feature selection methods in terms of the percent of selected features for each data set. Those feature selection methods which select a small percent of features while having a suitable performance, are more of interest. As can be seen, this property is true for the proposed feature selection method which makes it a good choice of action for feature selection on imbalanced data sets.

Feature Selection Method	# Selected Features	NB F1	NB AUC	1-NN F1	1-NN AUC
Baseline	34	0.829	0.935	0.857	0.822
CfsSubsetEval	14	0.92	0.958	0.885	0.852
ConsistencySubsetEval	7	0.872	0.926	0.875	0.849
SFS with Entropy	14	0.778	0.82	0.791	0.747
The Proposed Method	12	0.92	0.958	0.91	0.889

TABLE 2: Experimental results on Ionosphere data set in terms of the number of selected features and the AUC and F1 evaluation performances for NB and 1-NN classifiers.

Feature Selection Method	# Selected Features	NB F1	NB AUC	1-NN F1	1-NN AUC
Baseline	7	0.854	0.96	0.801	0.875
CfsSubsetEval	6	0.854	0.96	0.799	0.873
ConsistencySubsetEval	6	0.854	0.96	0.799	0.873
SFS with Entropy	6	0.791	0.947	0.766	0.854
The Proposed Method	6	0.854	0.96	0.799	0.873

TABLE 3: Experimental results on Ecoli data set in terms of the number of selected features and the AUC and F1 evaluation performances for NB and 1-NN classifiers.

Feature Selection Method	# Selected Features	NB F1	NB AUC	1-NN F1	1-NN AUC
Baseline	60	0.673	0.8	0.865	0.862
CfsSubsetEval	19	0.675	0.812	0.836	0.834
ConsistencySubsetEval	14	0.666	0.811	0.85	0.847
SFS with Entropy	14	0.557	0.658	0.659	0.657
The Proposed Method	14	0.652	0.769	0.88	0.878

TABLE 4: Experimental results on Sonar data set in terms of the number of selected features and the AUC and F1 evaluation performances for NB and 1-NN classifiers.

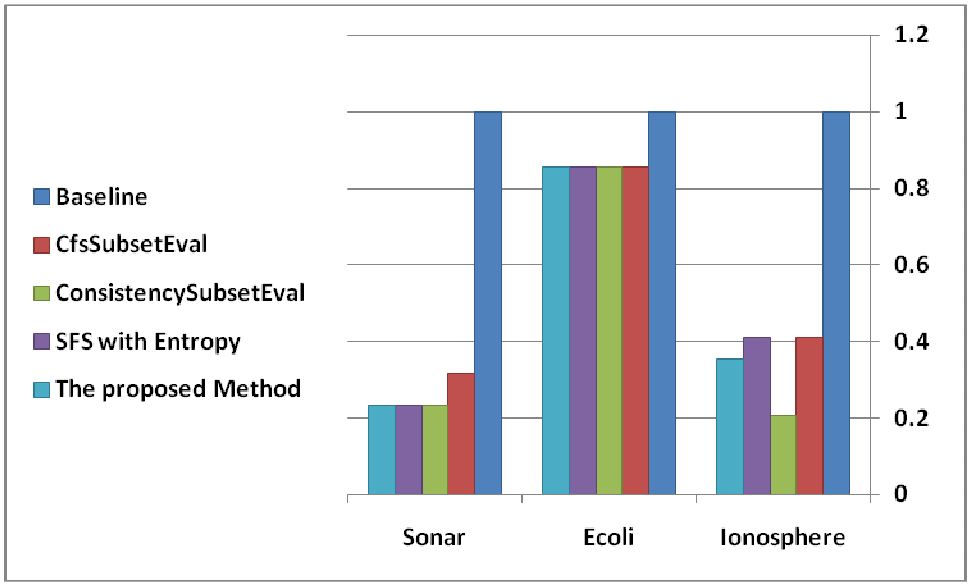

FIGURE 1: The percentage of selected features for each data set and in average for all data sets for the proposed feature selection method and other rival methods.

5. CONSLUSION AND FUTURE WORK

Feature selection techniques have a key role when encountering high dimensional data sets. Recently, filter based feature selection methods are of more interest because of their independence to any particular learning algorithm and their efficiency on high dimensional data sets. Since, most of the current feature selection methods perform poorly when fed on with imbalanced data sets, designing a feature selection method which is able to handle imbalanced data sets is recently mentioned by researchers.

Therefore, in this study, we proposed a new filter unsupervised feature selection scheme attributed to imbalanced data sets, which selects features based on the relation between probability density estimations of features. The main idea is that a feature, for which its distribution is more similar to the distribution of other features, is redundant because all or most of its information is repeated in those similar features. So, this feature can be removed from the original feature space with the least loss of information. Experimental results on a set of imbalanced data sets show that the proposed feature selection approach compared to the rival unsupervised feature selection method, can find a more informative subset of features which are more useful for classifying the instances of the minority classes. Also, the performance of the proposed method is comparable to the performance of two commonly used supervised feature selection frameworks in terms of both AUC and F1 evaluation measures.

For future work, it might be useful to apply this idea in the field of supervised feature selection and find the probability density functions per classes for each feature and find the similarity between features by considering their densities per classes.

6. ACKNOWLEDGEMENT

This work is supported by the Iran Tele Communication Research Center.

7. REFERENCES

1. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. *"From data mining to knowledge discovery in databases"*, AI Magazine, vol. 17, pp. 37–54, 1996

2. M. Lindenbaum, S. Markovitch and D. Rusakov. *"Selective sampling for nearest neighbor classifiers"*, Machine learning, vol. 54, pp. 125–152, 2004

3. A.I. Schein and L.H. Ungar, *"Active learning for logistic regression: an evaluation"*, Machine Learning, vol. 68, pp. 235–265, 2007

4. M.A. Hall. *"Correlation-based feature subset selection for machine learning"*, Ph.D. Dissertation, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999

5. I.K. Fodor. *"A survey of dimension reduction techniques"*, Technical Report UCRL- ID-148494, Lawrence Livermore National Laboratory, US Department of Energy, 2002

6. M.A. Hall. *"Correlation-based feature selection for discrete and numeric class machine learning"*, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2000

7. R. Bellman. *"Adaptive Control Processes: A Guided Tour"*, Princeton University Press, Princeton, 1961
8. H. Liu, J. Sun, L. Liu and H. Zhang, *"Feature selection with dynamic mutual information"*, Pattern Recognition, vol. 42, pp. 1330 – 1339, 2009

9. N. Pradhananga. *"Effective Linear-Time Feature Selection"*, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2007

10. M. Wasikowski and X. Chen. *"Combating the small sample class imbalance problem using feature selection"*, IEEE Transactions on knowledge and data engineering, 2009

11. G.H. John and P. Langley. *"Estimating Continuous Distributions in Bayesian Classifiers*". In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345, 1995
12. M.P. Narendra and K. Fukunaga. *"A branch and bound algorithm for feature subset selection"*, IEEE Trans. Comput. Vol. 26, pp. 917–922, 1997

13. P.A. Devijver and J. Kittler. *"Pattern Recognition: A Statistical Approach"*, Englewood Cliffs: Prentice Hall, 1982

14. M. Dash and H. Liu. *"Unsupervised Feature Selection"*, Proc. Pacific Asia conf. Knowledge Discovery and Data Mining, pp. 110-121, 2000

15. J. Dy and C. Btodley. *"Feature Subset Selection and Order Identification for Unsupervised Learning"*, Proc. 17th Int'l. Conf. Machine Learning, 2000

16. S.Basu, C.A. Micchelli and P. Olsen. *"Maximum Entropy and Maximum Likelihood Criteria for Feature Selection from Multi-variate Data"*, Proc. IEEE Int'l. Symp. Circuits and Systems, pp. 267-270, 2000

17. S.K .Pal, R.K. De and J. Basak. *"Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach"*, IEEE Trans. Neural Network, vol. 11, pp. 366-376, 2000

18. S.K .Das. *"Feature Selection with a Linear Dependence Measure"*, IEEE Trans. Computers, pp. 1106-1109, 1971

19. G.T. Toussaint and T.R. Vilmansen. *"Comments on Feature Selection with a Linear Dependence Measure"*, IEEE Trans. Computers, 408, 1972

20. H. Liu and R. Setiono. *"A probabilistic approach to feature selection - A filter solution"*. In: 13th International Conference on Machine Learning, pp. 319-327, 1996

21. K. Fukunaga. *"Introduction to Statistical Pattern Recognition"*, Academic Press, 2nd Ed. 1990

22. E. Frank, M.A. Hall, G. Holmes, R. Kirkby and B. Pfahringer. *"Weka - a machine learning workbench for data mining"*, In The Data Mining and Knowledge Discovery Handbook, pp. 1305-1314, 2005

23. M. Dash and H. Liu. *"Unsupervised Feature Selection"*, Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 110-121, 2000

24. P. Pudil, J. Novovicova and J. Kittler. *"Floating Search Methods in Feature Selection"*, Pattern Recognition Letters, vol. 15, pp. 1119-1125, 1994

25. R.O. Duda, P.E. Hart and D.G. Stork. *"Pattern Classification"*, Second Edition, Wiley, 1997

26. G. Forman. "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research, vol. 3, pp. 1289-1305, 2003