Information-Theoretic Feature Selection Algorithms for Text Classification

Jana Novovičová

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, Czech Republic
E-mail: novovic@utia.cas.cz

Antonín Malík

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague, Czech Republic
E-mail: antonin.malik@logicacmg.com

Abstract—A major characteristic of text document classification problem is extremely high dimensionality of text data. In this paper we present four new algorithms for feature/word selection for the purpose of text classification. We use sequential forward selection methods based on improved mutual information criterion functions. The performance of the proposed evaluation functions compared to the information gain which evaluate features individually is discussed. We present experimental results using naive Bayes classifier based on multinomial model, linear support vector machine and k-nearest neighbor classifiers on the Reuters data set. Finally, we analyze the experimental results from various perspectives, including precision, recall and F_1 -measure. Preliminary experimental results indicate the effectiveness of the proposed feature selection algorithms in a text classification.

I. INTRODUCTION

The goal of text document classification is to assign automatically a new document into one or more predefined classes based on its contents. An increasing number of statistical classification methods and machine learning algorithms have been explored to build automatically a classifier by learning from previously labelled documents including *naive Bayes*, *k-nearest neighbor*, *support vector machines*, *neural network*, *decision trees*, *logistic regression* (see e.g. McCallum and Nigam [1], Nigam et al. [2], Joachims [3], Yang [4], Yang et al. [5] and the references therein). The overview of Sebastiani [6] discusses the main approaches to text classification.

In text classification, usually a document representation using a *bag-of-words* approach is employed (each position in the feature vector representation corresponds to a given word). This representation scheme leads to very high-dimensional feature space, too high for conventional classification methods. Two approaches for dimensionality reduction exist, either selecting a subset of the original features, or transforming the features into new ones.

In text classification problem the dominant approach is feature selection using various criteria. Methods for feature subset selection for text document classification task use an evaluation function that is applied to a single word. All words are independently evaluated and sorted according to the assigned criterion. Then, a predefined number of the best features is taken to form the best feature subset. Scoring of individual words can be performed using some of the measures, for instance, *document frequency*, *word frequency*, *mutual*

information, *information gain*, *odds ratio*, χ^2 *statistic (chi-square)* and *term strength* [7], [8], [9]. Yang and Pedersen [7] and Mladenic and Grobelnik [8] give experimental comparison of the above mentioned measures in text classification. The information gain and a very simple frequency measures were reported to work well on text data. Forman in [9] presents an extensive comparative study of twelve feature selection criteria for the high-dimensional domain of text classification.

In this paper we propose to use sequential forward selection search based on *improved mutual information* criterion introduced by Battiti [10] and Kwak and Choi [11] for non-textual data. To our knowledge, this criterion have not yet been applied earlier to document classification task as a criterion for reducing vocabulary size. Furthermore, we modify these two criterion functions. We use the simple but effective naive Bayes based on multinomial model, linear support vector machine (SVM) and k-nearest neighbor (kNN) classifiers.

II. TEXT CLASSIFICATION

Let $\mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$ be the set of $|\mathcal{C}|$ predefined classes and let $\mathcal{D} = \{d_1, \dots, d_{|\mathcal{D}|}\}$ be the finite training document set. Let $\mathcal{V} = \{w_1, \dots, w_{|\mathcal{V}|}\}$ be the vocabulary set containing $|\mathcal{V}|$ distinct words occurred in training documents. Given a set of document vectors $\{d_1, \dots, d_{|\mathcal{D}|}\}$ and their associated class labels $c(d_j) = \{c_1, \dots, c_{|\mathcal{C}|}\}$, text classification is the problem of estimating the true class label of a new document. Text documents cannot be directly interpreted by a classifier. According to the bag-of-words representation, the document d can be represented by a feature vector consisting of one feature variable for each word in the given vocabulary set $\mathcal{V}$. A common characteristic of text data is its extremely high dimensionality. The number of potential features (several tens of thousands) exceeds the number of training documents.

III. DIMENSIONALITY REDUCTION

Dimensionality reduction is a very important step in text classification because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy. The number of features can be dramatically reduced by the domain dependent methods which include the elimination of stop words, stripping of

special characters as well as stemming algorithms or morphological analysis. For a further dimensionality the domain independent methods can be used.

There are two distinct ways of viewing dimensionality reduction, depending on whether the task is performed locally (a feature subset is chosen for each individual class) or globally (a feature subset is chosen for the classification under all classes). The second distinction may be drawn in words of the nature of resulting features: dimensionality reduction by *feature selection* (reduced feature set is a subset of the original feature set) or dimensionality reduction by *feature extraction* (the features in reduced feature set are not of the same type of the features in original feature set but are obtained by combinations or transformations of the original ones). In the so called *wrapper* approach to feature selection, the feature subset is identified by means of the same learning algorithm that will be used for classifier design. The alternative is the *filtering* approach that is, keeping the features that receive the highest score according to a function that measures the importance of the features for the classification task.

A. Feature Subset Selection

In text classification the dominant approach to dimensionality reduction is feature selection. Given a predefined integer $|\mathcal{V}'|$, methods for word selection attempt to select from the original set $\mathcal{V}$, the set $\mathcal{V}'$ of words with $|\mathcal{V}'| \ll |\mathcal{V}|$ that, when used for document representation, yields the highest effectiveness. Different methods for feature subset selection have been developed in pattern recognition and machine learning using different evaluation functions and search procedures.

1) *Feature Selection Evaluation Functions*: In this paper the global filtering approach to feature selection in text task is considered. The evaluation functions measures how good a specific subset is in discriminating between classes.

A simple and effective global word selection function is the *document frequency* [7] of a word w , that is, only the terms that occur in the highest number of documents are retained.

Information-theoretic functions for global word selection have been used in the literature, among them *mutual information* (MI), (Yang and Pedersen [7]), *information gain* (IG), (Caropreso et al. [12], Yang and Pedersen [7], Yang and Liu [13]), χ^2 *statistic* (Caropreso et al. [12], Yang and Pedersen [7], Yang and Liu [13]) and *odds-ratio* (Mladenec [14]). The mathematical definitions of the main functions used for dimensionality reduction in text classification task are summarized in Table 1 in the paper of Sebastiani [6]. All functions are specified locally to a specific class $c_k \in \mathcal{C}$. In order to assess the value of word w in a global class-independent sense, either the sum or the weighted sum, or the maximum of their class-specific values are usually computed.

2) *Feature Selection Search Methods*: Most of the feature subset selection methods used in machine learning or pattern recognition are not designed for tasks with a large number of features. Most methods for feature subset selection that are used on text are very simple compared to the other methods developed in machine learning and pattern recognition.

Best individual features (BIF) method [15] evaluates all the $|\mathcal{V}|$ words individually according to a given criterion, sorts them and selects the best $|\mathcal{V}'|$ words. Since the vocabulary has usually several thousands or tens of thousands of words, the BIF methods are popular in text classification because they are rather fast, efficient and simple. However, they evaluate each word separately and completely ignore the existence of other words and the manner how the words work together. The best subset of $|\mathcal{V}'|$ words need not contain the best $|\mathcal{V}'|$ single evaluated words. *Forward selection* algorithms start with an empty set of features and add feature one at a time until the final feature set is reached. *Backward elimination* algorithms start with a feature set containing all features and remove features one at a time. These feature selection procedures take note of dependencies between words as opposed to the BIF methods.

IV. PROPOSED ALGORITHMS

We consider the global filtering approach to feature selection in text document task. In this section novel algorithms for feature selection using mutual information are presented.

A. Feature Selection using Mutual Information

Our feature subset selection problem is formulated as follows: Given an initial set $\mathcal{V}$ with $|\mathcal{V}|$ features, find the subset $\mathcal{S}_s \subset \mathcal{V}$ with $|\mathcal{S}|$ features that maximizes the mutual information for text defined as mutual information for a set of words averaged over all classes given by the following formula:

$$MI(\mathcal{S}) = \sum_{k=1}^{|\mathcal{C}|} P(c_k) I(c_k, \mathcal{S}). \quad (1)$$

The mutual information for a feature/word w (word w occurred) averaged over all classes is defined as:

$$MI(w) = \sum_{k=1}^{|\mathcal{C}|} P(c_k) \log \frac{P(w|c_k)}{P(w)} = \sum_{k=1}^{|\mathcal{C}|} P(c_k) I(c_k, w). \quad (2)$$

Here $P(w)$ is the probability, that the word w occurred, $P(c_k)$ is the probability of the class c_k , $P(w|c_k)$ is the conditional probability of the word w given the class c_k , $I(c_k, w)$ is the mutual information between class c_k and word w . We can consider three strategy for solving our feature selection problem.

The optimal strategy generates all the word subsets $\mathcal{S}$ and compares their $MI(\mathcal{S})$. It is almost impossible for too many combinations. In the *backward elimination* strategy we remove the worst word from the complete set $\mathcal{V}$ one by one till the required number of words remain. This procedure has a lot of difficulties in computing $I(c_k, \mathcal{S})$.

The *forward selection* strategy using mutual information for text (1) can be realized as follows:

1) Initialization:

Set $\mathcal{S}$ = 'empty set' ($\mathcal{S}$ denotes the set of selected words),

set $\mathcal{U}$ = 'initial set $\mathcal{V}$ of all $|\mathcal{V}|$ words' ($\mathcal{U}$ denotes the set of unselected words).

2) *Pre-computation:*

For all $w \in \mathcal{V}$ compute $MI(w)$.

3) *First word selection:*

Find the word $w^* \in \mathcal{V}$ that maximizes $MI(w)$,
set $\mathcal{U} = \mathcal{V} \setminus \{w^*\}$, $\mathcal{S} = \{w^*\}$.

4) *Words selection step:*

Repeat until the desired number of words is selected.

(a) For $w \in \mathcal{U}$, compute $MI(w, \mathcal{S})$.

(b) Selection of the next word:

Find the word $w^+ \in \mathcal{U}$ that maximizes $MI(w, \mathcal{S})$, and
set $\mathcal{U} = \mathcal{U} \setminus \{w^+\}$, $\mathcal{S} \cup w^+$.

Computation of $MI(w, \mathcal{S})$ is difficult. So realization of the ideal forward selection algorithm is practically impossible.

To overcome this practical obstacle we inspired by the methods of Battiti [10] and Kwak and Choi [11], respectively developed for solving feature selection problem with non-textual data.

1) *MI and MaxMI Feature Selection Algorithm:* Using the idea of Battiti [10], we proposed two modifications of the ideal forward selection algorithm. Instead of computing $I(w, \mathcal{S})$, the mutual information between a candidate for newly selected word and already selected set $\mathcal{S}$, we used only $MI(w)$ and $I(w, s)$, where $w \in \mathcal{U}$ and $s \in \mathcal{S}$. In the first proposed algorithm (MI algorithm) the next selected word is the word w that maximizes $(MI(w) - \sum_{s \in \mathcal{S}} I(w, s))$. In the second algorithm (MaxMI algorithm) the next selected word is the word w that maximizes $(MI(w) - \max_{s \in \mathcal{S}} I(w, s))$.

The MI and MaxMI algorithm is the same as the ideal forward selection except the step 4). It will be replaced as it is presented in Section IV-A.2.

2) *Improved MI Feature Selection Algorithm:* The following relation holds

$$I(c, w_i, w_s) = I(c, w_s) + I(c, w_i | w_s). \quad (3)$$

$I(c_k, w_i | w_s)$ represents the remaining mutual information between class c and word w_i for given word w_s . For all words to be selected in the ideal feature selection algorithm, $I(c_k, w_s)$ is common and there is no need to compute this. The ideal algorithm now tries to find the word that maximizes $I(c, w_i | w_s)$. But in general to calculate $I(c_k, w_i | w_s)$ is practically impossible. We will approximate $I(c_k, w_i | w_s)$ with $I(w_i, w_s)$ and $I(c_k, w_i)$. The conditional mutual information $I(c_k, w_i | w_s)$ can be represented as

$$I(c, w_i | w_s) = I(c_k, w_i) - [I(w_i, w_s) - I(w_s, w_i | c_k)]. \quad (4)$$

The term $I(w_s, w_i | c)$ is the mutual information between already selected word w_s and the candidate word w_i for given class c . It holds:

$$\frac{H(w_s | c)}{H(w_s)} = \frac{I(w_s, w_i | c)}{I(w_s, w_i)}. \quad (5)$$

Here $H(w_s)$ and $H(w_s | c)$ is the entropy and the conditional entropy of w_s , respectively. Using the equation (4) and (5) we obtain

$$I(c, w_i | w_s) = I(c, w_i) - \frac{I(c, w_s)}{H(w_s)} I(w_s, w_i). \quad (6)$$

With this formula, we revise the step 4) of the ideal forward algorithm.

B. Feature Selection Algorithms using Mutual Information

Four proposed algorithms for feature selection in text document task can be described as follows:

1) *Initialization:*

Set $\mathcal{S}$ = 'empty set' ($\mathcal{S}$ is the set of selected words),
set $\mathcal{U}$ = 'initial set $\mathcal{V}$ of all $|\mathcal{V}|$ words' ($\mathcal{U}$ of unselected words).

2) *Pre-computation:*

For all $w \in \mathcal{V}$ compute $MI(w)$ $I(w, w')$.

3) *First word selection:*

Find word $w^* \in \mathcal{V}$ that maximizes $MI(w)$;
set $\mathcal{U} = \mathcal{V} \setminus \{w^*\}$, $\mathcal{S} = \{w^*\}$.

4) *Words selection step:*

Repeat until the desired number of words is selected.

(a) For all $s \in \mathcal{S}$ compute entropy $H(s)$.

(b) For all couples of words (w, s) with $w \in \mathcal{U}$, $s \in \mathcal{S}$ compute:

- MI: $I(w, s)$
- IMI: $I(w, s)MI(s)/H(s)$
- MaxMI: $I(w, s)$
- MaxIMI: $I(w, s)MI(s)/H(s)$

(c) *Selection of the next word:*

Find word $w^+ \in \mathcal{U}$ that maximizes

- MI: $MI(w) - \sum_{s \in \mathcal{S}} I(w, s)$
- IMI: $MI(w) - \sum_{s \in \mathcal{S}} I(w, s)MI(s)/H(s)$
- MaxMI: $MI(w) - \max_{s \in \mathcal{S}} I(w, s)$
- MaxIMI: $MI(w) - \max_{s \in \mathcal{S}} I(w, s)MI(s)/H(s)$

Set $\mathcal{S} = \mathcal{S} \cup w^+$, $\mathcal{U} = \mathcal{U} \setminus w^+$.

V. EXPERIMENTS AND RESULTS

A. Data

In our experiments we examined the common used the Reuters-21578 data set¹ for evaluation of our algorithms. Our text preprocessing included removing all non-alphabetic characters like full stops, commas, brackets, etc., lowering the upper case characters, ignoring all the words that contained digits or non alpha-numeric characters, removing word from a stop-word list. We replaced each word by its morphological root, removed all the words which had less than three occurrences. The resulting vocabulary size was 7487 words. The

¹<http://www.daviddlewis.com/resources/testcollections/reuters21578>.

ModApte train/test split of the Reuters-21578 data contains 9603 training documents and 3299 testing documents. There are 135 economic classes and we used only those 90 for which there exists at least one training and one testing document.

B. Feature Subset Evaluation Methods

In our comparison, feature selection is performed using the following three global criterions: *mutual information* given by (1), *information gain* (IG), *chi-square* (χ_{sum}^2) and *odds ratio* (OR_{sum}):

$$IG(w_i) = \sum_{k=1}^{|C|} P(c_k, w_i) \log \frac{P(c_k, w_i)}{P(c_k)P(w_i)} + \sum_{k=1}^{|C|} P(c_k, \bar{w}_j) \log \frac{P(c_k, \bar{w}_j)}{P(c_k)P(\bar{w}_j)},$$

$$\chi_{sum}^2(w_i) = \sum_{k=1}^{|C|} \frac{|\mathcal{D}| \cdot [P(c_k, w)P(\bar{c}_k, \bar{w}_i) - P(c_k, \bar{w}_j)P(\bar{c}_k, w_i)]^2}{P(c_k)P(\bar{c}_k)P(w_i)P(\bar{w}_i)},$$

$$OR_{sum}(w_i) = \sum_{k=1}^{|C|} \frac{P(w_i|c_k)[1 - P(w_i|\bar{c}_k)]}{[1 - P(w_i|c_k)]P(w_i|\bar{c}_k)}.$$

Here $|\mathcal{D}|$ is the number of training documents, $P(w_i)$ is the probability, that the word w_i occurred in a document, $\bar{w}_i$ means, that the word does not occur, $P(c_k)$ is the probability of the class c_k , $\bar{c}_k$ represents that a random document does not occur in the class c_k . $P(c_k, w_i)$ is the joint probability of the class c_k and the occurrence of the word w_i and $P(w_i|c_k)$ is the conditional probability of the word w_i given the class c_k .

For selecting salient words from the complete vocabulary set $\mathcal{V}$ we used best individual features method (BIF) based on information gain (BIF IG), chi-square (BIF Chi2) and odds ratio (BIF OR). Further, sequential forward selection (SFS) method based on the improved mutual information feature selection criteria and their modifications (SFS MI, SFS IMI, SFS MaxMI and SFS MaxIMI) presented in Section IV has been used in our experiments.

C. Classifiers

All feature selection methods were examined on three classifiers: A naive Bayes classifier, a linear Support Vector Machines classifier and a k -nearest neighbor classifier.

Naive Bayes. We use the multinomial model as described by McCallum and Nigam [1]. The predicted class for document d is the one that maximizes the posterior probability $P(c|d)$, which is proportional to $P(c) \prod_w P(w|c)^{n(w,d)}$, where $P(c)$ is the prior probability that a document belongs to class c , $P(w|c)$ is the probability that a word chosen randomly in a document from class c equals w , and $n(w, d)$ is the number of occurrences of word w in a document d ("word frequency"). We smoothed word and class probabilities using Bayesian estimate with word priors and a Laplace estimate, respectively.

Linear Support Vector Machines. The SVM method has been introduced in text classification by Joachims [3] and subsequently used in several other text classification works. The method is defined over the vector space where the classification problem is to find the decision surface that "best" separates the data points of one class from the other. In case of linearly separable data the decision surface is a hyperplane that maximizes the "margin" between the two classes.

k-nearest neighbor. Given an arbitrary input document, the system ranks its nearest neighbors among training documents, and uses the classes of the k top-ranking neighbors to predict the class of the input document. The similarity score of each neighbor document to the new document being classified is used as a weight if each classes, and the sum of class weights over the nearest neighbors are used for class ranking.

D. Performance measures

For evaluating the multi-label classification accuracy we use the standard multi-label measures: *precision*, *recall* and F_1 -*measure* both *micro-averaged* and *macro-averaged*.

– *Microaveraging:* precision, recall and F_1 -measure are obtained by summing

$$P_{mic} = \frac{\sum_{k=1}^{|C|} A_k}{\sum_{k=1}^{|C|} (A_k + B_k)}, \quad R_{mic} = \frac{\sum_{k=1}^{|C|} A_k}{\sum_{k=1}^{|C|} (A_k + C_k)},$$

$$F_{1mic} = \frac{\sum_{k=1}^{|C|} 2A_k}{\sum_{k=1}^{|C|} (2A_k + B_k + C_k)}.$$

Here, A_k , (B_k) is the number of documents correctly (incorrectly) assigned to the class c_k ; C_k (D_k) is the number of documents incorrectly (correctly) rejected to the class c_k .

– *Macroaveraging:* precision, recall and F_1 -measure are first evaluated for each class, and then the final measures are evaluated by averaging over the results of the all classes.

E. Experimental Results

In total we made 21 experiments and each experiment was performed for eleven different vocabulary sizes and measured by seven evaluation criterions. SFS are not usually used in text classification because of their computation cost due to large vocabulary size. However, in practice we can often either employ calculations from previous steps or make some pre-computations during the initialization. Since feature selection is typically done in an off-line manner, the computational time is not as important as the optimality of words subset or classification accuracy.

The time complexity of SFS algorithms is less than $O(|\mathcal{V}'||\mathcal{V}|^2)$ where $|\mathcal{V}'|$ is the number of desired words and $|\mathcal{V}|$ is the total number of words in the vocabulary. The required space complexity is $S(|\mathcal{V}|^2/2)$ because we need to store the mutual information for all couples of words (w, s) with $w \in \mathcal{U}$ and $s \in \mathcal{S}$.

SFS methods are more time consuming than BIF method but achieve significantly better results on the testing data. Since feature selection is typically done in an off-line manner, the

Fig. 1. **Naive Bayes**. Charts of micro-averaged F1 and macro-averaged F1 of Naive Bayes classifier on Reuters. The horizontal axis indicates the number of words.

Fig. 2. **SVM**. Charts of micro-averaged F1 and macro-averaged F1 of SVM classifier on Reuters. The horizontal axis indicates the number of words.

computational time is not as important as the optimality of words subset or classification accuracy.

The charts on Figures 1, 2, 3 show micro-averaged and macro-averaged performance criteria for various feature selection measures. In our experiments the best micro-averaged performance achieved by the new SFS methods using improved mutual information.

VI. CONCLUSIONS AND FUTURE WORK

This paper attacked the problem of feature selection for document classification. In this paper we presented methods based on novel improved mutual information measures. The proposed algorithms are new in the text classification field.

- Our experiments on the Reuters data indicate, that the SFS methods work significantly better than the commonly

used BIF methods on text classification. The proposed methods significantly overcome BIF method based on mutual information, information gain, χ^2 statistic and odds ratio on the testing data.

- The experimental results show that the SFS MaxIMI works with the highest classification evaluation measured by F_1 , precision and recall measures. The SFS MaxIMI modification of the sequential forward selection algorithms performs better than other tested variants on the Reuters data.
- In this paper we also present a comparative experimental study of three classifiers. SVM overcomes on average both Naive Bayes and k-nearest neighbor classifier.

Efficient search strategy is very important for feature selection considering high dimensional search space (about 5000

Fig. 3. kNN. Charts of micro-averaged F1 and macro-averaged F1 (right-side) of kNN classifier on Reuters. The horizontal axis indicates the number of words.

words). Future work could include:

- Investigation of sequential *floating* search methods [16] and *oscillating* algorithms [17] using the proposed new feature selection measures. These methods overcome effectively the "nesting" problem which sequential feature selection methods suffer from.
- Design of a new model for text document modelling based on modification of discrete distribution mixtures of factorized components to be able to solve simultaneously the problem of feature selection and classification [18], [19].
- Investigation feature clustering as an alternative to feature selection for reducing dimensionality of text data.

ACKNOWLEDGMENT

The work was supported in part by grants of the Grant Agency of the Academy of Sciences of the Czech Republic No A2075302, the Grant Agency of the Czech Republic No 102/03/0049 and the Complex research project of the Academy of Sciences of the Czech Republic K1019101.

REFERENCES

- [1] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, 1998.
- [2] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, pp. 103–134, 2000.
- [3] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the ECML'98*, 1998, pp. 137–142.
- [4] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol. 1, pp. 67–68, 1999.
- [5] Y. Yang, J. Zhang, and B. Kisiel, "A scalability analysis of classifier in text categorization," in *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.
- [6] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, March 2002.
- [7] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th ICML97*, 1997, pp. 412–420.
- [8] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive Bayes," in *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 258–267.
- [9] G. Forman, "An experimental study of feature selection metrics for text categorization," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.
- [10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, 1994.
- [11] N. Kwak and C. Choi, "Improved mutual information feature selector for neural networks in supervised learning," in *Int. Joint Conf. on Neural Networks (IJCNN '99)*, 1999, pp. 1313–1318.
- [12] M. Caropreso, S. Matwin, and F. Sebastiani, *Text Databases and Document Management: Theory and Practice*. Idea Group Publishing, Hershey, PA, 2001, ch. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization, p. 78102.
- [13] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 42–49.
- [14] D. Mladenic, "Feature subset selection using in text learning," in *10th European Conference on Machine Learning*, 1998, pp. 95–100.
- [15] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000.
- [16] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, November 1993.
- [17] P. Somol and P. Pudil, "Oscillating search algorithms for feature selection," in *Proceedings of the 15th IAPR Int. Conference on Pattern Recognition, Conference B: Pattern Recognition and Neural Networks*, 2000, pp. 406–409.
- [18] J. Novovičová, P. Pudil, and J. Kittler, "Divergence based feature selection for multimodal class densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 218–223, 1996.
- [19] P. Pudil, J. Novovičová, and J. Kittler, "Feature selection based on approximation of class densities by finite mixtures of special type," *Pattern Recognition*, vol. 28, pp. 1389–1398, 1995.