KYBERNETIKA — VOLUME 40 (2004), NUMBER 3, PAGES 293 304

TEXT DOCUMENT CLASSIFICATION BASED ON MIXTURE MODELS

JANA NOVOVIČOVÁ AND ANTONÍN MALÍK

Finite mixture modelling of class-conditional distributions is a standard method in a statistical pattern recognition. This paper, using bag-of-words vector document representation, explores the use of the mixture of multinomial distributions as a model for class-conditional distribution for multiclass text document classification task. Experimental comparison of the proposed model and the standard Bernoulli and multinomial models as well as the model based on mixture of multivariate Bernoulli distributions was performed using Reuters-21578 and Newsgroups data sets. Preliminary experimental results indicate the effectiveness of the proposed model in a text classification problem.

Keywords: text classification, multinomial mixture model

AMS Subject Classification: 62H30, 62G05, 68T10

1. INTRODUCTION

The objective of *text document classification* is to assign a free document into one or more predefined *classes* or *categories* based on its contents. It is a straightforward concept from supervised pattern recognition or machine learning. It implies the existence of the training data set of pre-classified documents, a way to represent the documents, and a statistical classifier trained using the chosen representation.

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. [4, 7, 10, 18, 19, 20] and the references therein).

In the 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). The number of potential words often exceeds the number of training documents. *Feature selection* is a very important step in the text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy. Well chosen features can improve classification accuracy and reduce the amount of training documents needed to obtain a required level of performance. Methods for feature subset selection for text document classification task use some evaluation function that is applied to a single feature. All features are independently evaluated, a score is assigned to each of them and the features are sorted according to the assigned score. Then, a predefined number of the best features is taken to form the best feature subset. Scoring of individual features can be performed using some of the measures, for instance, *document frequency*, *term frequency*, *mutual information*, *information gain*, *odds ratio*, χ^2 *statistic* and *term strength* [3, 9, 17]. Recently, Novovičová and Malík in [14] proposed to use sequential forward selection methods based on *improved mutual information* (introduced by Battiti [1] and Kwak and Choi [6] for non-textual data) as a criterion for reducing the dimensionality of text data. These feature evaluation functions take into consideration how features work together. The performance of these evaluation functions compared to the information gain which evaluate features individually has been discussed. The experimental results using naive Bayes classifier based on multinomial model on the Reuters data set have been analyzed by the authors from various perspectives, including F_1-measure, precision and recall. Preliminary experimental results indicate the effectiveness of the proposed feature selection algorithms in a text classification problem.

Forman in [3] presents an extensive comparative study of twelve feature selection criteria for the high-dimensional domain of text classification focusing on support vector machines and two class problem. The experimental results revealed the surprising performance of a new feature selection criterion called *bi-normal separation*. Yang and Liu [18] have shown that support vector machines perform favorably compared to competing procedures for the text classification. However, they did not consider the Bayes classifier based on the mixture models for class-conditional probability functions.

Recently, Ueda and Saito [16] proposed new types of mixture models for multiclass and multi-labelled text categorization, and efficient algorithms for both learning and classification. The authors empirically showed that their method could significantly outperform the conventional methods (naive Bayes, k-nearest neighbor, support vector machines, neural network) using real World Wide Web pages.

In this paper, we approach the text classification problem by using mixture model for class-conditional probability functions, which is a standard method in statistical pattern recognition. We focus on the application of the mixture of multivariate Bernoulli distributions (Bernoulli mixture model) and on the mixture of multinomial distributions (multinomial mixture model) for the task of text document classification. Bernoulli mixture model has been investigated by Juan and Vidal in [5]. The proposed approach is a generalization of naive Bayes that tries to properly model significant class-conditional dependencies by spreading them over different class mixture components. Maximum-likelihood estimation of mixture parameters is done by using the well-known expectation-maximization (EM) algorithm. Preliminary experimental results on Reuters-21578 and Newsgroups data sets indicate the effectiveness of proposed mixture models in a text classification task; an increase in classification accuracy has been achieved.

2. PROBABILISTIC FRAMEWORK FOR TEXT CLASSIFICATION

We consider classification of the text document into one of $|C|$ classes from the set of classes $C = \{c_1, \dots, c_{|C|}\}$ in a Bayesian learning framework with bag-of-words representation of the documents.

The framework defines a probabilistic model for the data and embodies two assumptions about the generation process [7, 10]: (a) the data are produced by a mixture model, (b) there is a one to one correspondence between mixture components and classes. In this formulation, every document is the vector generated according to a probability distribution defined by a set of parameters, denoted by θ. The probability distribution consists of a mixture of components $c_j \in C = \{c_1, \dots, c_{|C|}\}$. The data generation procedure for a document d_i can be understood as selecting a mixture component (a class) according to the mixture weights – *class prior probabilities*, $P(c_j)$, then having the corresponding mixture component generate a document according to its own parameters with *class-conditional probability function* $P(d_i|c_j;\theta_j)$. The *unconditional probability function* of generating document d_i independent of its class is given by

$$P(d_i;\theta) = \sum_{j=1}^{|C|} P(c_j)P(d_i|c_j;\theta_j) \tag{1}$$

where class prior probabilities $\theta_{c_j} = P(c_j)$, $0 \le \theta_{c_j} \le 1$, $\sum_{j=1}^{|C|} \theta_{c_j} = 1$ indicate the probabilities of selecting the different class mixture components, $P(d_i|c_j;\theta_j)$. Clearly, $\theta = \{(\theta_{c_j}, \theta_j) : j = 1, \dots, |C|\}$ is an unknown parameter set.

According to the bag-of-words representation, which ignores the order of word occurrence in a document, the document d_i can be represented by a feature vector consisting of one feature variable d_{it} for each word w_t in the given vocabulary $V = \{w_1, \dots, w_{|V|}\}$ containing $|V|$ distinct words; $t = 1, \dots, |V|$. The feature variables d_{it} can indicate either the presence or absence of the word w_t or can indicate some measure of the frequency of the word w_t in the document d_i.

There are two common models (we call them standard models) in the generative representation of text document (see e. g. [7]).

In the *multivariate Bernoulli* model based on the naive Bayes assumption (the words of the document are generated independently of the other words in the same document given the class label and furthermore, independent on its position in the document), the document is a binary vector over the space of words. Each feature variable of the document is either 0 or 1, indicating whether word w_t occurs at least once in the document. Each class can be represented as a multivariate Bernoulli distribution.

In the *multinomial* model, the document is represented by a feature vector, each feature variable is the number of times word w_t occurs in the document. In this model each document is drawn from a multinomial distribution over the set of all words in V with as many independent trials as the length of the document.

The estimate of complete set of model parameters for standard models can be found e. g. in the paper of McCallum and Nigam [7]. The authors clarify that the multinomial model has been found superior to the Bernoulli model in document classification.

3. FINITE MIXTURE MODELS

In a simple text document classification problem, there is a fixed, known number of classes and each class is modelled by a single component. A collection of class-labelled training data is used to estimate the class priors and class-conditional probability functions, following the conventional supervised statistical learning approach. Then a new document is classified by maximum posterior probability.

The usage of finite mixtures for class-conditional probability functions is a useful method in pattern recognition, because mixture models are able to represent arbitrarily complex probability functions (see e.g. [8]). Mixtures are flexible enough for finding an appropriate tradeoff between model complexity and the amount of the training data available. Usually, model complexity is controlled by varying the number of mixture components while keeping the same parametric form for all components. If each class is to be modelled by more then one component, we face an unsupervised learning problem.

Although mixture models are important for continuous data, some research shows, that these models perform, with discrete data in text classification, very well too [5, 13, 16].

Our approach to learning on text document is based on the fact that documents in the same class are often mixtures of multiple topics. Words within the document are not independent of each other. In our mixture approach to text document classification each class c_j is to be modelled by M_j ($M_j > 1$) components. It means we assume that the jth class-conditional probability function $P(d_i|c_j;\theta_j)$ is modelled as a finite mixture of the M_j probability functions with its own parameters. It can be expressed as

$$P(d_i|c_j;\theta_j) = \sum_{m=1}^{M_j} \alpha_{jm} P_m(d_i|s_{jm};\theta_{jm}) \,. \tag{2}$$

Here, M_j denotes the number of subclasses, say s_{jm}, in each class c_j, α_{jm} is the mixing proportion of the mth model component s_{jm} in class c_j, (also called "within-class" mixing proportion), $\alpha_{jm} \geq 0$, $\sum_{m=1}^{M_j} \alpha_{jm} = 1$, $j = 1, \ldots, |C|$. $P_m(d_i|s_{jm};\theta_{jm})$ denotes the probability of the document d_i in the mth subclass within the class c_j. Note, that in proposed approach each class c_j is divided into M_j artificial subclasses s_{jm}. In the sequel we suppose that $M_j = M$ for all $j = 1, \ldots, |C|$.

Substituting (2) into (1) yields the unconditional probability function $P(d_i;\theta)$ in the form of the mixture of mixtures

$$P(d_i;\theta) = \sum_{j=1}^{|C|} P(c_j) \sum_{m=1}^{M} \alpha_{jm} P_m(d_i|s_{jm};\theta_{jm}) \,. \tag{3}$$

3.1. Bernoulli mixture and multinomial mixture models

Bernoulli mixture model for class-conditional probability function is a mixture of the form (2) such that each component $P_m(d_i|s_{jm};\theta_{jm})$ has a multivariate Bernoulli probability function. Therefore, the probability function of the document given its class is approximated by the multivariate Bernoulli mixture

$$P(d_i|c_j;\theta_j) = \sum_{m=1}^{M} \alpha_{jm} \prod_{t=1}^{|V|} \theta_{t|jm}^{B_{it}} (1 - \theta_{t|jm})^{(1-B_{it})} , \tag{4}$$

where B_{it} is either 0 or 1, indicating absence or presence of the word w_t in the document d_i. Here, associated with each subclass s_{jm} is a word probability written $\theta_{t|jm} = P(w_t|s_{jm};\theta_{jm})$ for all words in the vocabulary $|V|$, $0 \leq \theta_{t|jm} \leq 1$. The parameter set $\theta_j = \{(\alpha_{jm},\theta_{jm}) : m = 1,\dots,M\}$, $j = 1,\dots,|C|$ with $\theta_{jm} = (\theta_{1|jm},\dots,\theta_{|V||jm})$ is unknown. The model (4) has been investigated by Juan and Vidal in [5].

We propose to use the *multinomial mixture model* (mixture of multinomial distributions) as a model for class-conditional probability function. It means that the probability function of the document d_i in the jth class has the form

$$P(d_i|c_j;\theta_j) = \sum_{m=1}^{M} \alpha_{jm} \frac{|d_i|!}{\prod_{t=1}^{|V|} N_{it}!} \prod_{t=1}^{|V|} \theta_{t|jm}^{N_{it}} \tag{5}$$

where N_{it} is the number of times word w_t occurs in the document d_i, $|d_i|$ is the length of d_i, $\theta_{t|jm}$ is the probability $P(w_t|s_{jm};\theta_{jm})$ of the word w_t in the subclass s_{jm}. The unknown parameter set is $\theta_j = \{(\alpha_{jm},\theta_{jm}) : m = 1,\dots,M\}$, $j = 1,\dots,|C|$ with $\theta_{jm} = (\theta_{1|jm},\dots,\theta_{|V||jm})$, $0 \leq \theta_{t|jm} \leq 1$, $\sum_{t=1}^{|V|} \theta_{t|jm} = 1$.

3.2. Model fitting with the EM algorithm

Let $\mathcal{D}_j = \{d_1,\dots,d_{|\mathcal{D}_j|}\}$ be a set of $|\mathcal{D}_j|$ independent and identically distributed training documents from class $c_j \in C$. Our problem consists of learning a mixture model for each class from training documents that are only labelled by the class they belong to. Estimation of the parameters of class-conditional probability function $P(d_i|c_j;\theta_j)$ given in (2), where $\theta_j = \{(\alpha_{jm},\theta_{jm}) : m = 1,\dots,M\}$, $j = 1,\dots,|C|$, is equivalent to solving the following optimization problem:

Maximize the log-likelihood function

$$L_j = \sum_{i=1}^{|\mathcal{D}_j|} \log \left[\sum_{m=1}^{M} \alpha_{jm} P_m(d_i|s_{jm};\theta_{jm}) \right] \tag{6}$$

under the constraints: $\alpha_{jm} \geq 0$, $m = 1,\dots,M$, $\sum_{m=1}^{M} \alpha_{jm} = 1$.

We are excluding the number of class-conditional components from the estimation problem. The estimate $\hat{\theta}_j$ cannot be found analytically. A possible approach is the *expectation-maximization* (EM) algorithm which is a general framework to incomplete data problems [2]. From this point of view, an observed document d_i can be regarded as being incomplete where the missing part is the true subclass labelling.

The EM algorithm alternates two steps: (1) E-step (an expectation step) where posterior probabilities are computed for the subclass variables, based on current estimates of the parameters, (2) M-step (maximization step) where parameters are updated based on so called expected complete data log-likelihood which depends on the posterior probabilities computed in the E-step.

The EM algorithm for models (4) and (5) is derived in Appendix A in [12]. Starting from an initial value $\theta_j^{(0)}$, the kth iteration is defined as follows:

E-Step: $j = 1, \ldots, |C|,\ m = 1, \ldots, M,\ i = 1, \ldots, |\mathcal{D}_j|,\ k = 0, 1, \ldots$

$$p_B^{(k)}(s_{jm}|d_i) = \frac{\alpha_{jm}^{(k)} \prod_{t=1}^{|V|} \left(\theta_{t|jm}^{(k)}\right)^{B_{it}} \left(1 - \theta_{t|jm}^{(k)}\right)^{1-B_{it}}}{\sum_{r=1}^{M} \alpha_{jr}^{(k)} \prod_{t=1}^{|V|} \left(\theta_{t|jr}^{(k)}\right)^{B_{it}} \left(1 - \theta_{t|jr}^{(k)}\right)^{1-B_{it}}} \tag{7}$$

in the case of Bernoulli mixture model (4) or

$$p_M^{(k)}(s_{jm}|d_i) = \frac{\alpha_{jm}^{(k)} \prod_{t=1}^{|V|} \left(\theta_{t|jm}^{(k)}\right)^{N_{it}}}{\sum_{r=1}^{M} \alpha_{jr}^{(k)} \prod_{t=1}^{|V|} \left(\theta_{t|jr}^{(k)}\right)^{N_{it}}} \tag{8}$$

in the case of multinomial mixture model (5).

M-Step: $j = 1, \ldots, |C|,\ m = 1, \ldots, M,\ t = 1, \ldots, |V|,\ k = 0, 1, \ldots$

$$\alpha_{jm}^{(k+1)} = \frac{1}{|\mathcal{D}_j|} \sum_{i=1}^{|\mathcal{D}_j|} p_B^{(k)}(s_{jm}|d_i) \tag{9}$$

and

$$\theta_{t|jm}^{(k+1)} = \frac{\sum_{i=1}^{|\mathcal{D}_j|} B_{it}\, p_B^{(k)}(s_{jm}|d_i)}{\sum_{i=1}^{|\mathcal{D}_j|} p_B^{(k)}(s_{jm}|d_i)} \tag{10}$$

in the case of model (4).
For mixture model (5) we obtain

$$\alpha_{jm}^{(k+1)} = \frac{1}{|\mathcal{D}_j|} \sum_{i=1}^{|\mathcal{D}_j|} p_M^{(k)}(s_{jm}|d_i) \tag{11}$$

and

$$\theta_{t|jm}^{(k+1)} = \frac{\sum_{i=1}^{|\mathcal{D}_j|} N_{it}\, p_M^{(k)}(s_{jm}|d_i)}{\sum_{r=1}^{|V|} \sum_{i=1}^{|\mathcal{D}_j|} N_{ir}\, p_M^{(k)}(s_{jm}|d_i)}\,. \tag{12}$$

The class priors can be estimated as

$$\hat{\theta}_{c_j} = \frac{|\mathcal{D}_j|}{|\mathcal{D}|} \tag{13}$$

where $|\mathcal{D}| = \sum_{j=1}^{|C|} |\mathcal{D}_j|$.

The EM algorithm produces a sequence of estimates $\{\theta_j^{(k)}, k = 1, 2, \ldots\}$ by alternatively applying E-step and M-step until some convergence criterion is met.

3.3. Classification

Given estimate $\hat{\theta} = \{(\hat{\theta}_{c_j}, \hat{\theta}_j) : j = 1, \ldots, |C|\}$ of the complete set of model parameters θ of the model (1) calculated from the training documents, classification can be performed on the test document d by calculating the posterior probability of each class by applying Bayes rule

$$P(c_j|d; \hat{\theta}) = \frac{\hat{\theta}_{c_j} P(d|c_j; \hat{\theta}_j)}{P(d; \hat{\theta})}, \quad \sum_{j=1}^{|C|} P(c_j|d; \hat{\theta}) = 1, \tag{14}$$

and simply select the class with the highest posterior probability.

4. EXPERIMENTAL RESULTS

This section provides empirical evidence that the Bernoulli mixture and the multinomial mixture models perform better than the corresponding standard models.

4.1. Data sets

For performance evaluation, we carried out the experiments with the *Rcuters-21578* data collection[1] and the *Newsgroups* data set[2].

Since the aim is classification in which each document has an exclusive category, we discarded documents with no label or with multiple labels from the Reuters data. Furthermore, the classes with less than twenty documents were removed. The resulting data set had 9159 documents with 33 document classes. After removing very short documents from the Newsgroups data, the resulting data set had 19958 documents in 20 classes.

The vocabulary was constructed by removing stop words, too infrequent words (words that had less than 3 occurrences per document) and the Porter stemming algorithm[3] was used. This resulted in a vocabulary of 7425 words in the case of Reuters data and of 21951 words in the case of Newsgroups data.

We randomly split data set for each class into two-third training set and one-third testing set. We repeated this random split twenty times.

Following traditional feature selection techniques for text classification, non-discriminative words were removed in accordance to the information gain criterion defined in [7]. It was computed from the training set for each word, and the $|V|$ most informative words were selected. Then each document was represented as a $|V|$-dimensional feature vector. As the dimension $|V|$ has an important effect in classification performance, several values of $|V|$ were considered in the experiments.

The effectiveness of the Bayes plug-in text classifier was measured by classification accuracy estimated as

$$accuracy = \frac{N_c}{N_{test}} 100\,\%$$

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578

[2] http://www.cs.cmu.edu/~textlearning

[3] http://www.tartarus.org/~martin/PorterStemmer

where N_c is the number of correctly classified documents from testing set and N_{test} is the total number of documents in the testing set (percentage of the test documents that were correctly classified). The average classification accuracy has been computed over all twenty testing sets.

4.2. Standard models versus mixture models

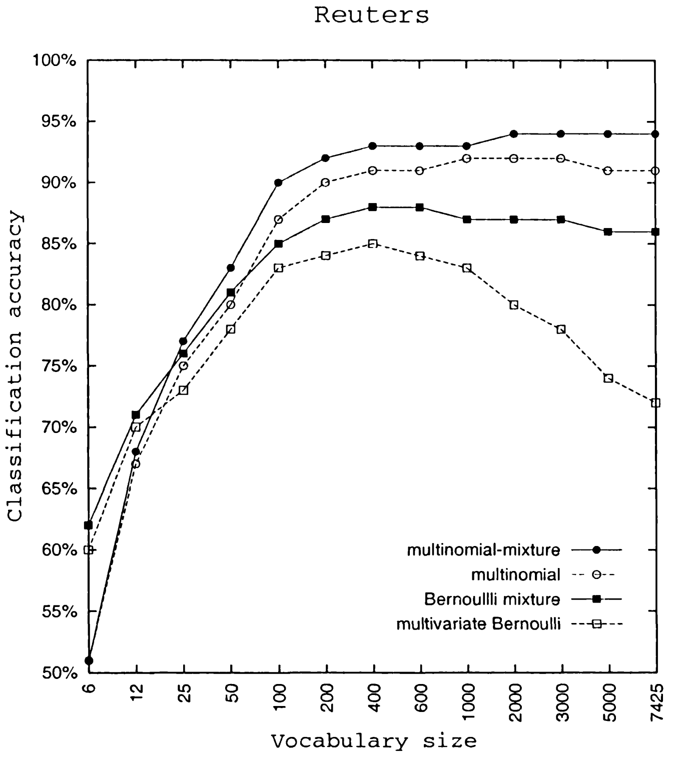

Fig. 1. Classification accuracy of the standard and the mixture models on the Reuters data set.

Figure 1 shows the performance of both the standard and the mixture models for several vocabulary sizes on the Reuters data set. We can see that the multinomial model performs better than the Bernoulli model. Good behavior of the Bernoulli model is observed for dimensions equal or smaller than 400, after this point the performance of the classifier degrades with an increase in vocabulary size. Accuracy of multinomial model improves monotonically.

The Bernoulli mixture is found to be better than the Bernoulli model on average of 5 %. The multinomial mixture achieves the highest accuracy 94.9 % and is on average 2 % better than multinomial model and on average 6 % better than Bernoulli mixture. The multinomial mixture performs the best results over all vocabulary sizes, except on very small number of words (6 and 12).

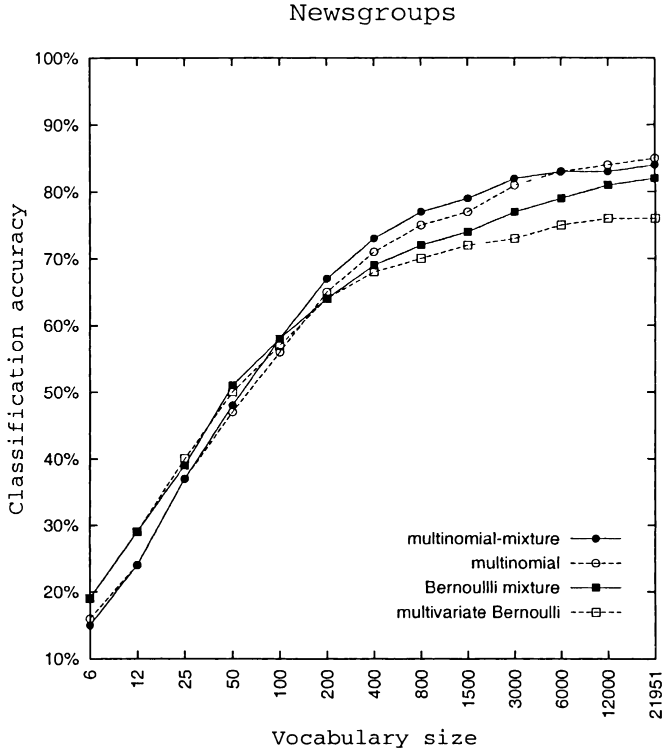

Fig. 2. Classification accuracy of the standard and the mixture models on the Newsgroups data set.

Figure 2 shows the behavior of the standard and the mixture models on the Newsgroups data set. The Bernoulli mixture achieves on average 2.1 % the higher accuracy than Bernoulli model. The multinomial mixture performs slightly better than the standard multinomial model (on average 0.6 %).

The number of components for each class-conditional mixture also has an important impact on performance. Good behavior was observed with three components per class for both mixture models on the Reuters data set. Six components per class were used on the Newsgroups data set.

The EM algorithm ran for each class separately. We used random initialization for parameters $\theta_{t|jm}$ and $\alpha_{jm}^{(0)} = 1/M$ for both Bernoulli mixture and multinomial mixture models. If the relative change in log-likelihood function is small, i. e. $|L_j^{(k+1)} - L_j^{(k)}|/L_j^{(k)} < \epsilon$ or the maximum number of iterations (100) is reached, we terminate the iteration process. The methods to train a classifier based on finite mixtures for class-conditional probability function are computationally more demanding than methods based on standard models.

5. CONCLUSIONS AND FUTURE RESEARCH

The following conclusions are reached from this paper:

- Bernoulli mixture and multinomial mixture models have been used as class-conditional models for the task of text document classification to relax the naive Bayes class-conditional independence assumption. This generalization of naive Bayes tries to properly model significant class-conditional dependencies by spreading them over different mixture components.

- An observation of the performance of Bayes classifier for text classification on Reuters-21578 and Newsgroups data sets suggests that learning methods based on Bernoulli mixture and multinomial mixture models for class-conditional probability functions of the documents performed better than the corresponding standard models.

- The multinomial mixture is promising model for class-conditional distributions in text classification task. The experimental results show that this model performs better than Bernoulli mixture model.

Many areas of future work remain. Ongoing work could include:

- 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 the optimal feature subset and the optimal number of mixture components [11, 15].

- Experimental comparison of Bayes classifier based on multinomial mixture model with other classifiers (e. g. support vector machines) for text document classification.

- Feature clustering as an alternative to feature selection for reducing the dimensionality of text data will be investigated.

ACKNOWLEDGEMENTS

The work has been supported 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.

(Received October 15, 2003.)

REFERENCES

[1] R. Battiti: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Networks 5 (1994), 537–550.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B *39* (1977), 1–38.

[3] G. Forman: An experimental study of feature selection metrics for text categorization. J. Mach. Learning Res. *3* (2003), 1289–1305.

[4] T. Joachims: Text categorization with support vector machines: Learning with many relevant features. In: Proc. 10th European Conference on Machine Learning (ECML'98), 1998, pp. 137–142.

[5] A. Juan and E. Vidal: On the use of Bernoulli mixture models for text clasification. Pattern Recognition *35* (2002), 2705–2710.

[6] N. Kwak and C. Choi: Improved mutual information feature selector for neural networks in supervised learning. In: Proc. Internat. Joint Conference on Neural Networks (IJCNN '99), 1999 pp. 1313 1318.

[7] A. McCallum and K. Nigam: A comparison of event models for naive Bayes text classification. In: Proc. AAAI-98 Workshop on Learning for Text Categorization, 1998.

[8] G. J. McLachlan and D. Peel: Finite Mixture Models. Wiley, New York 2000.

[9] D. Mladenic and M. Grobelnik: Feature selection for unbalanced class distribution and Naive Bayes. In: Proc. Sixteenth Internat. Conference on Machine Learning, 1999, pp. 258 267.

[10] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell: Text classification from labeled and unlabeled documents using EM. Mach. Learning *39* (2000), 103 134.

[11] J. Novovičová, P. Pudil, and J. Kittler: Divergence based feature selection for multimodal class densities. IEEE Trans. Pattern Anal. Machine Intell. *18* (1996), 218 223.

[12] J. Novovičová and A. Malík: Text Document Classification Using Finite Mixtures. Research Report No. 2063, Institute of Information Theory and Automation, Prague 2002.

[13] J. Novovičová and A. Malík: Application of multinomial mixture model to text classification. In: Pattern Recognition and Image Analysis (Lecture Notes in Computer Sciences 2652), Springer- Verlag, Berlin 2003, pp. 646–653.

[14] J. Novovičová, A. Malík, and P. Pudil: Feature selection using improved mutual information for text classification. In: Structural, Syntactic and Statistical Pattern Recognition (Lecture Notes in Computer Science), Springer–Verlag, Berlin 2004 (in press).

[15] P. Pudil, J. Novovičová, and J. Kittler: Feature selection based on approximation of class densities by finite mixtures of special type. Pattern Recognition *28* (1995), 1389 1398.

[16] N. Ueda and K. Saito: Parametric mixture models for multi-labeled text. In: Proc. Neural Information Processing Systems, 2003.

[17] Y. Yang and J. O. Pedersen: A comparative study on feature selection in text categorization. In: Proc. Internat. Conference on Machine Learning, 1997, pp. 412–420.

[18] Y. Yang and X. Liu: A re-examination of text categorization methods. In: Proc. 22nd Internat. ACM SIGIR Conference on Research and Development in Inform. Retrieval, 1999, pp. 42–49.

[19] Y. Yang: An evaluation of statistical approaches to text categorization. J. Inform. Retrieval *1* (1999), 67–88.

[20] Y. Yang, J. Zhang, and B. Kisiel: A scalability analysis of classifier in text categorization. In: Proc. 26th ACM SIGIR Conference on Research and Development in Inform. Retrieval, 2003.

Jana Novovičová, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8, Faculty of Management University of Economics, and Faculty of Transportation Sciences Czech Technical University in Prague. Czech Republic.

Antonín Malík, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8 and Faculty of Electrical Engineering Czech Technical University in Prague. Czech Republic.

e-mails: novovic, amalik@utia.cas.cz