ALGORITHM FOR CLASSIFICATION OF TEXTUAL DOCUMENTS REPRESENTED BY TANDEM ANALYSIS

Jasminka Dobša
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, Varaždin, Croatia
Tel: +385 42 390844; fax: +385 42 213413
e-mail: jasminka.dobsa@foi.hr

ABSTRACT

In this research is presented algorithm for classification of textual documents which are represented in the space of reduced dimension in respect to original bag of words representation. Algorithm is carried out in two steps: in the first step classification is conducted for documents represented in original bag of words representation, while in the second step classification is conducted for documents represented in the space of reduced dimension. Reduction of dimensionality is obtained also in two steps: in the first step documents are represented by usage of latent semantic indexing, while in the second step this representation is projected on the space of membership matrix defining a membership of documents in classes. Evaluation of algorithm is conducted on Reuters21578 collection of documents.

1 INTRODUCTION

In this paper is represented algorithm for classification of textual documents which is carried out in two steps. In the first step documents are represented using the *bag-of-words representation*, also referred to as *vector space model.* The vector space model is implemented by creating the *term-document matrix*, which can be explained as follows. Let the list of relevant terms for a certain collection be numerated from *1* to *m* and documents be numerated from *1* to *n*. The term-document matrix is an $m \times n$ matrix $A = [a_{ij}]$ where a_{ij} represents the weight of term i in document j. In the term-document matrix, documents are represented as column vectors which dimension is the number of relevant terms. The main characteristics of such text representation are high dimensionality of input space and sparseness of the term-document matrix.

Next step in presented classification algorithm is dimensionality reduction which is also carried out in two steps. In the first step original representations of textual documents are represented by method of latent semantic indexing (LSI). In the next step representations of documents obtained by LSI are projected on the space spread by columns of membership matrix which defines membership of documents in classes. Terms occurring in the document may not be the best representation of the document content, due to the problems of synonymy (different words with similar meaning) and polysemy (one word with more meanings). By dimensionality reduction we can represent a collection of documents in a more compact way, which could save memory space, speed up the main tasks and reduce the effect of noise in the data. The method of latent semantic indexing is introduced in [2]. Today, it presents benchmark in the field of representation of documents in the space of reduced dimensionality. According to some earlier investigations [6] the method of LSI has some disadvantages in fulfilling the task of classification since its application could remove some significant information concerning structures of the classes. The idea behind second step in dimensionality reduction is to stress the structure of the classes in the data by projection on columns of class membership matrix. The inspiration for such a step comes from the method of Factorial K-means introduced by Vichi and Kiers [9]. Therefore dimensionality reduction of the original representation of documents in term-documents matrix is obtained sequentially in two steps in procedure called *tandem analysis* [7] which is frequently used by practitioners, but is not applied in this form for the task of classification of textual documents yet. Similar approach is proposed in research of Dhillon and Modha [3]. In their work is proposed reduction of dimension of term-document matrix by concept decomposition - projection of documents on centroids of groups obtained by spherical k-means algorithm. It is shown that indexing of documents obtained by concept decomposition can improve performance of information retrieval of documents [4]. Also, concept decomposition applied in its supervised form by projection of documents on centroids of classes can improve classification performance [6].

The rest of the paper is organized as follows. In the second section is given description of used techniques for dimensionality reduction (latent semantic indexing and tandem analysis). Third section gives description of methods used for automatic classification of data. In forth section are

given results of experiments and the last section gives conclusion and discussion with directions for a further work.

2 DIMENSIONALITY REDUCTION TECHNIQUES

The idea of dimensionality reduction techniques is to represent documents by clustering them based on topic similarity regardless of indexing terms used. In the case of LSI the approximate representation of documents is accomplished using a truncated singular value decomposition (SVD) approximation of the term-document matrix. SVD of an arbitrary matrix **A** is given by

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}V^T \qquad (1)$$

where **U** is $m \times m$ orthogonal matrix where m is number of indexing terms, **V** is $n \times n$ orthogonal matrix where n is number of documents in collection and **Σ** is diagonal matrix on whose diagonal are singular values of matrix **A** in a decreasing order. For a purpose of representation of textual documents truncated SVD is used which has form

$$\mathbf{A} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T \qquad (2)$$

where $\mathbf{U_k}$ is $m \times k$ matrix whose columns consist of the first k columns of matrix **U,** $\mathbf{V_k}$ is $n \times k$ matrix whose columns consist of the first k columns of matrix **V**, and $\mathbf{\Sigma_k}$ is diagonal matrix on whose diagonal are the greatest singular values of **A** ordered in decreasing order. Representations of documents by LSI method are said to be representations in LSI space and are given by columns of matrix $\mathbf{V_k}^T$.

Procedure of Tandem analysis is performed by projection of matrix of document's representation by LSI on columns of membership matrix **M** in the sense of least squares. Membership matrix **M** is $n \times k$ matrix which defines a membership of documents into classes in a way that $m_{ik} = 1$ if the i^{th} document belongs to k^{th} class and $m_{ik} = 0$ otherwise. It is feasible that document belongs to multiple classes. Projection of LSI representations of documents onto column space of **M** is accomplished by solving the least square problem

$$\left\| \mathbf{V}^T - \mathbf{MZ} \right\| \to \min. \qquad (3)$$

It is known that solution of a set problem is given by

$$\mathbf{Z} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^T \qquad (4)$$

and representation of documents by Tandem analysis is given by transpose of $n \times k$ matrix

$$\mathbf{B} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}.^T \qquad (5).$$

3 CLASSIFICATION METHODS

Text classification is procedure of assigning a labels of previously defined classes to a unstructured textual documents [8]. Characteristic of classes are learned on the training set of documents and tested on the test set. If **A** is term-document matrix of a set of training documents then representation of training documents by method of LSI is given by matrix $\mathbf{V_k}^T$, while representation of matrix **A** by method of Tandem analysis is given by transpose of matrix **B** given by formula (5). If **T** is term-document matrix of a test set of documents then representation of matrix **T** in LSI space is given by matrix **C** where

$$\mathbf{C}^T = \mathbf{T}^T\mathbf{U}_k\mathbf{\Sigma}_k^{-1}. \qquad (6)$$

Representation of test documents by method of Tandem analysis is given by transpose of a matrix

$$\mathbf{D} = \mathbf{N}(\mathbf{N}^T\mathbf{N})^{-1}\mathbf{N}^T\mathbf{T}^T\mathbf{U}_k\mathbf{\Sigma}_k^{-1} \qquad (7)$$

where **N** is membership matrix of a test matrix. Since it is not allowed to use membership matrix of a test matrix before final evaluation, this matrix has to be approximated and here it will be applied the first step of classification. Classification in the first step will be conducted by the method of k-nearest neighbors or by support vector machines (SVMs), while classification in the second step will be conducted only by SVMs

K- nearest neighbor (k-nn) algorithm is a type of example-based classifiers. It observes class of nearest documents and assigns class c to a document if large enough proportion of nearest documents belong to that class [8]. SVMs is an algorithm that finds a hyperplane which separates positive and negative training examples with maximum possible margin [1,5]. This means that the distance between the hyperplane and the corresponding closest positive and negative examples is maximized. A classifier of the form $sign(w \cdot x + b)$ is learned, where w is the weight vector or normal vector to the hyperplane and b is the bias. Depending on which side of separating hyperplane the test example is, its prediction will be positive or negative.

4 EXPERIMENT

Experiments are conducted on the 10 largest classes of standard Reuters21578 collection using "ModApte" split having 9603 training and 3299 test documents. After stop words and words that occurred in less than 4 documents are removed, the list of 9867 terms is formed. Classification of documents is conducted by k-nn algorithm for k=10 or SVM algorithm in the first step and by SVM algorithm in the second step. LSI method is conducted for k=90, which means that documents are represented in LSI space by vectors of dimension 90. All other representation obtained by Tandem analysis also use LSI with k=90. We have treated the problem of classification for each category as a two-class problem, with members of that category being positive examples and all other documents being negative examples. Evaluation was performed using a commonly used combination of precision, recall, and the F_1 measure. Precision p is a proportion of documents predicted positive that are actually positive. Recall r is defined as a proportion of positive documents that are predicted positive. The F_1 measure is defined as $F_1 = 2pr/(p+r)$. Macroaverage is an average value of measure of evaluation for all observed classes. For classification of documents by SVM method

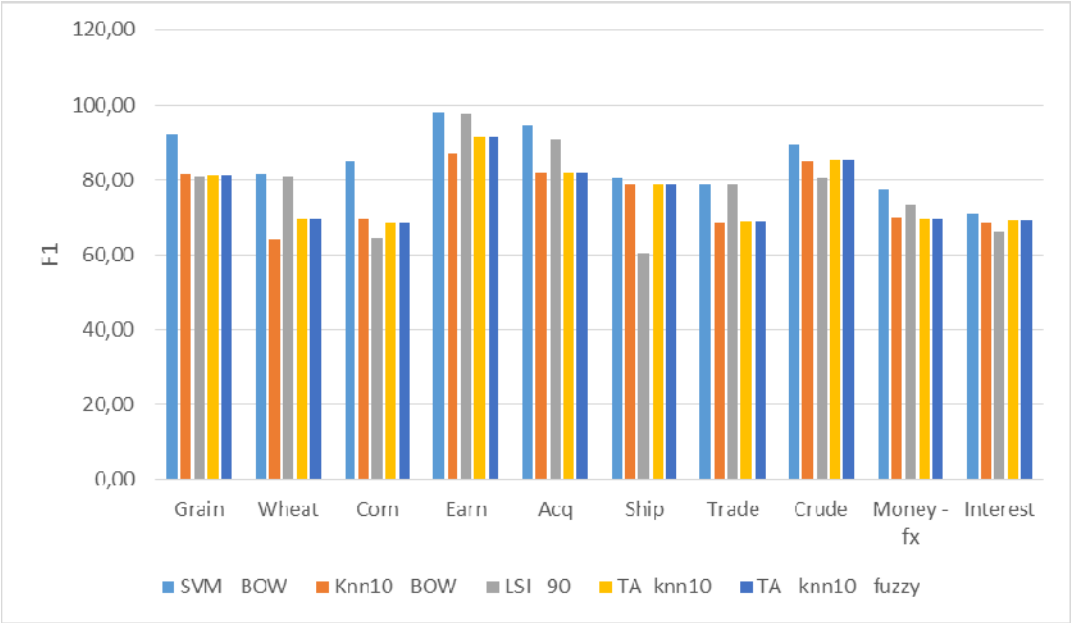

Figure 1: *F_1 measure for top 10 classes of Reuters21578 data set and for different representations of documents and used classification algorithms.*

Table 1*: Results of classification performance (precision, recall and F_1 measure) for top classes of Reuter21578 data set for different representations of documents. Classification in both steps of algorithm is conducted by SVMs.*

	Bag of words			***Tandem analysis***			***Tandem analysis modified***		
Class	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Grain	**97.04**	87.92	**92.26**	**97.04**	87.92	**92.26**	85.98	**94.63**	90.10
Wheat	**89.83**	74.65	81.54	**89.83**	74.65	81.54	78.75	**88.73**	**83.44**
Corn	**97.67**	75.00	**84.85**	**97.67**	75.00	**84.85**	81.82	**80.36**	81.08
Earn	**98.43**	97.79	**98.11**	**98.43**	97.79	**98.11**	93.36	**99.54**	96.35
Acq	**97.49**	91.66	**94.49**	**97.49**	91.66	**94.49**	89.14	**98.19**	93.45
Ship	**92.65**	70.79	80.26	**92.65**	70.79	80.26	75.00	**94.38**	**83.58**
Trade	**85.15**	73.50	**78.90**	**85.15**	73.50	**78.90**	62.21	**91.45**	74.05
Crude	**91.21**	87.83	**89.49**	**91.21**	87.83	**89.49**	76.39	**94.18**	84.36
Money - fx	**81.99**	73.74	77.65	**81.99**	73.74	77.65	72.32	**90.50**	**80.40**
Interest	**89.53**	58.78	70.97	**89.53**	58.78	70.97	73.33	**75.57**	**74.43**
Macroaverage	92.10	79.17	84.85	92.10	79.17	84.85	78.83	90.75	84.12

SvmLight v.5.0 software by Joachims (2002) with default parameters was used. For all other calculations it was used MATLAB v7.6.

On Figure 1 are shown results of classification performance in terms of F_1 measure. The first column shows F_1 measure for a classification obtained by bag of words representation with usage of SVM (SVM – BOW). The second column shows F_1 measure for classification performed by k-nn algorithm (k=10) for documents represented by bag of words representation (Knn10 – BOW). In third column are shown results of classification by SVM for LSI representation, while the fifth and the sixth column show F_1 measure for representation by Tandem analysis where classification is conducted by k-nn algorithm (k=10) in the first step and SVM algorithm in the second (TA – knn10 and TA – knn10 – fuzzy). Procedure denoted by TA – knn10 – fuzzy is a slight modification of procedure TA – knn. Namely, it can be seen from results of F_1 measure from Figure 1 that classification obtained by usage TA – knn10 representation did not improve much classification results by k-nn algorithm and bag of words representation. Since there is no need to decide categorically in the first step of classification about membership of documents to classes, the procedure TA – knn10 – fuzzy modifies

procedure TA – knn10 in a way that element n_{ik} of a membership matrix contains proportion of 10 nearest documents to i^{th} document contained in k^{th} class. For example, if there is 3 documents from k^{th} class among 10 nearest documents to i^{th} document then element of membership matrix for test set of documents is n_{ik} =0.3. In the case of TA – knn10 procedure n_{ik}=0, since decision that i^{th} document is in k^{th} class is made if there is at least 4 documents among 10 nearest documents to i^{th} document in the k^{th} class. From results shown in Figure 1 it can be seen that modification of TA – knn10 procedure did not result in significant improvement (there is improvement of macroaverage of approximately 1%) in terms of F_1 measure. Nevertheless, there are more differences in terms of precision and recall which is not elaborated here. Analysis of differences obtained by modifications of the predictions of classification obtained in the first step of algorithm will be discussed through results shown in Table 1. It shows results of classification performance (precision, recall and F_1 measure) for top 10 classes of Reuter21578 data set for three different representations of documents: bag of words representation, representation in reduced dimension space by Tandem analysis and representation in reduced dimension space by Tandem analysis with modifications of predictions obtained by classification in the first step. Classification is conducted by method of SVM in both steps. From Table 1 it can be seen that results of precision, recall and F_1 measure are exactly the same for a bag of words representation and representation obtained by Tandem analysis. Hence, by usage of Tandem analysis classification performance can be improved in comparison to LSI method (Figure 1), but apparently it is limited by approximation of membership matrix obtained in the first step of classification. In Tandem analysis with modification predictions obtained by SVM are modified in a following way: if prediction for i^{th} document belonging to a k^{th} class is greater than 0.6 then element of membership matrix for test set of documents is n_{ik} =1, if prediction is less than -0.6 then n_{ik} =0, otherwise n_{ik} =0.5. Such a modification resulted in significant improvement of classification recall, but precision of classification dropped at the same time resulting in a similar macroaverage of F_1 measure.

5 CONCLUSION AND DISSCUSION

In the paper is introduced a novel algorithm for a classification of textual documents represented in a space of reduced dimension obtained by Tandem analysis which consist of two steps. In the first step is performed LSI and in the second step representations in the LSI space are projected on a space spread by membership matrix of a train and test data set. Classification is performed twice, firstly to get approximate membership matrix of a test set of documents and secondly to classify documents represented by Tandem analysis. Although results of F_1 measure are the best for bag of words representation, it is shown that F_1 measure of classification is improved for 5 out of 10 top classes of Reuters21578 data set in comparison to LSI when k-nan algorithm is used in the first step of classification. Results of classification including precision, recall and F_1 measure when SVM algorithm is used in both steps are exactly the same for a bag of words representation and representation by Tandem analysis. This guides to a conclusion that classification performance is limited by the first step of classification. It is important to stress that in the case of bag of words representation documents are represented in space of dimension of almost 10 000 (number of indexing terms), while representation by Tandem analysis is of dimension 90 (dimension of LSI space). Hence, representation by Tandem analysis requires much less memory space.

In the further work method will be tested for a task of information retrieval and cross-lingual information retrieval. Also, algorithm of Tandem analysis will be compared with algorithm of simultaneous performance of both dimensionality reduction steps (reduction of variables/terms and reduction of objects/documents).

References

[1] Cristianini, N., Shave-Taylor, J., Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[2] Deerwester, S., Dumas, S., Furnas, G., Landauer, T., Harsman, R.., Indexing by latent semantic analysis, J. American Society for Information Science, 1990, 41: 391-407.

[3] Dhillon, I.S., Modha, D.S., Concept decomposition for large sparse text data using clustering, Machine Learning, 2001, 42(1): 143-175.

[4] Dobša, J., Dalbelo-Bašić, B. , Concept decomposition by fuzzy k-means algorithm, Proceedings of the IEEE/WIC International Conference on Web Intelligence, WI 2003, 2003, 684-688.

[5] Joachims, T., Text categorization with support vector machines: Learning with many relevant features, In Proc. of the European Conference on Machine Learning, 1998, Springer, 137-142.

[6] Karypis, G., Hong, E., Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval & categorization, Technical Report TR-00-016, Department of Computer Science, University of Minnesota, Minneapolis, 2000.

[7] Rocci, R., Gattone, S.A., Vichi, M., A new dimension reduction method: factor discriminant K-means, Journal of Classification, 28:210-226.

[8] Sebastiani, S., Machine learning in automated text categoriyation, ACM Computer Surveys, 2002, 34: 1-47.

[9] Vichi, M., Kiers, H.A. L., Factorial K-means analysis for two-way data, Computational Statistics and Data Analysis, 2001, 37:49-64.