2010 2nd International Conference on Industrial and Information Systems

Study of text classification methods for data sets with huge features

Guiying Wei， Xuedong Gao， Sen Wu
School of Economics & Management
University of Science &Technology Beijing
Beijing, China
weigy @manage.ustb.edu.cn

***Abstract*—Text classification has gained booming interest over the past few years. In this paper we look at the main approaches that have been taken towards text classification. The key text classification techniques including text model, feature selection methods and text classification algorithms are discussed .This work focus on the implementation of a text classification system based on Mutual Information and K-Nearest Neighbor algorithm and Support Vector Machine. The experimental results on Reuters collection are also presented. It shows that Mutual Information is a kind of efficient dimension reduction method for text data sets with huge features.**

Keywords-Text classification; feature selection; K-Nearest Neighbor; Mutual Information(key words)

I. INTRODUCTION

With the dramatic rise in the use of the Internet, there has been an explosion in the volume of online documents and electronic mail. Text classification (or text categorization), assigning free text to one or more predefined categories based on their content, is an important components in many real-time tasks like sorting of email into folders, topic identification and other information management tasks.

Text classification is now being applied in many contexts, ranging from document indexing based on a controlled vocabulary, to document filtering, automated metadata generation ， word sense disambiguation, population of hierarchical catalogues of Web resources, and in general any application requiring document organization or selective and adaptive document dispatching[1].

A major characteristic, or difficulty, of text classification problems is the high dimensionality of the feature space. The native feature space consists of the unique terms that occur in documents, which can be tens or hundreds of thousands of terms for even a moderate-sized text collection. This is prohibitively high for many learning algorithms. It is highly desirable to reduce the native space without sacrificing classification accuracy.

Feature selection for text classification is a well-studied problem; its goals are improving classification effectiveness, computational efficiency, or both. Aggressive reduction of the feature space has been repeatedly shown to lead to little accuracy loss, and to a performance gain in many cases.

Feature selection can be supervised with human support in labeling the data, or be unsupervised without any human involvement [2]. In supervised feature selection, a labeled training set is first trained to derive the model, which is then used to predict an unlabelled test set. Unsupervised feature selection does not need a pre-labeled dataset. Instead, heuristics are used for estimating the quality of the features [3].

In [4] a thorough evaluation of the five feature selection methods: Document Frequency Thresholding, Information Gain, χ2-statistic, Mutual Information, and Term Strength is given. At present, many researchers have studied and reached some new achievements. P. Barman gives a modified Mutual Information for text feature selection and classification [9]. Li-ping jing give a new TFIDF-based feature selection approach [5]. Jihong liu present a feature selection algorithm based on coalitional game, to select a sub-feature set in which the selected features are coalitional and relevant in order to obtain better classification performance [6]. Also rough feature selection for intelligent classifiers is present in [7][8]. Rough set theory offers a useful, and formal, methodology that can be employed to reduce the dimensionality of datasets. Shay Cohen and Eytan Ruppin present the Contribution-Selection algorithm (CSA) for feature selection. The algorithm is based on the Multiperturbation Shapley Analysis, a framework which relies on game theory to estimate usefulness [10]. Wenqian Shang design a novel Gini index algorithm to reduce the high dimensionality of the feature space based on Gini index theory [11].

The rest of the paper is organized as follows. In the next section, we review the key techniques in text classification: feature extraction, dimensionality reduction, methods for text classification. In section 3, we introduce the data sets used in this work, explain the related experimental work and give the experimental results and analysis. In section 4, we give the conclusion.

II. KEY TECHNIQUES IN TEXT CLASSIFICATION

A typical text classification process consists of the following steps: preprocessing, indexing, dimensionality

978-1-4244-8217-7/10/$26.00 ©2010 IEEE

reduction, and classification. In this work different approaches for all these steps have been discussed.

A. Preprocessing

The first step in text classification is to transform documents, which typically are string of characters, into a representation suitable for the learning algorithm and the classification task. The text transformation usually is of the following kind [12]:

(a) Remove HTML (or other) tags, (b) remove stopwords, (c) Perform word stemming. The stopwords are frequent words that carry no information (i.e. pronouns, preposition conjunctions etc.).

By word stemming we mean the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as walk, walker, walked, and walking. The Porter stemmer is a well-known algorithm for this task.

B. Indexing

The perhaps most commonly used document representation is the so called vector space model (VSM). In the vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by a word-by-document matrix A, where each entry represents the occurrences of a word in a document, i.e.

$$A = (a_{ik}),$$

where a is the weight of word i in document k . Since every word does not normally appear in each document, the matrix A is usually sparse. The number of rows, M, of the matrix corresponds to the number of words in the dictionary. M can be very large. Hence, a major characteristic, or difficulty of text classification problems is the high dimensionality of the feature space. In Section II C we discuss different approaches for dimensionality reduction.

There are several ways of determining the weight a of word i in document k , such as Boolean Weighting, Word frequency weighting, Word frequency weighting, tf×idf weighting, tfc-weighting, ltc-weighting, Entropy weighting [13]. but most of the approaches are based on two empirical observations regarding text:

(a) The more times a word occurs in a document, the more relevant it is to the topic of the document. (b) The more times the word occurs throughout all documents in the collection, the more poorly it discriminates between documents.

C. Feature selection methods

A major characteristic, or difficulty, of text classification problems is the high dimensionality of the feature space. Therefore feature selection or feature extraction is the primitive task and key step for text classification [14] .

Feature selection can reduce the dimensionality of feature space, decrease the computing complexity and improve the accuracy rate of classification. A number of Feature selection methods has been applied to text classification, including Document Frequency (short for DF), Information Gain (short for IG), Mutual Information (short for MI) , $\chi 2$ statistics method (short for CHI), Cross Entropy and Primary Component Analysis (short for PCA) ,etc.

1) Document Frequency Thresholding (DF)

The document frequency for a word is the number of documents in which the word occurs. In Document Frequency Thresholding one computes the document frequency for each word in the training corpus and removes those words whose document frequency is less than some predetermined threshold. The basic assumption is that rare words are either non-informative for category prediction, or not influential in global performance.

2) Information gain (IG)

Information Gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a word in at document.

Let $c_1, \mathrm{L}\ , c_K$ denote the set of possible categories. The information gain of a word w is defined to be:

$$\begin{aligned} IG(w) &= -\sum_{j=1}^{K} P\,(c_j)\log P(c_j) \\ &+ P(w)\sum_{j=1}^{K} P(c_j|w)\log P(c_j|w) \\ &+ P(\overline{w})\sum_{j=1}^{K} P(c_j|\overline{w})\log P(c_j|\overline{w}) \end{aligned} \tag{1}$$

Here $P(c_j)$ can be estimated from the fraction of documents in the total collection that belongs to class c_j and $P(w)$ from the fraction of documents in which the word w occurs. Moreover, $P(c_j|w)$ can be computed as the fraction of documents from class c_j that have at least one occurrence of word w and $P(c_j|w)$ as the fraction of documents from class c_j that does not contain word w.

The information gain is computed for each word of the training set, and the words whose information gain is less than some predetermined threshold are removed.

3) Mutual information (MI)

Mutual information is a criterion commonly used in statistical language modeling of word associations and related applications. If one considers the two-way contingency table of a term t and a category c , where A is the number of times t and c co-occur, B is the number of time the t occurs without c ,C is number of times c occurs without t ,and N is the total number of documents, then the mutual information criterion between t and c is defined to be:

$$I(t,c) = \log \frac{P_r(t \wedge c)}{P_r(t)\square P_r(c)} \tag{2}$$

And is estimated using

$$I(t,c) \approx \log \frac{A \times N}{(A+C)\square A+B} \tag{3}$$

$I(t,c)$ has a natural value of zero if t and c are independent. To measure the goodness of a term in a global feature selection, we combine the category-specific scores of a term into two alternate ways:

$$I_{avg}(t)=\sum_{i=1}^{m}P_r(c_i)I(t,c_i)$$

$$I_{\max}(t)=\max_{i=1}^{m}\{I(t,c_i)\}$$

(4)

The MI computation has a time complexity of $O(v_m)$, similar to the IG computation.

A weakness of mutual information is that the score is strongly influenced by the marginal probabilities of terms, as can be seen in this equivalent form:

$$I(t,c)=\log P_r(t|c)-\log P_r(t)$$

(5)

For terms with an equal conditional probability $P_r(t|c)$, rare terms will have a higher score than common terms. The scores, therefore, are not comparable across terms of widely differing frequency.

D. methods for text classification

A number of statistical classification and machine learning techniques has been applied to text classification, including regression models, K-Nearest Neighbor classifiers, Decision Tress, Bayesian classifiers, Support Vector Machines and Neural Networks.

In what follows we will describe Support Vector Machines and K-Nearest Neighbor which method will be used in the section 3.

1) Support Vector Machines

Support Vector Machines (SVMs) have shown to yield good generalization performance on a wide variety of classification problems, most recently text classification. The SVM integrates dimension reduction and classification. It is only applicable for binary classification tasks, meaning that, using this method text classification have to be treated as a series of dichotomous classification problems [15].

The SVM classifies a vector d to either -1 or 1 using

$$s=\sum_{i=1}^{N}\alpha_i y_i K(d,d_i)+b$$

(6)

Here $\{d_i\}_{i=1}^{N}$ is the set of training vectors as before and $\{y_i\}_{i=1}^{N}$ are the corresponding classes $(y_i \in -1,1)$. $K(d_i,d_j)$ is denoted a kernel and is often chosen as a polynom of degree d, i.e.

$$K(d,d_i)=(d^T d_i+1)^d$$

(7)

2) K-Nearest Neighbor

To classify an unknown document vector d, the K-Nearest Neighbor (KNN) algorithms ranks the document's neighbors among the training document vectors, and use the class labels of the k most similar neighbors to predict the class of the input document. The classes of these neighbors are weighted using the similarity of each neighbor to d, where similarity may be measured by for example the Euclidean distance or the cosine between the two document vectors.

KNN is a lazy learning instance-based method that does not have an off-line training phase. The main computation is the on-line scoring of training documents given a test document in order to find the k nearest neighbors. Using an inverted-file indexing of training documents, the time complexity is $O(L * N/M)$ where L is the number of elements of the document vector that are greater than zero, M is the length of the document vector, and is the number of training samples.

III. Experimental results & Analysis

A. Data sets

In our work each dataset is divided into training set and testing set. We use three datasets in our experiments, which are shown in Table 1.

Data set 3S and 3D are created from the famous Reuters-21578 text classification collection [16].We choose three similar classes from it, which are 'trade', 'money-fx' and 'earn' to create data set 3S, and then three different classes, 'earn', 'ship' and 'wheat', are used to create data set 3D. The features are the terms in the documents in these classes. The data are the term-frequency matrices, which are created based on the stemming algorithm in TMG toolbox [17]. The local and global term-frequency threshold is 2.

Table 1. Data sets

Data Set		3Similar Data	3 Different Data
Feature		1109	1161
Sample	total	735	735
	train	420	420
	test	315	315
Class		3	3

B. Method used

First of all, we do feature selection to realize the dimension reduction in our experiments. Then we classify the data sets in Table 1 with different classifiers.

Feature selection is performed here using Mutual Information (short for MI) algorithm which is the traditional and typical feature selection method.

The K-Nearest Neighbor (KNN) classifier and Support Vector Machines (SVM) are used in the classification work. The KNN method is a very simple approach and it also be seen that KNN showed good performance on text classification tasks in the various previous works.

C. Result & Analysis

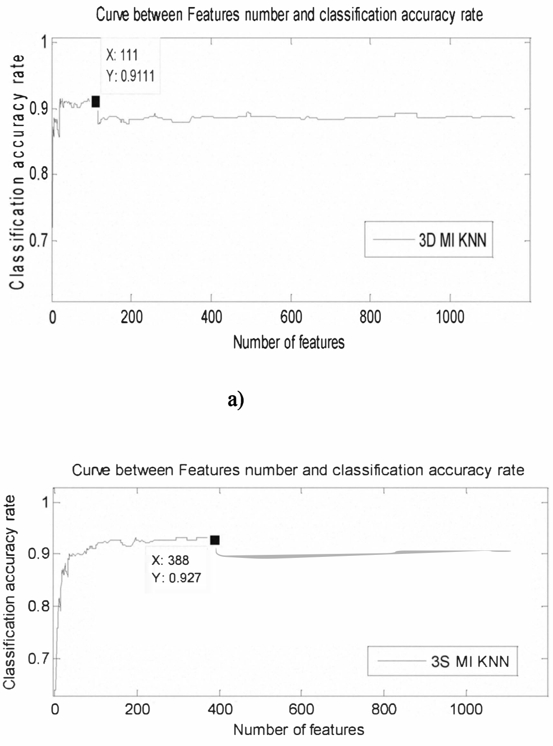
The experimental results are shown in Table 2 and Figure 1. It is shown that the number of native features are 1161, using MI when the number of features is 101, the Classification accuracy rates is 91.1% , obtains better performance than no feature selection, and also improve the classification rate. The SVM classifier gains the higher performance compared to KNN classifier in Figure 2.

TABLE2. Classification accuracy rates

Dataset	No Feature Selection KNN (%)	MI KNN (%)
3S	90.5 (1109)	93.0 (359)
3D	88.6 (1161)	91.4 (101)

IV. CONCLUSION

The aim of text classification is to build systems which are able to automatically classify documents into categories. In this paper we review the key text classification techniques including text model, feature selection methods and text classification algorithms in building a text classification system. Also we give an implementation of a text classification system based on Mutual Information and K-Nearest Neighbor algorithm and Support Vector Machine. Our experimental results show that Mutual Information is a kind of efficient dimension reduction methods for text data sets with huge features.



a)

b)

Figure 1. Classification rates curve. a) MI and KNN for 3D data set ,b) MI and KNN for 3S data set

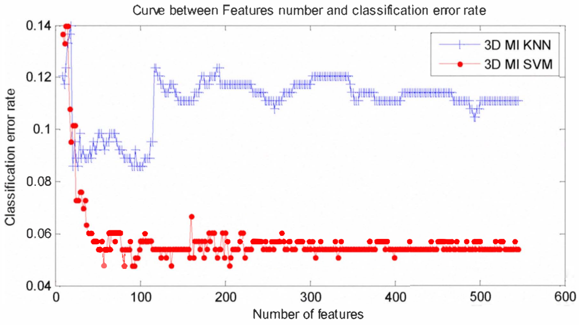


Figure 2. Classification rates curves with KNN and SVM classifier

REFERENCES

[1] F.Sebastiani.Machine Learning in Automated Text Categorization.ACM Computing Survey, 2002, 34(1):1 — 47..

[2] Tien Dung Do, Siu Cheung Hui and Alvis C.M. Fong, “Associative Feature Selection for Text Mining”, International Journal of Information Technology, Vol. 12 No.4, pp. 59-68, 2006

[3] H. Liu, M. Motoda, L. Yu, “Feature Extraction, Selection, and Construction”. In N. Ye (eds.):The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc. Publishers, pp. 409-423, 2003.

[4] Y.Yang and J.P.Pedersen, Feature selection in statistical learning of text categorization, In the 14th Int.Conf.on Machine Learning, pp.412~420, 1997.

[5] li-ping jing, improved feature selection approach tfidf in text minging.Proceedings of the first International Conference on machine learning and Cybernetics, Beijing, 4-5 November 2002

[6] Jihong Liu, Soo-Young Lee, "Study on Feature Select Based on Coalitional Game", IEEE Proceeding of 2008 International Conference on Neural Networks & Signal Processing, Zhenjiang, China, June 7-11, 2008

[7] Qiang Shen. Rough Feature Selection for intelligent Classifiers. T. Roughset 7, 2007, PP 244~255

[8] Zhang, Li-Juan and L.i. Zhou-jun. A Novel Rough Set Approach for Classification. IEEE GrC, 2006

[9] Paresh Chandra Barman, Soo-Young Lee, "Modified Mutual Information for Text Feature Selection and Classification", 9th China-India-Japan-Korea Joint Workshop on Neurobiology and Neuroinformatics, Jeju, Korea, July 5-7, 2007

[10] Cohen, S., Ruppin, E., Dror, G., “Feature selection based on the Shapley value”, Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 30 July - 5 August 2005, pp. 665-670, 2005.

[11] Wenqian Shang, Houkuan Huang, etc . A novel feature selection algorithm for text categorization. Expert Systems with Applications 33 (2007) 1–5

[12] K.Aas,L.Eikvil.Text Categorization:A Survey. Technical Report#941, Norwegian.

[13] G.Salton,M.E.Lesk.Computer Evaluation of Indexing and Text Processing.Journal of the ACM,1968,15(1):8 — 36.

[14] YimingYang,Jan O.Pedersen.A Comparative Study on Feature Selection in Text Categorization (1997).

[15] Thorsten Joachims.Text Categorization with Support Vector,2002.

[16] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html, 2007

[17] D. Zeimpekis and E. Gallopoulos, “TMG: A MATLAB toolbox for generating term-document matrices from text collections”. Grouping Multidimensional Data: Recent Advances in Clustering, J. Kogan, C. Nicholas and M. Teboulle, eds., pp.187-210, Springer, 2006.