An Evident Theoretic Feature Selection Approach for Text Categorization

UMAR SATHIC ALI
Research Scholar, Research and Development Centre
Bharathiar University
Coimbatore, India

JOTHI VENKATESWARAN
Associate Professor and Head
Department of Computer Science, Presidency College
Chennai, India

Abstract-With the exponential growth of textual documents available in unstructured form on the Internet, feature selection approaches are increasingly significant for the preprocessing of textual documents for automatic text categorization. Feature selection, which focuses on identifying relevant and informative features, can help reduce the computational cost of processing voluminous amounts of data as well as increase the effectiveness for the subsequent text categorization tasks. In this paper, we propose a new evident theoretic feature selection approach for text categorization based on transferable belief model (TBM). An evaluation on the performance of the proposed evident theoretic feature selection approach on benchmark dataset is also presented. We empirically show the effectiveness of our approach in outperforming the traditional feature selection methods using two standard benchmark datasets.

Keywords: Feature selection, Transferable Belief Model, Text Categorization.

I. INTRODUCTION

Text mining applications always need to deal with large and complex datasets of textual documents that contain much irrelevant and noisy information. One of the problems in text classification is high dimensionality of the feature space. Some features are commonly used terms, not specific to any category. These features may hurt the accuracy of the classifier. Moreover, the time required for induction increases as the number of features increases. That is, irrelevant features lead to an increase in training time. Feature selection aims to remove that irrelevant and noisy information by focusing only on relevant and informative data for use in text categorization [4]. Feature selection methods are used to achieve two objectives: to reduce the size of the feature set to optimize the classification efficiency; and to reduce noise found in the data to optimize the classification effectiveness [5]. Feature selection methods are used as a preprocessing step in the learning process. The selected features from the training set are then used to classify new incoming documents. Among the well-known feature selection methods are document frequency, information gain, expected cross entropy, the weight of evidence of text, odds ratio, term frequency, mutual information and chi-square statistic.

In this paper, we propose an evident theoretic feature selection approach for text classification. The approach is based on an assumption that evidence derived from existing feature selection metrics are combined together to support the partial evidence represented by them. Highly relevant features are selected using Transferable belief model [11] which can truthfully reflect the relations of the evidence that underlies in the existing feature selection metrics. An experiment has been conducted to evaluate the performance of the proposed evident theoretic feature selection approach.

II. RELATED WORK

Feature selection, an important step in text categorization, consists of two steps – preprocessing and classifier building. Preprocessing includes tasks such as feature extraction, feature selection, and document representation. After preprocessing, a document is represented as a vector of features in a vector space model [4] or "bag-of-words" in a probabilistic model; features are the components in a vector or "words". So, feature selection plays a very important role in later steps influencing overall system performance.

The two most common approaches to this problem in machine learning or data mining are the filter and the wrapper [2]. In the wrapper approach, the subset of features is chosen based on the accuracy of classifiers. Exhaustively trying all the subsets is not computationally feasible [6]. Technically, the wrapper is relatively

difficult to implement, especially with a large amount of data. The filtering is usually chosen because it is easily understood and has independent classifiers. The filter, as its name implies, chooses a subset of features by filtering based on scores assigned by specific weighting in text categorization, the filter is used often and features are selected by one of the following criteria.

A. Odds Ratio

The basic idea of using odds ratio[8] is to calculate the odds of a term occurring in the positive class (the category a term is related to) normalized by the odds of that term occurring in the negative class (the category a term is not related to). The odds ratio of a term t_k for a category c_i is defined using Equation 1:

$$OR(t_k, c_i) = \frac{P(t_k/c_i)[1 - P(t_k/\bar{c}_i)]}{[1 - P(t_k/c_i)]P(t_k/\bar{c}_i)}. \quad (1)$$

Odds ratio is known to work well with the naïve Bayes text-classifier algorithm.

B. Information Gain

Information gain [9] is commonly used as a surrogate for approximating a conditional distribution for text classification. In information gain, class membership and the presence/absence of a particular term in a given category are seen as random variables; one computes how much information about the class membership is gained by knowing the presence/absence statistics. If the class membership is interpreted as a random variable *c* with two values, positive (*c*) and negative ($\bar{c}$), and a word is likewise seen as a random variable *T* with two values, present (*t*) and absent ($\bar{t}$), then information gain is defined as Equation 2:

$$IG(t_k, c_i) = \sum_{c\in[c_i,\bar{c}_i]} \sum_{t\in[t_k,\bar{t}_k]} P(t/c)\, log\, \frac{P(t/c)}{P(t)P(C)}. \quad (2)$$

C. Chi-Squared

The $\chi 2$ test is used in statistics to test the independence between two events. In text classification, $\chi 2$ [13] is used to measure the association between a category and features. The $\chi 2$ measure of a term t_k for a category c_i is defined using Equation 3:

$$CHI(t_k, c_i) = \frac{P(t_k/c_i)P(\bar{t}_k/\bar{c}_i) - P(t_k/\bar{c}_i)P(\bar{t}_k/c_i)}{\sqrt{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}}. \quad (3)$$

Thus, the $\chi 2$ (t_k, c_i) score indicates the weight of term t_k with respect to category c_i. If a term is close to more categories, then the score of that term is higher. The score of each term t_k is calculated using Equation 3:

D. tfidf

In *tfidf* [1], *tf* refers to term frequency of a term in a given document. *idf* is defined as the inverse document frequency, i.e., the ratio of the total number of documents present in a dataset to the number of documents a given term appears in. A higher *idf* of a term indicates that the term appears in relatively few documents and may be more important during the process of text classification. *Tfidf* is a commonly used technique for term weighing in the field of information retrieval and is also used in text classification. The *tfidf* of a term t_k in document d_i is defined using Equation 4:

$$tfidf(t_k, d_i) = tf(t_k, d_i)\, log\, \frac{|D|}{df(t_k)}. \quad (4)$$

where |*D*| refers to the total number of documents in a dataset; *tf* (t_k,d_i) is the term frequency of a term t_k in document d_i; and *df* (t_k) refers to the number of documents in which term t_k appears.

III. TRANSFERABLE BELIEF MODEL (TBM)

TBM is a model developed to represent the quantified belief[11]. it describes two level mental models in order to distinguish between two aspects of belief, belief as weighted opinion and belief for decision making. Two levels are credel level, where beliefs are entertained and pignistic level, where beliefs are used to make decision. At the creedal level, beliefs are represented by a belief function; at the pignistic level, this belief function induces a probability function that is used to make decision.

TBM starts with the definition of all possible values, Ω, called a frame of discernment that a variable can take. An exact belief value is assigned to each subset of Ω and this represents the uncertainty that the value of the variable belongs to the set. For $A \subseteq \Omega$, *m*(*A*) is that part of Your belief that supports *A*, i.e., that the actual value belongs to *A*, and that, due to lack of information, does not support any strict subset of *A*. The degree of

belief *bel*(*A*) quantifies the total amount of justified specific support given to *A*. It is obtained by summing all basic belief masses given to subsets $X \subseteq \Omega$ with $X \subseteq A$ *as in Equation 5*. Indeed a part of belief that supports that the actual value is in *B* also supports that value is in *A* whenever $B \subseteq A$. So for all $A \subseteq \Omega$,

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B). \quad (5)$$

One of the key issues in the applications of Transaferable Belief Model is how to make a decision based on belief functions bel. Since there is no effective method to make decisions based on the belief directly yet, the conventional way is to transform belief to probability, which can be easily used in decision-making. Smets [11] proposed the Pignistic Probability Transformation (PPT), in which, picking a particular element of the set of compatibility measures BetP, known as the pignistic probability, and calculating the expected utility with respect to this. The pignistic transformation from m to a probability distribution BetP can be justified axiomatically: The pignistic probability measure BetP associated with mass function m is defined by

$$for\ A \subseteq \Theta,\ BetP(A) = \sum_{B \subseteq \Theta} m(B) \frac{|B \cap A|}{|B|}. \quad (6)$$

BetP is the result of distributing each mass m(B) equally over the elements of B.

IV. FEATURE SELECTION BASED ON TBM

Feature selection in text categorization is stated as follows: Given set *X* consisting of *n* features $x_1, x_2, x_3, \ldots\ldots x_n$, the problem in feature selection is to choose optimal subset *Y* of *X* (*Y*<<*X*) with highest effectiveness for the system. To solve this problem, we filter features based on the criteria we discussed in section 2. For each feature, we compute the term score according to a criterion. Thus we have *t* ways of representing documents with t criteria.

Initially, we describe the intuitive motivation behind our approach and then provide a formal definition of our method. In our approach, the existing feature selection criteria are considered as source of evidence. Each feature selection criteria such as information gain, odd ratio, chi-square predicts a set of features which are considered as an independence item of evidence E_i known as neighborhood. Thus each neighborhood consist of few hundred of features and these neighborhood may overlap such that some features may fall in all neighborhood thus playing an important role in representing the document. We combine these neighborhood of evidences to induce a mass function representing partial support by different neighborhood. While combining evidences, we take into account individual feature weight (tf-idf), only the feature which has substantial relevancy in terms of weight is considered. For this purpose, we split the term weight *R* into number of intervals $R_1, R_2, \ldots R_q$to which a term weight may belong to. Feature selection problem is stated mathematically as follows:

We take the frame of discernment to be Ω that is the collection of all possible informative words derived from the training set. Let $t \in \Omega$ be the informative for which we seek evidences from the existing metrics. For the sake of simplicity, we consider three neighborhoods $E_1, E_2\ and\ E_3$, to represent the set of features selected by IG, CHI and OR respectively.

Consider $E_i \in 2^{\Omega}$ and feature strength (tf-idf) $r \in R_j$. We are interested in the joint probability $P(E_i,r)$ –the probability that a randomly selected element *x* of Ω belongs to E_i and its feature strength falls in the interval *r*, i.e., $x \in E_i$ and $f(x) = r$. since the knowledge about the distribution *p* is unknown, we approximate $P(E_i,r)$ by applying the *principle of indifference* [12]

$$\bar{P}(E_i, r) = |E_i^r|/|\Omega|. \quad (7)$$

where $E_i^r = \{x \in E_i : f(x) = r\}$

Then we induce a mass function *m*[*t*], for *t* from the *h* neighborhoods, as a mapping $m[t]: 2^{\Omega} \rightarrow [0,1]$ such that, for $X \in 2^{\Omega}$ and $r \in R$,

$$m[t](X, r) = \begin{cases} \frac{\bar{P}(X,r)}{K}, if\ X = E_i\ for\ some\ i \\ 0, \qquad otherwise \end{cases} \quad (8)$$

Here K is a normalizing factor. It follows that $K = \sum_{i=1}^{h} \sum_{r \in R} \bar{P}(E_i, r)$. Note that by $m[t](X,r)$ we mean $m[t](X \cap \{x \in \Omega: f(x) = r\})$, which is similar to the interpretation of joint probability $P(X,r)$. Clearly $m[t]$ is a mass function. In particular $\sum_{r \in R} \sum_{X \in 2^{\Omega}} m[t](X, r) = \sum_{r \in R} \sum_{i=1}^{h} m[t](E_i, r) = 1$.

A. Decision Making based on Pignistic Probability

We propose to choose a feature through marginal pignistic probability. For this we specify the joint pignistic probability as $\overline{BetP} : 2^{\Omega} \to [0,1]$ such that, for $X \in 2^{\Omega}$ and $r \in R$

$$\overline{BetP}(X, r) = \sum_{i=1}^{h} m[t](E_i, r) \times \frac{|X \cap E_i|}{|E_i|}. \quad (9)$$

Since E_i is a collection of features so we have $t \in E_i$. We can understand t as a singleton set, therefore $t \cap E_i^r = \{t\}\{t\}$ is either 1 or 0 depends on its presence in E_i or not and $|\{t\}| = 1$. Then we have the following joint and marginal pignistic probabilities for $t \in \Omega$,

$$\overline{BetP}(t, r) = \sum_{i=1}^{h} m[t](E_i, r) / |E_i|. \quad (10)$$

$$\overline{BetP}(t) = \sum_{r \in R} \overline{BetP}(t, r). \quad (11)$$

Feature selection then proceeds using the following rule:

$$\arg\max \{\overline{BetP}(t)\}. \quad (12)$$

V. EXPERIMENTAL SETUP

To evaluate the effectiveness of the evident theoretic feature-selection method, we choose two benchmark data sets (Reuters-21578, WebKB), which are commonly used in text-classification evaluation. The details on these data sets are given in Table 3. We intentionally choose these datasets, which consist of news articles and web pages, to show the effects of our proposed feature selection method on different domains. With regard to text classification algorithm, we choose SVM and kNN text classifiers. SVM is commonly used, as it was shown to perform better in terms of effectiveness than other text classifiers such as naïve Bayes, kNN, C4.5, and Rocchio [0]. The kNN algorithm is,however, simple and more efficient than other algorithms [12].

A brief explanation about the benchmark datasets that are used in our experiments is given below.

A. Reuters 21578

The Reuters 21578 corpus [8] contains Reuters news articles from 1987. The documents range from being multi-labeled, single labeled, or not labeled. Reuters dataset consists of a total number of 135 categories (labels). However, ten of these categories have significantly more documents than the rest of the categories. Thus, commonly the top 10 categories are used for experimentations and to compare the accuracy of the classification results. The top 10 categories of Reuters 21578 are “earn”, “acq”, “money-fx”, “grain”, “trade”, “crude”, “interest”, “wheat”, “corn” and “ship”. We use Mod-Apte train-test split for Reuters 21578 dataset. There are 7,053 documents in training set and 2,726 documents in testing set. The total number of unique features in Reuters 21578 dataset is 19,249.

B. WebKD Dataset

The WebKB dataset is a collection of Web pages from four different college Web sites, namely Cornell, Texas, Washington, Wisconsin, and some miscellaneous Web pages. These Web pages are pre-classified into seven categories: student, faculty, staff, department, course, project, and other. WebKB contains 8,282 Web pages. The average document length in WebKB dataset is 130 terms.

C. Evaluation Measures

To evaluate the effectiveness of our approach and compare to the state of the art feature-selection research results, we use the commonly used evaluation metrics precision, recall, and F1 measure.

$$Precision\ (P) = \frac{true\ positive}{true\ positive + false\ positive}. \quad (13)$$

Precision (Equation 13) is defined as the ratio of correct classification of documents into categories to the total number of attempted classifications.

$$Recall\ (R) = \frac{true\ positive}{true\ positive + false\ negative}. \qquad (14)$$

Recall (Equation 14) is defined as the ratio of correct classifications of documents into categories to the total number of labeled data in the testing set. F1 measure (Equation 15) is defined as the harmonic mean of precision and recall.

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}. \qquad (15)$$

Hence, a good classifier is assumed to have a high F1 measure, which indicates that classifier performs well with respect to both precision and recall. We present the microaveraged results for precision, recall, and F1 measure. Microaveraging considers the sum of all the true positives, false positives, and false negatives.

VI. RESULTS & ANALYSIS

The purpose of each experiment is to select a list of terms with the corresponding scores by applying our proposed feature selection method. We simply sort the list of terms based on the scores and obtain the k relevant terms with the highest scores. To evaluate the goodness of each experiment, the k relevant terms are selected, then we evaluated these terms on text categorization task using kNN and SVM. We have conducted the experiment with a wide range of k for each classifier. The range of k has been set from 50 to 1000. We investigated the effectiveness our approach on various aspects such as precision, recall, and F1 measure. The results are shown in Figure 1 Figure 2 and Figure 3 and Figure 4. These figures shows the micro-averaged f1 measure for each of the features selection metrics with varying number of features

We presented the classification results for *SVM and kNN* algorithm using our proposed methods along with others on Reuters 21578 and WebKB datasets. The experimental results have shown that the evidence theoretic feature selection approach is better than the approach using IG , CHI and Odd Ratio when precision is strongly preferred over recall [5]. It can be explained that IG while focusing only on positive features, may miss some important negative features. The success of Evident theoretic feature method lies in combining evidence by selecting both positive and negative features considerably from these existing methods. However, for scalability reasons, one is limited to 50-100 features, the best metric that outperform others is IG. The study also reveals that if the precision is central goal, proposed method beat other methods by a smaller but significant margin.

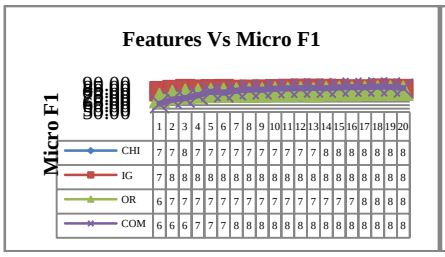

Figure 1: Compariosion of COM with others interms of Micro F1 on Reuters for voting k-NN

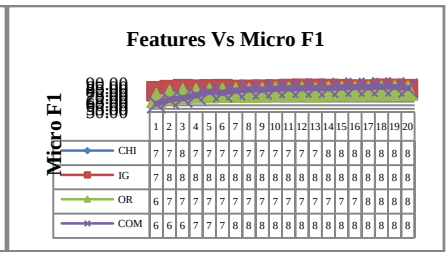

Figure 2: Compariosion of COM with others interms of Micro F1 on Reuters for SVM

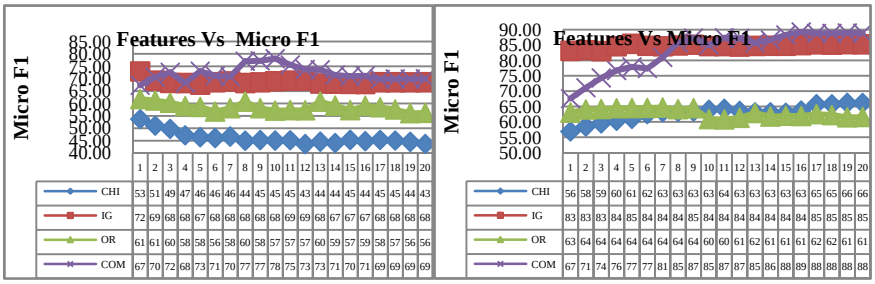

Figure 3: Compariosion of COM with others interms of Micro F1 on WebKB for voting k-NN

Figure 4: Compariosion of COM with others interms of Micro F1 on WebKB for SVM

VII. CONCLUSION

We explored an effective evident theoretic feature selection method; and we applied this on *SVM and kNN* text classification. With an ever-increasing number of digital documents, many traditional feature selection techniques fail to capture the potential feature present in the corpora due to their inherent mechanism. In this paper, we have shown that evidence theoretic feature selection method can capture the relevant and potential features without hurting the effectiveness of the classifier. We performed experiments on two standard benchmark datasets, Reuters 21578 and WebKB. We showed that our proposed method significantly better than the traditional feature selection algorithms on *SVM and kNN*. Furthermore, we provided analysis of how the F1 measure gradually increases as we add more features in the experiment.

REFERENCES

[1] Chih, H.,&Kulathuramaiyer,N. (2004).An empirical study of feature selection for text Categorization based on term weightage. In Proceedings of the 2004 IEEE/WIC/ACM International Conference onWeb Intelligence (pp. 599–602).Washington, DC: IEEE.

[2] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In International Conference on Machine Learning, 2001.

[3] Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. on Systems, Man and Cybernetics, 25(5), 804–813 (1995)

[4] Fabrizo Sebastiani, machine learning in automated text categorization ACM computing surveys, Vol.34,No 1,March 2002, pp.1-47.,2002.

[5] Forman, G. (2003). An extensive empirical study of feature-selection metrics for text classification. Journal of Machine Learning Research, 3, 1289–1305.

[6] D. Koller and M. Sahami. Toward optimal feature selection. In International Conference on Machine Learning, pages 284–292, 1996.

[7] D. Mladenic, Feature subset selection in text-learning. In Proc. of European Conference on Machine Learning (1998) 95-100.

[8] Mladeni´c, D., & Grobelnik, M. (1998, June). Feature selection for classification based on text hierarchy. Text and the Web. Paper presented at the Conference on Automated Learning and Discovery (CONALD-98), Pittsburgh, PA.

[9] Quinlan, J. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

[10] Shafer, G. (1976a). A mathematical theory of evidence. Princeton Univ. Press. Princeton, NJ.

[11] Smets, P. (1998b). The transferable belief model for quantified belief representation. In D. M. Gabbay & P. Smets (Eds.), Handbook of defeasible reasoning and uncertainty management systems (Vol. 1, pp. 267–301). Kluwer, Doordrecht, The Netherlands.

[12] Wang H., David Bell: Extended k-Nearest Neigbours based on Evidence Theory, The Computer Journal, Vol 47, pp 662-672(2004)

[13] Yang,Y. (1999).An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1–2), 69–90.

[14] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization. In Proc. of the 14th International Conference on Machine Learning (ICML-97), Morgan Kaufmann Publishers, San Francisco, US (1997) 412-420.

[15] Yang,Y., Zhang, J.,&Kisiel, B. (2003).A scalability analysis of classifiers in text categorization. In Proceedings of the 26thACMSIGIR Conference on Research and Development in Information Retrieval (pp. 96–103). New York: ACM.

AUTHORS PROFILE

Dr. C. Jothi Venkateshwaran received his Ph.D in Computer Science from Alagappa University, Karaikudi in 2006. He is currently an Associate Professor and Head, at the Department of Computer Science, Presidency College, Chennai, India.

P. Umar Sathic Ali is pursuing his doctoral research in Computer Science at Bharathiar University. His research interest includes Data mining and Neural Networks, Machine Learning, Text classification and feature selection.