Research Journal of Applied Sciences, Engineering and Technology 7(5): 981-985, 2014
DOI:10.19026/rjaset.7.347
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 31, 2013 Accepted: March 02, 2013 Published: February 05, 2014

Research Article

A Dempster-Shafer Model for Feature Selection in Text Categorization

[1]P. Umar Sathic Ali and [2]C. Jothi Venkateswaran
[1]MEASI Institute of Information Technology 87, Peters Road, Royapettah Chennai-600 014, Tamilnadu, India
[2]Department of Computer Science, Presidency College, Chennai-600 005, Tamilnadu, India

Abstract: In this study, we propose a feature selection method based on evident theoretic model for text categorization. The proposed model is formally expressed within the Dempster-Shafer Theory of Evidence. We discuss the way the theory is used to retrieve highly informative and relevant features from the document collection. The formal retrieval function is inferred from the said model and compared our proposed model with many of the conventional feature selection methods. Experimental evaluation on standard benchmark dataset has shown the effectiveness of the proposed method.

Keywords: Dempter-shafer theory, feature selection, text categorization

INTRODUCTION

Feature selection method focus on the problem of retrieving relevant features from document collection in order to represent the document for categorization (Sebastiani, 2002). In this study, we concentrate on developing a novel feature selection method for text-based categorization systems. Though traditional feature selection methods retrieve features from document collection to some extend but they are not capable of retrieving all possible potential features. Hence, they could not improve the classifier effectiveness.

The combination of traditional feature selection techniques used in TC attempts to overcome such shortcomings (Del Castillo and Serrano, 2004; Doan and Horiguchi, 2004). These combination techniques have been proven successful in improving the performance of the classifier substantially. They aim to extract possible potential features which are then used to represent documents, where features can take on various linguistic forms.

In this study, we use some of the most widely used feature selection techniques as source of evidences from which our proposed model retrieves highly relevant features. Our model is constructed based Dempster-Shafer (D-S) Theory of Evidence (Shafer, 1976). This is a mathematical theory of evidence which deals with uncertainty associated with available evidence (a set of hypothesis and their associated beliefs). The evidences here are the set of features generated by the conventional feature selection methods.

DEMPSTER-SHAFER'S THEORY OF EVIDENCE

Dempster Shafer theory also known as theory of evidence is a flexible framework for representing and reasoning with imprecise and uncertain data (Wang and David, 2004). We first describe some important measures which are ought to be used in our proposed model. Let Ω be a finite non-empty set of mutually exhaustive and exclusive events. The set Ω is called a frame of discernment. Let 2^{Ω} be the set of all subsets of the set Ω, including the empty set $\emptyset$; and Ω itself. Given a frame of discernment Ω, the function m: $2^{\Omega} \rightarrow [0, 1]$ is called a basic probability assignment (*bpa*) if it satisfies the following:

$$m(\phi) = 0 \quad and \quad \sum_{A \in 2^{\Omega}} m(A) = 1 \tag{1}$$

The *bpa* represents a source of evidence supporting various subsets A in 2^{Ω} with value, or "degree of support", $m(A)$. The subsets A of 2^{Ω} such that $m(A)>0$ are called focal elements. Given a bpa m: $2^{\Omega} \rightarrow [0, 1]$, a function Bel: $2^{\Omega} \rightarrow [0, 1]$, is called a belief function over Ω, is defined as:

$$Bel(A) = \sum_{B \subset A} m(B) \tag{2}$$

The measure $Bel(A)$ quantifies the strength of the total belief given to set A alone; but not any of its subsets. In contrast, $m(A)$ quantifies the exact belief committed to A. Unlike a probability theory, a salient

Corresponding Author: P. Umar Sathic Ali, MEASI Institute of Information Technology 87, Peters Road, Royapettah Chennai-600 014, Tamilnadu, India

characteristic of the evidence theory is that the belief in particular hypothesis does not necessarily imply that the remaining belief is associated to the negation of the hypothesis. Hence when there is no further evidence available regarding belief in negation of the hypothesis, the remaining belief is assigned to the entire frame of discernment (all the possible hypothesis), that represents the uncommitted belief or total ignorance (Smets and Kennes, 1994).

DESCRIPTION OF THE MODEL

This section looks into basic intuition on which our proposed model is built. The indexing features which act as a basic building blocks of text representation is described first followed by illustrating how document collection is represented as a frame of discernment. Then we describe the method by which features are represented within the frame. Feature retrieval rule is derived at last.

Indexing features: In order to process the document by the classifier, every document has to be converted into meaningful representation of its content (Sebastiani, 2002). Here the conversion of documents is obtained using the standard information retrieval technique known as bag-of-words approach, in which every document is represented as a group of words retrieved from that document.

We construct this representation based on word content only. Therefore, our approach ignores the word ordering and also ignores the concept of syntactic phrases in documents, thus treating every word equally. The purpose of our study is to achieve improvement in classifier effectiveness by extracting highly informative linguistic structures, and also to use them to construct more meaningful representation of document.

Frame of discernment: To have an insight into this model we define some terminologies regarding document collection and its associated features.

Definition 1: Let $C = \{D_1, D_2, \ldots, D_N\}$ be a document collection, where *N* is the number of documents. Let $S = \{s_1, s_2, \ldots, s_M\}$ be the resulting set of terms after doing all pre-processing tasks in the given document collection *C*, where *M* is the total number of single terms in the document collection.

Given a document collection *C*, we take the frame of discernment as the set *S* itself. Then elements of the frame are defined as mutually exclusive hypothesis derived from a power set of *S*.

Definition 2: For the set of single terms $S = \{s_1, s_2, \ldots, s_M\}$ of a document collection *C*, all the 2^S subset of *S* can easily be obtained using the terms $s \in S$.These subsets represent the elementary hypothesis of the constructed frame. It can be shown that the number of constructed elementary hypothesis is 2^S.

Table 1: Sample of elementary hypothesis of the frame *S*

e_0	Ø
e_1	{ Stake}
e_2	{ Merger }
e_3	{ Profit }
e_4	{ Acquire, Loss }
e_5	{ Loss, Stake}
e_6	{ Acquire, Stake}
e_7	{ Acquire, Stake, Loss }

Example 1: Let *S*= {Acquire, Loss, Stake, Merger, Share, Profit} and s_1 = Acquire, s_2 = Loss, s_3 = Stake, s_4 = Merger, s_5 = Share, and s_6 = Profit. We obtain 2^6 = 32 elementary hypothesis forming the frame of discernment. Some of the elementary hypotheses are shown in the Table 1:

Feature group representation: To retrieve highly relevant feature from the document collection, we selectively model each of the subset of the frame of discernment as feature groups. Each such feature group may consists of one or more features and combination of such feature group may resemblance the set of features generated by the conventional feature selection metrics (Rogati and Yang, 2002) such as information gain, chi-square or odd-ratio. This kind of resemblance to the existing feature selection metric is modeled as evidence through which our model select highly relevant features.

Focal and informative elements: In the D-S theory of evidence, an element with its associated positive evidence is considered as focal elements. Hence, set of focal elements can be grouped together as feature groups modeling the informative representation of a document collection. Given a document $D_i \in C$, these focal elements are defined upon the set S_i.

Definition 3: Every subset S_i of *S* of a document collection *C* defines a focal element. e.g., the hypothesis h_j. Furthermore, every super group $S_g \supseteq S_i$ also defines a focal element, the hypothesis $h_k = \cup h_l$, where each h_l is the hypothesis associated to single subset S_i of *S*. Θ_i is defined as the set that includes all the feature groups representing subset and super group of the document D_i.

Example 2: Let D_1 be the document with S_1 = {Acquire, Loss, Stake, Merger, Share, Profit}. The following feature groups contain the partial subset of features of S_1 and these feature may belongs to the set of features generated by the conventional feature selection methods such as Information Gain (IG), Chi-Square (CHI) or Odd-Ratio (OR):

f_{IG}	{ Acquire, Loss }
f_{CHI}	{ Stake, Merger }
f_{OR}	{ Share, Profit }
f_{IG+OR}	{ Acquire, Loss, Profit }

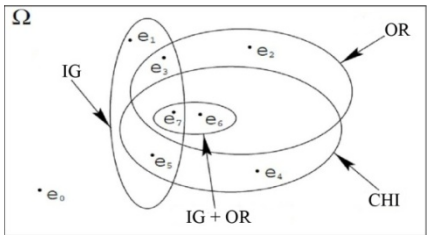

Fig. 1: An example of overlapped elementary hypothesis in a frame of discernment

The feature groups modeling informative elements must be defined in terms of the elementary subgroups defining the frame of discernment.

Definition 4: A feature group is represented as super group which is the union of elementary subgroups as follows:

$$\forall h_j \in \Theta,\ h_j = \bigcup \{e_k \in \Omega \mid e_k \subset h_j\} \tag{3}$$

Example 3: The feature group f_{IG+OR} in example 2 is defined in terms of the elementary sub groups defined in example 1 as $e_3 \cup e_4$. Similarly the hypothesis f_{CHI} is defined as $e_1 \cup e_2$.

The frame of discernment along with the feature groups modeling the informative elements of the document D_1 is shown schematically in Fig. 1. These feature groups overlap in some region so these features in this overlapped region play a important role in representing semantic of the document. Obviously, some feature groups have stronger evidence than others. This is represented in the D-S via the use of a *bpa*.

Basic probability assignment: A *bpa* must be defined for every feature in the document collection *C* to capture the exact belief that the various feature groups (focal elements) provides for good representation of the document content. We compute the *bpa* values from term statistical characteristics in documents.
The *bpa* formula considered is:

$$m_i(h_j) = \begin{cases} \#(t_k, d_j).\log \frac{|T_r|}{\#T_r(t_k)} & h_j \in S \\ 0 & h_j \notin S \end{cases} \tag{4}$$

where,

- $\#(t_k, d_j)$ = The number of times t_k occurs in document d_j.
- $\#T_r(t_k)$ = Number of documents in T_r in which t_k occurs at least once.
- T_r = The total number of training documents in the collection *C*.

The first part of above formula ($h_j \in S$) assigns a positive *bpa* value to hypothesis representing indexing elements of S_i. The same formula ($h_j \notin S$) assigns 0 to all other remaining hypothesis. Logarithmic value used in the above formulae ensures that the calculated total bpa value to be always one.

Feature retrieval: To estimate the degree of relevance of each feature term to a semantic of the document, we use the belief function of the D-S theory. To each features f_i with *bpa* m_i, we have an associated belief function Bel_i defined upon m_i. The degree of relevance of the feature term to a document is represented by the hypothesis q is formulated as:

$$Bel_i(q) = \sum_{q \subset h_i} m_i(q) \tag{5}$$

This measure encapsulates the evidence of all the feature groups used to describe the document content that imply the hypothesis q. If $Bel_i(q) = 0$, the feature doesn't imply any relevance to the semantic of document. For a document collection, we use the belief values $Bel_i(q)$ to rank the features according to their estimated relevance to the semantic of the document.

EXPERIMENTS AND EVALUATION

Three benchmark dataset have been chosen for evaluating the effectiveness of the proposed feature selection method. These datasets are Reuters-21578 (Lewis, 1997), WebKB and 20 News Groups, which are the most widely used text corpus in text-classification research. The details on these data sets are given in Table 2. Since these datasets contains news articles on various topics and to show the effects of our proposed feature retrieval method on different domains, these datasets are intentionally chosen. As for as text classification algorithm, we choose the following most promising algorithm in the domain: SVM and kNN text classifiers. SVM is the most common one, as it was shown to perform better in terms of effectiveness than other text classifiers such as naïve Bayes, kNN, C4.5, and Rocchio (Joachims, 1998). The kNN algorithm is chosen because of its simplicity and superior efficiency than other algorithms (Yang and Pedersen, 1997; Denoeux, 1995).

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 F_1 measure. Precision is defined as the ratio of correct

Table 2: Summary of the benchmark datasets used in our research

Dataset	No of documents	Avg. document length	No of categories	Size	Domain
Reuters 21578 (R8)	7674	193	10	30 MB	News articles
Web KB	4199	126	4	16 MB	Web pages
20 news groups	18821	304	20	56 MB	News articles

Table 3: Performance of kNN classifier on Reuters, Web KB and news group's datasets

Dataset	Metric	Precision	Recall	Micro.Avg.F_1
Reuters	IG	86.76	64.25	73.83
21578	OR	87.21	69.45	77.32
	CHI	87.88	62.87	73.30
	COM	87.34	69.7	77.53
WebKB	IG	82.57	67.32	74.17
	OR	84.49	57.27	68.27
	CHI	78.48	45.17	57.34
	COM	83.35	68.6	75.26
20	IG	86.67	64.35	73.86
Newsgroups	OR	90.12	79.87	84.69
	CHI	82.47	67.22	74.07
	COM	89.85	82.09	85.79

Table 4: Performance of SVM classifier on Reuters, Web KB, news groups datasets

Dataset	Metric	Precision	Recall	Micro. Avg.F_1
Reuters	IG	90.21	79.98	84.79
21578	OR	87.66	75.82	81.31
	CHI	88.16	76.7	82.03
	COM	91.42	79.45	85.02
WebKB	IG	89.82	82.12	85.80
	OR	77.1	61.28	68.29
	CHI	88.69	58.41	70.43
	COM	91.42	84.59	87.87
20	IG	83.23	68.56	75.19
Newsgroups	OR	91.38	79.37	84.95
	CHI	87.24	69.71	77.50
	COM	91.27	84.48	87.74

classification of documents into categories to the total number of attempted classifications. Recall is defined as the ratio of correct classifications of documents into categories to the total number of labeled data in the testing set. F_1 measure is defined as the harmonic mean of precision and recall. Hence, a good classifier is assumed to have a high F_1 measure, which indicates that classifier performs well with respect to both precision and recall. We present the micro averaged results for precision, recall and F_1 measure. Micro averaging considers the sum of all the true positives, false positives, and false negatives (Forman, 2003).

RESULTS AND DISCUSSION

We conducted several experiments using our model with various learning algorithms. The idea of each experiment is to generate potential features using the derived measure *Bel* (*q*). We simply sort the list of features based on the computed scores and obtain the list of *k* relevant terms with the highest scores. To evaluate the goodness of each such retrieved list of features, the *k* relevant terms are tested by the learning algorithm on measures such as precision and recall and compared to the prior reported work. We repeated this experiment with a wide range of *k* values for each classifier. The range of *k* value is from 50 to 1000. The results are summarized in Table 3 and 4.

The experimental results suggest that the proposed feature selection model called as COM performs better than the conventional feature selection method such as IG, CHI and Odd Ratio in terms of precision. This improvement in effectiveness resulted from the combination of evidence represented by different feature selection methods. However, in some applications due to scalability reason, if a situation warrants only a limited number of features, the best superior one that outperforms others is IG.

We presented the classification results for SVM and kNN algorithm using our proposed feature retrieval model on Reuters 21578, WebKB and 20 News Groups datasets. This series of experiments strongly recommend that that if the precision is central goal, proposed model defeats other traditional methods by a smaller but significant margin.

CONCLUSION

We constructed a Dempster Shafer model for feature selection in text categorization and we observed the model performance on two text classification algorithm namely SVM and kNN. With an enormous outburst digital documents on the World Wide Web, existing traditional feature selection techniques are found to be inadequate in capturing the potential features from the document collection. It has been shown that the proposed Dempter Shafer model could capture the relevant and potential features from the collection and thereby improved the effectiveness of the classifier. We performed experiments on two standard benchmark datasets, Reuters 21578, WebKB and 20 News Groups. We showed that our proposed model significantly perform well than the conventional feature selection methods on SVM and kNN.

REFERENCES

Del Castillo, M.D. and J. Serrano, 2004. A multistrategy approach for digital text categorization from imbalanced documents. SIGKDD Exp., 6(1): 70-79.

Denoeux, T., 1995. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. Syst. Man Cybernet., 25: 804-813.

Doan, S. and S. Horiguchi, 2004. An efficient feature selection using multi-criteria in text categorization. Proceeding of the 4th International Conference on Hybrid Intelligent Systems, Washington, DC, pp: 86-91.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3(Mar.): 1289-1305.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. Proceeding of the European Conference on Machine Learning, pp: 137-142.

Lewis, D., 1997. Reuters-21578 Text Categorization Test Collection, Dist. 1.0, Retrieved from: AT&T Labs-Research, http://www.research.att.com/lewis. Lewis.

Rogati, M. and Y. Yang, 2002. High-performing feature selection for text classification. Proceeding of the 11th ACM International Conference on Information and Knowledge Management, pp: 659-661.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comp. Surv., 34(1): 1-47.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press, Princeton, New Jersey.

Smets, P. and R. Kennes, 1994. The transferable belief model. Artif. Intell., 66: 191-234.

Wang, H. and B. David, 2004. Extended k-nearest neighbours based on evidence theory. Comp. J., 47: 662-672.

Yang, Y. and J.O. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceeding of the 14th International Conference on Machine Learning, Nashville, pp: 412-420.