International Journal of Soft Computing 2 (4): 488-493, 2007
© Medwell Journals, 2007

A Novel Approach to Concept Extraction Using Naïve Bayesian Classification Technique

[1]M. Sathya, [1]P.Venil and [2]M.S. Saleem Basha
[1]Department of CSE, RSMCS, Pondicherry University, Pondicherry, India
[2]Department of CSE, Mailam Engineering College, Mailam, India

Abstract: Most of the information resources are capable to provide concept dependent results for the biased query. But the retrieval mechanism could not provide the relevant documents for the query text due to the size of the information resources is dynamically growing as the new topics being added. This problem can be overcome by automatically generating wrappers for these hidden documents. We are proposing a novel approach for automatically generating wrappers for describing the content of the hidden documents using a co-occurrence based clustering algorithm and Naive Bayesian classification model. The initial stage is the learning stage, which clusters the document based on the distinct concepts present in that. The learning technique makes use of a thesaurus and builds a co-occurrence correlation model. Then the clustered document features are used to generate the concept description using Naive Bayesian classifier. The join and posterior probabilities are calculated using the greedy selection and joining algorithm to represent cluster. Our implementation was tested on the standard data set and shows a better performance.

Key words: Information extraction, naïve bayesian, classification, novel approach, technique

INTRODUCTION

Text classification is the assignment of predefined categories to text documents. Text classification has many applications in natural language processing tasks such as E-mail filtering (Sahami *et al*., 1998; Androutsopoulos *et al*., 2000), news filtering (Lang, 1995), prediction of user preferences (Pazzani and Billsus, 1997) and organization of documents (Koller and Sahami, 1997). Because of the variety of languages, applications and domains, machine learning techniques are commonly applied to infer a classification model from example documents with known class labels. The inferred model can then be used to classify new documents. A variety of machine learning paradigms have been applied to text classification, including rule induction (Cohen and Singer, 1999), Naive Bayes (McCallum and Nigam, 1998), memory based learning (Yang and Liu, 1999), decision tree induction (Mitchell, 1997) and support vector machines (Jie and Jintao, 2001).

Naive Bayes is often used in text classification applications and experiments because of its simplicity and effectiveness. However, its performance is often degraded because it does not model text well and by inappropriate feature selection and the lack of reliable confidence scores.

This study is concerned with the Naive Bayes classifier. Naive Bayes uses a simple probabilistic model that allows inferring the most likely class of an unknown document using Bayes' rule. Because of its simplicity, Naive Bayes is widely used for text classification (Sahami *et al*., 1998; Pazzani and Billsus, 1997; Koller and Sahami, 1997; Joachims, 1998; Androutsopoulos *et al*., 2000).

The Naive Bayes model makes strong assumptions about the data: It assumes that words in a document are independent. This assumption is clearly violated in natural language text: There are various types of dependences between words induced by the syntactic, semantic, pragmatic and conversational structure of a text. Also, the particular form of the probabilistic model makes assumptions about the distribution of words in documents that are violated in practice (Graven *et al*., 2000). Nonetheless, Naive Bayes performs quite well in practice, often comparable to more sophisticated learning methods (Katz, 1996; Kononenko, 1991). One could suspect that the performance of Naive Bayes can be further improved if the data and the classifier better fit together. There are 2 possible approaches: Modify the data, modify the classifier (or the probabilistic model).

Many researchers have proposed modifications to the way documents are represented, to better fit the

Corresponding Author: P.Venil, Department of CSE, RSMCS,Pondicherry University, Pondicherry, India

assumptions made by Naive Bayes. This includes extracting more complex features, such as syntactic or statistical phrases (Domingos and Pazzani, 1997) and exploiting semantic relations using lexical resources (Friedman, 1997). These attempts have been largely unsuccessful. Another way to improve the document representation is to extract features by word clustering (Mladeni and Grobelink, 1998) or by transforming the feature space (Gomez and Buenaga, 1997). These methods did show some improvement of classification accuracy. Instead of changing the document representation by using other features than words, it is also possible to manipulate the text directly, e.g., by altering the occurrence frequencies of words in documents (Dhillion *et al*., 2003). This can help the data to better fit the distribution assumed by the model. The most important way to better fit the classifier to the data is to choose an appropriate probabilistic model. Some researchers have also tried to improve performance by altering the way the model parameters are estimated from training data (Torkkola, 2001).

We propose a novel approach for text classification. In our approach, we first perform the segmentation of text document into smaller regions, followed by clustering of regions, before learning the relationship between concepts and region clusters using the set of training documents with pre-assigned concepts. The main focus of this paper is two-fold. First, in the learning stage, we perform clustering of regions into region clusters by incorporating pairwise constraints which are derived by considering the language model underlying the classes assigned to training documents. Second, in the classification stage, we employ a Naïve Bayes model to compute the posterior probability of concepts given the region clusters. Experiment results show that our proposed system utilizing these two strategies outperforms the state-of-the-art techniques in classifying large document collection.

To address the above problems, we first consider the use of a language model underlying the classes assigned to training document set to impose additional semantic pair-wise constraints when clustering the regions. Recently research on clustering shows that clustering with pair-wise constraints, a kind of realistic semi-supervised clustering method, performs considerably better than the unconstrained methods. Next, we formulate a Naïve Bayesian model to perform active classification. It aims to strike a good balance between the simplicity of naïve Bayesian model and the complexity of incorporating co-occurrence information of region clusters. Experimental results demonstrate that the combined approach utilizing both clustering with pair-wise constraints and Naïve Bayesian model outperforms the state-of- the-art systems in text classification.

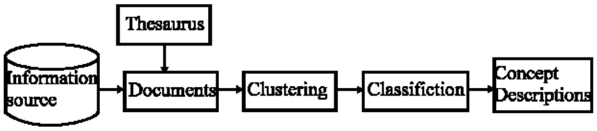

Fig. 1: Automatic concept extration

Our main contribution is two-fold. First, we develop a semi supervised region clustering method incorporating pair-wise constraints which are derived from language model. Second, we formulate a Naïve Bayesian model for concept prediction and inference. This study discusses the design and implementation of our system.

Figure 1 describes a two step process for automatic concept extraction. In the first step the document is clustered based on number of distinct concepts available, then in the second step the given document is classified by calculating the co-occurrence probabilities between the clusters and applied concepts.

FORMULATION OF PAIR-WISE CONSTRAINTS

Classification of documents reflects the semantics of the document as well as its regions and we would like to induce from the concepts that cannot-link and must-link relations between different regions. In general, it is easier to induce the cannot-link relationship from shared concepts but the must-link relationship is harder to deduce. We assume that the semantic irrelevance of 2 regions can be deduced by the irrelevance of all concepts between two regions. This assumption is reasonable because although classification of document is likely to be incomplete, it is always complete for those concepts that we care most.

Co-occurrence based correlation: In general, high co-occurrence concepts are likely to be used together to describe the same document. In other words, two concepts are likely to belong to the same conceptual group if they have high co-occurrence and vice visa. The co-occurrence-based correlation of 2 concepts c_1 and c_2 is computed as:

$$R_{co}(c_1, c_2) = df(c_1 \wedge c_2)/df(c_1 v\ c_2)$$

Where $df(c_1 \wedge c_2)$, $df(c_1\ v\ c_2)$ is the fraction of documents with concepts containing c_1 and (or) c_2.

Thesaurus based correlation: Word Net is an electronic thesaurus popularly used in research on lexical semantic acquisition. In WordNet, the meaning of a word is represented by a network of synonym (synset) and hypernym etc between words. The thesaurus based correlation between the two concepts c_1 and c_2 is computed as:

$$R_L(c_1, c_2)\begin{cases} 1(c_1 \text{ and } c_2 \text{ in the same synset, or } c_1 = c_2). \\ 0.8(c_1 \text{ and } c_2 \text{ have "antonym" relation}) \\ 0.5(c_1 \text{ and } c_2 \text{ have relation of "is a :,"part of",} \\ \text{or" member of").0(others)} \end{cases}$$

The relevance of 2 annotations Cp and Cq is defined as:

$$\text{Rel}\ (c_p, c_q) = \underset{Ci \in Cp,\ Cj}{\text{argmax}}\ (c_i, c_j)$$

Where the correlation definition R could be either Rco or RL. If the relevance of two annotations Rel (Cp, Cq) is smaller than a predefined threshold, then Cp and Cq and their corresponding document regions are regarded as irrelevant to each other.

CLUSTERING USING PAIR-WISE CONSTRAINTS

After the construction of pair-wise constraints between regions, we perform clustering to generate region clusters. K-Means is a popular clustering method. Since K-Means cannot directly handle pair-wise constraints, we adapt a variant of K-Means called Pairwise Constrains K-Means (PCK-Means) (Androutsopoulos *et al*., 2000) to perform the clustering. We formulate the goal of pair-wise constraint clustering as the minimization of a combined objective function, defined as the sum of the total squared distances between the regions and their region cluster centroids and the cost incurred by violating any of the pair-wise constraints.

Let $\{r_i\}_{i=1..N}$ be the whole set of regions, $\{\mu_h\}_{h=1..K}$ represent the centroids of K region clusters $\{R_h\}_{h=1..K}$, l(i) be the cluster assignment of a region r_i, where l(i) $\in$ {1,2,…k} and P be the cost incurred when the cannot-link pair-wise constraints are violated.

NAIVE BAYESIAN APPROACH

Bayesian text classification uses a parametric mixture model to model the generation of documents (McCallum and Nigam, 1998). The model has the following form:

$$P(d)=\sum_{J=1}^{|c|} p(c_j)p(d|c_j)$$

Where c_j are the mixture components (that correspond to the possible classes) and $p(c_j)$ are prior probabilities. Using Bayes' rule, the model can be inverted to get the posterior probability that d was generated by the mixture component c_j:

$$P(c_j|d)=\frac{p(c_j)p(d|c_j)}{p(d)}$$

To classify a document, the classifier selects the class with maximum posterior probability, given the document, where p(d) is constant and can be ignored:

$$c^*(d) = \text{argmax}_j\ p(c_j)\ p(d|c_j) \qquad (1)$$

The prior probabilities $p(c_j)$ are estimated from a training corpus by counting the number of training documents in each class c_j. The distribution of documents in each class, $p(d|c_j)$, cannot be estimated directly. Rather, it is assumed that documents are composed from smaller units, usually words or word stems. To make the estimation of parameters tractable, we make the Naive Bayes assumption: That the basic units are distributed independently. There are several Naive Bayes models that make different assumptions about how documents are composed from the basic units. The most common models are: the binary independence model, the Poisson Naive Bayes model and the multinomial model (MvCallun and Nigam, 1998; Rennie *et al*., 2003). The most apparent difference between these models is that the Poisson model and the multinomial model use word occurrence frequencies, while the binary independence model uses binary word occurrences. In this study we consider the multinomial Naïve Bayes model because it is generally superior to the binary independence model for text classification (MvCallun and Nigam, 1998; Rennie *et al*., 2003).

In the multinomial model, a document d is modeled as the outcome of |d| independent trials on a single random variable W that takes on values $w_t \in V$ with probabilities $p(w_t|c_j)$. Each trial with outcome w_t yields an independent occurrence of w_t in d. Thus a document is represented as a vector of word counts $d = \langle x_t \rangle_{t=1...|V|}$ where each x_t is the number of trials with outcome w_t, i.e., the number of times w_t occurs in d. The probability of d is given by the multinomial distribution:

$$p(d\,|\,cj)=p(|\,d\,|)\,|\,d\,|!\prod_{t=1}^{|v|}\ \frac{p(w_t\,|\,c_j)^{xt}}{x_t}$$

Here we assume that the length of a document is chosen according to some length distribution, independently of the class. Plugging this into Eq. 1 we get the following form:

$$c^*(d) = \arg\max_{c_j} p(c_j) \prod_{t=1}^{|v|} p(w_t \mid c_j)^{xt} \qquad (2)$$

The parameters $p(w_t|c_j)$ are estimated by counting the occurrences of w_t in all training documents in c_j, using a Laplacean prior:

$$p(w_t \mid c_j) = \frac{1 + n(c_j, w_t)}{|v| + n(c_j)}$$

Where $n(c_j|w_t)$ is the number of occurrences of w_t in the training documents in c_j and $n(c_j)$ is the total number of word occurrences in c_j.

WORD FREQUENCY

It is usually claimed that the multinomial model gives higher classification accuracy than the binary independence model on text documents because it models word occurrence frequencies (McCallum and Nigam ,1998; Rennie *et al*., 2003). Contrary to this belief, we show that word frequency hurts more than it helps and that ignoring word frequency information can improve performance dramatically.

The multinomial Naive Bayes model treats each occurrence of a word in a document independently of any other occurrence of the same word. In reality, however, multiple occurrences of the same word in a document are not independent. When a word occurs once, it is likely to occur again, i.e., the probability of the second occurrence is much higher than that of the first occurrence. This is called *burstiness* (Craven *et al*., 2003). The multinomial model does not account for this phenomenon. This results in a large underestimation of the probability of documents with multiple occurrences of the same word.

In Dhillon *et al.* (2003) a transformation of the form $x'_t = \log(1+x_t)$ was applied to the word frequency counts in a document in order to better fit the data to the probabilistic model. This does not eliminate word frequencies but has the effect of pushing down larger counts. An even simpler, yet less accurate method is to remove word frequency information altogether using the transform $x'_t = \min\{x_{t,1}\}$. This can be thought of as discarding all additional occurrences of words in a document. Instead of transforming the word counts, we can change the classification rule as in (3):

$$c^*(d) = \arg\max_{c_j} p(c_j) \prod_{t=1}^{|v|} p(w_t \mid c_j)^{\min\{x_t, 1\}} \qquad (3)$$

and the parameter estimation as in (4), where $d(c\ j, w_t)$ is the number of documents containing w_t in c_j:

$$p(w_t | c_j) = \frac{1 + d(c_j, w_t)}{|v| + \sum_{s=1..|v|} d(c_j, w_s)} \qquad (4)$$

CONFIDENCE SCORES

Sometimes it is desirable to have the classifier produce classification scores that reflect the confidence of the classifier that a document belongs to a class. For example, in binary classification problems where one class (the target class) contains examples that are relevant to some query, a document could be assigned to the target class only if its confidence score exceeds some threshold. In multi-label classification tasks (where each document can belong to zero, one or more classes), a document can be assigned to all classes for which the confidence is above the threshold. Such confidence scores must be independent of document length and complexity.

The posterior probabilities $p(c_j|d)$ computed by Naive Bayes are inappropriate as confidence scores because they are usually completely wrong and tend to go to zero or one exponentially with document length (Forman, 2003). This is a consequence of the Naïve Bayes independence assumption and the fact that the words in a document are not really independent. Note that the classification decision of Naive Bayes is not affected as long as the ranking of the classes is not changed (in fact it has been argued that the large bias can reduce classification error (Kononenko, 1991).

We follow the approach in (Joachims, 1998) to get better confidence scores for Naive Bayes. First, we replace the posterior scores with the KL-divergence scores.

$$\text{score}\,(c_j|d) = 1/|d| \log p(c_j) - \sum_{t=1..|v|} p(w_t|d) \log\,(p(w_t|d)/\,p(w_t|c_j))$$

This has 2 effects. By taking logarithms and dividing by the length of a document, instead of multiplying conditional probabilities (Eq. 2) we calculate their geometric mean and thus account for the impact of wrong independence assumptions under varying document lengths. Furthermore, by adding the entropy of (the probability distribution induced by) the document, we account for varying document complexities. Finally, to make the scores comparable across different documents, we normalize the scores such that they form a probability distribution over classes (i.e., the scores for all classes sum to one):

$$\mathrm{conf}(c_j|d) = \mathrm{score}(c_j|d)/\sum_{t=1\ldots|c|} \mathrm{score}(c_k|d)$$

We compare the posterior scores and the confidence scores on the Reuters-21578 dataset, using the ModApte split with only the 10 largest topics (Cover and Thomes, 2000). We remove all non-alphabetic characters and convert all letters to lower case. In addition, we map all numbers to a special token. For each topic, we build a binary classifier using all documents in that topic as relevant examples and all other documents as non-relevant examples. The threshold is set for each classifier individually such that recall equals precision (precision/ recall break-even point).

CONCLUSION

We have presented a novel Polynomial Naïve Bayesian approach incorporating clustering with pair-wise constraints for automatic text classification. The join and posterior probabilities are calculated using the greedy selection and joining algorithm to represent cluster. Our implementation was tested on the standard data set and shows a better performance. Future work includes incorporating different types of pair-wise constraints and classification of other document types.

REFERENCES

Androutsopoulos, I., G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos and P. Stamatopoulos, 2000. Learning to Filter Spam E-mail: A Comparison of a Naive Bayesian and a Memory-Based Approach. In: Zaragoza, H., P. Gallinari and M. Rajman (Eds.), Proc. Workshop on Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), Lyon, France, pp: 1-13.

Cohen, W.W. and Y. Singer, 1998. Context-sensitive learning methods for text categorization. ACM. Trans. Infor. Sys., 17: 141-173.

Cover, T.M. and J.A. Thomas, 2000. Elements of Information Theory. John Wiley, New York (1991) 25. Bennett, P.N.: Assessing the calibration of Naive Bayes' posterior estimates. Technical Report CMU-CS-00-155, School of Computer Science, Carnegie Mellon University.

Craven, M., D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, 2000. Learning to construct knowledge bases from the World Wide Web. Artificial Intelligence, 118: 69-113.

Dhillon, I.S., S. Mallela and R. Kumar, 2003. A divisive information-theoretic feature clustering algorithm for text classification. J. Machine Learning Res., 3: 1265-1287.

Domingos, P. and M. Pazzani, 1997. On the optimality of the simple Bayesian classifier under zeroone loss. Machine Learning, 29: 103-130.

Friedman, J.H., 1997. On bias, variance, 0/1-loss and the curse-of-dimensionality. Data Mining and Knowledge Discovery, 1: 55-77.

Gomez-Hidalgo, J.M. and M. de Buenaga Rodr´ýguez, 1997. Integrating a lexical database and a training collection for text categorization. In: ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp: 39-44.

Jie Ou and Li Jintao, 2001. The Personalized Index Service System in Digital Library, IEEE.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: Proc. 10th European Conference on Machine Learning (ECML). Volume 1398 of Lecture Notes in Computer Science. Heidelberg, Springer pp: 137-142.

Katz, S.M., 1996. Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2: 15-59.

Koller, D. and M. Sahami, 1997. Hierarchically classifying documents using very few words. In: Proc. 14th International Conference on Machine Learning (ICML)., pp: 170-178.

Kononenko, I., 1991. Semi-naïve Bayesian classifier, 6th European Working Session on Learning, pp: 206-219.

Lang, K., 1995. NewsWeeder: Learning to filter netnews. In: Proc. 12th International Conference on Machine Learning (ICML-95), Morgan Kaufmann, pp: 331-339.

McCallum, A. and K. Nigam, 1999. A comparison of event models for Naive Bayes text classification. In: Learning for Text Categorization: Papers from the AAAI Workshop, AAAI Press Technical Report WS., pp: 98-05.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill, New York.

Mladeni´c, D. and M. Grobelnik, 1998. Word sequences as features in text-learning. In: Proc. 17th Electrotechnical and Computer Science Conference (ERK), Ljubljana, Slovenia.

Pazzani, M. and D. Billsus, 1997. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27: 313-331.

Rennie, J.D.M., L. Shih, J. Teevan and D. Karger, 2003. Tackling the Poor Assumptions of Naïve Bayes Text classifiers. In Fawcett, T. and N. Mishra, (Eds.). Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, D.C., AAAI Press, pp: 616-623.

Sahami, M., S. Dumais, D. Heckerman and E. Horvitz, 1998. A Bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the AAAI Workshop, Madison Wisconsin, AAAI Press Technical Report WS-98-05, pp: 55-62.

Torkkola, K., 2001. Linear discriminant analysis in document classification. In: IEEE ICDM Workshop on Text Mining (TextDM), San Jose, CA.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. In: Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)., pp: 42-49.