A New Fine-Grained Weighting Method in Multi-Label Text Classification

Chang-Hwan Lee
Department of Information and Communications
Dongguk Univerity
Seoul, Korea
chlee@dgu.ac.kr

Abstract

Multi-label classification is one of the important research areas in data mining. In this paper, a new multi-label classification method using multinomial naive Bayes is proposed. We use a new fine-grained weighting method for calculating the weights of feature values in multinomial naive Bayes. Our experiments show that the value weighting method could improve the performance of multinomial naive Bayes learning.

Introduction

Correctly classifying the documents into particular category is still a challenging task because of large and vast amount of features in the dataset. In particular, multi-label text classification problems have received considerable attention, since each document may each be associated with multiple class labels. For example, documents may belong to multiple genres, such as entertainment and sports, or religion and culture. In this paper we explore the multi-label text classification problem, and there are two important issues in it.

The first is to improve the performance of multi-label text classification tasks. In multi-label text classification problem, multinomial naive Bayes (MNB) algorithm has been most commonly used. MNB classifier is an efficient and reliable text classifier, and many researchers usually regard it as the standard naive Bayes text classifier in recent years. However, their performances are not as good as some other learning methods such as support vector machines and boosting. In this paper, we mainly investigate the reasons behind the poor performance of MNB. Then to enhance the performance of MNB method, we propose a new paradigm of weighting method, called value weighing method, for MNB learning.

The second issue in multi-label classification is to effectively make use of the dependency relationships between class labels. One common approach to process these class dependencies is to treat each class as a separate binary classification problem; that is called binary relevance method(BR). (Read *et al.* 2009) When building the classifiers, BR does not directly model correlations which exist between labels in the training data. However, in many real-world tasks, labels are highly interdependent. Therefore, the key to successful multi-label learning is how to effectively exploit dependencies between different labels.

This paper presents a new weighting method for multinomial naive Bayes learning. In current multinomial naive Bayes, each word is associated with its frequency, and the frequency can be regarded as the importance of the word. Therefore, each word (feature) is given a weight (frequency). Furthermore we compare the performance of the proposed model with that of other state-of-the-art multi-label classifiers.

The rest of this paper is structured as follows. Section II shows the related works on naive Bayesian document classifier and multi-label problem. Section III discusses the multinomial naive Bayesian algorithm and the new value weighting method, and Section V shows the experimental results of the proposed methods. Finally Section VI summarizes the contributions made in this paper.

Related Work

Text classifiers based on naive Bayes have been studied extensively in the literature. Especially there have been many researches for using MNB model in text classification.

McCallum and Nigam (McCallum and Nigam 1998) compares classification performance between multi-variate Bernoulli model and multinomial model. In multi-variate Bernoulli model, a document is considered as a binary feature vector, and it expresses whether each word is present or absent. They show that the multi-variate Bernoulli model performs well with small vocabulary sizes, but the multinomial model usually performs even better at larger vocabulary sizes.

Rennie *et al.* (Rennie *et al.* 2003) introduced Complement Naive Bayes (CNB) for the skewed training data. CNB estimates parameters using data from all classes except the currently estimated class. Furthermore they demonstrated that MNB can achieve better accuracy by adopting a TFIDF representation, traditionally used in Information Retrieval.

Schneider (Schneider 2005) addressed the problems of naive Bayesian text classifier and shows that they can be solved by some simple corrections. He effectively removed duplicate words in a document to account for burstiness phenomena in text. And he proposed to use uniform priors to avoid problems with skewed class distributions when the documents are very short.

In recent years, there has been much study in multi-label classification problem as motivated from emerging applica-

tions. As mentioned above, the key to successful multi-label learning is how to effectively exploit dependencies between different labels.

Ghamrawi and McCallum (Ghamrawi and McCallum 2005) proposed two undirected graphical models that directly parameterize label dependencies in multi-label classification. The first is Collective Multi-Label classifier (CML) which jointly learns parameters for each pair of labels. The second is Collective Multi-Label with Features classifier(CMLF) which learns parameters for feature-label-label triples.

McCallum (McCallum 1999) defines a probabilistic generative model according to which, each label generates different words. Based on this model a multi-label document is produced by a mixture of the word distributions of its labels. The mixture models are trained by EM, selecting the most probable set of labels from the power set of possible classes.

Read et al. (Read *et al.* 2009) introduced classifier chain. The classifier chain is binary relevance-based methods and consist of binary classifiers which are linked in a chain. They again propose an ensemble of classifier chain combining several classifier chain by changing the order for the labels.

Improving the Performance of Multinomial Naive Bayes in Document Classification Tasks

In this paper, we assume that documents are generated according to a multinomial event model. Thus a document is represented as a vector $\mathbf{x} = (f_1, ..., f_{|V|})$ of word counts where $|V|$ is the vocabulary size and each f_t indicates how often t-th word W_t occurs in $\mathbf{x}$. Given model parameters $p(W_t|c)$ and $p(c)$, assuming independence of the words, the most likely class value c for a document $\mathbf{x}$ is computed as

$$c^*_{MNB}(\mathbf{x}) = \arg\max_c \; p(c) \prod_{t=1}^{|V|} p(W_t|c)^{f_t} \qquad (1)$$

where $p(W_{tj}|c)$ is the conditional probability that a word W_t may happen in a document $\mathbf{x}$ given the class value c and $p(c)$ is the prior probability that a document with class label c may happen in the document collections. The values of $p(W_t|c)$ and $p(c)$ are estimated from training documents using maximum likelihood estimation with a Laplacean prior: (Schneider 2004)

$$p(W_t|c) = \frac{1 + \sum_{\mathbf{x}_i \in c} f_{it}}{|V| + \sum_{t=1}^{|V|} \sum_{\mathbf{x}_i \in c} f_{it}}, \quad p(c) = \frac{|c|}{N} \qquad (2)$$

where f_{it} represents the t-th term frequency of i-th document, $|c|$ represents the number of class value c, and N is the number of training document, respectively.

MNB provides reasonable prediction performance and easy to implement. But it has some unrealistic assumptions that affect overall performance of classification. The first is that all features are equally important in MNB learning. Because this assumption is rarely true in real-world applications, the predictions estimated by MNB are sometimes poor. The second is that the importance of each word grows proportionally with its frequency. The more a word appears in the document, the more important it becomes. However, when a certain word appears for the first time, it is very important word in terms of the category of the document. For example, when a word 'computer' is encountered at first, it gives us a lot of information that this document is about technology. However, if the same word 'computer' already appeared 100 times, the subsequent 'computer' has virtually no importance.

In current MNB, we can not discriminate the importance of each frequency values of word. In fact, the importance of a word increases linearly with the frequency of the word. The performance of MNB can be improved by mitigating these assumptions. The following section describes these issues in detail.

A Fine-Grained Weighting Method in Text Classification

In classification learning, a classifier assigns a class label to a new instance. Naive Bayesian learning uses Bayes theorem to calculate the most likely class label of the new instance. A new instance d is classified to the class with the maximum posterior probability. In naive Bayesian learning, since all features are considered to be independent given the class value, the classification on d is defined as follows

$$\mathcal{V}_{nb}(d) = \text{argmax}_c \; P(c) \prod_{a_{ij} \in d} P(a_{ij}|c)$$

where a_{ij} represents the j-th value of the i-th feature.

Since the assumption that all features are equally important hardly holds true in real world application, there have been some attempts to relax this assumption in machine learning methods. The feature weighting in naive Bayesian approach is one approach for easing the independence assumption. Feature weighting assigns a continuous value weight to each feature, and is thus a more flexible method than feature selection. The naive Bayesian classification with feature weighting is represented as follows

$$c^*_{NB-FW}(\mathbf{x}) = \arg\max_c \; p(c) \prod_i p(a_i|c)^{w_i} \qquad (3)$$

In this formula, unlike the ordinary naive Bayesian approach, each feature i has its own weight w_i. The w_i can be any real number, representing the significance of feature i. The feature weighted naive Bayesian method involves a much larger search space than feature selection, and is generally known to improve the performance of naive Bayesian learning (Lee *et al.* 2011).

MNB classification is a special form of feature weighted naive Bayesian learning. The MNB classification with feature weighting is represented as follows.

$$c^*_{MNB-FW}(\mathbf{x}) = \arg\max_c \; p(c) \prod_{t=1}^{|V|} p(W_t|c)^{w_t} \qquad (4)$$

The basic idea of feature weighting is that the more important a feature is, the higher its weight is. In feature weighting naive Bayes, each word W_t has its own weight w_t.

In traditional MNB (Equation 1), the frequency(f_t) of each word W_t plays the role of the significance of the word.

Therefore, the basic assumption in MNB is that when a certain word appears frequently in a document, the word grows important in proportion to its occurrence. Each word is given a weight which is the frequency of the word in the document.

Kim *et al.* (Kim *et al.* 2006) proposed a feature weighting scheme using information gain. Information gain for a word given a class, which becomes the weight of the word, is calculated as follows:

$$w_t = f_t \cdot \{H(C) - H(C|W_t)\} \quad (5)$$
$$= f_t \cdot \sum_c \sum_{W_i \in \{W_t, \bar{W}_t\}} p(c, W_i) \log \frac{p(c, W_i)}{p(c)p(W_i)}$$

where $p(c, W_i)$ is the number of documents with the word W_i and class label c divided by the total number of documents, and $p(W_i)$ is the number of documents with the word W_i divided by the total number of documents, respectively.

In this paper, we think of a new method in which weights are assigned in a more fine-grained way. We are going to treat each occurrence of a word differently in terms of its importance. When a certain word appears for the first time in a document, it becomes very important with respect to the classification of the document. For example, when a word 'diabetes' appears for the first time in a document, it provides a significant implication that this document is about 'health'. However, when the same word 'diabetes' already appeared many times, say 100, the next occurrence has virtually no importance. The probability of the second occurrence is much higher than that of the first occurrence. Because MNB treats the significance of each occurrence of a word equally, the multinomial model does not account for this phenomenon. In this paper we will investigate whether assigning weights to each word count can improve the performance of classification.

In order to implement the fine-grained weighting, we first discretize the term frequencies of each word. The discretization task converts a continuous term frequency f_t to a categorical *word frequency bin* a_{tj}, which represents the j-th discretized value of the term frequency. In other words, instead of assigning a weight to each word feature(e.g. MNB), we assign a weight to each word frequency bin. After that, the weights of these word frequency bin are automatically calculated using training data.

We call this method as *value weighting* method. As we can see, unlike the current feature weighting methods, the value weighting method calculates a weight for each word frequency bin. The value weighting method in MNB can be defined as follows.

$$c^*_{MNB-VW}(\mathbf{x}) = \arg\max_c \; p(c) \prod_{a_{tj} \in \mathbf{x}} p(a_{tj}|c)^{w_{tj}} \quad (6)$$

where w_{tj} represents the weight of word frequency bin a_{tj}. You can easily see that each word frequency bin is assigned a different weight.

Calculating Value Weights This section describes the value weighting method for calculating weights of frequency bins. In this paper, we will use an information-theoretic method for assigning weights to each word frequency bin. The basic assumption of the value weighting method is that when a certain word frequency bin is observed, it gives a certain amount of information to the target word feature. The more information a word frequency bin provides to the target class, the more important the bin becomes. The critical part now is how to define or select a proper measure which can correctly measure the amount of information.

Table 1: Test datasets for the value weighting method.

Dataset	#(Data)	#(Label)	#(Feature)
New3	3204	6	13196
Ohsumed	1003	10	3183
Amazon	1500	50	10000

In this paper, we employ Hellinger measure (Beran 1977) in order to calculate the difference between the probability of a priori distribution and that of a posteriori distribution of the target class. The Hellinger measure (denoted as HW) for a word frequency bin a_{tj} is defined as

$$HM(C|a_{tj}) = \left(\sum_c \left(\sqrt{p(c|a_{tj})} - \sqrt{p(c)} \right)^2 \right)^{1/2} \quad (7)$$

The formula $HM(C|a_{tj})$ is the average mutual information between the events c and a_{tj} with the expectation taken with respect to a posteriori probability distribution of C. It can be used as a proper weighting function. So the value weight for a word frequency bin a_{tj} is defined as

$$\begin{aligned} w_{tj} &= \frac{1}{Z_t} HM(C|a_{tj}) \\ &= \frac{1}{Z_t} \left(\sum_c \left(\sqrt{p(c|a_{tj})} - \sqrt{p(c)} \right)^2 \right)^{1/2} \end{aligned} \quad (8)$$

where Z_t is a normalization constant given as $Z_t = \frac{1}{|a_t|} \sum_{j|t} w_{tj}$. The $|a_t|$ represents the number of word frequency bins in word feature t.

Experimental Evaluation

In order to evaluate the performance of each proposed method, we divide the experiment into two subsections. Firstly, performance of value weighting method is compared with MNB. Secondly, we compare co-training with other multi-label algorithms such as BR, CC, etc.

In this section, we describe how we conducted the experiments for measuring the effects of value weighting method. We then present the empirical results obtained using these methods.

We used 3 text datasets to conduct our empirical study for text classification. All these datasets have been widely used in text classification, and are publicly available. Table1 1 provides a brief description of each dataset.

'New3' (TunedIT 2012) dataset contains a collection of news stories and 'Ohsumed' (TunedIT 2012) is a dataset of medical articles. 'Amazon' (Frank and Asuncion 2010)

Table 2: Accuracies of the methods.

Dataset	NB	MNB	MNB-VW
New3	0.842	**0.935**	0.929
Ohsumed	0.918	0.951	**0.986**
Amazon	0.971	0.976	**0.994**

Table 3: Weights of Feature Value Bins

Dataset	1st	2nd	3rd	4th	5th
News3	0.0023	1.9674	1.2436	0.8458	0.7523
Ohsumed	0.0022	1.7143	1.4672	0.8310	0.4902
Amazon	0.1152	1.6844	1.2484	1.3550	1.0888

dataset is derived from the customer reviews in Amazon Commerce Website for authorship identification. The continuous features in datasets were discretized using equal distance method with 5 bins.

In this experiment, single-label classification was conducted. So the value weighting method(MNB-VW) was used without employing the co-training model. The proposed MNB-VW is compared with regular naive Bayes (NB) and multinomial naive Bayes (MNB). We used Weka software to run NB and MNB.

Table 2 shows the results of the accuracies of these methods. The numbers with bold letter mean they are the best accuracy among NB, MNB, and MNB-VW. The MNB-VW shows the best performance in 2 cases. And it always shows better performance than NB method. These results clearly indicate that assigning weights to each word frequency bin could improve the performance of the classification task of naive Bayesian in document classification.

We calculated the average value weights of all features in each dataset. Because the number of bins is 5, all features have 5 discretized word frequency bins. The average value weights used in above experiment are shown in Figure 3.

When term frequency is zero, it is discretized to first word frequency bin. The remaining term frequencies are discretized by equal distance and the discretized values are assigned in order of frequency. The highest term frequencies are discretized to the fifth word frequency bin.

Table 3 shows the weights of each frequency bin, and clearly shows that the low word frequency bins generally have higher weight than high word frequency bins. It means that higher weights are generally assigned to low term frequencies. Notice that the 1st bin means the term frequency is zero. Therefore, it reasonable that the 1st bin (TF=0) has very low weight value. Actually, the event that a word occurs once or less frequently has more significance than the event that a word occurs several times. These experiments demonstrate that each frequency bin has different importance in terms of document classification, and our fine-grained weighting method could clearly solves this issue.

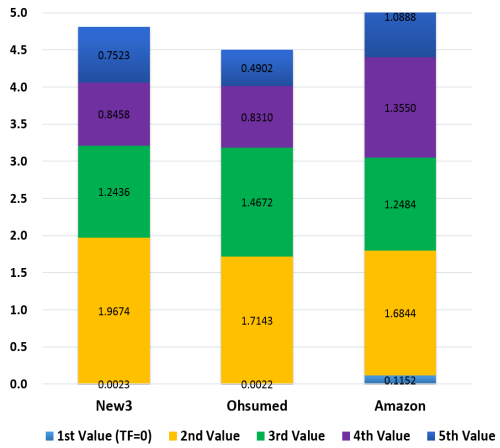

Figure 1: Average value weights of each dataset

Conclusions and Future Work

In the paper, multinomial naive Bayes algorithm, applied to multi-label document classification, is improved in a number of ways. A new paradigm of weighting method, called value weighting method, is proposed for multinomial naive Bayesian learning. The proposed method is a more fine-grained weighting method, and assigns a different weight to each feature value.

The experimental results show that the value weighting method shows better performance in most cases than its counterpart algorithms. As a result, this work suggests that we could improve the performance of multinomial naive Bayes in multi-label text classification by using value weighting approach. In the future, it will be interesting to apply this approach to some large scale document datasets in order to see the scalability of the proposed method.

Acknowledgments

This work was supported in part by National Research Foundation of Korea (NRF) (Grant number: 2011- 0023296).

References

R. J. Beran. Minimum hellinger distances for parametric models. *Ann. Statistics*, 5:445–463, 1977.

A. Frank and A. Asuncion. Uci machine learning repository. `http://archive.ics.uci.edu/ml`, 2010.

N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Inter. Conf. on Inform. and Know. Manage.*, 2005.

S. Kim, K. Han, H. Rim, and S. Myaeng. Some effective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 18(11):1457–1466, 2006.

Chang-Hwan Lee, Fernando Gutierrez, and Dejing Dou. Calculating feature weights in naive bayes with kullback-leibler measure. In *11th IEEE International Conference on Data Mining*, 2011.

A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

A. McCallum. Multi-label text classification with a mixture model trained by em. In *AAAI99 Workshop on Text Learning*, 1999.

J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *ECML/PKDD*, pages 254–269, 2009.

J. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *In Proceedings of the 20th international conference on Machine learning (ICML)*, pages 616–623,, 2003.

K. M. Schneider. A new feature selection score for multinomial naive bayes text classification based on kl-divergence. In *In Proceedings of the 42nd Meeting of the Association of Computational Linguistics (ACL)*, pages 186–189, 2004.

K. Schneider. Techniques for improving the performance of naive bayes for text classification. In *LNCS Vol 3406*, pages 682–693, 2005.

TunedIT. Tunedit data mining blog. `http://www.tunedit.org`, 2012.