Large Scale Text Classification using Semi-supervised Multinomial Naive Bayes

Jiang Su[1] JSU@SITE.UOTTAWA.CA
Jelber Sayyad-Shirabad[1] JSAYYAD@SITE.UOTTAWA.CA
Stan Matwin[1,2] STAN@SITE.UOTTAWA.CA

1 School of Information Technology and Engineering, University of Ottawa, K1N 6N5 Canada
2 Institute for Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

Numerous semi-supervised learning methods have been proposed to augment Multinomial Naive Bayes (MNB) using unlabeled documents, but their use in practice is often limited due to implementation difficulty, inconsistent prediction performance, or high computational cost. In this paper, we propose a new, very simple semi-supervised extension of MNB, called Semi-supervised Frequency Estimate (SFE). Our experiments show that it consistently improves MNB with additional data (labeled or unlabeled) in terms of AUC and accuracy, which is not the case when combining MNB with Expectation Maximization (EM). We attribute this to the fact that SFE consistently produces better conditional log likelihood values than both EM+MNB and MNB in labeled training data.

1. Introduction

Multinomial Naive Bayes (MNB) has been widely used in text classification. Given a set of labeled data, MNB often uses a parameter learning method called Frequency Estimate (FE), which estimates word probabilities by computing appropriate frequencies from data. The major advantage of FE is that it is simple to implement, often provides reasonable prediction performance, and is efficient.

Since usually the cost of obtaining labeled documents is high and unlabeled documents are abundant, it is desirable to leverage the unlabeled data to improve the MNB model learned form the labeled data. Numerous semi-supervised learning methods have been proposed to achieve this, and Expectation-Maximization (EM) (Dempster et al., 1977) is often used with MNB in semi-supervised setting.

Though the combination of EM+MNB is relatively fast and simple to use, past research has identified some inconsistencies with EM+MNB. Namely, depending on given dataset, EM may increase or decrease the prediction performance of MNB (Nigam et al., 2000). Additionally, (Chawla & Karakoulas, 2005) observed that an EM-based technique called Common Components underperforms naive Bayes in terms of AUC given moderately large labeled data. Thus, there is still a need for a semi-supervised learning method that is fast, simple to use, and can consistently improve the prediction performance of MNB.

This paper presents Semi-supervised Frequency Estimate(SFE), a novel semi-supervised parameter learning method for MNB. We first point out that EM's objective function, maximizing marginal log likelihood(MLL), is quite different from the goal of classification learning, i.e. maximizing conditional log likelihood (CLL). We then propose SFE that uses the estimates of word probabilities, obtained from unlabeled data, and class conditional probability given a word, learned from labeled data, to learn parameters of an MNB model. Our analysis shows that both SFE and EM learn the same word probability estimates from unlabeled data, but SFE obtains better CLL values than EM in labeled training data.

SFE can be easily implemented and does not require additional meta-parameter tuning. Our experiments with eight widely used text classification datasets show that SFE consistently improves the AUC of MNB given different number of labeled documents, and also generates better AUC compared to EM for most of

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

these datasets without any loss. Finally, while EM is one of the fastest semi-supervised learning methods, our computational cost comparisons for these datasets show that SFE can be as much as two orders of magnitude faster than EM and is potentially scalable to billions of unlabeled documents.

2. Related Work

Expectation Maximization (EM) is often chosen to make use of the unlabeled data for learning an MNB model (Nigam et al., 2000). The combination of EM+MNB produces a fast semi-supervised learning method. However, (Nigam et al., 2000) point out that EM may decrease the performance of MNB when the dataset contains multiple subtopics in one class. They proposed a Common Component(CC) method using EM to address this problem. As already mentioned, (Chawla & Karakoulas, 2005) observed that while CC may improve the AUC of naive Bayes given a small number of labeled data, it may significantly underperform naive Bayes given larger labeled data.

Though many semi-supervised learning methods have been proposed in recent years, there is no dominating method in this area. (Zhu, 2008) points out that the reason for this is that semi-supervised learning methods need to make stronger model assumptions than supervised learning methods, thus the performance of semi-supervised learning methods may be data dependent. (Mann & McCallum, 2010) proposed the Generalized Expectation method and observed that the classical EM+MNB outperforms it in text classification datasets.

3. Text Document Representation

In text classification, a labeled document d is represented as $d = \{w_1, w_2, \cdots, w_i, c\}$, where variable or feature w_i corresponds to a word in the document d, and c is the class label of d. The set of unique words w appearing in the whole document collection is called vocabulary V. Typically, the value of w_i is the frequency f_i of the word w_i in document d. We use the boldface lower case letters $\mathbf{w}$ for the set of word in a document d, and thus a document can also be represented as $\{\mathbf{w}, c\}$. We use T to indicate the training data and the d^t for the t_{th} document in a dataset T. Each document d has $|d|$ words in it. In general, we use a "hat" ($\hat{}$) to indicate parameter estimates.

Text representation often uses the *bag-of-words* approach. By ignoring the ordering of the words in documents, a word sequence can be transferred into a bag of words. In this way, only the frequency of a word in a document is recorded, and structural information about the document is ignored. In the *bag-of-words* approach, a document is often stored using the sparse format, i.e. only the non-zero words are stored. The sparse format can significantly reduce the storage space.

Text classification is often considered different from traditional machine learning because of its high-dimensional and sparse data characteristics. The high-dimensional data poses computational constraints, while the sparse data means that a document may have to be classified based on the values of a small number of features. Thus, finding an algorithm which is both efficient and can generalize well is a challenge for this application domain.

4. Multinomial Naive Bayes

The task of text classification can be approached from a Bayesian learning perspective, which assumes that word distributions in documents are generated by a specific parametric model, and the parameters can be estimated from the training data. Equation 1 shows Multinomial Naive Bayes (MNB) model (McCallum & Nigam, 1998) which is one such parametric model commonly used in text classification:

$$P(c|d) = \frac{P(c)\prod_{i=1}^{n} P(w_i|c)^{f_i}}{P(d)} \tag{1}$$

where f_i is the number of occurrences of a word w_i in a document d, $P(w_i|c)$ is the conditional probability that a word w_i may happen in a document d given the class value c, and n is the number of unique words in the document d. $P(c)$ is the prior probability that a document with class label c may happen in the document collections.

The parameters in Equation 1 can be estimated by a generative parameter learning approach, called maximum likelihood or *frequency estimate* (FE) , which is simply the relative frequency in data. FE estimates the conditional probability $P(w_i|c)$ using the relative frequency of the word w_i in documents belonging to class c.

$$\hat{P}(w_i|c) = \frac{N_{ic}}{N_c} = \frac{N_{ic}}{\sum_{j=1}^{|V|} N_{jc}} \tag{2}$$

where N_{ic} is the number of occurrences of the word w_i in training documents T with the class label c. N_c is the total number of word frequencies in documents with class label c in T, and can be estimated through N_{ic}.

For convenience of implementation, the FE parameter learning method only needs to update the word frequencies N_{ic}, which can be easily converted to $\hat{P}(w_i|c)$ using Equation 2. To compute the frequencies from a given training dataset we go through each training document, and increase the entry for N_{ic} in a word frequency table by 1 or a constant. By processing the training dataset once we can obtain all the required frequencies:

$$N_{ic} = \sum_{t=1}^{|T|} f_{ic}^t \tag{3}$$

where f_{ic}^t is the number of occurrence of a word w_i in the document d^t with the class label c. Once we have N_{ic} in hand, we can also estimate $P(w_i)$:

$$\hat{P}(w_i) = \frac{\sum_{c=1}^{|C|} N_{ic}}{\sum_{j=1}^{|V|}\sum_{c=1}^{|C|} N_{jc}} = \frac{N_i}{\sum_{j=1}^{|V|} N_j} \tag{4}$$

where N_i is the number of occurrence of a word w_i in dataset.

The FE method is a generative learning approach because its objective function, shown in Equation 5, is the log likelihood (LL):

$$LL(T) = \sum_{t=1}^{|T|} log\hat{P}(c|\mathbf{w}^t) + \sum_{t=1}^{|T|} log\hat{P}(\mathbf{w}^t) \tag{5}$$

In Equation 5, the first term is called conditional log likelihood (CLL), and measures how well the classifier model estimates the probability of the class given the words. The second term is marginal log likelihood (MLL), which measures how well the classifier model estimates the joint distribution of the words in documents.

Though MLL appears to be irrelevant to the classification, the maximization of MLL often leads to a relatively better classifier given insufficient labeled data. Previous research shows that generative learning may outperform discriminative learning that discriminatively maximizes CLL given a small number of data, but may underperform discriminative learning given large number of labeled data (Ng & Jordan, 2002). Our interpretation is that learning algorithm should firstly maximize CLL, and then maximize MLL if the labeled data does not provide sufficient information. Therefore, a learning algorithm that focuses on maximization of MLL but ignores CLL may have objective function mismatch problem in a classification task.

5. Semi-supervised Learning for MNB

In practice, it is often desirable to use unlabeled documents in order to partially compensate for the scarcity of the labeled documents. While the unlabeled documents only provide $P(\mathbf{w})$ information, the MLL term in Equation 5 provides an opportunity to utilize those documents in classification.

In this section, we use subscripts l and u to distinguish the parameters estimated from labeled data T_l, unlabeled data T_u and the combination of labeled and unlabeled data T_{u+l}. We also assume that $|T_l| << |T_u|$.

5.1. Expectation-Maximization

The classical semi-supervised method for MNB is Expectation-Maximization (EM), which is known to maximize the log likelihood (LL), and in doing so uses the $P(\mathbf{w})$ information from unlabeled documents. In its initial step, Frequency Estimate acquires the N_{ic} information from the labeled documents T_l, and then uses it to estimate the $\hat{P}(c|w_i)$ required for MNB. The resulting MNB model is used to assign a "soft" (weighted) class label to unlabeled documents in T_u using predicted probability value $\hat{P}(c|\mathbf{w})$. Subsequently, the MNB model is retrained on the mixed labeled and unlabeled documents. This process is iterated until the parameters of MNB are stable.

EM's frequency count formula is:

$$N_{ic} = \sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|\mathbf{w}^t) \tag{6}$$

where $\hat{P}(c|\mathbf{w}^t)$ is the prediction obtained from the MNB model trained on the previous iteration. Since we do not have the label c for unlabeled documents, we count each word w_i in an unlabeled document $|C|$ times with the weight $\hat{P}(c|\mathbf{w}^t)$. Note that even though $\sum_{t=1}^{|T_l|} f_i^t$ should be used in Equation 6, it can be ignored because $|T_l| << |T_u|$ and the influence of the frequencies from the labeled documents on N_{ic} will be insignificant[1]. Equation 6 shows EM's use of unlabeled documents. Plugging (6) into Equation (2), we can see that EM learns $\hat{p}(w_i|c)$ in the following way:

$$\hat{p}(w_i|c) = \frac{\sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|\mathbf{w}^t)}{\sum_{j}^{|V|}\sum_{t=1}^{|T_u|} f_j^t \hat{P}(c|\mathbf{w}^t)} \tag{7}$$

Replacing the N_{ic} in Equation 4 with the one in Equation 6, the parameter $\hat{P}(w_i)$ learned by EM can be

[1] Our implementation still use the T_l, and counts the labeled documents with 1 rather than $\hat{P}(c|\mathbf{w}^t)$.

estimated as follows:

$$\begin{aligned}\hat{P}(w_i) &= \frac{\sum_{c=1}^{|C|}\sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|\mathbf{w}^t)}{\sum_{j=1}^{|V|}\sum_{c=1}^{|C|}\sum_{t=1}^{|T_u|} f_j^t \hat{P}(c|\mathbf{w}^t)} \\ &= \frac{N_i}{\sum_{j=1}^{|V|} N_j} \end{aligned} \tag{8}$$

where N_i is the number of occurrences of a word w_i in T_u. The term $\hat{P}(c|\mathbf{w}^t)$ in Equation 8 can be dropped because of $\sum_{c=1}^{|C|}\hat{P}(c|\mathbf{w}^t) = 1$ for each document d^t, and thus EM learns the information $\hat{P}(w_i)$ in T_{u+l}. Since learning $\hat{P}(w_i)$ is not related to the term $\hat{P}(c|\mathbf{w}^t)$, it can be done in the first iteration and subsequent iterations will not change the estimation of $\hat{P}(w_i)$.

Although EM maximizes MLL in unlabeled data, we show in Section 6.3 that EM+MNB, when compared to MNB, often leads to relatively inferior conditional log likelihood (CLL) on labeled training data. As discussed in Section 4, we believe that an effective semi-supervised learning method should utilize the unlabeled data without decreasing the CLL score obtained on labeled training data.

5.2. Semi-supervised Frequency Estimate

As discussed above, our main idea is to augment an MNB model by using $\hat{P}(w_i)$ from unlabeled data while maintaining the CLL score in T_l since our objective function is no longer MLL.

Our new frequency estimation method combines the word frequency obtained from the unsupervised learning with the class prediction for that word obtained from the supervised learning (hence the name Semi-supervised Frequency Estimate, or SFE) :

$$N_{ic} = \sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|w_i)_l = \hat{P}(c|w_i)_l \sum_{t=1}^{|T_u|} f_i^t \tag{9}$$

Comparing Equation 6 to 9, the difference is that EM counts the frequency by $\hat{P}(c|\mathbf{w}^t)$ while SFE uses $\hat{P}(c|w_i)_l$ (again, in 9 we drop the frequency count from labeled documents as $|T_l| << |T_u|$).

Let $F = \sum_i^{|V|}\sum_{t=1}^{|T_u|} f_i^t$, be a normalization factor which will be canceled out in Equation 10. SFE learns $\hat{P}(w_i|c)$ in the following way:

$$\begin{aligned}\hat{P}(w_i|c) &= \frac{\sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|w_i)_l}{\sum_j^{|V|}\sum_{t=1}^{|T_u|} f_j^t \hat{P}(c|w_j)_l} \\ &= \frac{\frac{\hat{P}(c,w_i)_l}{\hat{P}(w_i)_l}\sum_{t=1}^{|T_u|} f_i^t}{\sum_j^{|V|}\frac{\hat{P}(c,w_j)_l}{\hat{P}(w_j)_l}\sum_{t=1}^{|T_u|} f_j^t} \\ &= \frac{\frac{\hat{P}(c,w_i)_l}{\hat{P}(w_i)_l}\hat{P}(w_i)_u F}{\sum_j^{|V|}\frac{\hat{P}(c,w_j)_l}{\hat{P}(w_j)_l}\hat{P}(w_j)_u F} \\ &= \frac{\frac{\hat{P}(c,w_i)_l}{\hat{P}(w_i)_l}\hat{P}(w_i)_u}{\sum_j^{|V|}\frac{\hat{P}(c,w_j)_l}{\hat{P}(w_j)_l}\hat{P}(w_j)_u} \end{aligned} \tag{10}$$

Similar to EM, SFE learns $\hat{P}(w_i)$ in unlabeled data as follows:

$$\begin{aligned}\hat{P}(w_i) &= \frac{\sum_{c=1}^{|C|}\sum_{t=1}^{|T_u|} f_i^t \hat{P}(c|w_i)_l}{\sum_{j=1}^{|V|}\sum_{c=1}^{|C|}\sum_{t=1}^{|T_u|} f_j^t \hat{P}(c|w_j)_l} \\ &= \frac{N_i}{\sum_{j=1}^{|V|} N_j} \end{aligned} \tag{11}$$

The term $\hat{P}(c|w_i)_l$ in Equation 11 can be dropped because of $\sum_{c=1}^{|C|}\hat{P}(c|w_i)_l = 1$, and thus SFE learns $\hat{P}(w_i)$ in unlabeled data in the same way as EM does. While there is no guarantee that SFE will maximize MLL, Section 6.3 shows that SFE+MNB often leads to relatively superior CLL score comparing to EM in real world datasets.

The following summarizes the properties of SFE:

1. If the unlabeled documents do not affect word probability estimations obtained from labeled documents, Equation 9 will appropriately learn a classifier identical to the one learned from labeled documents alone: when $\hat{P}(w_i)_l = \hat{P}(w_i)_u$, we have $\hat{P}(w_i|c)_l = \hat{P}(w_i|c)_u$. This is because $P(w_i)_u$ is canceled out in Equation 10, and thus $\hat{P}(c, w_i)_l$ from labeled documents will determine $\hat{P}(w_i|c)$. Therefore, unlabeled documents will not influence the MNB model.

2. If in the labeled documents a word is independent of the class, then substituting $P(c)_l P(w_i)_l$ for $P(c, w_i)_l$ in Equation 10 shows that the new frequency estimate preserves this independence:

$$\hat{P}(w_i|c)_u = \frac{\sum_{t=1}^{|T_u|} f_i^t}{\sum_j^{|V|}\sum_{t=1}^{|T_u|} f_j^t} = \hat{P}(w_i)_u \tag{12}$$

3. Estimating $\hat{P}(c|w_i)_l$ from the labeled documents can be done independent of estimating $\hat{P}(w_i)_u$ for the unlabeled documents, because $\hat{P}(c|w_i)_l$ is the same for w_i in each unlabeled document. Therefore we only need to process the unlabeled documents once to obtain $\hat{P}(w_i)_u$, and then use it in conjunction with $\hat{P}(c|w_i)_l$ to estimate $\hat{P}(w_i|c)$.

6. Experiments

Section 6.2 provides empirical comparisons of the prediction performance of SFE versus EM and MNB. All experiments were carried out on a machine with a Power 5 CPU running at 1.9GHZ and 16G of RAM.

6.1. Data Description

Table 1. Data Descriptions

Dataset	Source	Class	M	$\|V\|$	DocLen
New3	Whizbang	44	9558	26833	234
Ohscal	Ohsumed	10	11162	11466	60
20news	20-news	20	18846	25747	78
Sraa	UseNet	4	73218	63966	75
Economics	Rcv1	10	119299	58473	83
Market	Rcv1	4	203926	68604	70
Government	Rcv1	23	214721	131090	105
Corporate	Rcv1	18	379157	123853	61

We used 8 large multi-class datasets to conduct our empirical study for text classification. All these datasets have been widely used in large scale text classification tasks, and are publicly available. Table 1 provides a brief description of each dataset. The number of documents in the dataset is indicated by M, while $|V|$ and *DocLen* stand for the number of words in the vocabulary and the average document length, respectively. A good semi-supervised learning method should perform better given sufficiently large unlabeled documents.

We extracted four datasets from Reuters Corpus Volume I (RCV1), (Lewis et al., 2004)[2]:"Economics","Market","Government" and "Corporate". The datasets "Sraa" and "20news" are articles from newsgroups [3]. Additionally, two largest text datasets 'New3" and "Ohscal" in WEKA are also used in our experiments (Hall et al., 2009; Forman & Cohen, 2004)[4]. 'New3" dataset contains a collection of news stories and "Ohscal" is a dataset of medical documents.

All datasets except "Sraa" and "20news" are already preprocessed by the original authors, and thus are ready to use for text classification experiments. We preprocess "Sraa" and "20news" in a similar way as (Lewis et al., 2004), converting to lower case characters, and then applying tokenization, stemming, and punctuation and stop word removal.

[2]Available at http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

[3]Available at http://www.cs.umass.edu/ mccallum/code-data.html

[4]Available at http://www.cs.waikato.ac.nz/ml/weka/

6.2. AUC and Accuracy Comparisons

The following summarizes the abbreviations used in our experiments.

1. **MNB**: Weka's MNB classifier that only uses labeled documents.
2. **SFE**: Semi-supervised Frequency Estimate proposed in Section 5.2 and implemented using Weka framework.
3. **EM+MNB**: A basic EM algorithm which uses Weka's MNB as the base classifier. The number of iterations is set to 10 to assure convergence (Nigam et al., 2000). As in the earlier discussion we use EM as an abbreviation for EM+MNB as context permits.

In our experiments, most of the multi-class datasets have imbalance class distribution, and thus we use AUC to compare the prediction performance of learning algorithms. Multi-class AUC is calculated using the average of AUC scores on all pairs of two class AUC. However, we also include accuracy comparisons for a more comprehensive study.

We report the average results obtained from 30 runs of cross validation. In each run the datasets are split into three folds. Two folds are used as training data, and one fold is used as testing data. The training data is further randomly split into a smaller labeled dataset T_l and the remaining documents make up the unlabeled dataset T_u. We also experimented with different sizes of labeled set by setting $|T_l| = \{64, 128, 256, 512\}$. Note that 64 labeled documents for a multi-class problems constitutes a very small initial labeled set. MNB uses T_l while SFE and EM+MNB use both T_l and T_u to generate an MNB model. Performance of these models is then measured on the testing data. To compare the performance of different methods we use two-tailed t-tests with a 95% confidence interval.

Tables 2 and 3 provide the corresponding detailed AUC and accuracy results achieved by each algorithm on individual datasets with labeled documents $|T_l| = \{64, 128, 256, 512\}$. A summary of our observations follows:

1. The AUC comparisons show that SFE consistently improves MNB for most datasets. In Table 2 SFE outperforms MNB for 4 to 6 datasets and never underperforms it, and the average AUC improvement of SFE over MNB is around 8%. More importantly, SFE does not decrease the AUC of MNB in any dataset given different number of

Table 2. Comparisons of AUC

number of labeled documents T_l =64

Dataset	SFE	EM	MNB
New3	74.16±2.52	60.95±2.55 •	61.87±3.29 •
Ohscal	81.02±1.93	79.23±4.16	76.78±3.63
20news	77.49±1.73	75.11±3.96	72.83±3.15
Sraa	80.82±2.25	80.34±7.25	74.22±3.49 •
Economics	76.87±4.11	67.89±3.85 •	70.05±3.74
Market	93.20±1.58	88.53±4.28	86.02±5.98
Government	72.83±2.03	60.72±2.63 •	62.07±2.34 •
Corporate	69.55±2.03	53.59±3.25 •	58.26±2.08 •
Average	78.24	70.80	70.26

number of labeled documents T_l =128

Dataset	SFE	EM	MNB
New3	80.39±2.48	65.24±2.40 •	66.82±2.91 •
Ohscal	86.35±1.23	83.37±3.38	83.01±2.55
20news	83.14±1.62	81.52±4.13	78.31±2.94 •
Sraa	86.38±1.58	82.51±6.26	78.70±2.82 •
Economics	80.59±3.41	70.29±2.01 •	72.91±2.65 •
Market	95.20±1.19	89.61±2.73 •	88.52±4.13 •
Government	76.08±1.92	62.11±2.40 •	63.66±1.98 •
Corporate	73.60±1.59	54.29±3.18 •	59.46±1.92 •
Average	82.72	73.62	73.93

number of labeled documents T_l =256

Dataset	SFE	EM	MNB
New3	84.67±2.20	70.64±2.00 •	72.69±2.52 •
Ohscal	89.74±0.76	86.76±1.96	87.32±1.64
20news	88.57±1.12	86.95±3.42	85.03±1.86 •
Sraa	89.92±1.36	84.16±5.87	83.09±1.78 •
Economics	84.12±3.16	71.86±0.99 •	74.67±1.77 •
Market	96.26±0.68	91.00±1.38 •	90.92±2.40 •
Government	78.98±1.57	62.38±1.96 •	64.82±1.87 •
Corporate	76.78±1.36	55.24±2.71 •	60.86±1.72 •
Average	86.13	76.13	77.42

number of labeled documents T_l =512

Dataset	SFE	EM	MNB
New3	87.78±1.49	77.09±2.00 •	78.65±1.45 •
Ohscal	91.82±0.58	89.19±1.35 •	90.70±0.83
20news	92.18±0.80	90.65±2.01	90.38±1.24
Sraa	93.62±0.88	90.76±6.00	86.63±1.39 •
Economics	86.48±2.54	72.74±0.82 •	77.19±1.35 •
Market	97.06±0.40	91.39±1.03 •	93.28±1.18 •
Government	81.22±1.44	64.36±1.71 •	66.72±1.54 •
Corporate	79.19±1.11	56.81±2.76 •	62.26±1.36 •
Average	88.67	79.12	80.73

• worse, and ○ better, comparing to SFE

Table 3. Comparisons of Accuracy

number of labeled documents T_l =64

Dataset	SFE	EM	MNB
New3	39.81±3.37	24.42±4.51 •	20.35± 5.68 •
Ohscal	49.83±1.97	52.92±3.92	42.86± 3.99 •
20news	34.08±2.04	41.92±5.63	26.78± 4.22 •
Sraa	62.63±1.91	81.65±4.15 ○	65.13± 3.79
Economics	61.30±2.61	51.30±4.83 •	53.24± 6.16
Market	81.62±2.20	83.56±4.26	72.33±10.10
Government	49.30±2.34	46.87±4.17	38.01± 5.50 •
Corporate	46.67±1.92	47.25±3.39	43.28± 3.70
Average	53.16	53.74	45.25

number of labeled documents T_l =128

Dataset	SFE	EM	MNB
New3	48.78±3.19	29.88±4.21 •	25.84±4.64 •
Ohscal	57.64±1.68	57.24±2.42	50.82±3.82 •
20news	43.60±2.44	51.05±6.18	34.40±4.39 •
Sraa	71.85±1.42	82.47±3.78 ○	67.46±4.18
Economics	68.08±1.65	54.49±3.53 •	58.97±5.50 •
Market	86.24±1.68	85.75±2.64	76.57±7.48
Government	56.44±1.57	49.01±4.48 •	43.07±3.59 •
Corporate	53.09±1.41	49.70±3.24	43.70±3.77 •
Average	60.71	57.45	50.10

number of labeled documents T_l =256

Dataset	SFE	EM	MNB
New3	55.77±2.88	36.76±3.69 •	33.93±4.08 •
Ohscal	63.06±1.12	60.46±1.86	58.31±2.42 •
20news	53.61±1.67	58.94±4.87	45.17±3.84 •
Sraa	78.49±0.99	83.20±4.18	71.45±2.91 •
Economics	72.24±1.37	56.96±1.57 •	62.62±4.11 •
Market	89.18±1.02	87.13±1.35	81.63±4.59 •
Government	62.30±1.06	50.11±4.19 •	44.55±3.11 •
Corporate	58.67±1.06	51.99±2.30 •	45.07±3.56 •
Average	66.67	60.69	55.34

number of labeled documents T_l =512

Dataset	SFE	EM	MNB
New3	61.54±2.68	45.31±2.66 •	43.52±2.53 •
Ohscal	66.91±0.87	63.15±1.34 •	64.42±1.49 •
20news	62.69±1.71	66.29±3.41	56.64±2.76 •
Sraa	84.14±0.78	88.83±4.02	74.62±1.68 •
Economics	74.92±0.92	57.53±1.28 •	67.74±3.03 •
Market	91.27±0.72	87.82±0.93 •	85.41±2.03 •
Government	66.49±0.75	53.21±3.74 •	46.89±2.66 •
Corporate	63.05±0.87	53.32±1.62 •	46.76±3.52 •
Average	71.38	64.43	60.75

• worse, and ○ better, comparing to SFE

labeled documents. In contrast, EM decreases the AUC of MNB up to 6% in the "Corporate" dataset given 512 labeled documents. Also, in the dataset "Market", while EM does increase the AUC of MNB 2% given 64 labeled documents, it decreases the AUC of MNB by 2% given 512 labeled documents. (Chawla & Karakoulas, 2005) also observed that a variation of EM may perform worse given larger labeled documents.

2. When comparing accuracy results, Table 3 shows that SFE again consistently improves MNB without any loss. It outperforms EM given 512 labeled documents and performs competitively with EM given smaller number of labeled documents. The average accuracy improvement of SFE over MNB is 8% given 64 labeled documents, and is increased to 10% given 512 labeled documents. In contrast, the average accuracy improvement of EM over MNB is 8% for 64 labeled documents, but is decreased to 4% given 512 labeled documents. Also, while SFE does not significantly decrease the accuracy of MNB in any datasets, this is not the case for EM. In datasets "Economics", EM decreases the accuracy of MNB up to 10% given 512 labeled documents. The unreliable performance of EM has also been observed in (Nigam et al., 2000).

6.3. Conditional Log Likelihood in Training Data

As we discussed in section 4, classification performance is related to maximizing Conditional Log Likelihood(CLL). Table 4 presents the negative Conditional Log Likelihood(CLL) generated by each algo-

rithm from 512 labeled training documents. It is clear that SFE has better (smaller absolute values) CLL values than MNB, while EM has much worse CLL values. One cannot expect a reliable performance gain from a semi-supervised classification method that decreases the CLL score of MNB in labeled training data.

Table 4. Negative Conditional Log Likelihood on Training data

Dataset	SFE	EM	MNB
New3	1.79±1.46	94.30±10.79 •	10.72±3.01 •
Ohscal	0.13±0.08	10.63± 1.64 •	0.16±0.10
20news	0.06±0.12	15.43± 4.56 •	0.09±0.23
Sraa	0.05±0.05	6.66± 4.40 •	1.35±0.39 •
Economics	0.64±0.25	79.33± 7.16 •	7.88±1.68 •
Market	0.63±0.28	19.39± 4.67 •	1.78±0.79
Government	0.23±0.14	128.30±19.58 •	11.08±2.47 •
Corporate	0.19±0.13	95.74±11.32 •	7.31±1.82 •
Average	0.46	56.22	5.05

6.4. The Impact of Size of Unlabeled Data

This section presents the impact of increasing the number of unlabeled documents on the performance of SFE. Following the 3-fold 30 runs experimental protocol, we set the number of randomly selected labeled documents to $|T_l| = 512$, the largest in our experiments, and the number of randomly selected unlabeled documents $|T_u|$ to 10^3, and for larger datasets, up to 10^4, and 10^5, as applicable. Table 5 lists the AUC of SFE and EM given different T_u.

We now summarize our observations as follows:

1. It is clear that both semi-supervised learning methods require a large number of unlabeled documents to improve their performance. As shown in Table 5, 1000 or 10,000 unlabeled document may not significantly boost the performance of MNB for both SFE and EM. However, for the "Sraa" dataset, increasing the number of unlabeled documents from 1000 to more than 10,000 results in the AUC value of 93% for SFE+MNB and 90% for EM+MNB, while MNB achieves only 87%. Fortunately, very large datasets of unlabeled documents are often abundant in real world, making scalability of semi-supervised learning even more important.

2. As shown in Equation 10, one nice property of SFE is that it only affects MNB when $\hat{P}(w_i)_l \neq \hat{P}(w_i)_u$, which indicates that unlabeled documents provide extra information compared to labeled documents. Given insufficient number of unlabeled documents, where $\hat{P}(w_i)_l \sim \hat{P}(w_i)_u$, performance of an SFE-trained model is similar to that of a model trained on labeled documents only. In the "Sraa" dataset, when there are only 1000 unlabeled documents, SFE generates an AUC similar to MNB but EM decreases the AUC of MNB up to 4%. Note that for the same dataset EM may increase AUC of MNB by 4% given sufficient large set of unlabeled documents.

Table 5. AUC when different number of unlabeled documents T_u is used for training

	SFE			EM		
Dataset	10^3	10^4	$\leq 10^5$	10^3	10^4	$\leq 10^5$
20news	89.79	92.33	92.18	86.88	90.67	90.65
Sraa	86.98	91.73	93.62	82.36	82.28	90.76
Economics	77.52	81.97	86.48	72.73	74.56	72.74
Market	93.47	96.11	97.20	87.61	90.06	92.31
Government	66.34	70.14	79.77	58.06	58.46	64.24
Corporate	62.64	66.63	75.99	50.86	50.55	56.36

6.5. Computational Cost

Section 6.4 shows that semi-supervised learning methods require large unlabeled document datasets to significantly influence the performance of MNB. Hence, a practical semi-supervised learning method needs to scale well with the number of unlabeled documents. Table 6 shows the average training time for 30 runs of SFE and EM with a single iteration, and the reported times are in milliseconds. The reported results are for a P5 1.9GHz CPU and 16G of memory after data is loaded into memory.

To the best of our knowledge, EM combined with MNB, with a single iteration, is one of the fastest semi-supervised learning methods. Our results indicate that SFE is significantly faster than EM. For example, SFE is two orders of magnitude faster than EM in the case of "New3" and "Government" datasets. The major computational cost of EM is in computing of predictions on unlabeled documents, while SFE only requires simple frequency counting. We believe that the computational advantage of SFE, combined with its simplicity, further strengthens the case for semi-supervised learning if unlabeled data abounds.

A quick examination shows that SFE performs faster than EM given datasets with a larger number of classes. As shown in Table 1, "New3" and "Government" datasets both have a large number of classes. The time complexity analysis shows that SFE is $O(M \cdot DocLen)$, where M is the number of documents and $DocLen$ is the average document length. In contrast, the time complexity of EM is $O(M \cdot DocLen \cdot |C|)$, where $|C|$ is the number of classes. This analysis explains why SFE is significantly faster than EM for datasets with a large number of classes.

Another computational advantage of SFE is that its learning from the labeled and unlabeled documents can be done independently. Thus, $\hat{P}(w_i)_u$ can be learned from unlabeled documents once, and then used to augment different MNB models without revisiting the unlabeled documents. For example, while it is computationally expensive to learn $\hat{P}(w_i)_u$ from billions documents, one could use the Google N-gram data as $\hat{P}(w_i)_u$ to augment any MNB model.

Table 6. Training Time Comparison

Dataset	SFE	EM
New3	83	26826
Ohscal	27	1476
20news	71	7716
Sraa	285	5758
Economics	571	28318
Market	849	16025
Government	1278	212092
Corporate	1912	171793

7. Conclusions

In this paper, we introduced a simple and effective semi-supervised learning method, called Semi-supervised Frequency Estimate (SFE). We compared the performance and characteristics of SFE with EM+MNB, which uses well known Expectation-Maximization algorithm for semi-supervised learning. Our experiments show that SFE significantly and consistently improves the AUC and accuracy of MNB, while EM+MNB can fail to improve the AUC of MNB. We also showed that SFE consistently produces better conditional log likelihood (CLL) values than both EM+MNB and MNB trained on the same initial labeled training set, while EM+MNB can result in a model with worse CLL than MNB. Moreover, our analysis and empirical results show that SFE has a much lower computational cost than EM+MNB, which makes it a better choice in the presence of very large unlabeled datasets.

As future work, we will investigate the extensions of this research to Bayesian networks with a fixed structure. We will extend our experiments to the use of very large public text corpora to pre-compute word probabilities for SFE.

Acknowledgment

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) for this work.

References

Chawla, Nitesh V. and Karakoulas, Grigoris J. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *J. Artif. Intell. Res. (JAIR)*, 23:331–366, 2005.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society B*, 39:1–38, 1977.

Forman, G. and Cohen, I. Learning from little: Comparison of classifiers given little training. In *Proceeding of PKDD2004*, pp. 161–172. 2004.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

Lewis, David D., Yang, Yiming, Rose, Tony G., and Li, Fan. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Mann, Gideon S. and McCallum, Andrew. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984, 2010.

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

Ng, A. and Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Andrew. N. Ng and M. Jordan, in Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002.*, 2002.

Nigam, Kamal, McCallum, Andrew, Thrun, Sebastian, and Mitchell, Tom M. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2/3):103–134, 2000.

Zhu, Xiaojin. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2008.