International Journal of Scientific Research in Information Systems and Engineering (IJSRISE)
Volume 1, Issue 2, December-2015. ISSN 2380-8128

An Alternative Method for Sentiment Classification with Expectation Maximization and Priority Aging

Y Korkmaz Yengi, Kocaeli University Computer Engineering, Kocaeli, Turkey
E-mail: korkmaz.yeliz@gmail.com

M Karayel, Kocaeli University Computer Engineering, Kocaeli, Turkey
E-mail: mehmet.karayel@kocaeli.edu.tr

S İlhan Omurca, Kocaeli University Computer Engineering, Kocaeli, Turkey
E-mail: silhan@kocaeli.eud.tr

Abstract - Sentiment classification has been an active research topic in recent years due to its potential impact on semantic based text retrieval. In sentiment classification problem, semi-supervised techniques are more important because of the huge and nonsense unlabeled data in the texts. Since the class labels are manually as- signed by experts and the text data are usually difficult to distinguish positive la- beled ones from negatives, which of the unlabeled data points should be labeled before gaining importance. Basically in semi-supervised classification, data are partially labeled and the task is to label the remaining data. In this paper, a novel two-stage semi-supervised learning model for sentiment classification which is based on Ex- pectation Maximization (EM) and priority aging algorithms is proposed. In the pro- posed approach, the priority degrees initially assigned to unlabeled data and then these are labeled due to their priorities by EM. Namely, the most suitable unlabeled data points are joined primarily to the set of labeled data by using the priority aging algorithm. The effectiveness of the proposed approach is demonstrated by using the IMDB dataset. The experiments show that the more desired results are obtained with regard to the results of conventional EM algorithm.

Keywords *– Expec- tation and Maximization (EM), Naive Bayes (NB), Priority Aging, Sentiment Classification, Semi-Supervised Learning.*

1. INTRODUCTION

Today, very large amounts of complaints and recommendations on products are available in on- line user reviews. The information extracted from user reviews can eliminate any doubts potential customers may have about a product, or can help product selection. Online reviews provide users with information about products, services and several kinds of events based on the experiences of other users. To a better and overall understanding of user reviews, researchers have been actively investigating the problem of automatic sentiment classification ofreviews.

In machine learning applications traditionally, a large amount of unlabeled data can be found without difficulty, while labeled data are very time consuming and costly to obtain as human anno- tators are required [10]. In the literature of sentiment classification, supervised methods [2, 6] are proposed firstly. Subsequently, when the huge amount of reviews on the web is considered, the supervised learning becomes inapplicable. In a situation as such, a natural question is how to enhance accuracy of prediction in classification by using both unlabeled and labeled data. The problem of this sort is referred to as semi-supervised learning.

Although many semi-supervised learning methods have been proposed in recent years, there hasn't been an accepted dominating method in this area. The reason for this is pointed out that semi-supervised learning methods need to make stronger model assumptions than supervised learning methods, thus the performance of semi-supervised learning methods depends on initially labeled data [11]. Among the proposed methods for semi-supervised learning, the classical Expectation Maximization(EM) + Naïve Bayes(NB)

Corresponding Author
Y Korkmaz Yengi
Kocaeli University Computer Engineering, Kocaeli, Turkey
korkmaz.yeliz@gmail.com

outperforms it in text classification datasets [12]. In the literature review, we are more focused on the studies which use directly or indirectly Multinomial Naive Bayes and EM algorithms.In [13], the combination of NB and EM algorithm is proposed for constructing the generative semi-supervised model. Their studies show that classification with EM performs well when the number of labeled data is small. Additionally, Common Component (CC) method is used for im- proving classification performance. In [14], they present a comprehensive study including differ- ent techniques on diversified types and amounts of labeled and unlabeled data. It is emphasized that the common-mixture model which uses Naive Bayes achieves the maximum gain in compari- son to all other semi-supervised learning techniques at small number of labeled data. In [15], Semi-Supervised Frequency Estimate (SFE) method is proposed and claimed that SFE method achieves better conditional log likelihood values in comparison to the combination of NB and EM algorithm. In [5], Naive Bayes classifier for sentiment analysis problem is used for classifying labeled and also unlabeled data iteratively.

In this paper, we have built a semi-supervised sentiment classifier for user reviews. On the con- trary, collecting labeled documents, collecting unlabeled documents is easy and inexpensive in many text or web domains, especially those involving online sources as in user reviews [13] and [16]. In user reviews, a text may contain negative sentiments, however, explain a positive opinion; or a text may contain positive sentiments, however explain negative opinion. When these kinds of texts are considered, it is obviously seen that the classification of positive and negative data points is much harder. We give precedence to the reviews which are clearly labeled as positive or nega- tive while labeling process. The reviews which are sentiment-ambiguous documents remained. This motivates the general framework we are going to develop in this paper. Apart from the re- searches proposed in the literature, in this paper, we propose a new approach for the application of Expectation Maximization (EM) algorithm by combining it with priority-aging algorithm.

The rest of the paper is organized as follows: Section 2 explains the theoretical background and the proposed system. The experimental results presented in section 3. Finally, discussion and con- clusions for future work are summarized in Section 4.

2. THEORETICAL BACKGROUND AND PROPOSED METHOD

2.1. Expectation-Maximization

Expectation-Maximization is the commonly used algorithm in semi-supervised sentiment classifi- cation to improve the classification performance and eliminate the process of labeling unlabeled data. EM is firstly used in [6] as an iterative technique for handling unlabeled data for semi- supervised sentiment classification in the literature. Using the EM algorithm, unlabeled data are selected and used to improve the performance and robustness of the classifier which is constructed on the small set of labeled data. During the implementation of EM, the parameters of the classifier which is constructed for the labeled data are used as the initial values for an iterative model.

Locally maximum parameters of the model which is formed for the unlabeled data are found by using EM algorithm. One of the undesired case of the EM algorithm is that finding local maxi- ma instead of global.

The EM algorithm consists of the E-step in which the expected values of the missing sufficient statistics given the observed data and the current parameter estimates are computed, and the M- step in which the expected values of the sufficient statistics computed in the E-step are used to compute complete data maximum likelihood estimates of the parameters [6]. In our implementa- tion of the EM algorithm with the Naive Bayes classifier, the learning process using unlabeled data proceeds as follows:

i).Train the classifier using only labeled data.

ii). Classify unlabeled examples, assigning probabilistic labels to them.

iii). Update the parameters of the model. Each probabilistically labeled example is count ed as its probability instead of one.

iv).Go back to (2) until convergence.

2.2. Priority Aging

The priority scheduling is one of the CPU scheduling schemes in operating systems. In priority scheduling, each process is assigned a base priority defined by the users; they can also be changed dynamically to

prevent starvation. Equal priority processes are scheduled in first come first served order. However, the disadvantage of this scheme is that, it can easily starve processes. Thus to solve this problem, the priority-aging is used. Aging is a technique of gradually increasing or de- creasing the priority of processes that wait in the system for a long period of time [17].

In this paper, the aging approach is used for selecting the text which is going to be an input of EM algorithm. If a text has a high EM probability score then it is classified as positive or negative. However, if a text not has a high EM probability value then the priority of the text is decreased, and it is not labeled. In other words, it is given a change to be classified with a stronger prior prob- ability assumption.

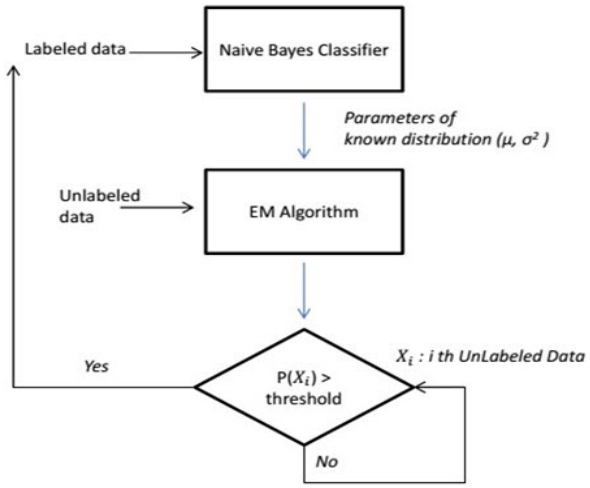

Figure. 1. The flowchart of proposed Priority Aging Method

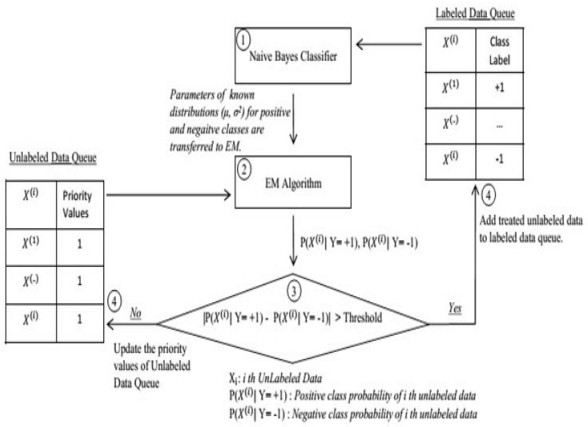

Figure. 2. The flowchart of proposed Priority Aging Method

2.3. EM with Priority Aging

In this research, a semi-supervised classifier is implemented because collecting unlabeled exam- ples or documents is easy and inexpensive from many text or Web page domains, especially those involving online sources [13] and [16].

To evolve the classical semi-supervised sentiment classification which is developed on the ba- sis of NB and EM, priority aging is implemented in the EM algorithm. Proposed priority aging method is presented in Fig.1 and Fig.2. In the priority aging method, the unlabeled data is added to the labeled data queue more consciously.

Let's assume that, the unlabeled data is represented by $X = \{x^1, x^2, \ldots, x^m\}$, where m is num- ber of the data points. To implement the priority aging method, first of all, the priority values ini- tially assigned 1 to unlabeled data and then these are labeled due to their priorities by EM. The priority values are selected between 0 and 1. Namely, 1 specifies the highest and the 0 determines the lowest priority. Together with EM algorithm, the priority values for each unlabeled data (x^i) are calculated by using equation 1. In other words, the absolute value of the difference between the probability values of positive and negative classes which are calculated by EM is used for priority value.

$$x^i_{priority} = | P(x^i \mid y = +1) - P(x^i \mid y = -1) | \qquad (1)$$

Where x^i represents "unlabeled data i", $P(x^i \mid y = +1)$ represents "positive class probability of unlabeled data i" and $P(x^i \mid y = -1)$ represents "negative class probability of unlabeled data i".

After the preprocessing of data, firstly, the NB parameters are estimated from the labeled doc- uments. Secondly, the classifier is used to assign probabilistically-weighted class labels to each unlabeled document by calculating expectations of the missing class labels. Consequently, the new classifier parameters are estimated using all the documents which are extended with the newly labeled documents. Unless the priority value of each unlabeled data exceeds the threshold value, the unlabeled one is added to the end of unlabeled data queue for regenerating priority value.

The most critical part of the proposed method is defining the the threshold value. The threshold value is used for adding the unlabeled data x^i to the labeled data queue ac-

cording to the priority value whether greater than threshold x^i ▯ri▯rity> threshold or not. The threshold value is deter- mined empirically and the empirical results are given in section 3.3.

After end of each EM iteration, the unlabeled data queue which has greater probability values than threshold value are added to the labeled data queue and NB is trained using extended data. Conversely, if the priority value of unlabeled data is less than the threshold value, the data point is added to the end of unlabeled data queue. Consequently, priority aging method provides that the most convenient unlabeled data points are primarily joined to the labeled data set.

3. Experimental Settings

3.1. Data Description and Transformation

In this study, the IMDB dataset [19] is used in experiments. The dataset consists of 1400 pre- classified movie reviews, 700 positive and 700 negative. This dataset is previously used for vari- ous classifiers (Maximum Entropy, SVM, and Naïve Bayes) to determine the correct polarity rat- ings for movie reviews [18]. At the beginning we select %10 of 1400 reviews that 70 positive sample, 70 negative samples for learn classifiers.

As the first step of the preprocessing, the review documents are cleaned from any HTML tags. The text was cleaned from non-alphabetic signs and abbreviations. As for stop words, a stop word list is constructed on the basis of several available standard stop word lists, with some changes related to the specific characteristics of the data. For example, the words such as film, movie, ac- tor, actress, and scene are non-informative in movie reviews data. These are considered as stop words because they are movie domain specific words.

Select words with specific part-of-speech, such as verb, adjective, and adverb are used in de- veloping the classifier [9]. In addition, stemming is performed on the documents to reduce the redundancy. Finally, the number of features was reduced from 23450 to 16614. After the term- document matrix is constructed, the term (or feature) weighting process is realized by applying term frequency–inverse document frequency (TF-IDF) method after that at the beginning separat- ed 1260 sample of reviews ordered according to length of words. As final we have matrix in size 16614 features and 1260 samples.

3.2. Performance Evaluation

The performance metrics used to evaluate the classification results are precision, recall and F- measure. Precision is a measure of the ability of a classification model to present only relevant items. Recall is a measure of the ability of a classification model to present all relevant items. F- measure is the weighted harmonic mean of precision and recall. The evaluation measures in equa- tions 2, 3 and 4 are defined due to the confusion matrix in Table 1.

Table 1. *Confusion Matrix*

	Classified positive	Classified negative
Actual positive	tp	fn
Actual negative	fp	tn

Where, tp (true positive): the number of correct classifications of the positive examples, fn (false negative): the number of incorrect classifications of positive exam-ples, fp (false positive): the number of incorrect classifications of negative examples, tn (true negative): the number of correct classifications of negative examples.

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$F-measure = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

3.3. The Threshold for Priority Aging Process

In this study, determining the threshold value is an important issue for the effectiveness of priority aging process. During the determination of threshold value, since the 0.5 is not illustrative for separating the positive and negative points, it is initially assumed that X^i score for each un- labeled data x^i must be

greater than 0.6. And then, 0.7, 0.8, 0.9 threshold values are implemented respectively. The classification accuracies at each threshold value are given in Table.2. It is con- cluded that the accurate classification results are obtained with 0.8 threshold value.

Table 2. *Classification performances at various threshold values.*

Threshold Value	Precision	Recall	F-Measure
0.6	0.5	0.5	0.5
0.7	0.57	0.57	0.57
0.8	0.73	0.73	0.73
0.9	0.67	0.66	0.66

3.4. Experimental Results

In this section, the results of several experiments are reported to assess the performance of the semi-supervised classification. The proposed classifier is ran on the movie reviews and the per- formance results are compared with the conventional EM algorithm.

Table 3 compares the classifier performances resulting from the classification on both conven- tional EM and EM with priority aging. The classification performance of priority aged EM is bet- ter than conventional EM algorithm. While implementing the proposed algorithm, it is observed that the presentation of data is an important issue. The better results are achieved when the data is presented to the algorithm due to the descending review length. The reason for this can be ex- plained as follows. The prior distribution of the data is a key part of Bayesian inference which is combined with the probability distribution of new unseen data to yield the posterior distribution. The posterior distribution is used for future inferences and decisions. Namely, the existence of a strong prior assumption for any problem can provides to construct more effective Bayesian infer- ence models. In this study, long documents which contain more terms are primarily presented to the NB classifier to construct more robust Bayesian model.

Table 3. *The performances of EM and EM with Priority Aging Algorithms*

	EM	EM + Priority Aging
Precision	0.49	0.8
Recall	**0.5**	**0.8**
F-Measure	0.49	0.8

ing method is more effective compared to the pure EM implementation in semi-supervised classification. While the conventional EM classification has nearly 0.5 accuracy rate, the Priority Aging with EM classification has 0.8 accuracy rate.

4. Conclusions

In semi-supervised learning, when the given data are usually difficult to distinguish positive la- beled ones from negatives as in sentiment classification problem; which of the unlabeled data points should be labeled before gaining importance due to labeled data must represent the class dis- tributions accurately. From this point of view, the priority degrees initially assigned to unlabeled data and then they labeled due to their priorities by EM iteratively. The priority aging algorithm provides that the most suitable unlabeled data points are primarily joined to the labeled data.

In this paper, an alternative semi-supervised classifier is proposed based on EM and priority aging algorithm. When EM with NB is implemented, the performance of the classification is not satisfactory. We compared the performance of EM with Priority Aging to conventional EM algo- rithm, which is used with NB. Our experiments demonstrate that priority aging significantly im- proves the accuracy of EM+NB classifier.

As a future work, we will investigate the performance of the proposed classifier with different datasets.

5. References

[1] V. Narayanan, I. Arora and A. Bhatia, "Fast and Accurate Sentiment Classification Using An Enhanced Naive Bayes model," in Intelligent Data Engineering and Automated Learning(IDEAL 2013), Lecture Notes in Computer Science, 2013, pp 194-201.

[2] K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell, "Text classification from labeled and unlabeled documents using EM. Machine Learning," in Special issue on information retrieval, 2000, pp 103-134.

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

[4] J. Su, J. Sayyad-Shirabadand, S. Matwin, "Large Scale Text Classification using Semi-Supervised Mul- tinominal Naive Bayes," in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011, pp. 97-104.

[5] P. Y. Shashi Kishore and K. Brahma Naidu, "Sentiment Analysis Using Semi-Supervised Naive Bayes Classifier," in International Journal

Of Innovatıve Technology And Research (IJITR), vol. 1, 2013, pp. 478 – 482.
[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," in Journal of the Royal Statistical Society, B. vol. 39, 1977, pp. 1–38.
[7] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning tech- niques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 2002, pp. 79-86.
[8] J. C. Na, H. Sui, C. Khoo, S. Chan and Y. Zhou, "Effectiveness of simple linguistic processing in auto- matic sentiment classification of product reviews," in I.C. McIlwaine (Ed.), Knowledge Organization and the Global Information Society, Proceedings of the Eighth International ISKO Conference, 2004, pp. 49-54. Wurzburg, Germany: Ergon Verlag.
[9] X. Zhu, Z. Ghahramani and J. Lafferty, "Semisupervised learning using Gaussian fields and harmonic functions," in ICML-03, 20th International Conference on Machine Learning, 2003.
[10] Z. Xiaojin, "Semi-Supervised Learning Literature Survey Computer Sciences," in TR 1530 University of Wisconsin – Madison Last modified, 2008.
[11] S. Mann, Gideon and A. McCallum, "Generalized expectation criteria for semi-supervised learning with weakly labeled data," in Journal of Machine Learning Research, vol. 11, 2010, pp. 955–984.
[12] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Learning to classify text from labeled and unlabeled documents", in AAAI-98, 1998.
[13] M. K. Chawla, J. F. Guzowski, V. Ramirez-Amaya, P. Lipa, K. L. Hoffman, L. K. Marriott and et al. "Sparse, environmentally selective expression of Arc RNA in the upper blade of the rodent fascia dentata by brief spatial experience," in Hippocampus vol. 15, 2005, pp. 579–586.
[14] J. Su, J. Sayyad-Shirabad, S. Matwin, "Large Scale Text Classification using Semi-supervised Multino- mial Naive Bayes," in Appearing in Proceedings of the 28 th International Conference on Machine Learn- ing, Bellevue, WA, USA, 2011.
[15] B. Liu, M. Hu and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web", in Proceedings of WWW-05, 2005, pp.342-351.
[16] Design and Implementation of Operating System, Er. Vivek Sharma, Er. Manish Varshney, Shantanu Sharma, Laxmi Publications, Ltd.
[17] E. Haddi, X. Liu, Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," in Information Tech- nology and Quantitative Management Procedia Computer Science vol. 17, 2013, pp. 26 – 32.
[18] Data sets : http://www.cs.cornell.edu/people/pabo/movie-review-data/
[19] S. Ang, D. Li, X. Song, Y. Wei, H. Li, "A feature selection method based on improved fisher's discrimi- nant ratio for text sentiment classification," in Expert Systems with Applications, vol.38, 2011, pp. 8696– 8702.