Volume 5, Issue 6, June 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Self Learning Naïve Bayes Multi Label Classifier for Analyzing Student Educational Interest

B. Narendra, T. Kavitha, P. Surya Chandra, T. Bala Krishna
Asst. Professors, Dept of CSSE, SVEC,
Tirupati, A.P, India

Abstract: In the technology revolution of world, social media is playing a major role in connecting people. People with same kind of thinking attitude may get connected through the social media easily, and they respond to different issues, and comment on others posts. Gathering all these posts and processing them to get conclusions on different issues happening in the society, is the challenging task. Here in this paper we tried to gather engineering student's opinions on their education. Traditional supervised machine learning algorithms are not appropriate for processing the large volumes of data. In this paper we propose, a new semi-supervised Naïve Bayes learning method for sentiment classification. However the semi-supervised learning methods are unreliable, they can easily classify the unlabeled data with very less labeled data set. Accuracy of the classifier is also more when compared with traditional methods.

Key Words: Semi-Supervised Learning, Bayes Rule, Apriori, Data Set, Feature Set.

I. INTRODUCTION

Machine learning is a wide subfield of artificial intelligence. It includes algorithms and methods that allow the computers to learn. Learning techniques play a key role in data mining process. In the fast growing technology world the amount of data producing is increasing from day to day. Social media is one of the sources of data. Traditional learning methods which are the supervised classifiers and unsupervised classifiers may work better with the available data. Many supervised learning algorithms are proposed for text data classification such as support vector machine (SVM), Naïve Bayes classifier, Decision tree method, K-nearest neighbor classifier etc.

Semi-supervised learning is the process of assigning more accurate class labels for the unlabeled data with small amount of training data set. This learning will yield good performance by dynamically training the data set, through which unlabeled data can be classified more accurately. The semi-supervised learning algorithms can be used in a variety of situations by identifying as opposed to specifying the relationship between labeled and unlabeled data. These can give improved accuracy of classification when the test data can reconstruct the optimal classification boundary. Some popular semi-supervised learning methods include: Self-training, Graph based models, Co-training, and Multi-view learning.

In this paper, we proposed an efficient semi-supervised Naïve Bayes classifier which effectively works on small set of training data set and large amount of test data set which is unlabeled. Here we have taken the case of student views, approaches and concerns about learning process. In this initially we find the frequent one item sets based on the data set which we can call as positive feature set.The algorithm is recursively applied to find multiple item sets which contains both labeled and unlabeled data items. While applying the algorithm we set certain limits for the minimum support and confidence. Now we will give this feature set to the Naïve Bayes classifier which will assign class labels to the different posts. Here we use a cooperative learning mechanism.

II. LITERATURE

Positive feature set generation: Apriori algorithm is one of the most powerful and efficient algorithm used for association rule mining. The apriori property itself says that item sets will be generated based on the prior knowledge about the data items it is about to use. The uses a tree structure to generate the candidate keys, and it follows breadth first mechanism. The algorithm is having two steps namely, the join step and pruning step. Here while applying the apriori algorithm to generate the positive feature set we maintain certain limits for confidence and minimum support.

Consider a transactional database D with, minimum support count 's'. C_l is the candidate set for level l. count[x] represents candidate set C.

Algorithm: Apriori(D,s)

```
$L_1 \leftarrow$ {large 1 item sets}
L $\leftarrow$ 2
While $L_{l-1} \neq 0$
$C_l \leftarrow \{a \bigcup \{b\} \mid a \in L_{l-1} \wedge b \in \bigcup L_{l-1} \wedge b \notin a\}$
For a transaction t $\in$ D
$C_t \leftarrow \{x \mid x \in C_l \wedge x \subseteq t\}$
```

For candidate sets $x \in C_t$

Count[x]← count[x]+1

$$L_l \leftarrow \{x \mid x \in C_l \wedge count[x] \geq \in\}$$

$$l \leftarrow l+1$$

$$return \bigcup_l L_l$$

Semi-supervised Classifier: Bayes rule is applied for assigning class labels to the unlabeled data items based on the training set, it assumes that the entire attribute are mutually and conditionally independent of each other. Bayes rule in general consider conditional probabilities rather than joint probabilities. It is quite simple to define P(A|B) without considering P(A,B).

$$P(A|B) = P(B|A)P(A)/P(B)$$

Consider the case A= (A_1,A_2)

Now $$P(A|B) = P(A_1, A_2 | B) = P(A_1 | A_2, B)P(A_2 | B) \\ = P(A_1 | B)P(A_2 | B)$$

This can be rewritten as $P(A_1, A_2, ...A_n | B) = \prod_{i=1} nP(A_i | B)$

Assume that here B is a discrete valued variable and $A_1,\ldots A_n$ are any discrete or real valued attributes, the probability that B will get the nth possible value is, $$P(B = b_n | A_1...A_n) = \frac{P(B = b_n)P(A_1...A_n | B = b_n)}{\sum_I P(B = b_i)P(A_1...A_n | B = b_i)}$$

Let A_i is conditionally independent on B then $$P(B = b_n | A_1...A_n) = \prod_i \frac{P(B = b_n)P(A_i | B = b_n)}{\sum_j P(B = b_j)\prod_i P(A_i | B = b_j)}$$

In our experiment we try to calculate both positive and negative probabilities about a particular measure. As there may be both positive and negative opinions about a particular issue it is convenient to consider both. This renders the method of Naïve Bayesian in which we assume for large data sets. We can compute the probability of a word by observing both the counts of positive estimation and negative estimation. From this we can observe that there is no linking between one word to other word, and this makes the trained samples to train other data items.

III. IMPLEMENTATION

Accumulation of learning: An ensemble learning approach is used here for assigning more appropriate class labels to the frequent item sets generated from the Apriori algorithm. An ensemble classifier is a supervised classification method, in which different classifiers are applied independently and finally the results are combined to get more appropriate results for the classification methods.

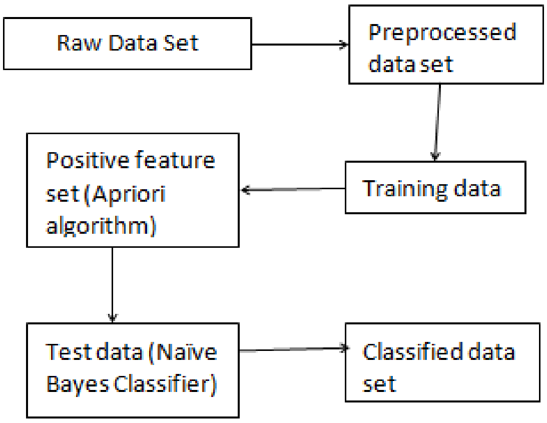
Raw Data Set
Preprocessed data set
Training data
Positive feature set (Apriori algorithm)
Test data (Naïve Bayes Classifier)
Classified data set

Fig: Architecture of the proposed method

Data set collection and Preprocessing: In general pupil will have different styles and views in expressing their feelings on a particular issue. The language style and spellings of words they used may also vary from person to person. Hence we first collect the different posts on a particular issue from the social media just by posting queries or by observing the different pols organized in the social media. Now the data we collected will obviously contain large volumes of unnecessary data and noisy data also. Next we have to preprocess the data set. Here in preprocessing we majorly concentrated on correcting misspelled words, removing extra characters in the key words, removing gestures and replacing with synonyms.

Constructing feature set: From the data set constructed after preprocessing, we first apply the apriori algorithm for constructing frequent one item sets for the related keywords, with a minimum support count and confidence values. Once we apply the Naïve Bayes classifier for the frequent one item set to assign class labels the apriori algorithm is reapplied for the classified data set to construct two item set the unlabeled data items are assigned class labels by using the semi-supervised naïve Bayes classifier.

```
Algorithm: Constructing Feature set
Input: labeled positive data set LPD, unlabeled data set UD
Output: positive feature set PF
    1. Initially PF= $\Phi$ , UFI= $\Phi$
    2. Apply Apriori algorithm with minimum support and confidence
    3. Generate FI
    4. From i=1 to Size(FI)
    5. Add FI to PF
    6. From i=1 to size(PF)
    7. For every UFI, if $PF_i \in$ UFI then add UFI to PF
    8. End
```

Labeling Training Set: Now we assign class Alabels to the unlabeled data set based on the training data set, here we use the frequency count of the words for assigning more relevant class labels for the unlabeled data set.

Constructing the final classifier: In the final step we give the newly labeled data set to the Naïve Bayes classifier. Here we use the ensemble mechanism for assigning the class labels. Repeatedly apply the Bayes classification algorithm for the unlabeled data set by adding 80% of the labeled data set. Now assign class labels using multi labels. Here one particular keyword may come under two different groups.

IV. EXPERIMENTAL ANALYSIS

Analysis of classification: For our experiment we have taken the data set from face book, we posted some posts related to engineering students and collected the comments from those posts. The comments are taken around more than 1000 engineering students from various engineering colleges.

Our main goal in this experiment is analyzing the engineering student's mental feelings by collecting their views about the present education system and problems they are facing. Here we collected the posts and from those we extracted keywords related to different feelings such as over work load, lack of entertainment, pressure during exams etc.

The data set contains the comments with a #tag, first we remove the #tags as an initial step of data preprocessing. Then remove all non-letter symbols, punctuations and gestures. Now apply the replacement of synonyms so that related words can be replaced with appropriate words.

Multi label classification: In simple way applying the Naïve Bayes classifier for assigning multiple labels means applying it for assigning single label for multiple times. Here the labels are assigned based on the binary relevance of the keywords.

Suppose there are N number of words W= $\{w_1, w_2, \ldots w_n\}$ and L number of class labels C= $\{c_1, c_2, \ldots c_l\}$. A particular word is with different probabilities in different classes. Then the word will be assigned a class label based on the likelihood value.

The probability of a particular word to have a certain class label is $P(W_n \mid C) = \dfrac{m_{w_n c}}{\sum_{n=1}^{N} m_{w_n c}}$

Now the probability of that particular word to be assigned other class label is $P(W_n \mid C^{|}) = \dfrac{m_{w_n c^{|}}}{\sum_{n=1}^{N} m_{w_n c^{|}}}$

Instance based evaluation: Assume that they are actually R number of classes but the data set is classified into S different classes by the classifier. Let there are a total of M data sets then,

$$Accuracy = \frac{1}{M}\sum_{i=1}^{M}\frac{R_i \cap S_i}{R_i \cup S_i}$$

$$\Pr ecision = \frac{1}{M}\sum_{i=1}^{M}\frac{R_i \cap S_i}{S_i}$$

$$\mathrm{Re}\,call = \frac{1}{M}\sum_{i=1}^{M}\frac{R_i \cap S_i}{R_i}$$

Class Label based evaluation:in the instance based evaluation we calculate the accuracy and precision based on each data item, here we can compute in another way by using the class labels available in the current instance of the data set.

$$Accuracu = \frac{FP + FN}{N}$$

$$\mathrm{Pr}\,ecision = \frac{\mathrm{Re} \cap Ri}{Ri}$$

$$\mathrm{Re}\,call = \frac{\mathrm{Re} \cap Ri}{\mathrm{Re}}$$

$$F - Score = 2 \times \frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{(\mathrm{Pr}\,ecision + \mathrm{Re}\,call)}$$

Experimental Results: For evaluating the performance of the classifier in our experiment we considered 500 data items.

Table 1: Performance of the classifier before replacing the keywords with synonyms.

Metric	Naïve Bayes	Ensemble Naïve Bayes
Precision	131/237=0.552	139/237=0.586
Recall	131/124=1.05	139/122=1.139
F-Score	0.7235	0.77385

Table 2: Performance of the classifier after replacing the keywords with Synonyms.

Metric	Naïve Bayes	Ensemble Naïve Bayes
Precision	134/237=0.565	148/237=0.624
Recall	134/126=1.063	148/126=1.174
F-Score	0.737	0.8148

V. CONCLUSION

Learning student’s problems in the current education system to resolve the issues and make them fell interested in education, we collected student’s opinions on over study work load, lack of entertainment, and work tension during exams time etc. The Multi label Semi-Supervised Naïve Bayes classifier given some good results in this analysis process, when compared with the simple supervised learning mechanism. The classification accuracy is improved with the ensemble classification mechanism.The semi supervised learning methods can be used for better assigning class labels to test data with less training data set.

REFERENCES

[1] R. Ferguson, “*The State of Learning Analytics in 2012: A Review and Future Challenges*,” Technical Report KMI-2012-01, Knowl- edge Media Inst. 2012.

[2] R. Baker and K. Yacef, “*The State of Educational Data Mining in 2009: A Review and Future Visions*,” J. Educational Data Mining, vol. 1, no. 1, pp. 3-17, 1, pp. 7-15, 1997.

[3] S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez, “*Microblogging in Classroom: Classifying Students’ Relevant and Irrelevant Questions in a Microblogging-Supported Classroom,*” IEEE Trans.

[4] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Law- son, “‘*I Can’t Get No Sleep’: Discussing #Insomnia on Twitter,*” Proc. ACM Ann. Conf. Human Factors in Computing Systems, pp. 1501-1510, 2012.

[5] M.J. Culnan, P.J. McHugh, and J.I. Zubillaga, “*How Large US Companies Can Use Twitter and Other Social Media to Gain Busi- ness Value*,” MIS Quarterly Executive, vol. 9, no. 4, pp. 243-259, 2010.

[6] M.E. Hambrick, J.M. Simmons, G.P. Greenhalgh, and T.C. Green- well, “*Understanding Professional Athletes’ Use of Twitter: A Content Analysis of Athlete Tweets,*” Int’l J. Sport Comm., vol. 3, no. 4, pp. 454-471, 2010.

[7] D.M. Romero, B. Meeder, and J. Kleinberg, “*Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags*, 2009. Learning Technologies, vol. 4, no. 4, pp. 292-300, Oct.- Dec. 2011.

[8] R. Bandari, S. Asur, and B.A. Huberman, “*The Pulse of News in Social Media: Forecasting Popularity*,” Proc. Int’l AAAI Conf. Weblogs and Social Media (ICWSM), 2012.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo, “*Earthquake Shakes Twit- ter Users: Real-Time Event Detection by Social Sensors*,” Proc. 19th Int’l Conf. World Wide Web, pp. 851-860, 2010.

[10] W. Zhao, J. Jiang, J. Weng, J. He, E.P. Lim, H. Yan, and X. Li, “*Comparing Twitter and Traditional Media Using Topic Models*,” Proc. 33rd European Conf. Advances in Information Retrieval, pp. 338- 349, 2011.

[11] D. Davidov, O. Tsur, and A. Rappoport, "*Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*," Proc. 23rd Int'l Conf.

[12] Ishtiaq Ahmed, Donghai Guan, Teachoon Chung "*A novel semi-supervised learning for SMS classification.*" Proceedings of the 2014 International Conference on Machine Learning and Cybernetics, Lanzhou, 13-16 July, 2014.

[13] RakeshAgrawal and Ramakrishna Srikanth " *Fast algorithms for mining association rules in large databases*" proceedings of the 20th International Conference on Very Large Data Bases, VLDB, 487-499, Santiago, Chile, September 1994.

[14] E. Boiy, P. Hens, K. Deschacht, M. – F. Moens, "*Automatic Sentiment Analysis in On-line Text*", Proceedings of the 11th International Conference on Electronic Publishing, pp. 349-360, 2007.