Large Data Set Mining With Naive Bayes Classifier and Association Rule

U Yin Moe Win

Lecturer,University of Computer Studies, Kalay

ABSTRACT

Nowadays, large data contains limitless business opportunities. Companies and organizations begin to analyze their stored data to predict their potential customers and business decisions using data mining tools as examples Naïve Bayes Classifier, Association Rule Mining, Decision Tree and other famous algorithms. An mining processing output produced accurate classification result it may help companies leading in its industry. Companies seek to find feasible business intelligences to obtain reliable prediction results. In this paper proposed method used an association rule mining to improve Naïve Bayes Classifier. Naïve Bayes Classifier is one of the famous algorithm in large data classification but based on an independent assumptions between features. Association rule mining is standard and useful for discovering relations between inputs in big data analysis. Proposed system use bank marketing data set to illustrate in this work. In general, this work is helpful to all the business data set.

Keyword: - *Business Data Set; Association Rule Mining; Naïve Bayes Classifier; Apriori algorithm.*

1. INTRODUCTION

Before making any business decisions, business intelligence is necessary and important. Business intelligence is the set of techniques for analyzing data and presenting actionable information. The benefit of Business intelligence include: accelerating and improving decision making; optimizing internal business processes; increasing operational efficiency; driving new revenues; and gaining competitive advantages over business rivals [1]. Business intelligence can also predict market trends, so companies need feasible. Business intelligence is to process their data and make decisions. . Business intelligence transfers raw data into meaningful and useful information for business analysis purposes. Since the raw data is for business processing, we call it business data. Business data save as many columns and each column represents an attribute. Proposed method use Naïve Bayes Classifier which is the most famous classification technology to evaluate business data and predict tendency. In many approaches, Naive Bayes Classifier always does surprisingly well, so it has been widely used in classification area. Naive Bayes has three popular models: Bernoulli model, Binarized model and Multinomial model. Bernoulli model is used when the absence of a particular word matters. Binarized model can be used when words don't play a significant role. Multinomial model is used when the multiple occurrences of words matters [2]. This paper tries to use association rule mining to improve the prediction accuracy of Naive Bayes Classification. Association rule finds relationship between seemingly unrelated data among a large set of data items. Therefore, association rule can always find a potential relationship between features and bring surprise to us. Apriori algorithm is one of the influential algorithms for frequent item sets mining and association rule learning. The classic example is the famous "Beer" and "Diaper" association problem that is often mentioned in data mining books and tutorials. The business data set we used is provided by UCI Machine Learning Repository. We use bank marketing data set, which is marketing campaigns of a Portuguese banking institution [3]. The purpose to this work is to predict the success of selling bank long-term deposits based on the given banking information. The training data set contains 20 input attributes and 40 thousand instances, the testing data set contains 20 input attributes and 4 thousand instances. Besides, we propose to use

Hadoop in our work. Hadoop is an open source framework for distributed processing of very large datasets. The reason why Hadoop is well suited to big data analysis is because Hadoop works by breaking the data into pieces and assigning each "piece" to a specific node for analysis [4]. Therefore, Hadoop framework is widely used in data training and big data analysis. And also, Hadoop Training and Certification Courses are becoming more and more popular with each passing year.

2. RELATED

Naive Bayes Classifier was wildly used in many data mining approaches. Most of these approaches focus on text based data set and classification. It's because Naive Bayes Classifier was suggested to give good results and used in many filtering software. The first mail-filtering program using Naive Bayes Classifier is Jason Rennie's ifile program. And then, the first scholarly publication on Bayesian spam filtering was proposed by Sahami et al. [5]. In 2002, Graham decreased the false positive rate to use as a single spam filter [6] [7]. Also a number of solutions use cluster as a part of text detection, such as KNN algorithm [8] and SVM classification [9]. Ketari's work illustrates the image spam filtering techniques [10]. Image spam filters use algorithms such as SIFT, TR-FILTER and NDD. Deshmukh proposed a spam filtering system using Sobel operators and AOCR [11] for filtering text and image based emails. Since research for spam needs a large dataset, some techniques and frameworks such as Hadoop, MapReduce and HDFS are become popular. Tran Ho proposed fingerprinting technique combines with sim-hash algorithm to detect spams. This is a novel similarity-based method that implements by using Hadoop framework [12]. We use the method of association rule into our work, and this is a frequency-based method that also implements on Hadoop framework. There are also many approaches using association rule in data mining. SM Kamruzzaman proposed solutions to classify text documents and design algorithm using Association Rule and Naive Bayes Classifier [13]. This combination helps Naïve Bayes Classifier to process some related data instead of independent data. Irina Rish proposed an analysis about Naive Bayes performance [14], which analyze attribute effects in Naïve Bayes. This approach shows an inspiration to obtain different performance by decreasing attributes. In our work, we use most of the above techniques and algorithms to solve Business data classification.

3. TECHNIQUE BACKGROUND

Our experiment uses three basic techniques: Hadoop training, Naive Bayes Classifier and Association Rule Mining. uses three basic techniques: Hadoop training, Naive Bayes Classifier and Association Rule Mining.

3.1 Hadoop & Mapreduce Training

The unique storage method in Hadoop is based on distributed file system called HDFS. Our training data are saved in HDFS and we propose to use MapReduce framework to handle the amount of words. MapReduce approach is to use <key, value> pairs and the groups that will be received in the reduce function will be grouped by the key.

Map: <k1, v1> → list<k2, v2>
Reduce□ <k2, list (v2) > → list (k3, v3)
Then, we can get fast and accurate results by this framework, and the procedure of MapReduce is shown in Fig.1.

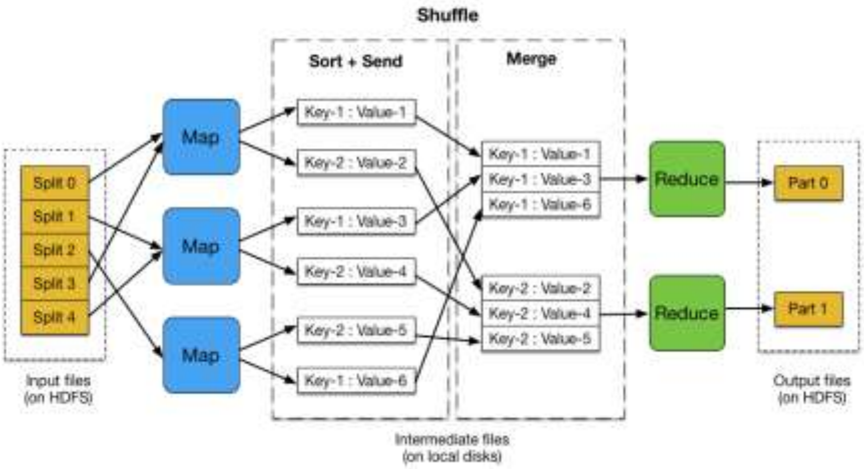

Fig. 1 Procedure of MapReduce

In training phase, we need to count every word in the data set. In order to process a large data set efficiently, Hadoop is a good technique to be employed.

3.2 Naïve Bayes Classifier

Naive Bayes Classifier is based on Bayes theorem and it has also exhibited high accuracy and efficiency when applied to large database. Naive Bayes Classifier assumes that the effect of each attribute on a given class is independent from the other attributes. Thus, it simplify the computation and considered as “naive”. Naive Bayes Classifier can be given by:

$$\mathrm{V}_{NB} = \arg\max P(\mathrm{C}_j)\prod P(\mathrm{x}_i \mid \mathrm{C}_j) \qquad (1)$$

The first term can be estimated based on the fraction of each class in the training data. The next term represent contribution probability for each class. The following equation is used to estimate contribution for text (non-continuous) variable:

$$P(\mathrm{x}_i \mid \mathrm{C}_j) = \frac{n_k + 1}{n + vocabulary} \qquad (2)$$

where nk is number of occurrences that variable is found among the attribute, n is the total number of variables in the attribute, vocabulary is the total number of distinct variables in the attribute.

If attribute contains a continuous variable, we use probability density rather than a probability. The following equation is normal distribution probability density function:

$$P(x_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad (3)$$

where μ is the mean of a set of continuous data and __ is variance for each variable in the attribute.

3.3 Association Rule Mining and Apriori Algorithm

Association rule mining helps uncover relationships between seemingly unrelated data in a relational database. It can be defined as following: most popular association rule is Apriori algorithm, which is used to extract frequent item sets from large database and get the association rule for discovering the knowledge. Apriori algorithm is an influential algorithm for mining frequent item sets. Since the Algorithm uses prior knowledge of frequent item set it has been given the name Apriori. Apriori uses a bottom-up approach, where frequent item sets are extended one item at a time. Each of frequent item sets will

occur at least as frequently as a pre-determined minimum support count. The support is the percentage of task-relevant data transactions for which the pattern is true. Following parameter are used in Apriori Algorithm: Support (s): support (s) is the ratio of transactions in the database that contain the item-set X to all transactions (i.e. Support(X) = 0.2). Confidence (c): For a given transactions, confidence (c) is the ratio that contains A which also contains B.
Confidence(A=>B) = Support(A ∪ B)/Support(A)

4. PROPOSED METHOD

In this work, using Apriori Algorithm to improve prediction accuracy of Naive Bayes Classifier is proposed. Our method combine Naive Bayes Classifier and Association rule. On the other hand, using only Naive Bayes Classifier may not perform a good result. It's because the independent assumptions to Bayes theorem. Thus, association rule plays an important role to identify related features. Flowchart for our proposed work is given in Fig. 2.

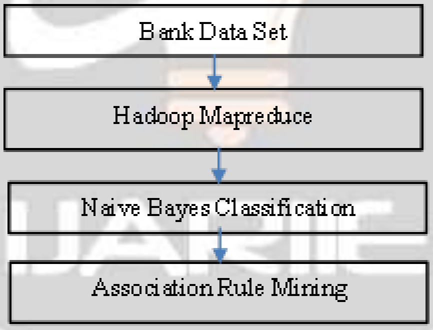

Fig -2 Flowchart of Proposed Work

4.1 Association Rule Mining and Apriori Algorithm

The data set saved as a comma-separated values (csv) file. It stores tabular data in plain text and each line is a data record. Features are separated by commas in one record. Our bank marketing data set contains 20 attributes includes: personal information, social and economic context and other contact information. Both numeric attributes and text attributes are listed in the data set. For text attribute, it contains different variables (i.e. variables in marital attribute include: divorced, married, single and unknown). If the input data set is not a csv file, then transfer the data set to csv format. Input: Data set, D. Output: Data set with csv format, Dcsv.

4.2 Apriori Algorithm

Each line in the data set is considered as an item set. Each variable of attributes is considered as a transaction. Using Apriori Algorithm, we generated maximum length sets that consist of several transactions, it's called frequent item set. Each class has its own frequent item set. These frequent item sets indicate relationships between transactions. In our work, the relationships derived from Apriori Algorithm are used to improve prediction accuracy in Naive Bayes Classifier. Input: Data set with csv format, Dcsv. Output: Frequent item sets for each class, Fclass.

4.3 Algorithm in Naïve Bayes Classifier

First, for text attributes, we use the probability values that matched the frequent item sets to calculate the probability for different class. The minimum support for each class is set to the fraction of each class. This probability result is one part of P(xi | Cj). And then, for other attributes that are not in the frequent item sets, we use the standard Naive Bayes Classifier to calculate the probability result. Input: Data set with csv format, Dcsv; Frequent item sets for each class, Fclass.
Output: Classification results and graphs.

5. EXPERIMENT AND RESULTS

We divide our experiment into two phases: phase and association phase. Classic phase shows the classification results using only Naive Bayes Classifier. Association phase performs the combination using both Association Rule Mining and Naïve Bayes Classifier.

5. 1Classic phase

There are two classes in the classification process, one class is the client accept the term deposit and another class is the client reject the term deposit. First, we train our bank marketing data set using Hadoop & MapReduce framework. The probabilities for two classes are
given by:

P (accept) = 4640/41188 = 0.113
P (reject) = 36548/41188 = 0.883

Based on the training result, we can obtain classification result using testing data set which contains 4119 instances.The result is shown in Table-1.

Table -1. Result for Classic phase

		Actual result	
		Accept	**Reject**
Predicted result	**Accept**	280 (TP)	393 (FP)
	Reject	171 (FN)	3275 (TN)

True positive rate: TP / (TP + FP) = 0.416
True negative rate: TN / (FN + TN) = 0.950
False positive rate: FP / (FP + TN) = 0.107
False negative rate: FN / (TP + FN) = 0.379

For classic Naive Bayes Classifier, we cannot accept this classification result because of the high false negative rate. The reason is that Naive Bayes Classifier is based on the independent assumptions. We cannot directly use Naive Bayes Classifier in business data set, because there are relationships between transactions. That's the reason we propose an association rule algorithm to combine related transactions to improve prediction accuracy.

5.2 Association phase

In this phase, we first run Apriori Algorithm to find relationships between transactions. We get frequent item sets for both accept and reject classes, and then, calculate probabilities for each class. Partial frequent item sets for both class are shown in Fig. 3 and Fig. 4. The number before each variable represent index of attribute. It works for duplicate transactions that come from different attributes.

Fig. 3 Partial frequent item sets for acceptance class

Fig. 4 Partial frequent item sets for rejection class

We use equation (2) to calculate the probabilities for each frequent item set. For attributes that are not in the frequent item sets, we directly use the probabilities from classic phase. The classification result using the same testing data set is shown in Table II.

Table - 2. Association phase results

		Actual result	
		Accept	**Reject**
Predicted result	**Accept**	406 (TP)	242(FP)
	Reject	45 (FN)	3426 (TN)

True positive rate: TP / (TP + FP) = 0.626
True negative rate: TN / (FN + TN) = 0.987
False positive rate: FP / (FP + TN) = 0.065
False negative rate: FN / (TP + FN) = 0.099

False positive rate and false negative rate are less than the classic phase, and true positive rate and true negative rate are greater than the classic phase. The chart is shown in Fig. 5.

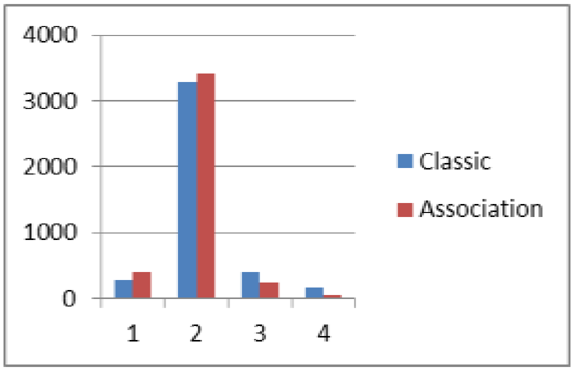

Fig. 5 Comparison Result

In this experiment, firstly, we use Naive Bayes Classifier to classify a business data set but result is not satisfactory. And then, we use association rule to decrease features by combining related attributes by its frequent item sets. Based on two result tables and the calculated proportions, the result met our expectations.

Association rules always bring surprise to us. The main purpose for our research is to reduce attributes by combining related attribute to fit the independent assumption in Naïve Bayes Classifier. In the future work, we can use other algorithm to combine related attributed. Apriori is just one typical method for association rule mining. We can also find solutions to improve Naive Bayes Classifier to fit into business data set.

6. CONCLUSIONS

In this experiment, firstly, we use Naive Bayes Classifier to classify a business data set but result is not satisfactory. And then, we use association rule to decrease features by combining related attributes. The purpose to use Apriori Algorithm is to combine related attributes by its frequent item sets. Based on two result tables and the calculated proportions, the result met our expectations. Association rules always bring surprise to us. The main purpose for our research is to reduce attributes by combining related attribute to fit the independent assumption in Naïve Bayes Classifier. In the future work, we can use other algorithm to combine related attributed. Apriori is just one typical method for association rule mining. We can also find solutions to improve Naive Bayes Classifier to fit into business data set.

7. REFERENCES

[1] R. Margaret, “Business Intelligence (BI) Definition”. http://searchdatamanagement.techtarget.com/definition/businessintelligence [Nov. 04, 2015]

[2] V. Vasilis, “Machine Learning Tutorial: The Naive Bayes Text Classifier”. http://blog.datumbox.com/machine-learning-tutorial-thenaive- bayes-text-classifier/ [Nov. 06, 2015]

[3] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Bank+Marketing [Nov. 06, 2015]

[4] P. Brien, “Hadoop clusters: Benefits and challenges for big data analytics”. http://searchstorage.techtarget.com/tip/Hadoop-clusters- Benefits-and-challenges-for-big-data-analytics [Nov. 07, 2015]

[5] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. 1998. “A Bayesian Approach to Filtering Junk E-mail”. In Proc. of the AAAI'98 Workshop on Learning for Text Categorization, pp. 1048–1054. [Nov. 07, 2015]

[6] Brian Livingston (2002), Paul Graham provides stunning answer to spam e-mails. http://www.infoworld.com/article/2674702/techologybusiness/ paul-graham-provides-stunning-answer-to-spam-e-mails.html [Nov. 07, 2015]

[7] Paul Graham (2003), Better Bayesian filtering. http://www.paulgraham.com/better.html [Nov. 09, 2015]

[8] Firte, L., Lemnaru, C. and Potolea, R. 2010. “Spam Detection Filter Using KNN Algorithm and Resampling,” in Intelligent Computer Communication and Processing (ICCP), 2010 IEEE International Conferenc , pp. 27 - 33.

[9] Kyriakopoulou, A. and Kalamboukis, T. 2006. “Text Classification Using Clustering,” in ECML-PKDD Discovery Challenge Workshop Proceedings.

[10] L.M. Ketari, L.M. Chandra, and M.A. Khanum. “A Study of Image Spam Filtering Techniques.” *4th IEEE Internet. Conf. Computational Intelligence and Communication Networks, 2012*.

[11] S.S. Deshmukh, P.R. Chandre, “Survey on: Naive Bayesian and AOCR Based Image and Text Spam Mail Filtering System”. *International Journal of Emerging Technoogy and Advanced Enginerring.*

[12] P.T. Ho, H.S. Kim,S.R. Kim," Application of Sim-Hash Algorithm and Big Data Analysis in Spam Email Detection System". Proc. 2014 Conference on Research in Adaptive and Convergent Systems, pp. 242- 246. [13] S. M. Kamruzzaman, H. Farhana and H. Ahmed, “Text Classification using Association Rule with a Hybrid Concept of Naive Bayes Classifier and Genetic Algorithm”. http://arxiv.org/ftp/arxiv/papers/1009/1009.4976.pdf [Nov. 09, 2015]

[14] R. Irina, H. Joseph and T. Jayram, “An analysis of data characteristics that affect naive Bayes performance”. ResearchGate, 2001. http://www.researchgate.net/publication/228560806_An_analysis_of_da ta_characteristics_that_affect_naive_Bayes_performance [Nov. 10, 2015]

[15]Tianda Yang, Kai Qian, Dan Chia-Tien Lo, Lixin Tao,” Improve the Prediction Accuracy of Naive Bayes Classifier with Association Rule Mining”, 2016 IEEE 2nd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing, IEEE International Conference on Intelligent Data and Security