DECISION SCIENCES INSTITUTE

A Heuristic Data Sampling Approach for Association Rule Classification under a Big Data Environment

Ching-Chin Chern
National Taiwan University
Email: cchern@ntu.edu.tw

Huai-De Peng
National Taiwan University
Email: r01725017@ntu.edu.tw

Ling-Chieh Kung
National Taiwan University
Email: lckung@ntu.edu.tw

ABSTRACT

This study proposes a heuristic data sampling approach to solve the problem of mining big data without changing the association rule classification method. The proposed heuristic data sampling approach consists of two parts. The first part is a heuristic sampling method applied in the initial phase, which samples representative data from a big data set with important, discriminative attributes. Then the second part deals with the incremental big data problem. Merging the sampled data from both the preliminary and incremental data sets and their classifiers, it is possible to apply them to verify the combined classifier.

KEYWORDS: Big Data, Incremental Algorithm, Association Rule Classification, Data Sampling, Attribute Selection

INTRODUCTION

In the past few years, a new concept, called “big data”, has emerged as an important issue of business decision making. The so-called “big data” includes three special characteristics, called “3Vs”: Volume, Variety and Velocity (Lamont, 2012; McAfee and Brynjolfsson, 2012; Stackpole, 2012). We need new solutions to solve the problem under the big data environment because previous ones are having troubles while facing big data problems. Techniques for reckoning with data at small scales can become insufficient or irrelevant at larger ones (McAfee and Brynjolfsson, 2012; Skinner, 2013). What make most big data big are repeated observations over time and/or space (Jacobs, 2009). Traditional data systems simply cannot handle big data very well, either because it is impossible to handle the variety of data – big data is much less structured due to the speed of evolution, or because such systems just cannot scale at the rate of ingesting data (Jackson, 2012).

The new solutions dealing with big data problems are expected to respond to three special characteristics of big data. For volume and velocity, it is no longer necessary to have a sophisticated model anymore, which may mean a shift in the way data mining is done (Collett, 2011). The solutions depend on breaking up data, sending out subsets for analysis and then regrouping the results to produce the output (Lamont, 2012). For variety, users have to put different forms of data together from many sources and make sense from them (Olavsrud, 2012).

Recently, Apache Hadoop is described as the core technology driving the adoption of big data and capable to quickly and economically process the huge datasets (Reid, 2012).

However, attributes are also dynamic in a big data problem (Jacobs, 2009). Although attributes do not change as significantly as records do most of the time, they still vary not only on values of attributes but also on the total number of attributes. Missing values of attributes in each record also need to be considered. The volume problem on records also happens on attributes. Since including all attributes into association rule classification method apparently slows down the processing speed, determining how many and which attributes should be selected are also issues in big data analytic.

To make the analysis of big data more feasible, the efficiency must be enhanced. Here, we are not focusing on the algorithm and the process of an association rule classification method but the data pre-process. By selecting proper data from a big data set and applying some treatment, the chosen data can be much more manageable than the volume of original data but still be representative of the original data. Thus, we can input these data to an association rule classification method with the expectation of generating patterns in a more efficient way.

PROBLEM DESCRIPTION

This study aims to solve the performance problem of an association rule classification method under big data environment by proposing a proper data preparation approach. Association rule classification is a popular way to extract patterns from several data records and then turn these patterns into knowledge for decision makers to support their decision (Liu et al., 1998). Though association rule classification methods have been developed for a long period, very few discussions have been focused on processing times unless the calculating of an association rule algorithm takes too much time (Thabtah, 2007). Most of the researches focus more on the accuracy and interpretation of an association rule classification method rather than the processing time issues.

A major reason of this situation is the limitation on the number of data records in the past due to data storage problem. However, by Moore’s Law (Moore, 1965), the total data volume is growing at an exponential rate. The fact is that terabytes of data with different features are created every day and the rate of data generation grows steadily nowadays, which is not yet expected by authors of previous data mining methods as well as users. This situation does not cause severe accuracy problem when applying original association rule classification methods, but rather it incapacitates these methods to handle the current characteristics of big data well. Hence, this study focuses on refining input of the general association rule classification methods based on existing procedures and trying to solve the problem without any adjustment on association rule classification methods themselves.

Most mechanisms dealing with big data problems are actually developed to solve the data collecting problem of unstructured data from various sources, e.g., MapReduce and Hadoop (Reid, 2012). These methods can be used as frameworks to collect and store data in several servers and then the results from different servers are computed and compared. Although they provide some simple ways to handle big data collection problem, these methods do not perform further analyses on the collected data. Our study focuses rather on volume and velocity than variety, and thus we just assume that the data is already collected and transformed to a structured database.

To be specific, we break down the performance problem of an association rule classification method caused by big data into three dimensions. First is the perspective of data records. Given that the data not only has extreme large size but also is incremental, the size of these records increases quickly and steadily. Thus, using all data to conduct data mining is almost impossible. Second, attributes of the dataset also have the potential to grow. Because of the variety of data features, diverse sources and different types of data result in possibly changes in dimensions of the data. Finally, the dynamicity makes analyses real-time impossible to make corresponding decisions in time. In the meantime, analytical result in previous period could be infeasible for making further decisions overtime. Moreover, growths on records and attributes make dataset harder to access. The increasing amount of data and properties causes it harder to analyze. These three sub-problems consists the difficulties of using big data.

The answers to the above three sub-problems can be broken down into three parts: determining how many and what records to be selected, determining how many and what attributes to be selected and dealing with the incremental nature of a big data problem. Our approach is to alter the original data preparing process without affecting the process and algorithm being used in data mining. Under this premise, our method will focus on the substances and features, selecting them from the original dataset, and then using them as input of upcoming data mining process, which means that the proposed data preparation process will not influence the following data mining steps.

Sampling is a possible solution to deal with big data. While the size of original dataset is extremely large, sampling can choose and analyze a small subset from the dataset and then infer the results to the entire dataset. Although sampling can reduce the size of data for data mining, it creates other problem, such as determining how many and what the records should be sampled. Choosing a smaller subset will significantly decrease the processing time on data mining, but may be insufficient to represent the whole dataset. There are trade-off between processing time and coverage of data. Thus, using sampling methods in data preparation of data mining should consider about the number of records to be sampled.

The strategy of stratification sampling is a valid way to handle the problem caused by the scale of data (Cano et al., 2005). The concept of stratification is to separate initial dataset into disjoint strata with same analogy of stratifying based on the distribution of the population. Hence, the representation of stratified sampled data is higher than the data by simple random sampling.

Facing the problem that the number of attributes is large and also increasing, the dimensions of data should be degraded to make sure processing time to be short so that the analysis result can be used on time (Shannon, 1948). Attribute selection is a viable way to deal with large amount of attributes under big data. In many attribute selection methods, entropy is used as a criterion to compare and choose corresponding features (Shannon, 1948).

Since prevailing data mining methods are not designed for big data, difficulties occur while using these original analyses to handle big data. Our goal in this study is to apply association rule classification methods on big data. Currently, most mechanisms proposed to solve problems caused by big data are mainly about the variety characteristic. In the other word, these solutions focused on how the data is collected rather than analyzed. However, this study is trying to work on the volume and velocity of big data. Hence, we have to make some proper adjustment and assumption so that we can aim on our target problem more clearly.

First, we assume the data is well collected in a structured database with extremely large size and high dimension. Thus, we are able to focus on the situations that are brought by volume and velocity instead of variety. Second, we assume that classes in original data or in incoming data are known before building classifiers. To enhance the efficiency of data mining under big data environment, we have to deal with both the size and the dimension problem. The method we proposed is using modified data preparation to refine the input of data mining. Therefore, the existing association rule classification methods can be adopted on analysis. We use sampling to solve the size problem and attribute selection for the dimension problem.

Dataset under big data environment is not static because new data objects are consistently adding to the dataset. The possibilities of growth on records, attributes and classes are all expected to happen. In this way, previous classifier could be unsuitable for incoming data with new elements. Thus, the time factor is considered to decide whether the model is fit or not. We have to choose between merging the new model with previous one or simply abandon previous model to use the new one. In the meantime, the processing time of these procedures is still an important criterion.

THE HEURISTIC BIG-DATA SAMPLING ALGORITHM (HBDSA)

The heuristic data selection algorithm proposed in this study includes two different phases: initial and incremental. The initial phase refers to the initial stage, where no valid sampled data exists before or a whole new dataset is sampled to regenerate the classifier. The second phase aims to handle the problem of using incremental data to update the classifier when there are already sampled data and a classifier generated before. Under this situation, we will make proper modification to the classifier because the original classifier is incapable to classify the newly arrived (incremental) data accurately. Separating the incremental stage from the initial stage mainly focuses on solving the velocity problem caused by big data. In a general case, the frequency of initial procedure runs should be much lesser than incremental procedure since the processing time of incremental phase is significantly less than initial stage. The incremental procedure attempts to modify the existing classifier to fit newly arrived data and thus the time to build up a valid classifier can be significantly reduced.

Dividing the job of constructing a valid classifier into these two phases reduces the scale of big data while transforming a big data into smaller datasets makes it both discriminative and analyzable. The first step in the initial phase samples data from the dataset in order to diminish the extreme scale on big data. Afterward, certain attributes are sorted and selected to generate the classifier. Both steps deal with volume and variety problems of a big data. Choosing data and features reduces the scales of database in both attributes and records, which in turn significantly enhances the efficiency and effectiveness of the association rule classification method applied on these data. The problem we have to face is the tradeoff between the efficiency and the accuracy or the coverage while reducing the scale of the dataset.

At the second phase, the incremental data is sampled first as well. Then, this sampled incremental data is tested by the original classifier. If the incremental data cannot be classified accurately, it is necessary to modify the preliminary classifier. Consequently, the inapplicable rules are removed and the new rules are generated to refine and thus increase accuracy of the classifier. In the following sections, we discuss and elaborate these two phases of the Heuristic Big-Data Sampling Algorithm (HBDSA) in detail based on the situations under which these two phases may face.

Under the big data environment, the data set contains considerably numerous data records and attributes. Since the data mining techniques screen out useless data and extract the discriminative data, scaling down the dataset is very important to ensure the success of analyzing big data. In the meantime, the scaled down dataset should still be representative enough for the original dataset so that the result of association rule classification is useful to classify the entire data set. The process of the Data and Attribute Sampling Algorithm (DASA) is based on these two aforementioned principles.

DASA first samples data before selecting attributes for two reasons. First is about the feasibility. Attribute selection requires calculation, ordering and sorting, which will be largely affected by the amount of records no matter which method it takes. Sampling data can narrow down the size of data, and hence speed up the process of attribute selection. Second is to retain the characteristics of original data. By ensuring that the sampled data is approximately similarly distributes as the original big data, the results of further analyses applied on the sampled data can be inferred to the original big dataset. Hence, after sampling the data, DASA selects attributes to extract needed features for building up a classifier.

After generating the preliminary classifier, it should be available to classify data for a period of time. However, as time passes, the incoming data are possibly changed. The scope of data variation not only includes changes in value, but also affect changes in classes and attributes. These diversifications could make original classifier infeasible to classify new data inflow, which leads to a lower accuracy rate and possible misclassification risks. Hence, the preliminary classifier should be modified to adapt changes on data. We first separate the problem into two possible scale-wise problems, which are the changes on attributes and classes, and to handle them sequentially. Afterward, we use actual data to verify the combined classifier and to rank rules in the classifier to generate modified classifier.

We discuss both scale-wise problems individually and solve the problems case by case. Still, the real situation could include at most both problems, i.e. adding and removing both attributes and classes simultaneously. Thus, we only need to decide the sequential of the problems to be handled. Simplifying the problem will make the process easier to settle.

First, the characteristics of both scale-wise problems need to be clarified. Removing attributes and classes requires the removal of inappropriate rules in the preliminary classifier. This action will not influence the combination of both original and new classifier. Nevertheless, adding attributes and classes results in some specific rules that should be retained from new classifier in the final modified classifier. The procedure of addition is more complicated than removal.

In a different view, adding or removing classes may only cause removing and creating corresponding rules. However, adding or removing attributes will alter the schema of a classifier. That is, the change of attributes will possibly impact on other rules, further affects the efficiency of the process. Thus, the attribute problem should be considered before class problem.

As a result, dealing with mixed problems of attribute and class variations, we should solve these scale-wise problems by the following order: (1) removing and adding attributes; and (2) removing and adding classes. The process of verifying a modified classifier requires both previous and incremental data.

After dealing with these problems, we then combine the preliminary and incremental classifier. Since the combined classifier still needs to be ranked to establish the priority of the classification rules, it requires some data to verify the classifier. Both previous and incremental sampled data are similar to the distribution of population, so we only have to choose enough data from them, which is the minimum sample size we calculate before in order to satisfy required condition of Chi-squared goodness-of-fit test, all expected frequency is no lesser than 5. Comparing the size of original and new sampled data, we sample them again by proportions to obtain the minimum sample size of data and use them to compute support rate and confidence. For the extracted rules with new attributes or classes, we use their support rate and confidence in the incremental classifier because previous data cannot verify them. Next, we can further prune the rules that do not reach the threshold. After all, the result is the final modified classifier.

COMPUTATIONAL ANALYSIS

The dataset used in following experiments to verify the validity and efficiency of the HBDSA is originally from KDD Cup 1999, which is a competition held by SIGKDD in conjunction with the annual conference every year. This dataset, retrieved from UCI Machine Learning Repository (KDD, 1999), contains 4,898,431 connection records originally, which can be used to detect intrusion. Furthermore, the dataset includes 43 attributes of both continuous and discrete data type. These attributes represent different parameters and features of connection records, depending on various classes of data. In this study, we use FID (Fayyad and Irani, 1993) to part the values of continuous attributes into intervals, mapping these intervals to sequential integer values.

In order to demonstrate the real condition in big data, we have to consider not only increasing the number of objects but also changing in attributes and classes. Thus, we further partition the original dataset into several groups with possibly different numbers of attributes and numbers of classes. We separate the dataset into six smaller datasets, one for initial stage and other four for incremental scenarios. Hence, we divide the data with high frequency class into 4 classes so that the class distribution will be closer to uniform. Let dataset 0 be the dataset for the initial phase, dataset 1 to 4 be the datasets for the incremental phase. Moreover, the class distribution in original dataset has one extremely high and one extremely low class. The original data set will be used as a representation of long tail class distribution. Let dataset 5 be the dataset for the initial phase, dataset 6 to 9 be the datasets for the incremental phase.

For the purpose of validating the feasibility of the DASA, we separate two different types of class distributions, uniform and long tail. In each distribution, we first compare the differences between the sampled data and the population in certain criteria, including accuracy and runtime. In addition, we will discuss the misclassification cost issue in long tail class distribution and comparing the result of original classifier after sampling. Afterward, we show the effect of changing number of attributes selected on the efficiency and effectiveness of the DASA.

We use dataset 0, consisting of 1,574,501 records, 43 attributes and 11 classes, in the experiments for data with a uniform class distribution. The experiments of data with long tail distribution use dataset 5, including 1,574,501 records, 43 attributes and 8 classes. We further divide data into a training set and a test set, with the proportion of their sizes as 10:1. The accuracy is calculated by the ratio of correctness in classifying objects in the test set. All the experiments set up thresholds in CBA, with minSupp = 0.01 and minConf = 0.5. The minSupp stands for that a rule should be happened at least 1% in all training data so that it can be considered as a frequent rule. The minConf means the probability of classifying an object right when it matches a corresponding association rule, thus there is at least 50% possibility to correctly classify the object if it fits one of the rules in the classifier. We set these thresholds relatively low in order to obtain more rules in classifier.

At the beginning, we would like to verify the validity and accessibility of the sampled data comparing with population. However, the population, either dataset 0 or dataset 5, is too large in both size and number of attributes so that the CBA is unable to generate the classifier and analysis result. Since proving that the sampling method does not distort the result of building classifier is necessary, we have to use another approach instead of using original population of dataset 0 and dataset 5. We further select 20 attributes and 157,451 records by random selection from dataset 0 and dataset 5 to be populations, which is barely applicable for the CBA. Moreover, we also select 10 attributes randomly in these records from dataset 0 and dataset 5 in order to show that no matter how the number of attribute changes in the population, the sampling method is still available and representative.

After defining the populations, we further apply the DASA to sample attributes and records from these populations. The sample size is calculated by the pilot test in each population using Good-Turing estimation method, thus the size to be sampled in each population is fixed. We have to use Kolmogorov-Smirnov test to test whether difference occurs in the class distribution between every pair of sample and population. Since the populations from the same dataset have the same number of records and are only different in attributes, we need to conduct one test for each data source. The critical value of the Kolmogorov-Smirnov test, D_n, and the test values of samples from dataset 0 and dataset 5 are then computed. Because the test values are lesser than D_n, there is not enough evidence to infer that there are differences between these samples and populations.

After verifying that the class distributions are similar in sampled data and populations, we further compare the result of CBA using sampled data and populations. The number of attributes to be selected increases to show the effect. From the results we can see that the accuracies of the classifiers generated by sampled data can be close to one by population although difference happens because of sampling biases. Moreover, the more attributes selected will possibly result in a higher accuracy but also a longer runtime. Furthermore, the processing time is significantly different between sampled data and population. Hence, we are convinced that using the DASA does not sacrifice too much accuracy but rather increases the efficiency significantly.

First, we compare populations from the same class distribution. From comparing results of the CBA in populations 0 with 1 and 2 with 3, we can see that increasing the number of attributes are not necessarily lead to higher accuracy since these attributes are randomly selected. Although populations 0 and 2 have 12 attributes which is more than 7 attributes in population 1 and 3, the accuracy in population 0 is lower than population 2 while the accuracy in population 1 is higher than population 3. However, increasing the number of attributes will definitely require a longer runtime. Runtimes in populations 0 and 2 needs hours to be done while populations 1 and 3 only need several minutes.

In every population, we further compare the effect of increasing the number of attributes selected from each class. We increase the numbers of attributes through 1 to 5 to see their influences on the classification results and runtimes. Like the results in population, increasing the number of attributes will result in more processing time. CBA processes all samples that selected an attribute from each class within 1 second. As the numbers of attributes grows, the runtimes also increase. In the meantime, the accuracy has potential to be higher while more attributes are selected. In populations 0 and 2, the accuracy grows when more attributes are selected. In addition, the number of attributes selected in population 1 and 3 will not make difference on the accuracy.

At last, we can observe that the runtimes are significantly decreased after applying the DASA when comparing samples and populations. The shortest runtime of population to apply the CBA is 8 minutes and 4 seconds, but all of its samples required only 1 second. Moreover, the accuracies of the classifier built by the sampled data can still be retained around the ones built by population. The largest difference in accuracy between the population and the sample is in population 3, which is 0.6352 in population and 0.5741 in all sampled data. Although greater difference occurs in population 2 and sampled data with 1 attribute selected from a class, it can be improved by selecting more attributes.

CONCLUSIONS

This study proposes the Heuristic Big-Data Sampling Algorithm (HBDSA) to deal with the performance problem of an association rule classification method under a big data environment. The algorithm we proposed focuses mainly on refining the input and integrating the output of association rule classification methods. HBDSA is divided into two phases: the initial stage and the incremental stage. At the initial stage, we use the data and attribute sampling algorithm (DASA) to sample proper amount of data objects that still maintain the characteristics of the population. Afterward, we compute the information gain for each attribute and rank the importance of attributes based on the information gain. Extracting certain number of attributes from each class on these sampled records, the dataset we apply our association rule classification method is relatively small comparing to the population. From the related experiments, we know that with a proper number of attributes, the result of association rule classification method can be very efficient without sacrificing much on accuracy. Thus, we can infer that the DASA is an effective way to deal with the volume problem under a big data environment both in object and attribute perspectives. In addition, increasing the number of selected attributes has the chance to increase the accuracy, but definitely require more processing time.

At the incremental phase, we proposed the Incremental Classifier Modifying Algorithm (ICMA) to handle the fact that data is increasing and changing. The inflow of data has different characteristics in attributes and classes. We use the characteristics based on the incremental data in order to predict the data that will arrive in the foreseen future. Combining the sampled data by the DASA from both original and incremental datasets and retaining the size of this dataset, it will be the dataset to verify the classification rules in the original and incremental classifiers. The improper rules, which contain removed classes or attributes in the incremental data, will be deleted while the ones with newly adding attributes and classes will be preserved in the final result of the modified classifier. The validity of applying the ICMA to generate a classifier is proved when comparing with using all sampled data by the DASA. Moreover, the ICMA solved the problem that data will be incremental in both perspectives of objects and attributes. Hence, the ICMA is a proper way to deal with the variety and velocity problems of an association rule classification method under a big data environment.

Nevertheless, there are many classification techniques to be used for classification. In this study, we use the classification based on the association rule (CBA) method. We choose to use the CBA not only because it is one of the simplest method of classification, the rules is relatively observable comparing to other methods. However, there are still other classification techniques that are popular and more common in the reality like decision tree, logistic regression or neural network, etc. We believe that the concept of dealing the problem with smaller scale and representativeness can be applied to any other techniques but requires only proper transformation for input and output data. Future researches should apply the concept in different classification methods under the big data environment.

At last, more and more frameworks are provided by the public because the big data problem is recently a hot issue among academia and industries. The most two well-known frameworks to apply on big data are MapReduce and Hadoop. These frameworks are still currently focus more on collecting data rather than analyzing them. Nonetheless, we are still looking forward to developing a more effective and efficient data analyzing method especially for big data in these frameworks.

ACKNOWLEDGEMENTS

This research was sponsored by the Ministry of Science and Technology in Taiwan, project number MOST 103-2410-H-002-099-MY3.

REFERENCES

Cano, J. R., Herrera, F. & Lozano, M. (2005). Stratification for scaling up evolutionary prototype selection. *Pattern Recognition Letters,* 26(7), 953-963.

Collett, S. (2011). Why Big Data Is a Big Deal? *Computerworld*.

Fayyad, U. M. & Irani, K. B. (1993). Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *International Joint Conferences on Artificial Intelligence*, 1022-1029.

Jackson, J. (2012). The Big Promise of Big Data. *CIO*.

Jacobs, A. (2009). The Pathologies of Big Data. *Communication of the ACM*.

KDD. (1999). Cup 1999 Data Set UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data.

Lamont, J. (2012). Big data has big implications for knowledge management. *KMWorld*.

Liu, B., Hsu, W. & Ma, Y. (1998). Integrating Classification and Association Rule Mining. *Knowledge Discovery and Data Mining*, 80-86.

McAfee, A. & Brynjolfsson, E. (2012). Big Data: the Management Revolution. *Harvard Business Review*, 90(10), 60-66.

Moore, G. E. (1965). Cramming More Components onto Integrated Circuits. *Electronics*, 38(8), 114-117.

Olavsrud, T. (2012). How to Be Ready for Big Data. *CIO*.

Reid, C. (2012). Can Big Data Fix Book Marketing. *Publishers Weekly*.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3), 379-423.

Skinner, D. (2013). Big Data: It’s not Just Big. *HPC Source*, ISC’13 Special Edition.

Stackpole, B. (2012). Five things IT should do to prepare for big data. *Computerworld US*.

Thabtah, F. A. (2007). A Review of Associative Classification Mining. *Knowledge Engineering Review*, 22(1), 37-65.