Big Data without Map Reduce

Yanan Xiao, *Student Member, IEEE,* Aziza Al Sawafi

***Abstract*—We present our approach to the big data problem put forward in KDD Cup 2009. We carry out a more technical analysis of our initial plan. And we describe our trials and errors with the large dataset. Till now, our team has not come up with any practical solution with large dataset in a limited time. We present our results with the small dataset in the end.**

***Keywords*—Customer relationship prediction, ensemble selection, classification, data mining**

I. Introduction

CUSTOMER relationship management (CRM) is an essential model for managing the interactions between the company and its current and future customers. It takes up-to-date technologies to organize, automate, and synchronize CRM and marketing information system [1]. CRM applications that use data mining are called Analytic CRM, which provides valid predictions from customer data collected and stored with various attributes. There are a lot of data mining tools and methods to extract and analyze data generally, and customer data specifically. A naive step in doing so is to summarize the statistical attributes of the data (such as means and standard deviations) and use charts and graphs to review it visually [2]. However, in many situations customer relationship data volume is vast and massive. Therefore, more sophisticated methods are created and evaluated. In this field, we found the following techniques are most frequently employed: decision trees, support vector machines, artificial neural networks and Bayesian classifiers.

Data mining has powerful capability in processing and analyzing data; its key technologies that applied in CRM are categorized into three main categories; clustering, classification and forecast, and association rules [3]. Classification and forecast analysis classifies unknown data into the most proper predefined class based on category description that is obtained by training a set of data using certain algorithm. Key classification techniques are; decision making tree, Bayesian statistics, BP neural networks, Genetic Algorithm, rough set theory, fuzzy set theory and so on. Classification methods in CRM can predict new customers behaviors and activities.

As mentioned in [4], classification analysis is the one that is widely used in classifying CRM data. It can be processed in two steps; learning phase and training phase [5]. In the learning phase the classification algorithm analyzes the training data set and learns it, then in the second phase the accuracy of the classifier will be estimated using the test data set. After that, the classifier can be used to predict and classify new data set. In order to obtain better accuracy, some preprocessing and filtering techniques can be applied to the data before going through the classification phases. Those techniques are; data cleaning, data discretization, and feature selection. The common challenges in knowledge discovery from CRM data are the high-dimensionality, and imbalanced corrupted records. Besides, researchers roved that customer classification and prediction is cost sensitive in nature. For example, if a valuable customer predicted as loyal but then that customer churns, the cost is higher than if a loyal customer is classified as one who will churn [6].

In this paper, we present various tools that we tried and several plans that were carried out when dealing with big data problem like this. The rest of the paper is organized as follows. Section II displays what our thoughts were before tackling the real large dataset.

II. Initial Proposal

In this section we describe what we proposed as out initial plan.

III. Revised Proposal

IV. Result and Conclusion

Well, the result shall come out. Eventually.

Acknowledgment

The authors would like to thank Dr. Wei Lee for giving high quality data mining lectures and selecting this challenging but rewarding topic as this semester's project. They would like give more gratitude to Masdar Institute for the studying and researching environment.

References

[1] Wikipedia. (2013) Customer relationship management. [Online]. Available: https://en.wikipedia.org/wiki/Customer_relationship_management

[2] A. Al-Mudimigh, Z. Ullah, and F. Saleem, "Data mining strategies and techniques for crm systems," in *IEEE International Conference on System of Systems Engineering*, 2009, pp. 1–5.

[3] K. Wu and F. Liu, "Application of data mining in customer relationship management," in *Management and Service Science (MASS), 2010 International Conference on*, 2010, pp. 1–4.

[4] N. Shahrokhi, R. Dehzad, and S. Sahami, "Targeting customers with data mining techniques: Classification," in *User Science and Engineering (i-USEr), 2011 International Conference on*, 2011, pp. 212–215.

[5] I. Guyon, V. Lemaire, M. Boulle, G. Dror, and D. Vogel, "Analysis of the kdd cup 2009: Fast scoring on a large orange customer database," in *Proceedings of KDD-Cup 2009 competition*, Paris, France, Jun. 2009, pp. 23–34.

Y. Xiao, A. Sawafi are 1st year master students with the Department of Electrical Engineering and Computer Science, Masdar Institute of Science and Technology, Masdar City, Abu Dhabi, UAE. P.O. 54224 Email: {yxiao,aalsawafi@masdar.ac.ae}.

One team member quit Masdar Institute not soon after the project. But we managed to get things done by ourselves.

[6] M. Lobur, Y. Stekh, and V. Artsibasov, "Challenges in knowledge discovery and data mining in datasets," in *Proceedings of VIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, 2011, pp. 232–233.

Yanan Xiao A first year master student as well as IEEE student member in CIS program, Masdar Institute. He loves programming when all the coursework is finished. When he feels tired of programming, he would read some books.

Aziza Al Sawafi First year Computing and Information Science student at Masdar Inst., got a bachelor degree in Network Engineering (United Arab Emirates University). Sport, drawing, designing, blogging, reading poems, photography, and riding horse/bicycle are my interests beside all things that are related to networking and computer science.