Getting The MAX out of Ensemble Selection

Yanan Xiao, *Student Member, IEEE,* Aziza Al Sawafi, and Mansoor Al Dosari

***Abstract*—We present a detailed literature review of data mining techniques for customer relationship prediction (CRP). Papers are selected from various sources with different method discussed. Some main technical challenges are identified after reviewing even more papers. We find that Microsoft, SAP and some less famous research groups have conducted years of research in the area of customer relationship management. Elementary experiments demonstrate that ensemble selection is a good method to implement for next phase. We propose a set of well-defined research outlines and summarized each team member's contributions in the end.**

***Keywords*—Customer relationship prediction, preprocessing, ensemble selection, 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]. The KDD[1] Cup 2009 was organized to solve a marketing problem using data mining techniques that can proficiently build predictive models and use them to score new entries on a large database, respectively [2]. It was an opportunity to work on large customer databases provided by Orange company (a French Telecom company) to predict customer behavior probabilities by evaluating three variables: churn, appetency and up-selling. Churn is the measure of clients that terminate their commitments to a service over a given time. Appetency is the client tendency to buy a new service or product. Up-selling is the attempt to persuade the client into inquiring a product to make additional profit for the business.

KDD Cup 2009 had two challenges: the Fast challenge and the Slow challenge. The participants were given at most 5 days to submit their results after the labels on the training set was released, as with the Fast challenge. For the Slow challenge, an extra month was available for participants to submit their prediction.

The data set was composed of 100,000 instances, spilt randomly into a pair of 50,000 instances each. There were 15,000 variables for use to make predictions, including 260 categorical ones. The majority of categorical variables as well as 333 numerical ones had missing values. All variables were scrambled due to confidential issues of customers. No description could be found of each variable's meaning.

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

MAX serves the group name at the same time.

[1]ACM SIGKDD: Association for Computing Machinery, Special Interest Group, Knowledge Discovery from Data

For the slow track, there was a much smaller number of variables, i.e. 230 provided, including 40 categorical ones. The scrambling of small data set was different from that of large one, and it was reported later that uncovering the links between them provided little help, if any [3].

The structure of the rest of paper is as follows. Section II gives a broad discussion of problem addressed as well as solutions proposed in the area of data mining for customer relationship prediction. Section III analyzes technical challenges shown in customer relationship prediction by way of data mining. Section IV presents a much more detailed discussion of papers that arouse our interests. Section V proposes a course of research that we would use in the evaluation part. Finally section VI concludes this paper by summarizing our work up to now.

II. Related Work

Implementing data mining in CRM will be the key to success of CRM, besides understanding customer consumption behavior patterns that helps in making marketing decisions and improving revenue. Data mining is used to analyze and classify various aspects in customer relationship management (e.g. customer groups, background, satisfaction, credit, churn, benefit, etc.), and the main target of data mining here is to find the hidden knowledge from customer data with huge dimensionality. In [4], the author enlisted the following practices as the main functions: clustering analysis; automatically prediction of behaviors and trends; concept description; correlation analysis and error detection.

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 [5]. 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 [6]. Clustering is to group similar and create subgroups based on meaningful relation between the grouped objects. It aims to minimize the distance within the group and maximize it between different groups. In CRM, customers are clustering based on their different background, payments activities and habits, etc. That will

significantly help in making more efficient marketing decisions and improve the customer-enterprise relationship. K-Means, K-Mediods, BIRCH, CURE, DBSCAN, and OPTICS are examples of clustering methods.

Classification and forecast analysis classifies unknown data into the most proper pre-defined 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.

According to [6], association analysis aims to find out the most repeated pattern within the data set. The association is created when the values of two or more variables have certain rule. Example of such pattern is 70% of customer buys commodity B together with commodity A in one shopping. The association analysis realizes marketing strategy which can help to retain customers and improve their loyalty.

In [7], the researchers stated that 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 [2]. 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.

III. Technical Challenges

Customer behavior classification and prediction is one of the most important issues in customer relationship management where organizations and companies cluster customers into predefined groups with similar behavior patterns. Business can market the right products to the right segments at the right time through the right delivery channels [8].

In [9], researchers stated the following challenges, which acted as good indicators, when carrying data mining on customer behavior data: the definition of appropriate distances between objects, the choice of appropriate clustering algorithm and the appropriate evaluation criteria of final results. They evaluated their framework on the KDD Cup 1998 dataset and demonstrated that it outperformed the methods of Cup winner. After further validating their methodology on a real dataset from a large Chinese mobile telecommunication company, they concluded that the genetic weighted k-means algorithm (a hybridization of a genetic and weighted k-means algorithms) performed better than the k-means in terms of quality and sensitivity to initial partitions. Additionally, the team used data distribution information (an equi-depth algorithm) to normalize all attributes' range in order to solve the problem of data imbalance distribution.

In a recent paper [10], researchers pointed out some common challenges in knowledge discovery from data. They proved 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.

Another interesting classification model mentioned in [11] was tested on a high-dimension data set, which is composed of highly imbalanced, corrupted and inaccurate records. Their proposed method dealt with the mentioned problems effectively. They adopted CFS (Correlation-based Feature Selection) method for preprocessing to remove the redundant and irrelevant features based on feature-class and inter-feature correlation. The proposed method showed great flexibility and provided accurate result with highest AUC and sensitivity claiming that this method is computationally efficient.

IV. Further Analysis

In this part, we followed the proceedings of KDD Cup 2009 and found it helped us a lot to figure out our course of research. We described our exploring process here.

Guyon *et. al* published their summary of KDD Cup in [12]. They stated that the challenge started on March 10, 2009 and ended on May 11, 2009. It attracted over 450 participants from 46 countries and the results of this challenge were discussed at the KDD conference (June 28, 2009). Their key conclusions are that ensemble methods are very effective. Moreover, from reports of winners they generalized that to predict customer relationship out of data that contains large numbers of instances with a great deal of attributes, mixed types of variables and missing values, ensemble of decision trees outperformed other methods. Particularly, as an throughout analysis of this challenge, some messages the researchers conveyed are enlightening:

- KDD Cup 2009 could be regarded as a classification problem, and after years of research there are mature methods at hand which deal with relatively large data like this efficiently. It is confirmed by competitors that this problem is solvable either with compute clusters or individual desktop, laptops.
- There seems to be an upper limit with performances of this problem. It's found that there was not a significant increase in results just after the first day.
- There are many model evaluation methods, but cross-validation was found to be used by all top ranking participants. The standard process of k-fold cross-validation is to randomly partition the initial data into $D_1, D_2, \ldots, D_k$, k mutually exclusive subsets or "folds" of approximately equal size. The cross-validation method differs from the handout and subsampling methods in that each sample or "fold" is used the same number of times for training and once for testing. This served as a good explanation that top participants in KDD Cup 2009 did not overfit the validation set.
- There was not all about ensemble methods in the challenge. A popular technique, i.e. logistic regression, which belongs to linear classifiers proved good performance at the expense of increasing computational requirements. The same went with non-linear kernel methods like support vector machine (SVM).

A keynote from [12] is as well-illustrated by figure as by words. KDD Cup 2009 dataset was provided by Orange Telecom company, which had an authentic industrial flavour. Even if as diversified as performance, process automation , training and development time criteria should be considered in an industrial setting, the sorted final scores are very astonishing. From this figure we could infer that even if at the cost of a huge performance deterioration of the other criteria, it is still hard to achieve significant improvement of prediction accuracy via state-of-the-art techniques. Confirmation from figure that the top 50% values lie on an almost horizontal line.

The winning solution came from an IBM research group. Generally speaking their strategy is ensemble selection and more details are described in [3]. The IBM team used ensemble selection [13] to generate and ensemble model from large heterogeneous libraries of classifiers for each problem. The robustness and performance evaluation can be found in [14] and [15]. In addition, it is an anytime method where ensemble can be generated very fast using the available classifiers in the library at that time without over-fitting. Their article is organized chronically.

- Fast track. As mentioned above, it's almost impossible to apply any algorithm directly on complex data like this. The IBM group did some fairly standard preprocessing and in order to make the most of ensemble selections, they normalized all the features, discarding merely 1,564 really waste features. When finished the "textbook" ensemble selection, they created 20 additional features for each existing feature and trained a decision tree of limited depth to directly predict the target.
- Slow track. It is recommended in [16] generally to use a stratified 10-fold cross-validation for estimating accuracy, since it will provide relatively low bias and variance. In the fast track, Niculescu-Mizil *et. al* only carried 2 out of 10 due to the lack of 10. For the slow challenge, two extra folds were trained. More importantly, a combination of even more features helped pull the average accuracy rate from 0.8443 to 0.8509. The following features were constructed: a number of features which distinct from others to be handled isolately; pairs of attributes used as inputs of the specific decision tree to obtain two-way non-additive interactions; a fast probabilistic bi-clustering algorithm [17] run to identify bi-clusters.

The IBM research group built a large library of 500-1,000 base classifiers, and this number was just for each classifier. It is stated that this huge library contributed a lot to their winning.

A paper clearly and repeatedly mentioned in IBM and some other groups' article is [13], which served as a "beginner's guide" in ensemble selection. In this paper step by step illustrated what "ensemble selection from libraries of models" was, and what this novel technic could achieve. Here we discussed this paper in depth.

- A naive ensemble selection would proceed in this way. Repeatedly adding model that maximizes ensemble's performance to the error metric, and returned the ensemble from the nested set of ensembles that performs best on given metric. The process was done on a hillclimb validation set.
- One aspect particular about ensemble selection in that the validation dataset was used for both parameter training and model selection. To improve its performance, Caruana *et. al* introduced some techniques: selection with replacement, which flattened the performance curve past the peak and enabled the weighting of models; sorted ensemble initialization, which prevented overfitting when ensembles were small; bagged ensemble selection, which minimizes the likelihood of selecting model combinations that are overfitting, as the hillclimb set increases.
- Researchers carefully selected as many as 7 datasets from different repositories to test their ensemble selection method. Those datasets were large enough to remain some data for a large final test, which later strengthened their findings.
- An case study was carried out to support their work. It was done to classify sub-atomic particles. Performance was measured against the Stanford Linear Accelerator Center (SLAC) Q-score: $SLQ = \varepsilon(1-2w)^2$. The outcome was an increase of 6% SLQ, compared with the best bagged trees.

With a great deal of proof at hand, the researchers concluded that "using many different learning methods and parameter settings is an effective way of generating a diverse collection of models", which in turn could find ensembles outperforming all other models. A lively plot of ensemble methods is shown in figure.

In [14], which published 4 years earlier than Caruana's findings above, a more mathematical analysis of then popular ensemble methods was done. Dietterich reviewed error-correcting output coding, bagging and boosting, and provided reasons as varied as statistical, computational and representational in his discussions.

After listing so many benefits brought up by ensemble methods, it was no wonder for us to find and figure out why it was so pervasively used in KDD Cup 2009. For report by Lo *et. al* in [18], they built an ensemble of three classifiers, namely, expanded linear model, heterogeneous boosting and selective naive Bayes. In [19] researchers from the Hungarian Academic of Science partially took ensemble methods into their strategies, with a logistic boost as their approach, accompanied by an ADTree classifier.

Alternatively, a team from ID Analytics Inc. used a combination of boosting and bagging and they achieved the fast scoring on a large database [20]. Xie *et. al* considered the three tasks (churn, appetency, and up-selling) as binary classification problems. They believed that ensemble learning schemes are widely used to improve the overall performance of a single classifier by combining predictions from multiple classifiers. The team used TreeNet stochastic gradient decision tree as the main classifier and since the three tasks are binary classification problems, the log likelihood function was chosen. They combined bagging and boosting, and for each task they bagged 5 boosted tree models and took the average as the final prediction result. Bagging and boosting both decreased

error rate of decision tree learning, and the question came as which one would be better to use as well as under what kind of circumstances.

A comparative study was conducted by Khoshgoftaar, Hulse and Napolitano to compare between boosting and bagging techniques with noisy and imbalanced data [21]. They evaluated four algorithms of boosting and bagging (SMOTEBoost, RUSBoost, Exactly Balanced Bagging, and Roughly Balanced Bagging) in a comprehensive suite of experiments for nearly four million classification models were trained. The results were tested for statistical significance via analysis-of-variance modeling. They recommended the use of bagging technique since it performed better than boosting when data were noisy and imbalanced. However, when the data were clean and imbalanced the difference was less significant. On the contrary, Jain and Kulkarni stated that boosting is more accurate than bagging. They published a paper in 2012 reviewing the state of the art group learning techniques for the imbalanced data sets. In addition, they proposed a new group learning algorithm called Incorporating Bagging in to Boosting (IB). Moreover, the final results showed that Incorporating Bagging into Boosting (IB) was more stable than boosting and it is on average more accurate than bagging, also its average error rate was lower than the average error rate of boosting [22]. These results emphasized the advantages of combining bagging and boosting techniques to get high quality performance and accuracy, and were in return confirmed by the results of researchers from ID Analytics in KDD Cup 2009.

V. Our Blueprint

We reviewed throughout of each report in the proceedings [23] to find their future work and proposed our course of research as followed. To be short, we would do an ensemble selection of 3 to 5 base classifiers, which depends on the time allocated to the second phase of this project. Furthermore, we planned to add 1 to 2 features generated from in-use features, to test if this would improve our ensemble selection's performance. In the end, when time permitting we would replace standard classifiers with newly-discovered classifiers to test their validity.

A. Dataset

Due to the lack of time allocated to the second part of this project, we made an decision to work only on large dataset in the beginning. We considered experimenting our methods on the small dataset to validate them.

B. Preprocessing

As described in section III, the dataset itself remains to be a big challenge. Therefore, we agree to take the following steps to tackle it.

Missing values. We would consider missing categorical values as a separate value. We would take a standard approach which calculates the mean of the feature to impute missing values. And as proven an effective technique by [3], we decide to add an extra indicator variable to indicate "missingness" for every one of the 333 variables with missing values. We planned to do this because some linear models in our base classifiers could then estimate the optimal constant instead of merely relying on the means to replace the missing value with.

Categorical values. Since categorical values are not easily handled by many learning algorithms, we decide to recode categorical values using the same way as by IBM Research. For different values a categorical attribute could take, we would generate corresponding indicators. As an good example shown in IBM's paper, limiting the number of values encoded would greatly reduce the number of features, which may be from variables with a enormous vocabulary.

Clean up. We would normalize each feature by dividing up by their range. And we would clean the data by eliminating redundant features, which are either constant, or duplicate of other features.

C. Base Classifiers

As stated by Dietterich [14], "A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse." The winning team of this KDD Cup employed as many as 10 base classifiers–random forests, boosted trees, logistic regression, SVM, decision trees, TANs, Naive Bayes, Sparse Network of Windows and k-NN. Building libraries would be expensive, what's more is that time is limited to train all of them in the next phase. To maintain the accuracy and diversity of our next-phase base classifier library, we decide to build it using the classifiers as follows, boosted decision trees, decision trees, linear regression, SVM and k-NN.

D. Ensemble Selection

An ensemble is a collection of models. To make predictions, we just need to calculate weighted average or "voting" based on those models. One important reason for us to work with ensemble selection is that it can be optimized to any easily computed metric. Same as described before, for an ensemble selection classifier used to improve classification accuracy, one thing is the base classifiers in the library are accurate. This is not hard to achieve with a great literature at hand. The other characteristic is diverse. Due to the fact that our computation power is limited, we planned to build 100 to 200 base classifiers in our library.

E. More Features

Not so many prize winning teams reported to have used carefully-treated features, namely artificially selected features to produce even better results. However, the top 1 team IBM Research built more features from those massive already-given ones, in both fast track and slow track. As a good example shown by them, we planned to add one more feature after the first round of ensemble selection. The feasible approach our team members unanimously choose is to use decision tree to identify the optimal splitting points [3]. The output, i.e. probabilistic prediction, of this decision tree would help at least to some degree express nonlinear relationships in a linear model.

F. More Base Classifiers

As shown by some researchers that in the end KDD Cup 2009 challenge is merely a binary classification problem with three unique probabilities to estimate [18]. Therefore we did a massive search for recent papers that report good results of new/updated base classifiers. Based on our findings, we planned to add one of those two classifiers to our library if possible (time sufficient after the first round). Paper [24] presented a large linear classifier that performed great on big data classification with computer memory limitation, they provided their code at this site [2]. Researchers at Cornell university presented new models for classification and regression [25]. Their innovative method is based on tree ensembles, which "manages" its leaves in an adaptive way. This is well worth trying since we also planned to include some classic decision tree models. We would like to see how much improvement this patched tree generation methods would have in our study.

G. Objectives

For the evaluation of our next-phase implementation, we propose two research questions here.

1) ***RQ1: How great would the probability be before and after the ensemble selection step****:* In [16] it is clearly stated that "An ensemble for classification is a composite model, made up of a combination of classifiers. The individual classifiers vote, and a class label prediction is returned by the ensemble based on the collection of votes." Therefore, we expect to see some probability calibrations, as shown in [13].

2) ***RQ2: How great would the performance be if we use newly developed models****:* It is witnessed that many newly developed or updated models arose for recent ACM SIGKDD annual meetings. Apart from a few we discussed above, there are new model for nonlinear classification [26], new classifier based on composite hypercubes [27] etc. We planned to select one or two of them out of all those appeared in KDD and ICDM, to replace models of same characteristic (linear/nonlinear). We would like to see the performance change after this add-up as a validation of new models.

VI. Conclusion

We have conducted a broad search of related literature in the field of customer relationship prediction. We proposed to use ensemble selection of several base classifiers for the next phase. We conclude each team member's contributions as follows.

- Yanan Xiao. He finished half of the literature review and took the responsibility to typeset this report in LaTeX2e.
- Aziza Al Sawafi. She contributed the other half of this report, especially the section II and III.
- Mansoor Al Dosari. He did review some related papers. Later we found it hard to contact him.

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/exp.html

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] ACM SIGKDD. (2009) KDD Cup 2009: Customer relationship prediction. [Online]. Available: http://www.kdd.org/kdd-cup-2009-customer-relationship-prediction

[3] A. Niculescu-Mizil *et al.*, "Winning the KDD Cup orange challenge with ensemble selection," in *Proceedings of KDD-Cup 2009 competition*, Paris, France, Jun. 2009, pp. 23–34.

[4] L. Zhang, "Data mining application in customer relationship management," in *International Conference on Computer Application and System Modeling (ICCASM)*, 2010, pp. 171–174.

[5] 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.

[6] 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.

[7] 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.

[8] H. Gao, "Customer relationship management based on data mining technique: Naive bayesian classifier," in *International Conference on E -Business and E -Government (ICEE)*, 2011, pp. 1–4.

[9] J. Pan, Q. Yang, Y. Yang, L. Li, F. Li, and G. Li, "Cost-sensitive-data preprocessing for mining customer relationship management databases," *Intelligent Systems, IEEE*, vol. 22, no. 1, pp. 46–51, 2007.

[10] 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.

[11] Y. Tu, Z. Yang, and Y. Benslimane, "Towards an optimal classification model against imbalanced data for customer relationship management," in *Seventh International Conference on Natural Computation (ICNC)*, vol. 4, 2011, pp. 2401–2405.

[12] 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.

[13] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from library of models," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, Jun. 2004.

[14] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, J. Kittler and F. Roli, Eds. New York: Springer Verlag, 2000.

[15] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1-2, pp. 105–139, 1999.

[16] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Berlington, USA: Morgan Kaufmann, 2012, ch. 8-9.

[17] J. Xiao, L. Wang, X. Liu, and T. Jiang, "An efficient voting algorithms for finding additive biclusters with random background," *Journal of Computational Biology*, vol. 15, no. 10, pp. 1275–1293, Dec. 2008.

[18] H. Lo *et al.*, "An ensemble of three classifiers for kdd cup 2009: Expanded linear model, heterogeneous boosting, and selective naive bayes," in *Proceedings of KDD-Cup 2009 competition*, Paris, France, Jun. 2009, pp. 23–34.

[19] M. Kurucz, D. Siklosi, I. Biro, P. Csizsek, Z. Fekete, R. Iwatt, T. Kiss, and A. Szabo, "Kdd cup 2009 @ budapest: feature partitioning and boosting," in *Proceedings of KDD-Cup 2009 competition*, Paris, France, Jun. 2009, pp. 65–75.

[20] J. Xie, V. Rojkova, S. Pal, and S. Coggeshall, "A combination of boosting and bagging for kdd cup 2009 - fast scoring on a large database," in *Proceedings of KDD-Cup 2009 competition*, Paris, France, Jun. 2009, pp. 35–43.

[21] T. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, no. 3, pp. 552–568, 2011.

[22] K. Jain and S. Kulkarni, "Incorporating bagging into boosting," in *Proceedings of the 12th International Conference on Hybrid Intelligent Systems (HIS)*, Beijing, China, Aug. 2012, pp. 443–448.

[23] G. Dror, M. Bollue, I. Guyon, V. Lemaire, and D. Vogel, Eds., *Proceedings of KDD-Cup 2009 competition*. Paris, France: JMRL,SPARC, 2009.

[24] H. Yu, C. Hsieh, K. Chang, and C. Lin, "Large linear classification when data cannot fit in memory," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, pp. 23:1–23:23S, Feb. 2010.

[25] Y. lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, Aug. 2012, pp. 150–158.

[26] Z. Wang, N. Djuric, K. Crammer, and S. Vucetic, "Trading representability for scalability: Adaptive multi-hyperplane machine for nonlinear classification," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, California, USA, Aug. 2011, pp. 24–32.

[27] L. Wikinson, A. Anand, and D. N. Tuan, "Chirp: A new classifier based on composite hypercubes on iterated random projections," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, California, USA, Aug. 2011, pp. 6–14.

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.