A Case Study of Predicting Banking Customers Behaviour by Using Data Mining

Xujuan Zhou
School of Management and Enterprise
University of Southern Queensland
Springfield, Australia
xujuan.zhou@usq.edu.au

Ghazal Bargshady
School of Management and Enterprise
University of Southern Queensland
Springfield, Australia
ghazal.bargshady@usq.edu.au

Moloud Abdar
Département d'informatique
Université du Québec à Montréal
Montréal, Canada
m.abdar1987@gmail.com

Xiaohui Tao
Faculty of Health, Engineering and Sciences
University of Southern Queensland
Springfield, Australia
xtao@usq.edu.au

Raj Gururajan
School of Management and Enterprise
University of Southern Queensland
Springfield, Australia
raj.gururajan@usq.edu.au

KC Chan
School of Management and Enterprise
University of Southern Queensland
Springfield, Australia
kc.chan@usq.edu.au

***Abstract*— Data Mining (DM) is a technique that examines information stored in large database or data warehouse and find the patterns or trends in the data that are not yet known or suspected. DM techniques have been applied to a variety of different domains including Customer Relationship Management CRM). In this research, a new Customer Knowledge Management (CKM) framework based on data mining is proposed. The proposed data mining framework in this study manages relationships between banking organizations and their customers. Two typical data mining techniques - Neural Network and Association Rules - are applied to predict the behavior of customers and to increase the decision-making processes for recalling valued customers in banking industries. The experiments on the real world dataset are conducted and the different metrics are used to evaluate the performances of the two data mining models. The results indicate that the Neural Network model achieves better accuracy but takes longer time to train the model.**

Keywords*— *Customer Relationship Management, Customer Knowledge Management, Data Mining, Neural Networks, Association rules

I. INTRODUCTION

Customer Relationship Management (CRM) is a modern management tool that, employs information technology such as database management, data analysis, and data mining to understand, target, and attract customers, with the objective of satisfying and retain them [1].

The Customer Knowledge Management (CKM) model has drawn attention recently through the convergence of both the technology-driven and data-oriented approach in CRM and the people-oriented approach in Knowledge Management (KM), with a view to exploiting their interaction potential [2,3,4] The expectation of this effort is to derive more comprehensive knowledge for customers; knowledge about customers, and knowledge from customers.

Data mining is defined as a process that uses mathematical, statistical, artificial intelligence and machine learning techniques to extract and identify useful information and subsequently gain knowledge from databases. Data mining algorithms have been widely used in range of research fields such as healthcare and medicine [5-7], sentiment analysis [8, 9], education [10] etc. The purpose of applying data mining in bank industry is to use the available data to retain its best customers and to identify opportunities sell them additional services.

Information technology tools, and explosion in banks' customer data has improved and changed the managing relationships between banks and their customers. Data mining can be used in banks and finance organizations for decision making and forecasting. One of the most common learning models in data mining that predicts the future customer behaviours is classification. The prediction is done by the classification of database records into several predefined classes based on certain criteria. Neural networks, decision trees, naive Bayes, logistic regression, association rule, and SVM are the common tools used for classification [11].

In this research, a new Customer Knowledge Management (CKM) framework based on data mining is proposed. The neural networks and association rule mining are used as two classification techniques in this study. Furthermore, for the CRM applications in the banking domain, customer segmentation, prospecting and acquisition, security, profitability, risk analyses, strengths and weaknesses have been taken into consideration to show the performance of classification models. The Saman Bank customers dataset are used in the experiments. K-fold cross validation technique used to evaluate the predictive methods. The comparison of the performances of multilayer perception neural network and association rule classifier are provided in terms of effectiveness and efficiency.

The rest of the article are organized as follows. Section 2 is about related works from the literature. Section 3 describes the proposed data mining based bank CKM framework. section 4 covers experiments and results analysis, and finally section 5 concludes the article.

II. RELATED WORKS

A. Customer Relationship Manegnment

The term CRM emerged in the information technology (IT) vendor community and practitioner community in the mid-1990s [12]. It is often used to describe technology-based customer solutions, such as sales force automation (SFA). In the academic community, the terms "relationship marketing" and CRM are often used interchangeably [13]. However, CRM is more commonly used in the context of technology solutions and has been described as "information-enabled relationship marketing" [14]. Zablah *et al.* [15] suggest that

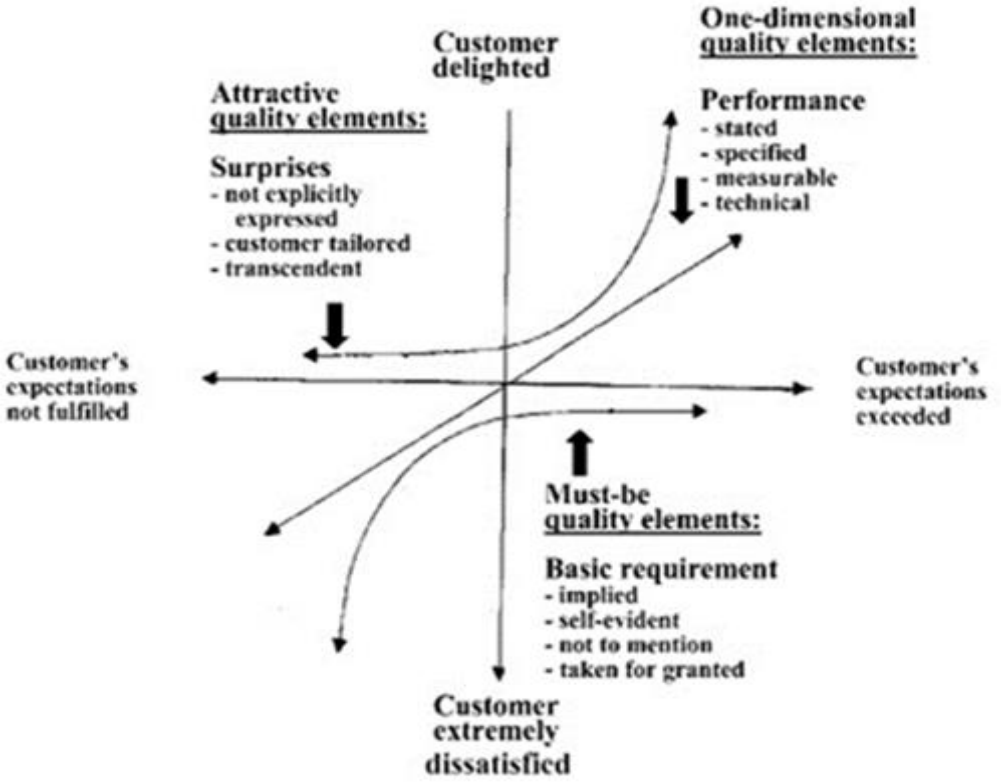

Fig. 1. Kano's model of customer satisfaction [18]

CRM is "a philosophically-related offspring to relationship marketing which is for the most part neglected in the literature," and they conclude that "further exploration of CRM and its related phenomena is not only warranted but also desperately needed." [16].

CRM is a strategic approach that is concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments. CRM unites the potential of relationship marketing strategies and IT to create profitable, long-term relationships with customers and other key stakeholders. CRM provides enhanced opportunities to use data and information to both understand customers and co-create value with them. This requires a cross-functional integration of processes, people, operations, and marketing capabilities that is enabled through information, technology, and applications [12,16,17].

In Kano's model, one-dimensional quality elements refer to conventional ideas about quality: the customer satisfaction is proportional to the functionality of the product, where less function leads to less satisfaction and vice versa. However, the quality elements that generate only customer satisfaction and no dissatisfaction are categorized as 'attractive quality elements', corresponding to the 'motivator' or 'satisfier' in Herzberg's Motivation-Hygiene theory; while the quality elements that generate only customer dissatisfaction and no satisfaction are categorized as 'must-be quality elements', corresponding to the 'hygiene factor' or 'dis-satisfier' in Herzberg's Motivation-Hygiene theory. The 'must-be' quality elements may coexist with 'attractive quality elements' in a certain product [18,19].

A Kano-CKM framework was proposed to address the improvement of issues in the management of customer knowledge as shown in Fig. 2. In the Kano-CKM model, the CKM process comprises four stages bolstered by Kano's Method, and the implementation scheme is described as follows [19].

B. Data Mining for CKM

Data Mining (a.k.a. Knowledge Discovery from data) is a technique that examines information stored in large database or data warehouse and find the patterns or knowledge in the data that are not yet known and discovered patterns and knowledge can be used to predict the meaningful trends and relationships. In this regard, customer knowledge obtained via a CRM system is a valuable intellectual asset for a company to develop or improve products and services in order to meet or even exceed customers' expectations. CRM

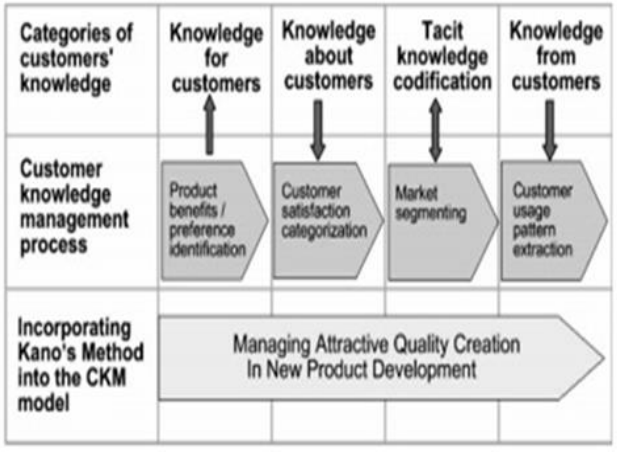

Fig. 2. The Kano-CKM Model as applied in customer knowledge management for attractive quality creation [19]

systems that collect information for customer knowledge are classified into three main categories [1]:

1. Operational CRM systems enhances the efficiency of a CRM process through service center management and marketing-automation like database marketing.
2. Analytical CRM systems evaluates knowledge of an individual customer's attitude, needs, and values for cluster analysis. Data mining is a typical technique in this category.
3. Collaborative CRM systems synchronizes customer communication time through channels such as e-mail, the Internet, and/or the telephone.

In the literature most studies on KM and CRM are treated in separate research domains. However, lately their mutual synergy potential has drawn the attention of researchers in the field. By employing KM in an effort to help CRM to transcend from its original technology-driven and data oriented approach into a more people-oriented 'customer knowledge management' model or CKM model, it has already invoked a convergence of the two [3, 4] The CKM model emphasizes a bi-directional communication channel. This interaction with customers and customer knowledge management, set up strategies for how a company can develop attractive innovative products, or improve its services to win the satisfaction of its customers.

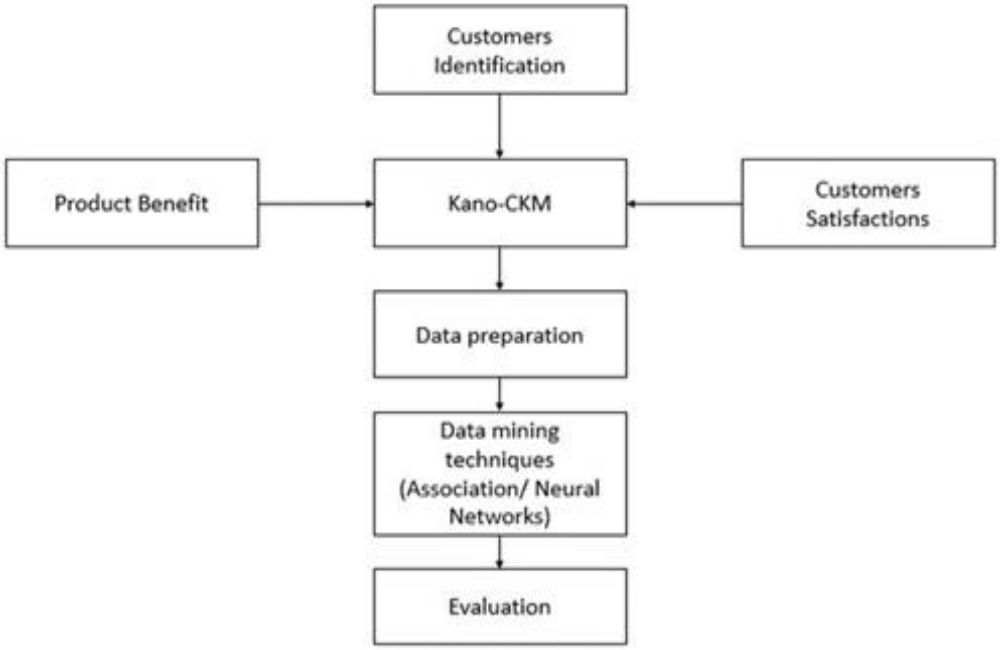

Fig. 3. Proposed CKM framework based on data mining

III. Proposed Data Mining based Banking CKM Framework

A. Proposed Framework

The proposed data mining based CKM framework is shown in Fig.3. Understanding the product benefits, customers identifications, and their satisfactoriness forms the initial phase of any problem in data mining. A close study and management of customer identification and their knowledge will help to identify attract and retain effective customers in the domain. The next phase of data preparation helps in preparing the data by the processes of cleaning, attribute selection, data transformation for further building up of models and their evaluation. Model construction in the CKM framework is a major step in which effective model that satisfy the business requirements are constructed. These models help in predicting the behaviour of the customers. Model evaluation measures the effectiveness of the model for enhancing their performance.

B. Neural Network and Association Rule Mining

Neural networks (NNs) are multi-layer networks of neurons that are used to classify things or make predictions. Among the input, output and hidden layers of neurons[5] the actual computations of the network are performed in the hidden layer, where each neuron sums its input attributes x_i after multiplying them by the strengths of the respective connection weights $w_{ij.}$ The activation function (AF) of this sum gives the output y_j and sigmoid function[16] is the AF used in the experiment [20].

$$y_j = f\left(\sum w_{ij}, x_i\right) \qquad (1)$$

Back-propagation (BP) learning is the most common training technique used for NNs. The sum of squared differences between the desired and asset value of the output neuron's E is defined as:

$$E = {}^{1}/_{2} \sum j(y_{dj} - y_j)2 \qquad (2)$$

Where y_j is the output of a neuron j whose desired value is y_{d_j}. Weights $w_{i,j}$ in equation (1), are adjusted to finding the minimum error E of equation (2) as early as possible. The difference between the network outputs and the desired ones is reduced by the application of weight correction by BP. The neural networks help in learning and reducing the future errors. Good learning ability, fast real-time operation, less memory demand, analysis of complex patterns is some of the advantages of NNs and the disadvantages include high-quality data requirement of the network, and careful selection of parameters.

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Association Rule Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [21,22]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories.

Let $I = I_1, I_2, \cdots, I_m$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called item sets, and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support (***s***) and confidence (***c***). The two basic parameters of Association Rule Mining (ARM) are: support and confidence. Support (s) of an association rule is defined as the percentage/fraction of records that contain $X \cup Y$ to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account.

Support (s) is calculated by the following:

$$Support(XY) = \frac{Support\ count\ of\ XY}{Total\ number\ of\ transaction\ in\ D} \qquad (3)$$

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule $X \Rightarrow Y$ can be generated.

$$\text{Confidence}(X|Y) = \frac{Support(XY)}{Support(X)} \qquad (4)$$

IV. EXPERINMENTS AND RESULTS

A. Study Dataset

The dataset used for experiments in this paper, contains results of the Saman bank customers data. Saman Bank is one of the Iranian private banks which estabilised recently [23].

TABLE I. DATASET FEATUTERS DESCRIPTIONS

Feature Number	Feature Name	Type
1	age	Numeric
2	average yearly balance	Numeric
3	job type	Numeric
4	marital status	Numeric
5	education	Numeric
6	type of account	Numeric
7	gender	Categorical
8	type of loan	Categorical
9	satisfactory of electronic bank services	Categorical
10	satisfactory of security	Categorical
11	satisfactory of bank facilities	Categorical
12	satisfactory of complain reports	Categorical
13	satisfactory of banking service	Categorical
14	satisfactory of bank informativity.	Categorical

Data includes answers of 124 Saman Bank customers and were collected by questionnaires. From the overall 255 filled questionnaires only 124 questionnaires were completely answered all questions and then were selected to use in this study. The questionnaire includes 21 questions which cover the bank services and customers satisfaction rates. The variables of the questionnaire are: age, average yearly balance in Iranian Rials, job type, marital status, gender, education, type of account, type of loan, satisfactory of electronic bank services, satisfactory of security, satisfactory of bank facilities, satisfactory of complain reports, satisfactory of banking service, satisfactory of bank informativity. All these variables are used as features for prediction of customer behaviour in two classes (satisfactory/non-satisfactory). The details of the features are described in Table I.

B. Experinment Setup

All experiments were performed using weka tool and were conducted in windows 10 with Intel Core i7 processor. The Waikato Environment for Knowledge Analysis (Weka) is a machine learning toolkit used extensively for research, education and projects. Weka is introduced by Waikato University, New Zealand and is open source software written in Java (GNU Public License). It consists of collection of machine learning algorithms and tools for data mining tasks such as data pre-processing or data preparation, classification, association rules, clustering, regression, forecasting and visualization and is well suited for developing new machine learning schema.

In this work, we build two distinct DM classifier models: Neural Networks (NN) and Association Rule (AR). For all the two models test mode of tenfold cross validation was used. NN uses back propagation to classify the instances. The nodes are all sigmoid except for when the class is numeric. We set

TABLE II. CONFUSION MATRIX

	Predicted: No	Predicted: Yes
Actual: No	TN	FP
Actual: Yes	FN	TP

the number of hidden layers using the heuristic a = round (M/2) where *M* is the sum of attributes and classes. Other

network parameters were set as follows: learning rate *0.2*, momentum *0.1*, training time 100ms.

During the modelling phase we successfully tested the two models, NN and AR using the Weka tool. Based on the response, two classes were obtained, those which responded positive and those responded negative.

C. Evaluation Metrics

Evaluating of a machine learning algorithm is an essential part of any project. Most of the time classification accuracy is used to measure the performance of a model, however it is not enough to truly judge a model. Classification accuracy and validation accuracy are the ratio of number of correct predictions to the total number of input samples in both training and testing sets. Different evaluation metrics were measured in this research to measure the effectiveness of the proposed method and hypotheses such as ROC, accuracy, TPF, FPR, F-measure, precision, recall, sensitivity, and specificity.

A confusion table as shown in Table II has two rows and two columns that reports the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Accuracy: the percentage of instances classified correctly into a given category in relation to the total number of instances tested:

Accuracy = (TP + TN) (TP + TN + FP + FN)

The F score is used to measure a test's accuracy, and it balances the use of precision and recall to do it. The F score can provide a more realistic measure of a test's performance by using both precision and recall. F-measure and precision are calculated based on FP, TN, FN, and TP.

ROC area (area under the ROC curve): In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

D. Results and Discussion

The results of our experiment for automatically classifying a given dataset are summarized in Table III. Which shows values for two different classifiers. For each method the classification accuracy (amount of correctly classified instances), true positive rate (TPR) (the proportion of actual positives which are correctly identified as such), false positive rate (FPR) (incorrectly classified positive), ROC and the time taken to build the classifier model is shown.

TABLE III. RESULTS OF COMPARISON OF CLASSIFIERS, AVERAGE OVER 10 RUNS.

Classifier	Accuracy (%)	TPF	FPR	ROC	Time (s)
NN	97%	0.87	0.042	0.92	1500
AR	94.3%	0.84	0.052	0.90	12

TABLE. IV. F-MEASURE, PRECISION, RECALL, SENSITIVITY, SPECIFICITY METRICS RESULTS.

Classifier	F-Measure	Precision	Recall	Sensitivity	Specificity
NN	88%	0.83%	85.5%	84%	87%
AR	82%	78.5%	81%	79%	83.4%

NN classifier model shows better accuracy (97%) and ROC (92%) among the two models experimented. The time taken to build the model is very high for NN (1500 s). In this work three measures namely classification accuracy, ROC and confusion matrix are used to evaluate the performance of the classification models.

Also, the F-measure, precision, recall, sensitivity, and specificity measures are calculated, and results are shown in Table IV and Fig. 4. As it is shown, the f-measure, precision, recall, sensitivity, and specificity of NN is higher than AR in this experimental.

Out of the 124 instances in the dataset, the instances classified correctly, and instances classified incorrectly for each model is as shown in Table V. NN classified 120 instances correctly whereas AR classified 116 instances correctly.

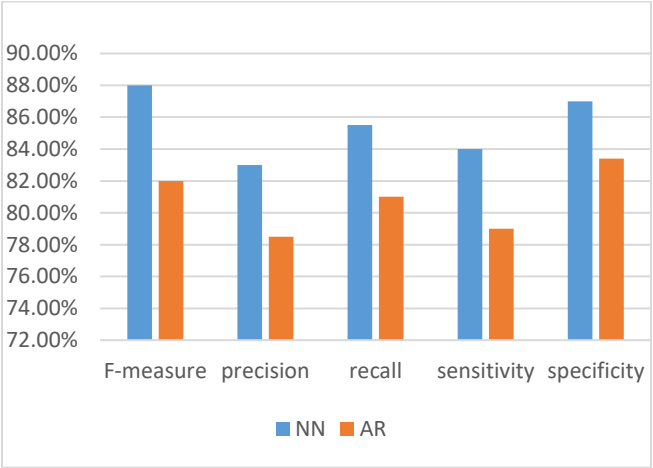

Fig. 4. The comparison between different evaluation metrics for NN and AR.

TABLE. V CLASSIFICATION OF 124 INSTANCES IN THE DATASET

Classifier	Correctly Classified Instants	Incorrectly Classified Instants
NN	120	4
AR	116	8

V. CONCLUSION AND FUTURE WORK

In this paper, an efficient CKM-data mining framework for the prediction of customer behavior was proposed. Two representative data mining techniques - Neural network and association rules – were applied in the new CKM base data mining framework for bank customer behavior predicting in the real world. Our findings indicate that both models attain good predicting performance and data mining techniques can be used to improve customer behavior understanding and predicting. The best model that achieves high predictive performance was Neural Network with accuracy rate of 97%. However, Neural Network takes longer time to train the model.

In the future, this work can be extended to use other new models like Neuro fuzzy classifiers, Ensemble of classifiers and so on in order to improve the predicting capacity. Also, the other large banking datasets should be used for testing the proposed new framework.

REFERENCES

[1] J. Dyche, "The CRM handbook: A business guide to customer relationship management." Addison-Wesley Professional, 2002.

[2] H. Gebert, H., M. Geib, L. Kolbe, and W. Brenner, "Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts [1]", *Journal of knowledge management*, *7*(5), pp.107-123, 2003.

[3] M. García-Murillo, and H. Annabi, "Customer knowledge management", *Journal of the Operational Research society*, *53*(8), pp.875-884, 2002.

[4] T.H Davenport, J.G Harris, and A.K Kohli, "How do they know their customers so well?", *MIT Sloan Management Review*, *42*(2), p.63, 2001.

[5] M. Abdar, E. Nasarian, X. Zhou, G. Bargshady, V.N. Wijayaningrum, and S. Hussain, "Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach". In: 4th IEEE International Conference on Computer and Communication Systems (ICCCS 2019), 23-25 Feb 2019 , Singapore.

[6] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P.D. Barua, R. Gururajan. "A new nested ensemble technique for automated diagnosis of breast cancer". Pattern Recognition Letters. 2018 Nov 4.

[7] X. Zhou, Y.Wang, G. Tsafnat, E. Coiera, F. T. Bourgeois, A. G. Dunn, Citations alone were enough to predict favorable conclusions in reviews of neuraminidase inhibitors, Journal of clinical epidemiology 68 (1) (2015) 87–93.

[8] X. Tao, X. Zhou, J. Zhang, J. Yong, Sentiment analysis for depression detection on social networks, in: International Conference on Advanced Data Mining and Applications, Springer, 2016, pp. 807–810.

[9] X Zhou, X Tao, MM Rahman, J Zhang, Coupling topic modelling in opinion mining for social media analysis, in proceedings of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 533-540.

[10] M Abdar, M Zomorodi-Moghadam, X Zhou, "An ensemble-based decision tree approach for educational data mining", In 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC 2018), 12-14 Nov 2018, Kaohsiung, Taiwan.

[11] T. F. Bahari and M. S. Elayidom, "An efficient CRM-data mining framework for the prediction of customer behaviour", Procedia computer science, vol. 46, pp. 725-31, 2015.

[12] A. Payne and P. Frow, "A strategic framework for customer relationship management", *Journal of marketing*, vol. 69, no. 4, pp. 167-76, 2005.

[13] A. Parvatiyar and J. N. Sheth, "Customer relationship management: Emerging practice, process, and discipline", *Journal of Economic & Social Research*, vol. 3, no. 2, 2001.

[14] L. Ryals and A. Payne, "Customer relationship management in financial services: towards information-enabled relationship marketing", *Journal of strategic marketing*, vol. 9, no. 1, pp. 3-27, 2001

[15] A. R. Zablah, N. B. Danny and J. J Wesley, "Customer relationship management: an explication of its domain and avenues for further inquiry", *Relationship Marketing, Customer Relationship Management and Marketing Management: Co-Operation–Competition–Co-Evolution*, pp. 115-24, 2003.

[16] M. J. Lanning and E. G. Michaels, "A business is a value delivery system", McKinsey staff paper, vol. 41, no. July, 1988.

[17] N. Bendapudi and R. P. Leone, "Psychological implications of customer participation in co-production", *Journal of marketing*, vol. 67, no. 1, pp. 14-28, 2003.

[18] K. Matzler and H. H Hinterhuber, "How to make product development projects more successful by integrating Kano's model of customer satisfaction into quality function deployment", Technovation, vol. 18, no. 1, pp. 25-38, 1998.

[19] Y-H. Chen and C-T. Su, "A Kano-CKM model for customer knowledge discovery," Total Quality Management & Business Excellence, vol. 17, no. 5, pp. 589-608, 2006

[20] I.H. Witten, E. Frank, M.A. Hall and C.J. Pal, *Data Mining: "*Practical machine learning tools and techniques", Morgan Kaufmann, 2016.

[21] R. Agrawal, T. Imieliński and A. Swami, "Mining association rules between sets of items in large databases", In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM, 1993.

[22] Q, Zhao and S.S, Bhowmick, "Association rule mining: A survey.", *Nanyang Technological University, Singapore*, 2003.

[23] *Saman bank*, 2010, http://www.cms.sb24.com/fa/aboutbank/index.html>.