Customer Relationship Management Classification by Hybridizing Genetic Algorithm and Fuzzy K-Nearest Neighbor

Jashandeep Kaur
Punjabi University
Regional Centre for IT and
Management Mohali,
160062(India)

Rekha Bhatia
Punjabi University
Regional Centre for IT and
Management Mohali,
160062(India)

ABSTRACT

Data mining is the procedure of extraction of data from different datasets on the premise of various attributes. In the CRM, various relational attributes are accessible in the dataset. Information about relations of the customer with the enterprise is available in the dataset. The dataset must be secured utilizing rules for extraction of information. Basically Churn, appetency, up selling and score are the significant entities which will be considered in the proposed work. To eliminate the problems of CRM database a new hybrid algorithm is introduced which will be the combination of GA and Fuzzy KNN classification.

Keywords

CRM, Types of CRM, Data Mining, Genetic Algorithm, Fuzzy KNN.

1. INTRODUCTION

1.1 CRM (Customer relationship management)

CRM stands for Customer Relationship Management, that help to maintain customized relationship with customers for creating higher level customer satisfaction and offering the finest customer service

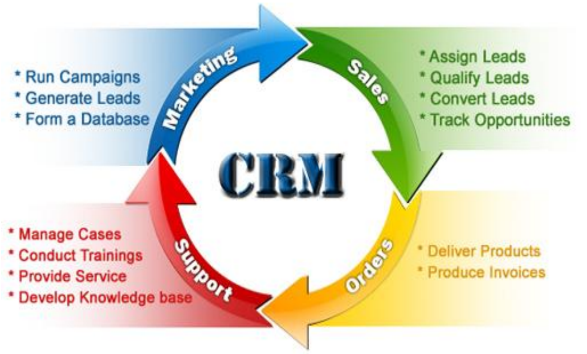

Fig 1.1: CRM

The relationship covers the relationship among customers and also among its attributes. For example products, hobby, revenue, etc. and from this relationship CRM team can use it to reach their objectives. This information can also be used to profile the customer to access their characteristic and behavior. Traditional CRM (Customer Relationship Management) contains 3 modules, marketing to gather information that will be delivered as lead, sales to follow up the lead to become revenue for the company, support to provide after sales service for the customer and gather information from customer that will also has the possibility to become lead and to order the product. Thus, the company's revenue depends on the marketing and support performance to retrieve and extract information from any source that has the possibility to become lead. The lead itself can be further categorized into several level of possibility. For example hot lead reflects a high possibility a lead can become revenue and cold lead reflects a low possibility a lead can become revenue.

1.2 Types of CRM

1.2.1 Operational CRM:

The primary goal of CRM systems is to integrate and automate sales, marketing, and customer support. Therefore, these systems typically have a dashboard that gives an overall view of the three functions on a single page for each customer that a company may have. The dashboard may provide client information, past sales, previous marketing efforts, and more summarizing all of the relationships between the customer and the firm. Operational CRM is made up of three main components: sales force automation, marketing automation, and service automation.

1.2.2 Analytical CRM*:*

The function of analytical CRM frameworks is to analyze customer information gathered through various sources and present it so that business administrators can make more precise decisions. Analytical CRM systems use procedures, for example, data mining, correlation, and pattern recognition to analyze the customer data. These analytics enhance customer service by discovering little issues which can be resolved perhaps, by advertising to various parts of a consumer audience differently. For instance, through the analysis of a customer base purchasing behavior, an organization may see that this customer base has not been purchasing a considerable amount of items recently. After scanning through this data the company might think to market to this subset of consumers differently, in order to best communicate how this company's products might benefit this group specifically.

1.2.3 Strategic CRM:

Strategic CRM is a type of CRM in which the business puts the customers first. It collects, segregates, and applies information about customers and market trends to come up with better value proposition for the customer**.** The business considers the customer’s voice important for its survival. In contrast to Product-Centric CRM (where the business assumes customer requirements and focuses on developing the product that may sometimes lead to over-engineering), here the business constantly keeps learning about the customer requirements and adapting to them.

1.2.4 Collaborative CRM:

The third primary aim of CRM systems is to incorporate external stakeholders such as suppliers, vendors, distributors and share customer information across organizations. For example, feedback can be collected from technical support call which could help provide direction for marketing products and services to that particular customer in the future.

1.3 Data Mining

Data mining is pivotal for extracting and recognizing helpful information from a lot of data that is the reason retailing organizations operate purchase databases in a long way, such that all transactions are put away in organized order. A record-of-transaction database contains the transaction date and the items purchased over the span of a given transaction. Typically, every record contains shopper ID particularly when it is purchased using credit card or a regular buyer card. Therefore, the purchasing sequence of an e-shopper in the database that has made repetitive purchase can easily be determined. This purchase sequence provides a description of the changes in an e-shopper's preferences over time, because purchase sequence can reveal the changes of e-shopper preferences overtime.

1.4 Algorithm Used:

1.4.1 Fuzzy KNN Algorithm

Fuzzy KNN is a simple, effective and non parametric classification technique that is broadly utilized in text classification. However there is a common issue, when the density of training data is uneven it might reduce the accuracy of classification.If we only consider the sequence of first k nearest neighbours but do not consider the differences of distances. To resolve this issue, we need to us the theory of fuzzy sets, developing new membership function rely on document similarities. A correlation between the proposed technique and other existing KNN methods is made by tests. The experimental output shows that the algorithm established on the theory of fuzzy sets (F KNN) can introduce the precision and recall of text categorization to a certain level.

1.4.2 Genetic algorithm

The genetic algorithm is established on natural selection method which is utilized for both constrained and unconstrained optimization problems. The procedure that drives biological development. The genetic algorithm repeatedly evolves a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. You can apply the genetic algorithm to solve a variety of optimization problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, non differentiable, stochastic, or highly nonlinear. The genetic algorithm can address problems of mixed integer programming where some components are restricted to be integer-valued. The genetic algorithm uses three main types of rules at each step to create the next generation from the current population.

2. REVIEW OF LITERATURE

S.Ummugulthum Natchiar et al [1] "**Customer Relationship Management Classification Using Data Mining Techniques**" conclude Customer Relationship Management fulfill the needs of customers and increases the revenue. Data quality and integrity have been improved by data mining techniques and classify the data into different classes. In this a new feature selection method has been proposed. The SVM, Naive Bayes, J48 and KNN algorithms were used and the performance is analyzed by using various parameters.

Nedaabdelhamid et al [2] "Emerging trends in associative classification data mining" In this paper an association rule discovery has been used for enhancing the predictive performance of classifiers. The IF-THEN rules was developed for finding the hidden relation between the several attribute values and maintain the quality of rules. In the previous Associative algorithm provide better classification of data sets but there are some emerging trends in the associative rule discovery.

Sankaranarayanan, S. et al [3] "Diabetic Prognosis through Data Mining Methods and Techniques" The various data mining techniques has been used for the prediction of the diseases. These techniques can select and expired the hidden patterns from the database. These patterns are determined and give the decision to the clinical researchers that this medicine could be effective to the patient according to their disease. The FP growth and Apriori algorithms was introduced for the medical domain. These algorithms induce the association rules and classify the patients.

Wang, Guoyin et al [4] "Granular computing based data mining in the views of rough set and fuzzy set" The knowledge in the database are already available but there is no use of this type of knowledge, if they are not in the human understandable format. The transformation process can take place in order to convert into a human understandable format. This can be done by fuzzy set and rough set for converting the knowledge at the granular level. It could convert the fine granularities into coarse granularities.

Tzung-Pei Hong et al [5] "Using divide-and-conquer GA strategy in fuzzy data mining" In this paper and Genetic algorithm based framework has been used for the mining of membership function and induces the fuzzy association rules. The GA based framework has been utilized for finding of the membership function from the transaction data. In the previous work the membership function was assumed. The membership function could be encoded into the string representation because a GA has to be applied. The framework has number of populations then from each population, the best item of membership has collected and which mine the association rules. The framework saves the acquisition time while collecting the data and human expert's intervention

Tzung-Pei Hong et al [6] "GA based item partition for data mining" Due to the memory limitation, the mining was not to be done on the whole database. The various algorithms were introduced. All the Algorithms have some drawbacks.to overcome the problems and GA has been proposed. It saves the time while searching the item sets. In this an FP growth algorithm and FP tree data structure has been used. The big items divided into the different partitions and this partition has not more cross over between the different sub groups.

CRM Dataset: In the customer relationship management dataset, different relational attributes are available in the dataset. This dataset contains the information about the relations of the customer with an enterprise. The dataset has to be classified using rules for extraction of information. Mainly Churn, appetency, up selling and score are the major entities which will be considered in the proposed work. In CRM dataset various instances are available that are combination of different features.

3. METHDOLOGY

Data mining is a process for extraction of valuable information from raw data. In the process of data mining raw information has been pre-processed and used for extraction of different relationship between different attributes. In the purposed research CRM data has been classified for enhancing satisfaction level of different customers on the basis of their demands.

In the purposed research classification of CRM dataset has been done using fuzzy membership based classifier that has been optimized using genetic approach. In the purposed research classification has been done by using hybrid genetic approach with fuzzy based classification approach.

Genetic algorithm is a nature based inspired approach that utilizes different selection, mutation, and crossover and replacement process for selection of best centroid of a particular class. On the basis of fitness evaluation of centroid selection fuzzy based KNN assign membership weight to a single instance according to class. Simple KNN approach assigns attribute vector to a single class.

In FKNN a membership weight has been assigned to vector rather than that vector belongs to a particular class. In this process a single instance can have multiple weightage vectors that can be used for classification of the instance to different classes. In the process of classification nearest neighbour has been finding on the basis of distance computation from other instances available in the dataset.

In this research Euclidian distance has been used for computation of distance between different instances available in the dataset from the class centroid. Centroid has been used for selection of nearest neighbour from the dataset instances.

X=[X1, X2, X3……., Xn] are the centroid available in the dataset from all the instances. Y = [Y1, Y2, Y3,,,,,,,,,,,,,,,,,,,,,,,Yk] instance available in the dataset. On the basis of distance these instances these instances have define fuzzy weightage that has been used for classification of the dataset instance to a single class.

$$d = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_i - Y_j\right)^2} \qquad (1)$$

On the basis of above define equation (1) distance between different instances from centre point has been measured and fuzzy weightage membership has been assigned to instance on the basis of all classes. The instance that contain highest weightage value belongs to a particular class has assigned that class label. Genetic approach has been used for optimization of predicted classes on the basis of nature inspiration approaches. In the purposed research various parameters have been evaluated that has been explained in results section. On the basis of these parameters classification approach accuracy has been measured.

4 RESULTS

In the process of CRM classification class labels has been predicted for different classes available in the dataset. In this process training dataset has been loaded for extraction of features from the dataset that can be used for prediction of class label.

In the purposed work center point has been selected from the training dataset that has been used for computation of minimum distance between class center points to dataset instance. On the basis of distance and nearest neighbor different dataset class labels have been predicted. In the purposed work various performance evaluation parameter have been measured for classification

These parameters are accuracy, sensitivity, specificity, precision recall and f-measure. On the basis of these parameters performance has been evaluated for purposed system.

- **Confusion Matrix:** A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix.
- **Accuracy**: the proportion of the total number of predictions that was correct.
- **Precision** the proportion of positive cases that were correctly identified.
- **Negative Predictive Value:** the proportion of negative cases that were correctly identified.
- **Sensitivity or Recall**: the proportion of actual positive cases which are correctly identified.
- **Specificity:** proportion of actual negative cases which are correctly identified.
These above described parameters are essential for performance evaluation of purposed system on the basis of these parameters classifier performance can be measured.

Table 4.1Perfromance evaluation for classification of dataset into two different classes

Classifiers/metrics	GAF	SVM	KNN	FKNN
Accuracy	0.98	0.96	0.87	0.93
Sensitivity	0.92	0.9	0.8	0.9
Specificity	0.16	0.18	0.2	0.1
Precision	0.92	0.87	0.73	0.83
Recall	0.95	0.89	0.65	0.85
F-measure	0.97	0.88	0.79	0.89
G-mean	0.85	0.75	0.73	0.89

This table represents classification parameters that have been measured for performance evaluation of purposed classifier. These parameters have been evaluated on dividing whole dataset into two different classes.

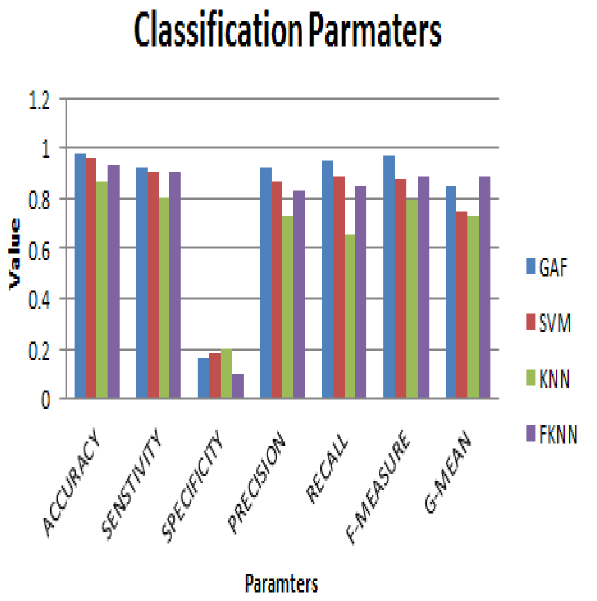

Fig 4.1 Data classification parameters for two different classes

This figure represents classification performance parameters for two distinct classes division of dataset. In this classification whole dataset has been divided into two classes and classification accuracy, specificity, sensitivity and precision has been measured using testing labels and predicted class labels. By analysing these parameters value we can say that hybrid classifier outperform over SVM, KNN and simple fuzzy KNN classifier.

After division of dataset into different classes three classes has been labelled in the testing and training dataset. After this process testing dataset has been classified for prediction of three different class labels according to training dataset.

Table 4.2 Performance evaluation for classification of dataset into three different classes

Classifiers/metrics	GAF	SVM	KNN	FKNN
Accuracy	0.86	0.80	0.66	0.76
Sensitivity	0.76	0.70	0.45	0.55
Specificity	0.24	0.30	0.55	0.35
Precision	0.75	0.70	0.7	0.73
Recall	0.85	0.75	0.62	0.72
F-measure	0.86	0.78	0.75	0.75
G-mean	0.85	0.60	0.67	0.71

This table represents classification accuracy and different performance evaluation parameters of the dataset using different classifier. On the basis of different classification approaches Genetic with fuzzy KNN performs better.

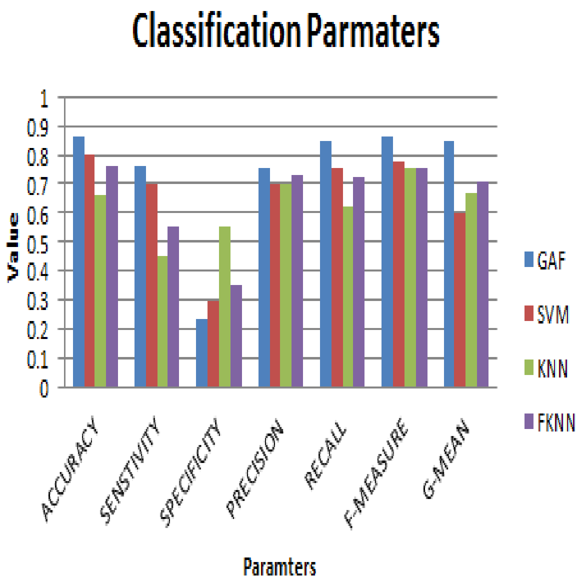

Fig 4.2 classification parameters for three different classes

This figure represents classification performance parameters for three distinct classes division of dataset. In these three different classes has been divided for testing dataset that creates on other testing labels of testing dataset. On the basis of these classes some instance also belongs to third class out of two classes and performance evaluation parameters have been represented in graphical format.

Table 4.3 Performance evaluation for classification of dataset into four different classes

Classifiers/metrics	GAF	SVM	KNN	FKNN
Accuracy	0.70	0.64	0.52	0.6
Sensitivity	0.66	0.56	0.52	0.51
Specificity	0.36	0.44	0.48	0.49
Precision	0.75	0.65	0.54	0.62
Recall	0.70	0.68	0.63	0.69
F-measure	0.69	0.67	0.61	0.64
G-mean	0.68	0.61	0.55	0.59

This table represents classification parameters that have been measured for performance evaluation of purposed classifier. These parameters have been evaluated on dividing whole dataset into four different classes.

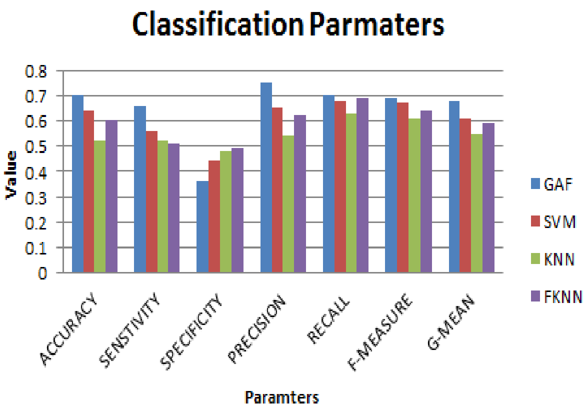

Fig 4.3 Classification parameters for four different classes

This figure represents classification performance parameters for four distinct classes division of dataset. In these four different classes has been divided for testing dataset that creates on other testing labels of testing dataset. On the basis of these classes some instance also belongs to fourth class out of three classes and performance evaluation parameters have been represented in graphical format.

5. CONCLUSION AND FUTURE SCOPE

In the purposed work the fuzzy based membership function has been used for classification. This approach assigns the weight age to dataset attributes on the basis of fuzzy membership rules. After the assignment of weight age the data set distance has been computed using Euclidian classifier and the function has been used that use both weight age and distance factor for prediction of class label to a single dataset. After this classification genetic algorithm has been used for optimization of predicted label. Genetic algorithm use different operators like crossover, mutation, and selection for prediction of best classification labels to a data samples. After this the actual and predicted class labels have been used for prediction of various parameters that have been used for validation of purposed work. The purposed work has been compared with various approaches like KNN, Fuzzy KNN and SVM by analyzing various parameters like accuracy, precision, recall, f-measure, sensitivity, specificity and G-mean. It concludes that the Fuzzy and Genetic Algorithm based approach provide better classification rather than that of the simple Fuzzy classification and KNN classification.

In the future reference this classier model can be used for classification other huge dataset. In future reference one can research for extraction of association rules between different attributes of CRM dataset that can be used for best classification to support business planning.

6. REFERENCES

[1] S.Ummugulthum Natchiar "*Customer Relationship Management Classification Using Data Mining Techniques*", International Conference on Science, Engineering and Management Research, 2014, pp 223-234.

[2] Nedaabdelhamid, Aladdin Ayesh and FadiThabtah "*Emerging trends in associative classification data mining*" International journal of electronics and electrical engineering Volume 3, Issue 1,_ Feb 2015.

[3] Sankaranarayanan, S. "Diabetic Prognosis through Data Mining Methods and Techniques", *International Conf. on Intelligent Computing Applications (ICICA),* 2014, pp. 162 – 166.

[4] Wang, Guoyin "Granular computing based data mining in the views of rough set and fuzzy set" *IEEE Conf. on Granular Computing*, 2008, pp. 67.

[5] Tzung-Pei Hong "*Using divide-and-conquer GA strategy in fuzzy data mining*" IEEE Conf. on Computers and Communications, 2004, pp. 116 - 121 Vol.1.

[6] Tzung-Pei Hong "GA-*based item partition for data mining*" IEEE Conf. on Systems, Man, and Cybernetics (SMC), 2011, pp. 2238 – 2242.

[7] Jo-Ting Wei "Customer relationship management in the hairdressing industry: An application of data mining techniques*", IEEE Conf. on Expert Systems with Applications, 2013*, pp Pages 7513–7518.

[8] Wen-Yu Chiang "Applying data mining with a new model on customer relationship management systems: a case of airline industry in Taiwan", *Conf. on Data Mining*, 2014, pp 89-97.

[9] Alexander Tuzhilin "*Customer relationship management and Web mining: the next frontier*", Springer conf. on CRM & WM, 2012, pp 584-612.

[10] Siavash Emtiyaz "Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management", *Advances in information Sciences and Service Sciences (AISS)*, 2011, PP 56-67.

[11] Farnoosh Khodakarami "*Exploring the role of customer relationship management (CRM) systems in customer knowledge creation*", Conf. on CRM, 2014, PP 56-70.

[12] Paresh Tanna "Using Apriori with WEKA for Frequent Pattern Mining", *International Journal of Engineering Trends and Technology (IJETT),* 2014, pp. 127-131.

[13] Shrey BavisiÀ "*A Comparative Study of Different Data Mining Algorithms*", International Journal of Current Engineering vand Technology, 2015, pp. 3248-3252.

[14] Manjari Anand "Customer Relationship Management using Adaptive Resonance Theory", *International Journal of Computer Applications*, 2013, pp. 43-47.

[15] Ms. Saranya, **"***Decision Support System for CRM in Online Shopping System***",** International Journal of Advances in Computer Science and Technology, 3(2), February 2014, 148, 2014, pp. 148-151.