International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Data Mining of CRM dataset using GAF Algorithm

[1]Er. Priyanka Aggarwal, [2]Er. Navpreet Rupal
[1]MTech (CSE)
Shaheed Udham Singh College of Engg. & Technology, Tangori
[2]Asst. Prof
Shaheed Udham Singh College of Engg. & Technology, Tangori
[1]prinka_aggarwal@yahoo.co.in, [2] er.nrupal@gmail.com

Abstract: *Data mining is the process of extraction of information from various datasets on the basis of different attributes. Mining has to be done to extract hidden relationship between various database entities. On the basis of these entities, different types of decisions are taken for the extraction of different relationships. In the customer relationship management, different relational attributes are available in the dataset. In the paper, To overcome the problems of CRM database a new hybrid algorithm is introduced which is the combination of GA and fuzzy KNN classification.*

Keywords: *CRM,KNN, Genetic algorithm ,mining, k-NN, fuzzy k-NN.*

1. INTRODUCTION

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"
This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

1.1 The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining

algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

1.2 CRM Dataset

CRM software is like understanding the difference between a data nerd and a data evangelist. A data nerd, as I was back in 1989, likes to rummage through the data to understand what's going on. A data evangelist, which is what I wanted to become, wants to interpret that data and use it to move forward.

A database is just that – a base full of data. You can drill down into it, you can extract items of data, you can create pivot tables… and that's about it.
A CRM is more of a process than a piece of software. It takes the database as its foundation and lets you collect, interpret and use the data for business improvement.
Let's create an example: you are collecting data on customers. You have all of their personal information, all of their product information, their spend, their interactions with you, their level of satisfaction with you… you name it, you've collected it. That's in your database.
A CRM is what allows you to collate that data in one simple view – a dashboard, if you like. It allows you to create processes around the data. For example, you may want to know when a specific customer spends more than £1,000, and create an alert that says "get in touch". You may want to create an alert around customers who stop spending, so that you can get in touch and revive their engagement with you. You've taken the manual deep-dive into a database and made it automatic, intelligent and business-driven.

1.3 GENETIC ALGORITHM

There are no known polynomial time algorithms to solve many real-world optimization problems making them hard to solve. A number of heuristics have been designed to solve the hard problems. These heuristics may provide sub-optimal but acceptable solution in a reasonable computational time. A number of meta-heuristics such as simulated annealing, evolutionary algorithms, and artificial neural networks derived from natural physical and biological phenomena have also been used to solve these problems. Genetic Algorithms (GAs) are adaptive procedures derived from Darwin's principal of survival of the fittest in natural genetics. GA maintains a population of potential solutions of the candidate problem termed as individuals. By manipulation of these individuals through genetic operators such as selection, crossover and mutation, GA evolves towards better solutions over a number of generations. Genetic algorithms start with randomly created initial population of individuals that involves encoding of every variable. A string of variables makes a chromosome or individual.

1.4 FUZZY K-NEAREST NEIGHBOR ALGORITHM (FUZZY K-NN)

The k-nearest neighbor (k-NN) algorithm is one of the oldest and simplest non parametric pattern classification methods. In the k-NN algorithm a class is assigned according to the most common class amongst its k nearest neighbors. In 1985, Keller proposed a fuzzy version of k-NN by incorporating the fuzzy set theory into the k-NN algorithm, and named it as" fuzzy k-NN classifier algorithm"(Fuzzy k-NN). According to his approach, rather than individual classes as in k-NN, the fuzzy memberships of samples are assigned to different categories according to the following formulation.

Let $W=\{x_1,x_2,.....x_n\}$ be the set of *n* labelled samples. Also let $\mu_i(x)$ be the assigned membership of vector x(to be computed), and U_{ij} be the membership in the *i*th class of the *j*th vector of the labelled sample set.

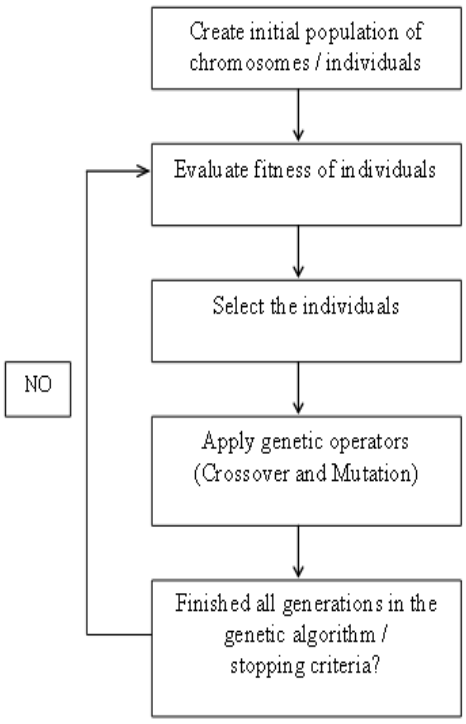

Figure-1: Flowchart of a genetic algorithm

The algorithm is as follows:

Steps :

BEGIN
Step 1. Input x, of unknown classification.

Step 2. Set k, 1<= k < n.
Step 3. Initialize i=1.
Step 4. DO UNTIL(k-Nearest Neighbors to x found).
Step 5. Compute distance from x to x_i.
Step 6. IF(i<=k)THEN
Step 7. Include x_i in the set of k-Nearest Neighbors.
Step 8.ELSE IF(x_i is closer to x than any previous NN) THEN
Step 9. Delete the farthest of the k-Nearest Neighbors.
Step 10. Include x_i in the set of k-Nearest Neighbors.
Step 11. END IF
Step 12. END DO UNTIL
Step 13. Initialize i=1.
Step 14. DO UNTIL(x assigned membership in all classes).
Step 15.Compute $\mu_i(x)$
Step 16. Increment i.
Step 17. END DO UNTIL
Step 18. END.
Step 19. Where $\mu_i(x)$ $\frac{\sum_{j=1}^{k} u_{ij}(1/|y-xj|^{2/(m-1)}}{\sum_{j=1}^{k} u_{ij}(1/|y-xj|^{2/(m-1)}}$

END

The value of i is initialized until k-Nearest Neighbors to found. Compute distance between unknown classification and k-Nearest Neighbors. If i is less than or equal to value of k then include x_i in the set of k-Nearest Neighbors else x_i is closer to x than any previous NN. Then delete the farthest of the k-Nearest Neighbors and again Include x_i in the set of k-Nearest Neighbors and do until x assigned membership in all classes and compute $\mu_i(x)$ with increment value of i.

2. PROPOSED WORK

In the classification of customer relationship management dataset various attributes that has to be mined on the basis of decision. Various stages that has to be carried out for development of proposed model:

Phase 1

In this phase, data is acquired using various acquisition tools this data has to be stored in structured format later on mining is performed on the basis of defined rules.

Phase 2

In this phase, structured data has to be normalized and pre –processed. This database will be classified into different classes on rules of GA and Fuzzy KNN classification. After classification various attributes will be extracted from classified data and various parameters are analysed for performance evaluation.

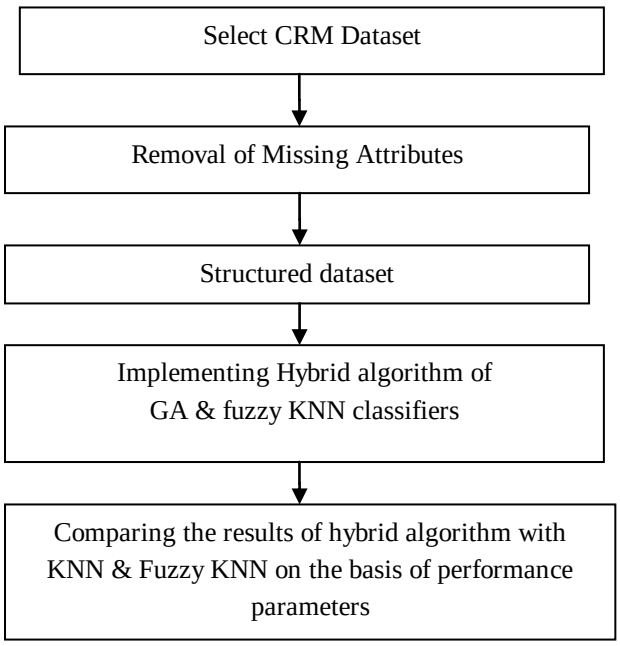

Fig 2: Flow of Work

3. RESULTS AND DISCUSSION

In the purposed work different parameters have been used for evaluation of the performance of the purposed work. In this purposed work CRM dataset have been used that has 50000 instances and 231 attributes.

This dataset has been used for classification using GAF approach this approach use membership function and genetic operation to classify the data into different classes. To validate the purposed work different algorithm have been used for comparison of the purposed approach. The purposed work has been compared with Fuzzy-KNN classification and KNN classification approach on the basis of different parameters.

The dataset have been classified into different classes on the basis of the value of K. the parameters have been computed by using these different values of K.

Classifiers/metrics	GAF(hybrid)	KNN	FKNN
ACCURACY	0.98	0.87	0.93
SENSTIVITY	0.92	0.8	0.9
SPECIFICITY	0.16	0.2	0.1
PRECISION	0.92	0.73	0.83
RECALL	0.95	0.65	0.85
F-MEASURE	0.97	0.79	0.89
G-MEAN	0.85	0.73	0.89

Table 1. Parameters for classification in two classes

This table represents the classification of the CRM dataset values when classification has been done on the basis of two classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset.

Classifiers/metrics	GAF(hybrid)	KNN	FKNN
ACCURACY	0.76	0.86	0.76
SENSTIVITY	0.65	0.45	0.55
SPECIFICITY	0.25	0.55	0.35
PRECISION	0.65	0.7	0.73
RECALL	0.79	0.62	0.72
F-MEASURE	0.87	0.75	0.75
G-MEAN	0.65	0.67	0.71

Table 2 Parameters for classification in three classes

This table represents the classification of the CRM dataset values when classification has been done on the basis of three classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset. As the number of classes increases in the prediction of the dataset accuracy gets decrease due to availability of actual dataset into two classes.

Classifiers/metrics	GAF(hybrid)	KNN	FKNN
ACCURACY	0.64	0.52	0.6
SENSTIVITY	0.56	0.52	0.51
SPECIFICITY	0.56	0.5	0.4
PRECISION	0.65	0.54	0.62
RECALL	0.59	0.63	0.69
F-MEASURE	0.69	0.61	0.64
G-MEAN	0.61	0.55	0.59

Table 3 Parameters for classification in four classes

This table represents the classification of the CRM dataset values when classification has been done on the basis of four classes. These approaches classified the testing data into two different classes on the basis of distance of different samples from the training dataset. As the number of classes increases in the prediction of the dataset accuracy gets decrease due to availability of actual dataset into two classes.

3.1 Comparison of Evaluation Parameters

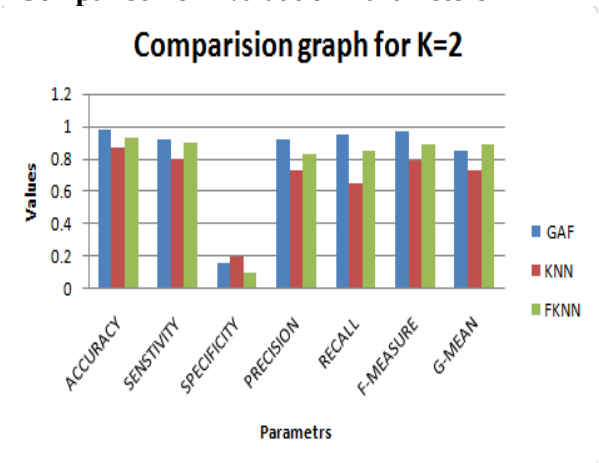

Fig Comparison of different approaches for K=2

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into two different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

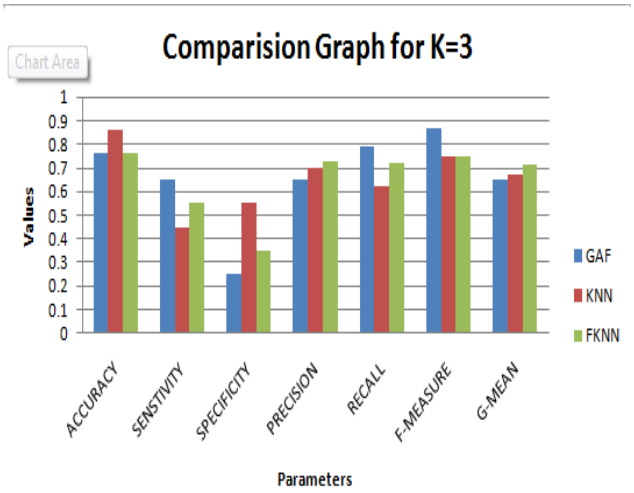

Fig 4.2 Comparison of different approaches for K=3

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into three different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

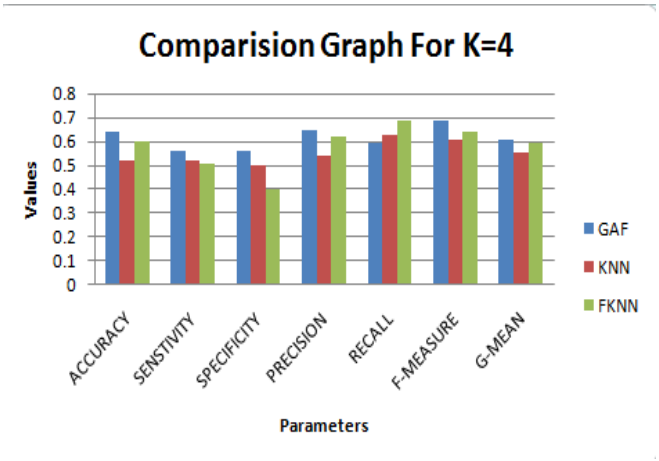

Fig 4.3 Comparison of different approaches for K=4

This figure represents the comparison between different classification approaches on the basis of different parameters. These approaches have been used for classification of data into four different classes and the parameters have been analyzed for all the approaches. On the basis of these parameters one can validate and optimized best approach for classification.

CONCLUSION

The different attributes have been used for classification on the basis distance between dataset attributes of training data and testing data. By analyzing various parameters like accuracy, precision, recall, f-measure, sensitivity, specificity and G-mean one can conclude 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 the data can be used for classification by using other artificial intelligence approaches for optimization of predicted labels.

REFERENCES

[1] Balaji Padmanabhan,"Data Mining Overview and Optimization Opportunities" Microsoft Research Report MSR-TR-98-04, January 2013.

[2] Chandra, S "Creation of an Adaptive Classifier to enhance the classification accuracy of existing classification algorithms in the field of Medical Data Mining", International Conf. of Computing for Sustainable Global Development (INDIA Com), 2015, pp 376 – 381.

[3] C Namrata Mahender "Text Classification and Classifiers a Survey" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[4] Ghareb, A.S "Text associative classification approach for mining Arabic data set", International Conf. on Data Mining and Optimization (DMO), 2012, pp 114 – 120.

[5] Jian Cheng1, 2, 3 and Yongheng Zheng1,"Object-oriented Classification of High-resolution Remotely Sensed Imagery" IPCSIT Vol. 47,pp. no.123,july 2010.

[6] Misra, B.B. G "Simplified Polynomial Neural Network for classification task in data mining", International Conf. on Evolutionary Computation, 2007, pp 721 – 728.

[7] Mahender. Dalla Mura, A. Villa, J. A. Benediktsson, J.Chanussot, and L. Bruzzone, ―Classification ofhyperspectral images by using extended morphologicalattribute profiles and independent component analysis,"IEEE Geosci. Remote Sens. Lett., Vol. 8, pp.no 542–546, May 2011.

[8] Mooney,"A survey of association rule mining using genetic algorithm", International journal of computer application & information Technology Vol.1, IssueII, ISSN: 2278, pp. no.1-8,Auguest 2012.

[9] Maria Vargas Vera "Knowledge Extraction by using an Ontology based Annotation Tool" Knowledge Media Institute (KMi), The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom.

[10] P.Pooja, J.Jayanthv and S. Koliwad, "Classification of RS data using Decision Tree Approach," International Journal of Computer Applications,Vol. 23(3), pp. no.7-11, Febuary2011.

[11] Pang Ning Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining" ,Addison Wesley, pp. no.769,June 2005.

[12] Raymond J. Mooney "Mining Knowledge from Text Using Information Extraction" Department of Computer Sciences University of Texas at Austin 1 University Station C0500, Austin, TX 787120233 Volume 7, Issue 1, pp 10, 2010.

[13] Suneetha K.R "Data Preprocessing and Easy Access Retrieval of Data through Data Ware House" Proceedings of the World Congress on Engineering and Computer Science, 2009, Vol I, October 20-22, 2009, San Francisco, USA.

[14] S .subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22 2013.

[15] Shambharkar, P.G "Automatic classification of movie trailers using data mining techniques: A review", International Conf. on Computing, Communication & Automation (ICCCA), 2015, pp 88 – 94.

[16] Usama Fayyad, Gregory Piatestsky-Shapiro Smith, "Knowledge Discovering and Data Mining: towards a Unifying Framework", In Proc. Of KDD-996, Vol. 2, pp. no.20―30, July.

[17] Vargas, J., Yu, H. and Jiang, X. Privacy preserving SVM classification. Knowledge and Information Systems,Vol.14,pp. no.161-178,September 2007.

[18] Weiguo Fan, Linda Wallace,"Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, Vol. 1, pp. no. 240―330, September 2005.

[19] Xiaolin Zhang "Research on privacy preserving classification data mining based on random perturbation", IEEE Conf. on Information Networking and Automation (ICINA), 2010, pp V1-173 - V1-178.

[20] Yu Qiao,HuiPingLiua, MuiBai, XiaoDong Wang, XiaoLuo Zhou, "The decision tree algorithm of urban extraction from multisourceimage data".Vol. 3, pp. 301―308, June 2005.