Volume 3, No. 5, May 2012

Journal of Global Research in Computer Science

RESEARCH PAPER

Available Online at www.jgrcs.info

ISSN-2229-371X

CLUSTER DETECTION USING GA-KNN CONJUNCTION APPROACH

Manoj Kumar Rawat*[1] and Dr. D.C.Upadhyay[2]

[1] Research Scholar, Singhania University,Pacheri Beri, Rajasthan, India
rawat1420@gmail.com

[2] Professor,RPS Engineering College Mohindergarh,Haryana, India
upadhyaydc@yaho.com

Abstract: This Paper provides insights into data mining solution for mining customer's information from customer opt-in database of mCRM. The basis of approach is to use a K nearest neighbor algorithm to learn how to classify samples within different clusters of interest.

Therefore a new approach using Genetic Algorithm is followed in this paper to overcome some of the shortcomings of the K nearest neighbor algorithm, by allowing the system to learn to warp the n-dimensional feature space so as to maximize the clustering of individuals within a class, and at the same time maximize the separation between classes.

The Output of the Genetic Algorithm is acting as input to the K nearest neighbor algorithm And finally the global clusters are being formed and the customization for a particular Customer is done seeing in which Cluster a particular customer falls.

The main result of this paper indicates that GA-KNN Conjunction may be an effective element to mCRM. Data mining from the customers' database, stores can offer their customers interesting services via the mobile medium (SMS/MMS) and can retain customers with different ways and maintain fruitful relations with their customers based on trust.

Keywords: Mobile customer relationship management (mCRM), K nearest nighbour Algorithm (KNN), Genetic Algorithm (GA)

INTRODUCTION

The main focus of this paper is on data mining techniques and mobile technology (as a channel for CRM); both of them, according to this definition, are CRM enablers.

mCRM, which stands for Mobile CRM. It is eCRM with wireless tools, such as mobile phone, PDAs or laptop computers. The aim of mCRM is to enable two-way interactivity between the customer and the enterprise continuously at anywhere.

As in the present scenario Data Mining tools are very costly and only few well off companies can afford them. The technique being used for analysis is automatic cluster detection. As doing on line analysis the algorithm to be used should be fast that is in mathematical term say it should not be compute intensive but still should give good results, so a technique has to be chosen that is time optimal, is not computationally intensive and forms globally optimal distinct clusters. Among all the algorithms that are being for Data Mining K Nearest Neighbor Algorithm is the fastest one, so our approach for Cluster detection and for globalizing our search used Genetic algorithm with K Nearest Neighbor Algorithm for mobile customer relationship management (mCRM).

Interacting with customers is also not as simple as it has been in the past. Customers and prospective customers want to interact on their terms, meaning that one needs to look at multiple criteria when evaluating how to proceed. One will need to automate:

a. The Right Offer
b. To the Right Person
c. At the Right Time
d. Through the Right Channel

THE K NEAREST NEIGHBOR ALGORITHM

The K nearest neighbor algorithm (KNN) is a method for classifying objects based on closest training examples in the feature space. KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It can also be used for regression.

The K nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its K nearest neighbors. K is a positive integer, typically small. If K = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose K to be an odd number as this avoids tied votes.

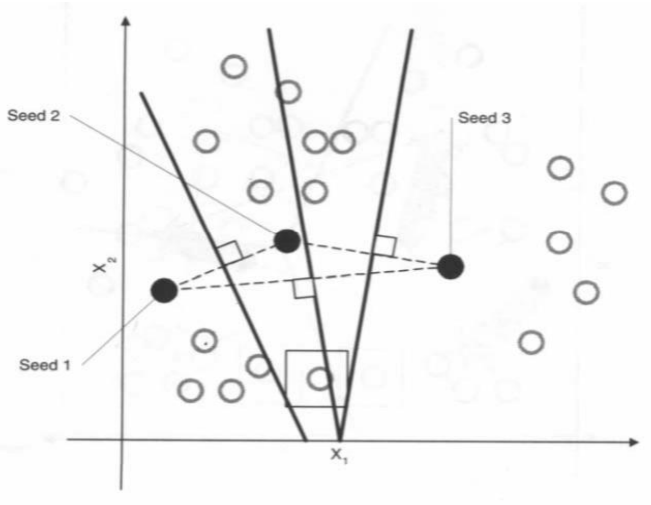

Figure 1: Showing Initial Cluster Boundaries.

Flowchart for The K Nearest Neighbor Algorithm:

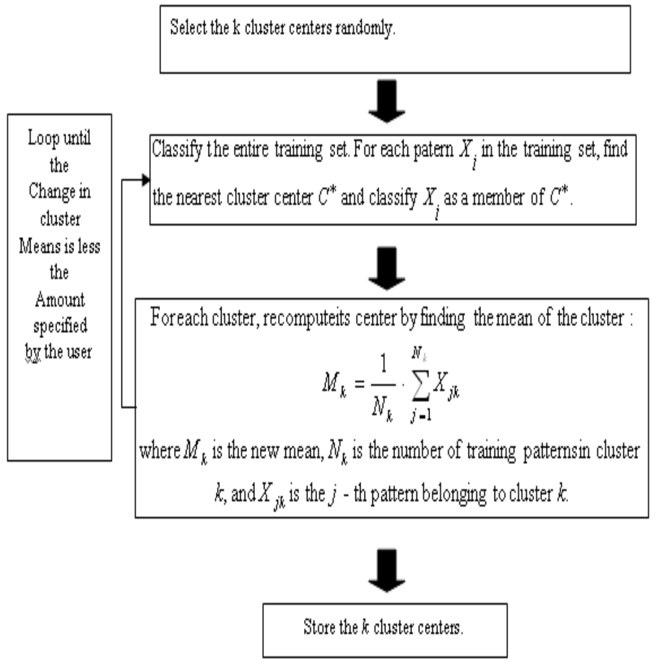
Select the k cluster centers randomly.
Loop until the Change in cluster Means is less the Amount specified by the user
Classify the entire training set. For each pattern X_i in the training set, find the nearest cluster center C^* and classify X_i as a member of C^*.
For each cluster, recompute its center by finding the mean of the cluster :
$M_k = \frac{1}{N_k} \cdot \sum_{j=1}^{N_k} X_{jk}$
where M_k is the new mean, N_k is the number of training patterns in cluster k, and X_{jk} is the j - th pattern belonging to cluster k.
Store the k cluster centers.

Figure 2: Flowchart for KNN.

The basic problem with KNN was that it was not searching globally it was a local search so the clusters that were being formed were not global clusters. This was found out by giving different initial points to the KNN.

GENETIC ALGORITHM

The Genetic Algorithms are defined as: "… search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of artificial creatures (strings) is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance" [1&2]

The algorithm uses the principals of Darwinian natural selection principle of biological evolution to search for the optimal, or near optimal, solution in a multi-dimensional feature space. GA is population based parallel search strategy. In nature, competition among individuals for scant resources results in fittest individuals dominating among weaker ones. GA uses the same concept. This search was to be globalized using some global minimization algorithms.

GENETIC ALGORITHM – PROPOSED APPROACH

To make our search global Genetic Algorithms have been proposed.

Genetic Algorithms generate a sequence of populations by using a selection mechanism, and use crossover and mutation as their search mechanism .The structured recombination through crossover operator results in highly exploitative global search that combines solid theoretical analysis with carefully controlled computational experiments .GA also differs from traditional search methods, e.g. Gradient Descent, in some fundamental points .The main differences are GAs work with coding of parameter set ,not the parameters themselves.

Following points regarding GAs are noteworthy:

a. GAs starts their search from a set of multiple points, not a single point .This increases probability to find optimum.
b. GAs use payoff (objective function) information, not derivatives or other auxiliary knowledge
c. GAs use probability transition rules, not deterministic ones. They use random choice as a tool to guide the search space with better performance [3,4&5]

The algorithm wil1 find the optimal weighting for each feature in order to minimize a cost function which is employed to determine if offspring have reached a specified survival value. The weights discovered in this manner are then used to calculate distance in the kNN algorithm. Each individual in the population contains n floating numbers representing the weights for n features. The values of the weights in the population were initially generated using random numbers (lying between 0 and

(i). The values of the cost function were then calculated and ranked in ascending order. The reproduction, cross-over and mutation operators were then applied to create two new individuals which replaced the two least optimal ones. To achieve this goal, populations of vectors representing randomly-valued weights for the features were selected. New offspring were generated from the initial population of vectors using the following the three genetic operators: Selection, Crossover and Mutation.
GA comprises a set of initial random population and biologically inspired operators like selection, crossover and mutation. A typical GA cycle consists of following steps:

a. Creation of 'population' of strings
b. Evaluation of each string
c. Selection of 'best' strings
d. Genetic manipulation to create new population of strings

The population *comprises* a group of chromosomes. Initially, population is generated randomly. Each member of population represents a potential solution to the problem. A chromosome is usually expressed in a string of variables, each element of which is called *gene.* The variable can be represented by binary, real number or other forms and its range is usually problem-specified. Bit- string encoding is the classical approach used by GA researchers due to its simplicity and tractability. [3&4]

The evaluation function (objective function) is a main source to provide mechanism for evaluating the status of each chromosome. This is an important link between GA and the system. The fitness values of all chromosomes are evaluated by calculating the objective function in a decoded form with respect to the constrains imposed by the problem. For example, Table 1 shows fitness evaluation (objective function: $f(x) = x^2$ of the population of four individuals.

The selection operator emulates nature's survival of fittest policy. Based on fitness values, selection operator selects the parent for mating process. It is expected that a fitter chromosome receives a higher number of offspring and thus has a higher chance of surviving in subsequent generation.

Table 1: Example-fitness evaluation.

#Strin g	Chromosome	Decoded Chromosome	Fitness $f(x)=x^2$	% Fitness
1	01101	13	169	14.4
2	11000	24	576	49.2
3	01000	8	64	5.5
4	10011	19	361	30.9
Total			1170	100.00

The combined evaluation and selection process is called reproduction. The genetic manipulation process uses standard 'crossover' and 'mutation' operators to produce new population of individuals (offspring) by manipulating the 'genetic information' referred to as genes possessed by members (parents) of the current population. Crossover is a new recombination operator that combines subparts of two parent chromosomes to produce offspring that contain subparts of both parents, genetic material. A probability term Pc is set to determine operation rate.

Table 2 shows, after the crossover point has been randomly chosen, how portions of the parents are swapped to produce offspring. *Mutation operator introduces* variations into the chromosome. This variation can be global or local .The operation
Occurs occasionally (usually with small probability p_m) but randomly alters the value of string position.

A bit is flipped at randomly generated location if probability test is passed. The new population thus generated after crossover and mutation operations replace the old population. This completes the GA cycle. As the generation advances, overall fitness of the population increases

Table 2: Genetic Manipulation.

Mating Pool (Parents)	# String	Crossover	Offspring	Mutation	New Pop	Fitness F(x)=x2	% Fitness
01101(169)	1	01[101]	01000(64)	[0]1000	11000	576	24.6
11000(576)	2	11[000]	11101(841)	1110[1]	11100	784	33.4
11000(576)	3	1100[0]	11001(625)	110[0]1	11011	729	31.1
10011(361)	4	1001[1]	10010(324)	100[1]0	10000	256	10.9
Total						2345	100

IMPLEMENTATION DETAILS

Genetic Algorithm Implementation:

a. Coding of the data set i.e. convert each and every point to binary equivalent, developing procedures for the same
b. Initial population is generated on random basis. It consist of 62 bit (x1, y1, z1, f1, x2, y2, z2, f2) where x1, x2 indicates age and are of seven bits, y1and y2 indicates credit and are of ten bits, z1and z2 indicates income and are of ten bits, f1and f2 indicates no of family members and are of four bits
c. Selection is implemented using Roulette wheel
   a) The chromosome with maximum fitness survives with maximum probability
   b) The chromosome with minimum fitness dies out
d. Selection will be based on two fitness functions
   a) Hamming Distance[7]
   b) Euclidean Distance
e. Two point crossover will be implement
f. For every ten selection, there are nine crossovers and two mutation i.e. Ps (probability of selection) = 1.000

Pm (probability of mutation) = 0.133
Pc (probability of crossover) = 0.866

The output of GA will act as input to the K Nearest neighbor algorithm

K Nearest Neighbor Algorithm Implementation:

Step1: Take initial points from the table initial points from the database
Step2: Put these points into respective Clusters.
Step3: Select each point from the table customer information
Step4: Find distance from each point from initial points.
Step5: Put the point in cluster of initial points from which the distance is least. Step6: For each Cluster now formed find the new centeroid or the initial point for
the next iteration
Step7: This new centroid will be averaged mean of the points in the cluster
Step8: Now the process starts again from these new centroids and repeat the process all over again
Step9: This process is repeated till the distance between old and new centroid formed is less than 5 units.

How Euclidean Distance Works As Fitness Function:

Objective Function: To find the best points that can input to the Kth Nearest neighbor algorithm i.e. far away points.
Taking an example for two dimensions only
And let $0<x_1<3$ $0<y_1<3$ $0<x_2<3$ $0<y_2<3$

Our chromosome be

x_1	y_1	x_2	y_2

Our fitness function is

$F (x) =\surd(x1-x2)^{2} + (y1-y2)^{2}$

SELECTION

The ‘steady state’ approach was used to generate off-springs at each generation. At each iteration, only two individuals in the population were selected to produce two offspring for the next generation with the rest of the population being retained in the next generation. These two off- springs replaced the two least optimal individuals in the population. The two new individuals were created through crossovers and mutations from the two parent individuals. The selection of the two parent individuals will be random with the fitter ones having a higher probability of being selected than the less fit ones. This was achieved by ranking the whole population of offspring in ascending order in terms of the value of their cost (fitness) function.

Two bits will be there for each x_1, y_1, x_2, y_2 in our chromosome initially random population of four strings

Table 3: Selection-1.

Chromosome	Converted	Fitness Function	Probability	Selection
0000 0110	(0,0,1,2)	$\surd(2-0)^{2}+(1-0)^{2} =\surd5$	√5/8.07=.0.277	1
1001 0000	(2,1,0,0)	√5	0.277	1
0011 0000	(0,3,2,0)	√13	0.447	2
1111 1111	(3,3,3,3)	0	0	0

Sum = 8.07

This selection procedure guarantees that the fitter individuals have a higher chance of being selected for reproduction than the less fit ones.

CROSSOVER

The crossover operator was used to select the parameter values from the two parents. The n weight values from the father individual and the n-n1 weight values from the mother individual were selected. The n1 was chosen randomly.

```
0000 0110    0001 0110
0011 1000    0010 1000
1001 0000    1011 0000
0011 1000    0001 1000
```

CONCLUSION

We have proposed an algorithm that hybridizes the classification power of KNN algorithms with the search and optimization power of the genetic algorithm. The result is an algorithm that requires computational capabilities above that of the KNN algorithm, but achieves improved classification performance in a reasonable time. We anticipate that extensions to the research will improve the algorithm's performance and there are a number of issues that we plan to address in further work.

To globalize this search Genetic algorithms were employed in conjunction with KNN. This ‘hybrid learning’ approach gives advantage of global search at a high speed.

REFERENCES

[1]. Asem Omari, Alexander Hinneburg and Stefan Conrad. Temporal Frequent Itemset Mining. In Proceedings of the Knowledge Discovery, Data Mining and Machine Learning workshop (KDML 2007), Halle, Germany, September 2007. Poster.

[2]. Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001.

[3]. Ruey-Shun Chen, Ruey-Chyi Wu, and J. Y. Chen. Data Mining Application in Customer Relationship Management of Credit Card Business. In 29th Annual International Computer Software and Applications Conference COMPSAC, pages 39–40, 2005.

[4]. David E Goldberg Genetic algorithm in search, optimization and machine learning AWL Publications

[5]. K.K.Shukla Neuro-Computer Optimization Publication House

[6]. Michael J.A.Berry & Gordon Lindoff Data Mining Techniques for Marketing, Sales and Customer Support Wiley Eastern Publication

[7]. Alex Berson & Stephen Smith Building Data Mining Applications for Customer Relationship Management Tata McGraw Hill Publications.

[8]. Peter Adrians Data Mining AWL publications

[9]. Min Pei & Ying Ding Classification & Feature Extraction of High Dimensionality Binary Pattern using Genetic Algorithm to evolve rule Genetic Algorithm Research and Application Group (GARAGe), Michigan University

[10]. D.Doval, S.Mancoridis Automatic Clusterisation of Subsystem using Genetic Algorithm Department of Mathematics and Computer Science, Drexel University, Philadelphia

[11]. Terrence W Dexter Optimization of Genetic Algorithm and within a Genetic Algorithm 2D ^{1}layout problem, Michigan University