RESEARCH ARTICLE OPEN ACCESS

An Efficient System for Heart Risk Detection using Associative Classification and Genetic Algorithms

P Vinay Bhushan[1], Ch Narasimha Chary[2], V Nithesh[3], Dr Rp Singh[4]

[1,2](Research Scholar, Sssutms, Bhopal, Madhya Pradesh, India),
[3](Asst Prof,Dept of Cse,Vignan Instute of Mangnt & Tech for Womens,Rr Dist, Telangana,India),
[4](Research Supervisor, Sssutms, Bhopal, Madhya Pradesh,India).

----------**********----------

Abstract:

Associative classification is an ongoing and compensating system which incorporates affiliation govern mining and classification to a model for forecast and accomplishes greatest exactness. Associative classifiers are particularly fit to applications where most extreme exactness is wanted to a model for expectation. There are many spaces, for example, medication where the most extreme exactness of the model is wanted. Heart illness is a solitary biggest reason for death in created nations and one of the fundamental supporters of malady load in creating nations. Mortality information from the enlistment center general of India demonstrates that heart illness is a noteworthy reason for death in India, and in Andhra Pradesh, coronary heart malady causes around 30%of passings in rustic territories. Consequently there is a need to build up a choice emotionally supportive network for foreseeing heart sickness of a patient. In this paper, we propose efficient associative classification algorithm using genetic approach for heart ailment expectation. The primary inspiration for using a genetic algorithm in the revelation of abnormal state expectation decides is that the found tenets are profoundly intelligible, having high prescient exactness and of high intriguing quality qualities. Trial Results demonstrate that a large portion of the classifier rules help in the best forecast of heart sickness which even helps specialists in their finding choices.

Keywords — Associative classification, Genetic algorithm, Gini Index, Z-Statistics.

----------**********----------

1. Introduction

The real reason that the information mining has pulled in a lot of consideration in the information business in the ongoing years is because of the wide accessibility of gigantic measures of information and the requirement for transforming such information into valuable information and learning. The information picked up can be utilized for applications ranging from business management, generation control, and market analysis to rising plan and science investigation and wellbeing information analysis [1]. Affiliation administer mining and classification are two fundamental functionalities of information mining. Affiliation administer mining is utilized to discover affiliations or relationships among the thing sets. It is an unsupervised realizing where no class trait is engaged with finding the affiliation run the show. Then again, classification is an administered realizing where a class characteristic is associated with the development of the classifier and is utilized to order or foresee the information obscure example.

Associative classification includes two phases.

1)Generate class based affiliation rules from a preparation dataset

2)Classify the test informational index into predefined class names.

There is developing proof that blending classification and affiliation lead mining together can deliver more efficient and exact classification system than customary classification strategies.

Genetic algorithms are regularly utilized for issues that cannot be comprehended efficiently with conventional methods.

Genetic algorithms appear to be valuable for seeking exceptionally broad spaces and optimization issues.

Coronary heart sickness is plague in India and one of the significant reasons for illness weight and passings. Information from enlistment center general of India demonstrates that heart illnesses are a noteworthy reason for death in India, concentrates to decide the exact reason for death in urban Chennai and provincial regions of A.P have uncovered that CVD cause around 40% of the passings in urban and 30% in country regions [2].

1.1 Associative Classification

Classification is a standout amongst the most important assignments in information mining. Analysts are focusing on outlining classification algorithm to construct exact and efficient classifiers for expansive datasets. Relate classification accomplishes high precision, its guidelines are interpretable and it gives certainty likelihood while ordering objects which can be utilized to tackle classification issue vulnerability. Therefore, it turns into a hot subject as of late [3].

Associative classification is an extraordinary instance of affiliation manage mining in which class trait is considered in the govern's ensuing. For instance in an administer AB, B must be a class property. A classifier is of the form A1, A2, - An⊇B, where Ai is a property and B is a class. Decide thing that fulfills mins up is called visit lead things, while the rest are called rare manage things. Associative classification is to gather runs in preparing informational collection D, organize them in a specific request to form a classifier. At the point when given an unlabelled question, the classifier chooses the administer as per the request whose condition coordinates the articles and relegates class names of the manage to it.

Table 1.Training data set

Sl.no	A1	A2	A3	CLASS
1	a 11	a 21	a 31	C1
2	a 12	a 24	a 32	C2
3	a 13	a 23	a 33	C0
4	a 11	a 21	a 31	C1
5	a 12	a 22	a 32	C2

1.2 Genetic Algorithm

Genetic algorithms are computing methodologies constructed in analogy with the process of evolution [4].It closely resembles the natural process of regeneration, reproduction, inheritance evolution. Genetic algorithms are typically used for problems that cannot be solved efficiently with traditional techniques. Genetic algorithms seem to be useful for searching very general spaces and optimization problems. Each solution generated in Genetic algorithms is called a chromosome (individual).Each chromosome is made up of genes, which are the individual elements (alleles) that represents the problem. The collection of chromosomes is called a population. The internal representation of the chromosomes is known as its genotype. This can be either bit strings or gray codes or hexadecimal codes. The external manifestation of the genotype or the real world representation of the genotype is known as the phenotype [5].Basically there are three genetic operators are used for generating new strings. The functions of genetic operators are as follows:

1) Selection: selection deals with the probabilistic survival of the fittest in that, more fit chromosomes are chosen to survive.

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point. Various types of crossover operators are a) single point b) two point c) uniform d) half uniform e) reduced surrogate crossover f) shuffle crossover g) segmented crossover [6].

3) Mutation: mutation alters the new solutions so as to add stochasticity in the search for better solution. The most common method way of implementing mutations is to select a bit at random and flip (change) its value. There are 2 types of mutations use in genetic network programming 1) mutating the judgment node 2) mutating the value of the judgment node. In associative classification attributes and their values are taken as judgment nodes and class values as processing nodes.

Fitness function: Ideally the discoved rules should have a) high predictive accuracy b) be comprehensible c) be interesting. The accomplishment of a genetic algorithm is directly linked to the accuracy of the fitness function.

1.3 Heart Disease

Coronary heart disease is a narrowing of the small blood vessels that supply blood and oxygen to the heart. This is also called as coronary artery disease. Coronary heart disease is usually caused by a condition called atherosclerosis, which occurs when fatty material and a substance called plaque builds up on the walls of arteries. This causes them to get narrow. As the coronary arteries narrow, blood flow to the heart can slow down or stop, causing chest pain,shorteness of breath, heart attack, and other symptoms. Men in their 40's have higher risk of Coronary heart disease than women, but as women gets older, their risk increases so that it is almost equal to a man's risk. Major risk factors for Coronary heart disease are 1) Diabetes 2) High blood pressure 3) High LDL (bad) cholesterol 4) low LDL (good) cholesterol 5) Not getting enough physical activity6) Obesity7) Smoking.

India is undergoing rapid epidemiological transition as a consequence of economic and social change, and cardiovascular disease is becoming an increasingly important cause of death. India's disease pattern has undergone a Mmajor shift over the past decade. As per WHO report, at present out of 10 deaths in India, eight are caused by non communicable diseases, such as cardio vascular diseases, and diabetes in urban india.In rural India, 6 out of every 10 deaths is caused by NCD'S [7].Data from registrar general of India shows

that heart attacks are major cause of deaths in india.in Andhra Pradesh 30% of rural population died due to CHD.

There is an urgent need for development and implementation of suitable primordial, primary, and secondary prevention approaches to control this epidemic. An urgent and sincere bureaucratic, political, and social will to initiate steps in this direction is required.

2 Related Work

Large no. Of work is carried out in finding efficient methods of medical diagnosis for various diseases. Our work is an attempt to predict the cardiac disease in Andhra Pradesh using data mining.

Carlos implemented efficient search for diagnosis of heart disease comparing association rules with decision trees [8].A novel technique to develop the multi-parametric feature with linear and non linear characteristics of HRV was proposed by Hean Gyu lee et al.[9].A model intelligent heart diseases prediction system built with the aid of data mining techniques like decision trees, naive bayes and neural network was proposed by sellappan palaniappan et al[10]. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos

Ordonez [11].MA.jabbar, Priti Chandra, B.L.Deekshatulu proposed evolutionary algorithm for heart disease prediction. They used genetic algorithm to predict the heart disease for Andhra Pradesh population [1].Enhanced prediction of heart disease with feature subset selection using genetic algorithm was proposed by M.Ambarasi et al [12].Intelligent and effective heart attack prediction system using data mining and AINN was proposed by [13].They employed the multilayer perception neural network with back propagation as the training algorithm. Graph based approach for heart disease prediction was proposed by MA.jabbar, B.L.Deekshatulu, and Priti Chandra [14].Their method is based on maximum clique and weighted association rule mining. Associative classification for heart disease prediction was proposed by MA.jabbar, B.L.Deekshatulu, and Priti Chandra [15].They used Gini index based classification to predict the heart disease. Cluster based association rule mining for heart attack prediction was proposed by MA.jabbar, B.L.Deekshatulu, and Priti Chandra [16].Their method is based on digit sequence and clustering. The entire data base is divided into partitions of equal size and association rule will be mined from each partition.

In this paper we propose efficient association classification for heart disease prediction for Andhra Pradesh population. We used Gini index to produce a compact rule set and filter rules further by applying Z-Statistics and genetic algorithm.

3 Proposed Method

Most of the associative classification algorithms adopt the exhaustive search method presented in the famous APRIORI algorithm to discover the rules and require multiple passes over the data base. Furthermore, they find frequent items in one phase and generate the rules in a separate phase consuming more resources such as storage and processing time. Moreover, since rule ranking plays an important role in classification and the majority of the associative classifiers select rules mainly in terms of their confidence levels. Even after pruning infrequent items, the APRIORI association rule generation procedure, produces a huge no. of association rules .If all the rules are used in the classifier then the accuracy of the classifier would be high but the building of classification will be slow. In order to improve the accuracy of associative classification we propose an informative attribute entered rule generation and hypothesis testing Z- statistics for heart disease prediction. The class association rules are represented as chromosomes and Michigan approach is used to encode the rules.

3.1 Proposed Algorithm

STEP 1: find Gini index of each attribute. The attribute with minimum Gini index is selected for class association rule generation. These class association rules are known as initial population and represented as chromosomes.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [\, p(i / t)]^2$$

STEP 2: Evaluate fitness of rule using Z statistics Z=S(X)-Minimum support/SQRT (min sup*(1-minsup))/N Where **S(X)** is support of pattern and **min sup** is user defined threshold

STEP 3: Prune the rules based on Z statistics. After rule evaluation the rules having highest fitness are stored in a pool. Then apply genetic functions on these rules.

STEP 4: Perform single point cross over. Judgement nodes are selected for crossover.

STEP 5: Perform mutation by mutating the value of judgment node. This process will be repeated till last generation reached.

STEP 6: Build classifiers using the generated Rules

STEP 7: Predict the rules on test data

STEP 8: Find the accuracy of the data set

$$\textbf{Accuracy} = \frac{\text{Number of objects correctly Classified}}{\text{Total No. of objects in the test set.}}$$

3.2 Explanation of Algorithm

A) Attribute selection based on Gini Index

An informative attribute centred rule generation produces a compact rule. **Gini** index is used as filter to reduce the no. of candidate item sets. It is used to select the best attribute. Those attributes with minimum Gini index are selected for rule generation.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [\, p(i / t)]^2 \qquad (1)$$

Let us consider a sample medical training data set given in table 2.

Table2: Example Medical Training data

No.	mcv	alkphos	sgpt	sgot	gammagt	drinks	selector
1	85.0	92.0	45.0	27.0	31.0	0.0	1
2	85.0	64.0	59.0	32.0	23.0	0.0	2
3	86.0	54.0	33.0	16.0	54.0	0.0	2
4	91.0	78.0	34.0	24.0	36.0	0.0	2
5	87.0	70.0	12.0	28.0	10.0	0.0	2
6	98.0	55.0	13.0	17.0	17.0	0.0	2
7	88.0	62.0	20.0	17.0	9.0	0.5	1
8	88.0	67.0	21.0	11.0	11.0	0.5	1
9	92.0	54.0	22.0	20.0	7.0	0.5	1
10	90.0	60.0	25.0	19.0	5.0	0.5	1

After calculating Gini index of each attribute sgpt has the lowest Gini index. So sgpt would be the better attribute.

The rules generated like the following are considered for classifier.

1. sgpt='(-inf-19.1] ==> selector=2
2. sgpt='(-inf-19.1]' gammagt='(-inf-34.2]' ==> selector=2
3. sgpt='(19.1-34.2]' sgot='(20.4-28.1]' ==> selector=2
4. sgpt='(19.1-34.2]' gammagt='(-inf-34.2]' ==> selector=1
5. sgpt='(19.1-34.2]' drinks='(-inf-2]' ==> selector=1

B) Z-Statistic (Hypothesis Testing)

Hypothesis Testing is a statistical inference procedure to determine whether a conjecture or hypothesis should be accepted or rejects based on the evidence gathered from data [17]. In our proposed approach we use Z-Statistic to verify the quality of pattern or rule.Various steps involved in testing of hypothesis'-Statistic is preferred if the sample size is greater than 30.

1) **Null Hypothesis**: Define a null hypothesis H_O taking into consideration the nature of the problem and data involved.

2) **Alternative Hypothesis**: set up alternative hypothesis H_1 so that we could decide whether we should use one tailed and two tailed test.

Level of significance: Select the appropriate level of significance(α)

Test statistics: compute Z statistics

Z=S(X)-Minimum support/SQRT (min sup*(1-minsup))/N

5) **Conclusion** :compare the computed value of Z statistics with the critical value of Z_α (given in Z-Statistic Table) at given level of significance(α)

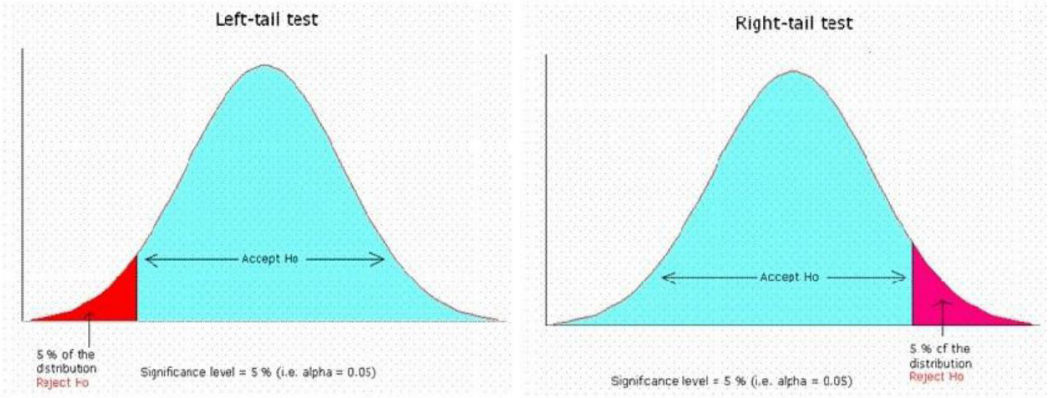
If $|Z| < Z_\alpha$ accept the null hypothesis

If $|Z| > Z_\alpha$ reject null hypothesis [18].Fig 1 shows left tail and right tail test.

Table 3: Critical values of Z

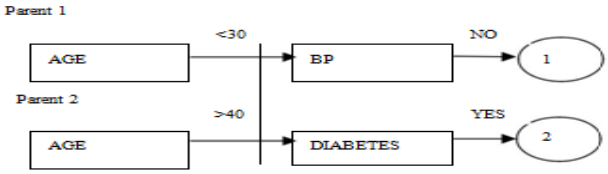
Level of significance	*1%*	*5%*	*10%*
Two tailed test	$\|Z_\alpha\|=2.58$	$\|Z_\alpha\|=1.96$	$\|Z_\alpha\|=1.645$
Right tailed test	$Z_\alpha=2.33$	$Z_\alpha=1.645$	$Z_\alpha=1.28$
Left tailed test	$Z_\alpha=-2.33$	$Z_\alpha=-1.645$	$Z_\alpha=-1.28$

Example: Let N=10000 S(X) =11% Minimum Support =10% Z-Statistic under null hypothesis is Z=3.33.Suppose level of significance (α) =0.001 sets up a rejection region with Z_α =3.09. Since Z> Z_{α}, the null hypothesis is rejected and the pattern is considered statistically interesting.

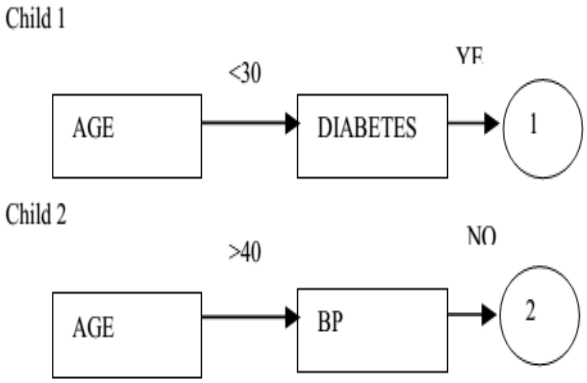


C) Crossover and Mutation

Crossover operator forms off springs by combining judgment nodes which are selected as crossover nodes. We used single point crossover in our approach



After crossover



D) Accuracy Computation

Accuracy measures the ability of the classifier to correctly classify unlabelled data.

$$\text{Accuracy} = \frac{\text{Number of objects correctly Classified}}{\text{Total No. of objects in the test set.}}$$

Conclusion

In the ongoing years, India and other creating nations have seen a quickly raising plague of cardiovascular infection (CVD). It is anticipated that by 2020 coronary heart malady will lead reason for death in grown-up Indians, and Andhra Pradesh is at risk of more passings because of CVD. The need to contain the pestilence of cardiovascular infection and in addition battle its effect and minimize its toll on Andhra Pradesh is clear and pressing. Thus a choice emotionally supportive network is proposed to distinguish a risk score for anticipating the heart illness. In this paper, we proposed a system for heart malady expectation using information mining procedures. In our element work, we plan to diminish no. of ascribes and to decide the property which contributes towards the determination of sickness using the genetic algorithm.

References

[1] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: An evolutionary algorithm for heart disease prediction, ICIP, CCIS 292 PP 378-389, Springer-Verlag (2012)

[2] Rajeev Gupta.: Recent trends in coronary heart disease epidemiology in India, Indian heart journal, pp B4-B18 (2008)

[3] Zhonghua Tang and Qin liao.: A new class based associative classification algorithm, IAENG, IJAM 36:2(2007)

[4] Goldberg DE.: Genetic Algorithm in search, optimization and machine learning, Addison Wesley (1989)

[5] S.P Syed Ibrahim et al.: An Evolutionary approach for rule set selection in a class based associative classifier.Europian journal of scientific research, vol 50 no 3pp417-425(2011)

[6] Picek, S., Golub, M.: On the Efficiency of Crossover Operators in Genetic Algorithms with Binary Representation. In: Proceedings of the 11th WSEAS International Conference on Neural Networks (2010)

[7]The Times of India.14th august 2011

[8] Carlos Ordonez.: Comparing association rules and decision trees for heart disease prediction, ACM, HICOM (2006)

[9] Hean Gyu Lee et al.:Mining bio signal data :CAD Diagnosis using linear and non linear features of ARV,LNAI 4819 pp 56-66(2007)

[10] Sellappan Palaniappan et al.: Intelligent heart disease prediction on system using data mining techniques.IJCSNS Vol 8 no 8(Aug 2008)

[11] Carlos Ordonez.: Improving Heart Disease Prediction using constrained association Rule, seminar presentation at TOKYO (2004)

[12] M.Ambarasi etc al.:, Enhanced Prediction of Heart Disease with Feature subset selection using Genetic Algorithm, IJESI, Vol 2(10) (2010)

[13] Shantakumar B patil etc all.: Intelligent and effective heart attack prediction system using data mining and artificial neural network,Europian journal of scientific research vol 31,No 4(2009)

[14] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: Graph based approach for heart disease prediction. In: proceedings of ITC 2012, Bangalore, Springer-Verlag (2012)

[15] MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: Knowledge Discovery using Associative Classification for Heart Disease Prediction. In: International symposium on Intelligent Informatics (ISI 2012) (Springer)

[16] MA.Jabbar, Priti Chandra, B.L.Deekshatulu..:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no 2,(Oct 2011)

[17] Ping Ning tan, Steinbach, vipin Kumar. : Introduction to Data Mining, Pearson Education, (2006)

[18] Krishna Gandhi: Probability and Statistics, S.Chand (2011)

[19] http://www.sgi.com/tech/mlc/db

[20] UCI Machine Learning Repository http://archive.ics.uci.edu/ml

[21] Koza, J.R, Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press (1992)

[22] S.P Syed Ibrahim et al.:Efficient associative classification using genetic network programming, IJCA,Vol 29,No 6 sep(2011)