Chaid and Genetic Algorithm-Based Hybrid Classifier for Pattern Classification

S.R.SWARNALATHA
Lecturer, Dept. of MCA
SCT Institute of Technology,
Bangalore, India.
swarni_r@yahoo.co.in

Dr. G.M KADHAR NAWAZ
Director, Dept. of MCA
Sona College of Technology,
Salem, India.

***Abstract*-In the context of pattern classification, one of the major issues discussed by most of researchers is 'curse of dimensionality' problem which occurs in data classification because the data processed in most of the application is high-dimensional feature space. So, the essential consideration here is that irrelevant features should be identified which causes less classification accuracy and the main motto is to find a minimum set of attributes from the initial set of data helping to make the patterns easier to understand along with improved classification accuracy and reduced learning time. Therefore, the selection of feature set is the process to search for an optimal feature subset from the initial data set without compromising the classification performance and efficiency in generating classification model. In this paper, we develop a hybrid classifier by combining CHAID and genetic algorithm. Initially, the genetic algorithm with ABC operator will extract the best attributes and based on the extracted attributes the CHAID will generate the decision tree. We analyze the performance with different datasets and compare the analysis with the existing technique.**

Keywords: Pattern Classification, Hybrid Classifiers, CHAID

I. INTRODUCTION

A well-recognized Data Mining task is classification and it has been studied widely in the fields of statistics, pattern recognition, and decision theory, machine learning literature, neural networks and more. Classification process usually uses supervised learning techniques that induce a classification model from a database. The task of classification is to allocate a new object to a class from the given set of classes derived from the attribute values of the object.

The classification algorithm learns from the training set and generates a model and this model is used to classify fresh objects. Numerous classification algorithms have been recommended in the literature, such as decision tree classifiers, rule-based classifiers, Bayesian classifiers, support vector machines (SVM), artificial neural networks, Lazy Learners, and ensemble methods. Basically, decision tree is an eminent classification model in machine learning, artificial intelligence and data mining because they are: Practical, with a wide range of applications, Simple and easy to understand, and Rules can be extracted and executed manually. Currently, research on data classification primarily focuses on certain data, in which precise and definite value is habitually assumed.

Decision trees are tree-shaped arrangement that stand for sets of decisions. The decision tree approach can engender rules for the classification of a data set. Specific decision tree techniques comprise Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree approaches used for classification of a data set. They provide a set of rules that can be applied to a new (unclassified) data set to envisage which records will have a given outcome. CART generally requires smaller amount data preparation than CHAID.

Chi-square automatic interaction detection (CHAID) is an automated process used to show the relationships amid independent and dependent variables that can advance or optimize the performance of a customer acquisition campaign. CHAID modeling is an exploratory data examination model used to study the relationships amid a dependent measure and a large series of possible predictor variables those themselves may interact. The dependent measure could be a qualitative (nominal or ordinal) one or a quantitative indicator. For qualitative variables, a series of chi-square analyses are conducted amid the dependent and predictor variables. For quantitative variables, examination of variance methods is used where intervals (splits) are determined optimally for the independent variables so as to maximize the ability to detail a dependent measure in terms of variance components.

II. PROPOSED HYBRID CLASSIFIER FOR PATTERN CLASSIFICATION

One of the primary issues discussed by many researchers in the context of pattern classification is ‘curse of dimensionality’. This issue occurs in the data classification because the data processed in most of the application is high dimensional feature space. So the major consideration is reducing the irrelevant features which cause less classification accuracy. So, the main goal is to identify a minimum set of attributes from the initial set of data helping to make the patterns easier to understand along with improved classification accuracy and reduced learning time. So, the selection of feature set is the process to search for an optimal feature subset from the initial data set without compromising the classification performance and efficiency in generating classification model.The Fig.1 shows the overall process of our recommended technique.

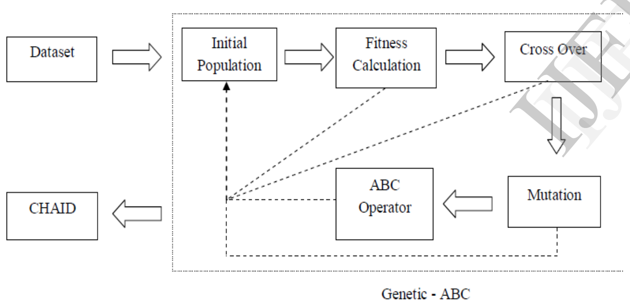

Fig.1 Overall Process

III. GENETIC ALGORITHM WITH ABC

The genetic algorithm with ABC operator contains the following process as solution encoding, initial population, fitness calculation, crossover, mutation and ABC operator.

A. *Solution Encoding*

Definition: It is the process of generating different set of necessary attributes randomly from the attributes in our dataset.

The dataset contains the class field and number of attribute fields. We have to choose the attributes randomly which are necessary to consider for our technique. Similarly, we have to generate different set of attributes based on the number of attribute fields we have in the dataset.

C	A1	A2	A3	A4	A5
C1	A1V1	A2V1	A3V1	A4V1	A5V1
C1	A1V2	A2V2	A3V2	A4V2	A5V2
C1	A1V3	A2V3	A3V3	A4V3	A5V3
C2	A1V4	A2V4	A3V4	A4V4	A5V4
C2	A1V5	A2V5	A3V5	A4V5	A5V5
C2	A1V6	A2V6	A3V6	A4V6	A5V6

Fig.2 Sample Dataset

The Fig.2 shows the sample dataset that contains a class field and five different attribute fields. The column denoted by ‘C’ is the class field and the columns denoted by ‘A1’, ‘A2’, ‘A3’, ‘A4’ and ‘A5’ are different attribute fields. In the sample dataset the class field has two different classes as ‘C1’ and ‘C2’.

B. *Initial Population*

Definition: It is the process of choosing the set of necessary attributes randomly generated by selection encoding for further process.

To process the genetic algorithm, we have to give a set of inputs in the form of chromosomes becausegenetic algorithm is based on the process of natural evolution. So we represent the initial population as chromosomes. The set of inputs are chosen randomly from the different necessary attributes generated by the solution encoding.

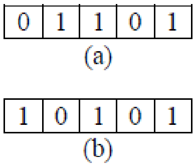

Fig.3 Sample set of attributes

The Fig.3 shows the sample set of attributes generated by the selection encoding. We select ten set of attributes randomly generated by the selection encoding.

C. *Fitness Calculation*

Definition: It is the process of calculating the fitness of the chromosomes chosen to find the best solution for our technique.

After choosing the initial population randomly from the solution encoding, we have to calculate the fitness for all the chromosomes separately. The formula to calculate the fitness value is as follows:

$$f = \frac{1}{N}\sum_{i=1}^{n} A_i \qquad (1)$$

D. *Crossover*

Definition: It is the process of combining two parent chromosomes to form two different child chromosomes by interchanging the set of gene values in the parent chromosomes.

After calculating the fitness values for each chromosome separately, the chromosomes are applied for crossover process by selecting two by two chromosomes. The two parent chromosomes give two child chromosomes after crossover process.

E. *Mutation*

Definition: It is the process of altering a gene value in a chromosome.

After the cross over process, we have to apply the mutation process on every chromosome separately. A sample mutation process is shown in the Fig.4.

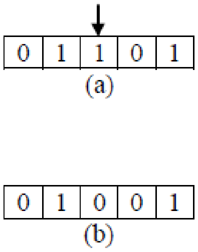

Fig.4 Sample Mutation Process

F. *ABC Operator*

Definition: It is the operator used in the Artificial Bee Colony algorithm to generate a new solution.

IV. CHAID

The Chi-square Automatic Interaction Detector is denoted as CHAID. CHAID is a technique of decision tree that is used to discover the relationship among the variables. CHAID examination identifies how the variables combine to detail the result in given dependent variables. The categorical or ordinal data is used in CHAID examination. The CHAID technique is the best tool to answer the survey related queries. In CHAID analysis we generate a decision tree or classification tree. The analysis initiated with the identification of target variable or dependent variable. The CHAID algorithm approves only the nominal or ordinal categorical predictors. When the predictors are nonstop, they are transformed into ordinal predictors before using it. The CHAID algorithm consists of three stages; they are merging, splitting and stopping. A tree is grown by repeatedly using the three stages on each node initiated from the root node.

A. *Binning Continuous Predictors*

For a given set of break points $a_1, a_2, a_3, \ldots\ldots, a_{K-1}$ which are arranged in ascending order, the given X is represented as category $C(X)$ as follows:

$$C(x) = \begin{cases} 1 & ; x \le a_1 \\ k+1 & ; a_k < x \le a_{k+1}, k = 1, \ldots., K-2 \\ K & ; a_{K-1} < x \end{cases} \quad (2)$$

B. *Merging*

Merge non-significant categories for each predictor variable X . If X is used to split the node, each final category of X will result in one child node.

The calculation of the p-value depends on the type of dependent variable. The merging step of the CHAID needs the p-value for a pair of X categories and occasionally needs the p-value for each category of X . When the p-value for a pair of X categories is essential, only part of data in the current node is related. Let D represents the relevant data and suppose in D there are I categories of X and J categories of Y (if Y is categorical). The p-value computation using data in D is as follows:

Suppose a predictor variable initially has I categories and reduced to r categories after the merging process. The Bonferroni multiplier B is the number of possible ways that I categories can be combined into r categories for $B = 1, r = I$ .For $2 \le r < I$, use the formula mentioned below:

$$B = \begin{cases} \binom{I-1}{r-1} & \textit{Ordinal Predictor} \\ \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!(r-v)!} & \textit{Nomin al Predictor} \\ \binom{I-2}{r-2} + r\binom{I-2}{r-1} & \textit{Ordinal with a missin g category} \end{cases} \quad (3)$$

If the dependent variable of a case is missing, it will not be used in the analysis. If the entire predictor variables of a case are missing, this case is ignored and if the case weight is missing zero or negative, the case is ignored and if the frequency weight is missing, zero or negative, the case is ignored. Otherwise, the missing values will be considered as a predictor category. Using all the existing information from the data, the algorithm initially creates the best set of categories for ordinal predictors. The algorithm then identifies the most similar category to the missing category. Eventually, the algorithm decides whether to combine the missing category with its most akin category or to keep the missing category as a separate category.

C. *Splitting*

The best split is identified in the combining step for each predictor. The splitting step is exploited to adopt the predictor to split the node. The selection is done by comparing the adapted p-value connected with each predictor.

V. CONCLUSION

In this paper we have recommended a hybrid classifier technique based on CHAID and genetic algorithm. The genetic algorithm is used with the ABC operator to extract the best necessary attributes from a set of attributes. The genetic algorithm takes the input from solution encoding as chromosomes. The fitness calculation in genetic algorithm is used to find the fitness values of each chromosome to decide which chromosomes are better and the chromosomes are applied for the crossover process, mutation process and eventually with ABC operator. After each process in genetic algorithm, we have calculated the fitness values separately and the whole process of genetic algorithm is repeated until we get a best solution. The CHAID is then used to generate the decision tree using the best solution. The performance of our technique is analyzed using two different datasets and it is compared with the existing techniques that take all the attributes from the dataset for classification using CHAID and showed our proposed technique is better in terms of accuracy.

References

[1] Thearling, K, Information About Data Mining and Analytic Technologies, http://www.thearling.com/text/dmwhite/dmwhite.htm , accessed July, 2009.

[2] Koyuncugil, A, S “Fuzzy Data Mining and Its Application to Capital Markets,” PHD. Thesis, Ankara University, 2006.

[3] Lee, S.J and Siau, K “A Review of Data Mining Techniques,” Industrial Management & Data Systems, vol.101,no.1,pp. 41-46,2001.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Vol. 17, pp. 37-54, 1996

[5] A B M Shawkat Ali, Saleh A. Wasimi (2009), “Data Mining: Methods and Techniques”, CENGAGE Learning, India.

[6] Ali, S. (2005), “Automated Support Vector Learning Algorithms”, Ph. D Thesis, Monash University, Australia.

[7] Bauer, E., Kohavi, R. (2004), “An empirical comparison of voting classificationalgorithms: Bagging, Boosting, And Variants, Machine Learning”, 36(1-2), pp 105-139.

[8] Boser, B. E., Guyon, I. M., Vapnik, V. N. (1992), “A training algorithm for optimal margin classifiers”. In D. Haussler, Editor.Proceeding of the 5th Annual Workshop on COLT, pp 144-152, ACM Press, Pittsburgh.

[9]Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984), “Classification and Regression Trees”, Wadsworth: Belmont, CA.

[10]M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, IJCST ISSN: 2229- 4333, vol. 2, no. 2, **(2011)** June.

[11] J. Soni, U. Ansari, D. Sharma and S. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, **(2011)**.

[12] C. S. Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques”, **(2012)**.

[13] K. Srinivas, B. Kavihta Rani and Dr. A.Govrdhan, “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, International Journal on Computer Science and Engineering, vol. 02, no. 02, **(2010)**, pp. 250-255.

[14] D. S. Kumar, G. Sathyadevi and S. Sivanesh, “Decision Support System for Medical Diagnosis Using Data Mining”, **(2011)**.

[15] W. L. Zuoa, Z. Y. Wanga, T. Liua and H. L. Chenc, “Effective detection of Parkinson’s disease using an adaptive fuzzy k-nearest neighbor approach”, Biomedical Signal Processing and Control, Elsevier, **(2013)**,

[16] T. Balasubramanian and R. Umarani, “An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique”, Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, **(2012)** March 21-23.

[17] J. Escudero, J. P. Zajicek and E. Ifeachor, “Early Detection and Characterization of Alzheimer’s Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means”, 33rd Annual International Conference of the IEEE EMBS Boston, assachusetts USA, **(2011)** August 30-September 3.

[18] H. Chipman and R. Tibshirani, “Hybrid hierarchical clustering with applications to microarray data”, Biostatistics, vol. 7, no. 2, **(2009)**, pp. 286-301.

[19]A. Rajkumar and G. S. Reena, Diagnosis of Heart Disease Using Datamining Algorithm”, Global Journal of Computer Science and Technology, vol. 10, no. 10, **(2010)**.

[20] K. Srinivas, B. K. Rani and A. Govrdhan, “Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks”, International Journal on Computer Science and Engineering (IJCSE), vol. 02, no. 02, **(2010)**, pp. 250-255.

[21] Y. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat and T. Naenna, “Data Mining of Magneto cardiograms For Prediction of Ischemic Heart Disease”, EXCLI Journal, **(2010)**.