Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYSIS OF OPTIMIZATION TECHNIQUES IN CHI-SQUARED AUTOMATIC INTERACTION DETECTION

[1]SWARNALATHA S.R, [2]G.M. KADHAR NAWAZ
[1]lecturer Faculty of Master of Computer Applications,
SCT IT, VTU University, Bangalore
[2]Director, Department of Computer Applications Periyar University, Salem
E-mail: [1]swarnalatha0280@gmail.com

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: *Optimization, Chi-Square, Automatic Interaction, Detection*

1. INTRODUCTION

Data are gathering from diverse fields and are accumulated at a dramatic pace. There is an imperative requirement for a fresh generation of computational theories and tools to help humans in deriving useful information (knowledge) from the hastily growing volumes of digital data. These theories and tools are the topic of the field which is becoming known for knowledge discovery in databases (KDD) [4].The Knowledge Discovery in Databases process contains a small number of steps leading from raw data collections to some form of fresh knowledge. The iterative process consists of the following steps [5] which are Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation and Knowledge representation. Among the diverse Knowledge Discovery in Databases process, data mining is a becoming known domain that combines statistical analysis, machine learning and database technology to derive hidden patterns and relationships from large databases [6].

A well-recognized Data Mining task is classification and it has been studied widely in the fields of statistics, pattern recognition, decision theory, machine learning literature, neural networks and more. Classification process usually uses supervised learning techniques that induce a classification model from a database [7]. 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 [8]. The classification algorithm learns from the training set and generates a model and this model is used to classify fresh objects [9]. Numerous classification algorithms have been recommended in the literature, such as decision tree classifiers [10], rule-based classifiers [11], Bayesian classifiers [12], support vector machines (SVM) [13], artificial neural networks [14], Lazy Learners, and ensemble methods [15].

Basically, decision tree is an eminent classification model in machine learning, artificial intelligence and data mining because they are: (1) Practical, with a wide range of applications, (2) Simple and easy to understand, and (3) Rules can be extracted and executed manually. Traditionally, numerical values are handled as precise and definite point-values. In many applications, however, data uncertainty arises naturally because of: (i) Measurement errors caused by equipment limitations, (ii) Data staleness caused by

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

transmission bandwidth, and (iii) Repeated measurements. Currently, research on data classification primarily focuses on certain data, in which precise and definite value is habitually assumed.

Data mining techniques can yield the advantages of automation on prevailing software and hardware platforms, and can be applied on fresh systems as existing platforms are upgraded and fresh products developed. When data mining tools are applied on high performance parallel processing systems, they can evaluate massive databases in minutes. In data mining, the techniques which are used in general are [1] [2] artificial neural networks, Decision trees, Genetic algorithms, nearest neighbor method, Rule induction. 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 [3].

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 [1].

In this paper, we recommend a hybrid classifier by combining the CHAID with the genetic algorithm. Here, the genetic algorithm with ABC operator is used to extract the best attributes from a set of attributes. Initially, a set of inputs are taken after solution encoding and give the inputs to the genetic algorithm. The processes involved in the genetic algorithm are fitness calculation, crossover and mutation. We include the ABC operator after the mutation process in the genetic algorithm to improve the performance. The process of the genetic algorithm with ABC operator is repeated until we get the best attributes to consider. After we get the best attributes, it is given to CHAID to produce a decision tree that can provide satisfactory classification performance for the given input data.

This paper is organized as follows: the second section portrays a short review of related work and the third section explains our recommended technique and the fourth section shows the outcome of our technique and compared it with the existing technique and the fifth section shows the conclusion.

2. RELATED WORKS: A SHORT REVIEW

Many Researchers have developed different techniques for Chi-squared automatic interaction detection and decision tree. Among them, a handful of significant researches are presented in this section. Mu-Chen Chen *et al.* [16] have recommended an algorithm of decision trees, Chi-squared automatic interaction detection (CHAID) to generate a classifier for predicting breast cancer and fibroadenoma. The outcomes of their technique demonstrated that the decision tree technique was more reliable than logistic regression in terms of rule accuracy and knowledge transparency to physicians. Moreover, the medical classifier can help the inexperienced physicians to prevent from misdiagnosis.

A technique to compute the confidence bounds around the predictions from chi-squared automatic interaction detection (CHAID) was developed byMerel van Diepen and Philip Hans Franses [17]. These bounds were effective for choosing the ex ante target as some clusters may or may not overlap. They were also useful for ex post calculation of the predictive quality of CHAID. Their technique was illustrated using simulated data. Yih-Jeng Lin *et al.* [18] have suggested chi-square automatic interaction detector (CHAID) to find the solution for the polysemy issues in a Chinese to Taiwanese TTS system. The precise pronunciation of a word impacts the clarity and fluency of Taiwanese speech. They have solved the polysemy problems using language models with outstanding outcomes. They have used the CHAID which had been found to be an optional technique for getting meaningful segments that are predictive

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of a K-category (nominal or ordinal) criterion variable to the polysemy problem.

Rather than abstracting unclear data by statistical derivatives (such as mean and median), Smith Tsang *et al*.[19] have discovered that the accuracy of a decision tree classifier was much advanced if the "complete information" of a data item (taking into account the probability density function (pdf)) was utilized. To deal the data tuples with the uncertain values, they extended the classical decision tree building algorithm. More researches have been conducted that showed the resulting classifiers were more precise than those using value averages. The CPU demand for the construction of decision tree for uncertain data was more than the certain data since the processing of pdfs was more expansive for computation than processing single values. So, they proposed a series of pruning models to manage those issues and the construction was improved efficiently.

Sample selections in fuzzy decision tree induction in terms of maximum ambiguity have been suggested by Xi-Zhao Wang *et al.* [20]. Contrast with the prevailing sample selection techniques, their mechanism selects the samples based on the principle of maximal classification ambiguity. The primary privilege of their mechanism is while including the selected samples to the training set, the adjustment of the fuzzy decision tree was minimized. The privilege was affirmed by the theoretical analysis of the leaf-nodes frequency in the decision trees. The decision tree created from the selected samples typically has better performance contrast to the decision tree created from the original database. Moreover, their experimental outcomes showed that the generalization ability of the tree based on their selection mechanism was more superior than that based on random selection mechanism.

The socioeconomic and demographic factors connected with the digital divide in the province of Cordoba were analyzed by Diaz, C and Jones, C [21]. Distinctively they saw to what extent factors such as gender, age, education level, socioeconomic status, industry, combined and were connected with e-inclusion. Data were collected from a survey of 4026 residents of Cordoba in 2009, as branch of a research project approved by the Ministry of Science and Technology, National University of Cordoba. They used the CHAID decision tree taxonomy technique for analysis. Their outcomes suggest that the level of education and age were the main factors connected with the e-inclusion, the first being the most critical.

3. 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. Here, a hybrid learning procedure will be developed to select the most suitable sub set of features with the help of genetic algorithm and it combined with CHAID to recursively learn the decision rules.Initially, Genetic algorithm obtains a set of initial population that evolves subsets of discriminatory features for pattern classification. Thereafter the genetic algorithm search the population space of all possible subsets to select optimal subset based on the designed fitness function and it is given to CHAID to produce a decision tree that can provide satisfactory classification performance for the given input data.The Fig.1 shows the overall process of our recommended technique.

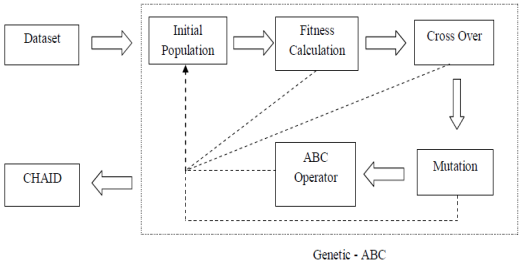


Fig.1 Overall Process of our proposed technique

3.1. 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. The processes are detailed as follows:

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1.1.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'.

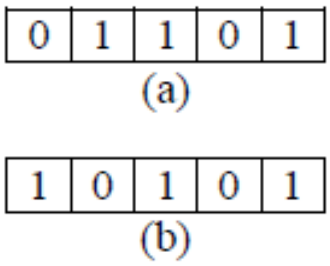


Fig.3 Sample set of attributes

The Fig.3 shows two set of sample attributes we considered for our technique. The Fig.3(a) shows that we have considered only the attributes 'A2', 'A3' and 'A5' and the Fig.3(b) shows that we have considered only the attributes 'A1', 'A3' and 'A5'. Similarly, we have to generate different attributes based on the attributes in our dataset.

3.1.2.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 because genetic 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. 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.

3.1.3.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:

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

Where,

f → Fitness

N → Total number of attributes considered

A_i → Values for each attributes

The evaluations of each attributes are as follows: Initially we have to split the values of each attributes into four parts based on the values in the attributes and give the part name to the specified value. For instance, split the values from 0 to 100 as four parts and assign the part value to the specified value. Take the values from 0 to 25 as first part and assign the value as 1. Similarly, from 26 to 50 as second part and assign the value as 2 and from 51 to 75 as third part and assign the value as 3 and from 76 to 100 as fourth part and assign the value as 4. The formula for calculating the values of each attribute is as follows:

$$A_i(j) = \frac{\begin{array}{c}diff\ between\ total\ no.of\ 'j'\ in\ first\ type\\ of\ class\ and\ total\ no.of\ 'j'\ in\ second\ type\ of\ class\end{array}}{Total\ no.of\ sets\ in\ the\ dataset}$$

Where,

j → Assigned value after splitting the values in the attributes and j varies from 0 to 4

3.1.4.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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

applied for crossover process by selecting two by two chromosomes. The two parent chromosomes give two child chromosomes after crossover process.

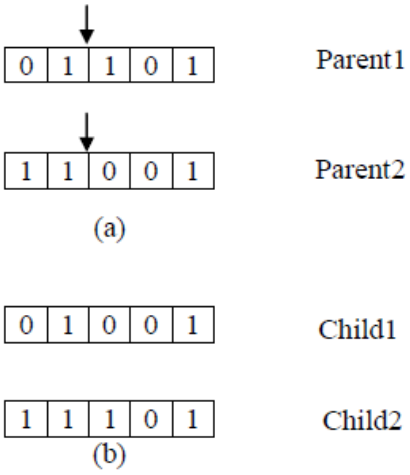


Fig.4 Sample Crossover Process

The Fig.4 explains the sample crossover process. The Fig.4(a) shows the two parent chromosomes we considered for crossover process and the Fig.4(b) shows the child chromosomes generated after crossover process and the arrow marks in the Fig.4(a) shows the point to make the crossover process on the parent chromosomes. After the crossover process, we have to calculate the fitness value for each chromosome separately.

3.1.5.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.5.

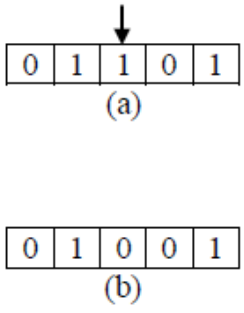


Fig.5 Sample Mutation Process

The Fig.5(a) shows the chromosome we considered for mutation process and the Fig.5(b) shows the chromosome after mutation process. After applying the mutation on every chromosome, we have to calculate the fitness for each chromosome separately.

3.1.6.ABC Operator

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

The ABC operator [22] is applied on every chromosome after the mutation process to improve the performance. The addition function involved in the ABC operator is binary addition. The ABC operator is denoted by the equation below:

$$v_{i,j} = x_{i,j} + \Phi_{i,j}(x_{i,j} - x_{k,j})$$

Where,

$x_{i,j}$ → Gene value of the chromosome after mutation

$x_{k,j}$ → Gene value of the randomly generated chromosome

$\Phi_{i,j}$ → Random value which would be in 0 or 1

After applying the ABC operator, we need to find the fitness value for each chromosome individually. From the whole process we have to select a chromosome that has best fitness value and store it separately and we need to do the same process repeatedly by taking best ten chromosomes based on the fitness values from the whole process i.e. initial population, crossover, mutation and ABC operator. The iteration will go on until we find the best solution i.e. chromosome with best fitness value.

3.2. 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 amid 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. During examination the CHAID technique alters the continuous data into ordinal data. In CHAID technique, we can visually see the connection amid the variable and the associated related factor with a tree. In most surveys, the solution is a categorical value as an alternative of a continuous value. Identifying the connection amid

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the categorical values is a difficult job. 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.

3.2.1. Binning Continuous Predictors

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

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

3.2.2. 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 merging step also evaluates the adjusted p-value that is to be used in splitting.

1. Stop and set the adjusted p-value as 1, if X has only one category.
2. If X has two categories, go to step 8.
3. Else find the allowable pair of categories of the predictor variable that is least considerably different (i.e. most similar). The acceptable pair of categories for ordinal predictor is two contiguous categories and for nominal predictor is any two categories. The most similar pair is the pair that has largest value of the p-value with respect to the dependent variable Y.
4. For the pair having largest p-value, check if its p-value is greater than the user specified alpha level α_{merge}. If its p-value is greater than the user specified alpha level, this pair is merged into a single compound category. Thereafter, a new set of categories of X is formed. If its p-value is not greater than the user specified alpha level, go to step 7.
5. Find the binary split within the compound category that has least p-value, if the recently produced compound category has three or more original categories. This is an optional step.
6. Go to second step.
7. The category that has very few observations as compared with a user specified minimum segment size is merged with the most similar other category as measured by the largest p-values.
8. The adjusted p-value is calculated for the merged categories by applying Bonferroni adjustments.

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:

$$B = \begin{cases} \binom{I-1}{r-1} & Ordinal\,\text{Pr}edictor \\ \sum_{v=0}^{r-1} (-1)^v \frac{(r-v)^I}{v!(r-v)!} & Nominal\text{Pr}edictor \\ \binom{I-2}{r-2} + r\binom{I-2}{r-1} & Ordinal\,with\,a\,mis\sin g\,category \end{cases}$$

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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2.3. 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. The adapted p-value is obtained in the merging step.

1. Adopt the predictor that has the least adjusted p-value.
2. Then check the adjusted p-value and the user specified alpha level α_{merge}. Split the node using the predictor, if the adjusted p-value is less than or equal to the alpha level α_{merge} or else don't split the node and the node is considered as a terminal node.

3.2.4.Stopping

The stopping step examines if the tree growing process should be stopped according to the stopping rules mentioned below:

1. If the node becomes pure i.e. all cases in a node have the similar values of the dependent variable, the node will not be split.
2. The node will not get split, if the entire cases in a node have identical values for each predictor.
3. The tree growing process will get stop, if the current tree depth arrived the limit value of the user specified maximum tree depth.
4. If the size of the node is less than the minimum node size set by the user, the node will not get split.
5. If the split of the node ensues in a child node whose node size is less compared to the user specified minimum child node size, the child nodes that have very few cases as compared with this minimum will get combine with the most similar child node as calculated by the largest p-values. The node will not get split, if the ensuing number of child nodes is one.

4. RESULT AND DISCUSSION

This section details the outcome we obtained for our recommended technique. Here, we have used two datasets which are Hepatitis and Multiple Features for our recommended technique and the outcome is evaluated using sensitivity, specificity and accuracy.

4.1. Experimental setup and Dataset Description

Our recommended method is implemented in java (jdk 1.6) that has the system configuration as i3 processor with 2GB RAM. The Hepatitis dataset and the Multiple Features dataset are taken from the UCI Machine Learning Repository. The performance of our technique is calculated using some evaluation metrics.

4.2. Metrics used for Evaluation

Our technique is evaluated using sensitivity, specificity and accuracy. The evaluations are as follows: Sensitivity is the ratio of true positive to the sum of true positive and false negative. It is denoted by an equation below:

$$Sensitivity = \frac{TP}{TP + FN}$$

Here, TP represents true positive and FN represents false negative. The specificity is defined as it is the ratio of true negative to the sum of true negative and false positive. It is denoted by a formula below:

$$Specificity = \frac{TN}{TN + FP}$$

In the above formula, TN represents true negative and FP represents false positive. The accuracy is the ratio of sum of true negative and true positive to the sum of true negative, true positive, false negative and false positive.

$$Accuracy = (TN + TP)/(TN + TP + FN + FP)$$

4.3. Performance Analysis

This section shows the performance of our technique based on the evaluation metrics using the two datasets we chosen i.e. Hepatitis dataset and Multiple Features dataset. We compare the performance of our proposed technique with the existing techniques [23, 24] that takes all the attributes in the data set for classification using CHAID. But in our proposed technique we remove the less significant attributes from the data set using genetic algorithm with ABC operator and thereafter we use CHAID for classification.

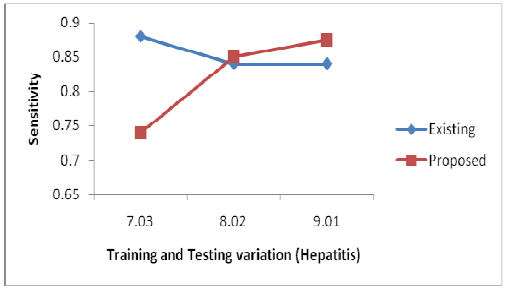
4.3.1.Hepatitis Dataset

The metrics we mentioned above are calculated for the Hepatitis Dataset using three variations of training and testing data. The variations are seventy percentages of data for training and thirty

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

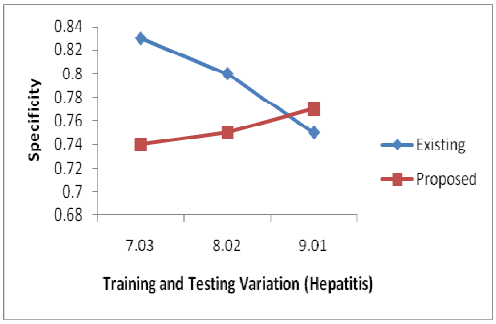
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

percentages of data for testing, eighty percentage of data for training and twenty percentage of data for testing and ninety percentages of data for training and ten percentages of data for testing and the variations are mentioned as 7.3, 8.2 and 9.1 in the x-axis.



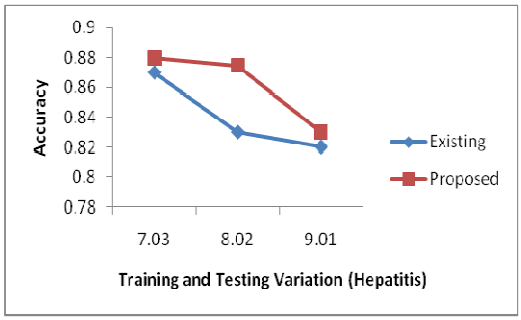
Graph.1 Sensitivity Using Hepatitis Dataset

The Graph.1 shows the comparison of sensitivity values we obtained for the existing technique and our recommended technique using Hepatitis dataset for different variations of training and testing data. The Graph.1 explains as follows: the sensitivity of existing technique is 0.88 and the sensitivity of our recommended technique is 0.74 when we use seventy percentages of data in the dataset for training and thirty percentages of data in the dataset for testing. When we set eighty percentages of data in the dataset for training and twenty percentages of data in the dataset for testing, the sensitivity value is 0.84 for the existing technique and 0.85 for our recommended technique. The sensitivity values for existing and proposed technique is 0.84 and 0.875 respectively when we use ninety percentages of data in the dataset for training and ten percentages of data in the dataset for testing.



Graph.2 Specificity Using Hepatitis Dataset

The Graph.2 shows the specificity values we gained for the existing and our recommended technique using Hepatitis Dataset. The Graph.2 details as follows: when we use seventy percentages of data for training and thirty percentages of data for testing, the specificity value we obtained for the existing technique is 0.83 and the specificity is 0.74 for our recommended technique. While using eighty percentages of data for training and twenty percentages of data for testing, we obtained the specificity as 0.8 for existing technique and 0.75 for our recommended technique. The specificity of the existing and our recommended technique is 0.75 and 0.77 respectively when we use ninety percentage of data in the dataset for training and ten percentage of data in the dataset for testing.



Graph.3 Accuracy Using Hepatitis Dataset

The Graph.3 shows the accuracy of the existing and our recommended technique for different variations of training and testing data using Hepatitis Dataset. The explanation of Graph.3 is as follows: when the training data used as seventy percentage of the dataset and the testing data used as thirty percentage of the dataset, the accuracy obtained for the existing technique is eighty seven percentages and the accuracy is eighty eight percentages for our recommended technique. The accuracy is eighty three percentages for the existing technique and eighty seven percentages for our recommended technique when we use the variation 8.02 and when we use the variation 9.01, the accuracy we obtained is eighty two percentages for the existing technique and eighty three percentages for our recommended technique. The Fig.6 shows the decision tree generated for the Hepatitis Dataset using our technique.

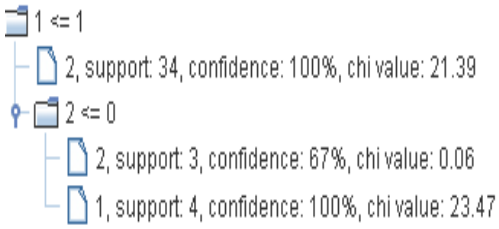


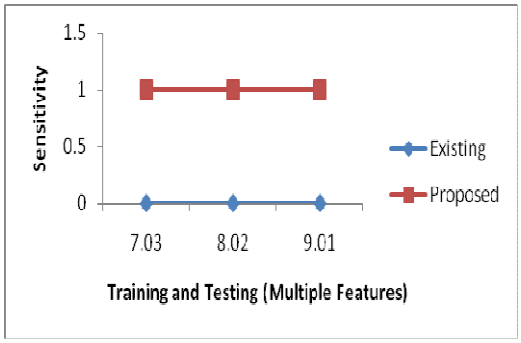
Fig.6 Decision Tree for Hepatitis Dataset

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

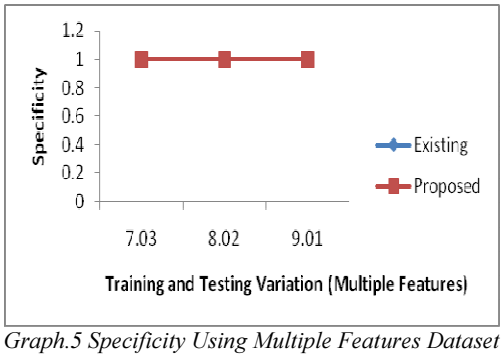
4.3.2.Multiple Features Dataset

The evaluation metrics are calculated for the existing and our recommended technique using Multiple Features dataset for different variations of training and testing data. The different variations are seventy percentages of data in the dataset are used for training and thirty percentages of data in the dataset are used for testing, eighty percentages of data in the dataset are used for training and twenty percentages of data in the dataset are used for testing, and ninety percentages of data in the dataset are used for training and ten percentages of data in the dataset are used for testing and the variations are represented by 7.3, 8.2 and 9.1 in the x-axis.



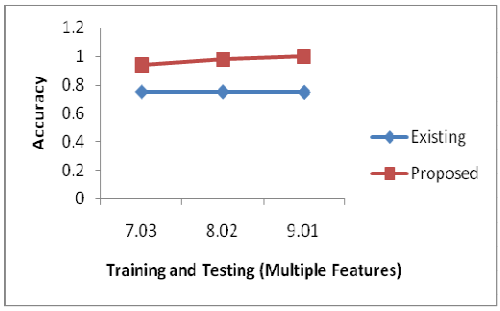
Graph.4 Sensitivity Using Multiple Features Dataset

The sensitivity for the existing and proposed technique is shown in the Graph.4 using Multiple Features dataset. The sensitivity value is zero for the existing technique for all the variations and the sensitivity value is one for all the variations of our recommended technique.



Graph.5 Specificity Using Multiple Features Dataset

The Graph.5 shows the specificity values for the existing and the proposed technique using Multiple Features dataset. Here, the specificity values for both the existing and proposed technique is one for all the variations.



Graph.6 Accuracy Using Multiple Features Dataset

The accuracy we obtained for the existing and the proposed technique is shown in the Graph.6 using the Multiple Features dataset. Here, the accuracy of existing technique is seventy five percentages and the accuracy of the proposed technique is ninety four percentages when the training data used is seventy percentages of the dataset and the testing data used is thirty percentages of the dataset. When the variation is 8.02, the accuracy we obtained is seventy five percentages for the existing technique and ninety eight percentages for our recommended technique. The accuracy for the existing and the recommended technique is seventy four percentages and hundred percentages respectively when the training data used is ninety percentages of the dataset and the testing data used is ten percentages of the dataset. The decision tree created for the Multiple Features dataset is shown in Fig.7 using our recommended technique.

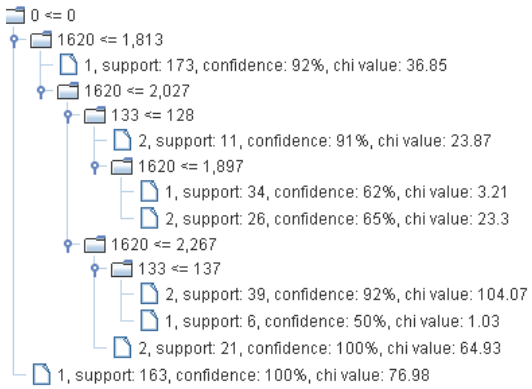


Fig.7 Decision Tree for Multiple Features dataset

5. 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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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] Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", *IT Professional*, pp: 28-31, 2000.

[6] Osmar R. Zaiane, "Chapter I: Introduction to Data Mining", *CMPUT690 Principles of Knowledge Discovery in Databases*, 1999.

[7] Fayyad U., Piatetsky-Shapiro G., Smyth P., From data mining to knowledge discovery: an overview, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, *Menlo Park, CA, AAAI/MIT Press*, 1996, pp: 1-36.

[8] Romero C., Ventura S., Espejo P.G., and Hervas C., Data Mining Algorithms to Classify Students, proceedings of the 1st Int'l conference on educational data mining, *Canada*, 2008, pp: 8-17.

[9] Zhang J., Mani I., kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction, In Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), *Workshop on Learning from Imbalanced Data Sets II,* August 21, 2003

[10] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufman Publishers, 1993.

[11] W. W. Cohen, "Fast effective rule induction," *in Proc. of the 12th Intl. Conf. on Machine Learning*, 1995, pp. 115–123.

[12] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," *in National Conf. on Artigicial Intelligence,* 1992, pp. 223-228.

[13] V. Vapnik, The Nature of Statistical Learning Theory. *Springer Verlag*, 1995.

[14] R. Andrews, J. Diederich, and A. Tickle, "A survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge Based Systems*, vol. 8, no. 6, pp. 373–389, 1995.

[15] T. G. Dietterich, "Ensemble methods in machine learning," *Lecture Notes in Computer Science,* vol. 1857, pp. 1–15, 2000.

[16] Mu-Chen Chen, Hung-Chang Liao; Cheng-Lung Huang, "Predicting Breast Tumor via Mining DNA Viruses with Decision Tree,"*IEEE International Conference on Systems, Man and Cybernetics*, vol.5, pp.3585-3589, 2006.

[17] Merel van Diepen,Philip Hans Franses ,"Evaluating chi-squared automatic interaction detection, "*Journal Information Systems*,Vol.31,no.8,pp.814-831, December 2006.

[18] Yih-JengLin,Ming-Shing Yu ; Chin-Yu Lin, "Using Chi-Square Automatic Interaction Detector to Solve the Polysemy Problems in a Chinese to Taiwanese TTS System, "*Eighth International Conference on Intelligent Systems Design and Applications*,vol.1,pp.362-367, 2008.

[19] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho and Sau Dan Lee, "Decision Trees for Uncertain Data", *IEEE Transactions On Knowledge And Data Engineering*, vol. 23, no.1, pp. 64 - 78, January 2011.

[20] Xi-Zhao Wang,Ling-Cai Dong ; Jian-Hui Yan ," Maximum Ambiguity-Based Sample Selection in Fuzzy Decision Tree Induction,"*IEEE Transactions on Knowledge and Data Engineering*, vol.24,no.8,pp.1491-1505,2012.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[21] Diaz, C. Jones, C. "e-Inclusion in the province of Cordoba: Relations between social and digital divide," *6th Euro American Conference on Telematics and Information Systems (EATIS)*, pp.1-4, 2012.

[22] Dervis Karaboga and Celal Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied Soft Computing, vol. 11, pp. 652-657, 2011.

[23] Evgeny Antipov and Elena Pokryshevskaya, "Applying CHAID for logistic regression diagnostics and classification accuracy improvement", *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 18, pp. 109-117, 2010.

[24] Mingqi Zhao and Zhijun Sun, "Research on CHAID Decision Tree Model Based on Rating of China's Small Enterprises", *4th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1-5, 2008.