Middle-East Journal of Scientific Research 23 (6): 1197-1203, 2015
ISSN 1990-9233
© IDOSI Publications, 2015
DOI: 10.5829/idosi.mejsr.2015.23.06.93199

Identifying Genetic Defected Dna by Using C4.5 Based Binary Decision Tree Classifier

[1]B. Lakshmipathi and [2]G. Kousalya

[1]Department of CSE, University College of Engineering Kancheepuram, Kanchipuram, India, 631552
[2]Department of CSE, Coimbatore Institute of Technology, Coimbatore, India, 641014

Abstract: Data mining is the investigation of experimental datasets to extract tendency and relationships which will be substantive for the user. In genetic studies, these techniques have disclosed attention-grabbing findings, particularly within the hereditary predisposition to contract specific diseases. In this, Call Trees plays major role and this can be a choice support tool that contains tree like graph of choices and also the potential consequences. They need ordinarily been utilized in totally different universe eventualities starting from research to classifying mintage. The aim of this paper is to explain the final framework that we tend to adopted within the application of call tree algorithms to the analysis of deoxyribonucleic acid knowledge set associated with cases of genetic defect. The four algorithms have compared interms of their accuracy, exactness and recall on the data sets taken from world protein data bank. The time taken for learning the decision tree by every algorithmic program has a dditionally compared in this paper.

Key words: Classification · Decision Tree · CART · ID3 · CHAID · DNA Data · Medical Data Mining

INTRODUCTION

Data mining applications has wealthy focus as a result of its significance of classification algorithms. The comparison of classification formula may be a complicated and it's an open downside. First, the notion of the performance has outlined in some ways like accuracy, speed, cost, dependability, etc. Second, an appropriate tool is important to quantify this performance [1]. Third, a regular methodology should choose to match with the measured values.

The selection of the best classification algorithm for a given dataset is a very widespread problem. In this sense, it requires to make several methodological choices [2]. The various data mining techiques have been employed for classification process in the diagnosis of diseases [3]. Among them, this paper focuses on the decision tree algorithms from classification methods, which has used to assess the classification performance and to find the best algorithm in obtaining qualitative student data. The decision tree algorithm is very useful and well known for their classification. It has an advantage of easy to understand the process of creating and displaying the results [4]. Given a data set of attributes together with its classes, a decision tree produces sequences of rules that have used to recognize the classes for decision-making. The decision tree method has gained popularity due to its high accuracy of classifying the data set [5]. The most widely used algorithms for building a decision tree are ID3, C4.5, CHi-squared automatic interaction detector (CHAID) and classification and regression trees (CART).

Algorithms, such as ID3, often use heuristics that tend to find short decision trees [6], however, finding the shortest decision tree is a hard optimization problem [7]. The real world process of evolution inspires genetic Algorithms (GAs) [8]. GAs has used to construct short and near-optimal decision trees. In order to utilize genetic algorithms, decision trees must have represented as chromosomes on which genetic operators such as mutation and crossover has applied. Genetic algorithms have used in two ways for finding the near-optimal decision trees.

Corresponding Author: B. Lakshmipathi, Department of CSE, University College of Engineering Kancheepuram, Kanchipuram, India, 631552, Tel: +91 8220008850.

One way is that they had used to construct decision trees in a hybrid or preprocessing manner. The other way is to apply them directly to decision trees [9]. In this paper, we implement decision trees using C4.5 algorithm as well as ID3, CHAID and CART. The comparison of performance among the algorithms has done in this paper.

MATERIALS AND METHODS

Chi-Squared Automatic Interaction Detector (CHAID): CHAIS algorithm selects a set of predictors and their interactions and predicts the optimal value of the dependent variable. In the end, what we get is a classification tree. The dependent variable could be a qualitative variable or a quantitative variable.

The CHAID model or a CHAID diagram had thought of as an inverted tree trunk, which splits into different branches and sub branches. Initially the "Tree Trunk" is the totality of all the participants in the study [10]. A series of predictor variables has studied to see if splitting the sample based on these predictors leads to a statistically significant discrimination in the dependent variable. For this Chi square test and F tests are done and their P values are calculated. If the p values are not statistically significant, then the algorithm merges the respective predictor variables (or categories in case of categorical data). If a statistical significance has observed, then a split has made. This becomes the first branching of the tree [11]. Then for each of the groups, we face the question whether they had further split into subgroups so that there are significant differences in the dependent variable. The program will actually compute F-tests. Specifically, the algorithm proceeds as follows:

Preparing Predictors: The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) have "naturally" defined.

Merging Categories: The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a Chi-square test (Pearson Chi-square); for regression problems (where the dependent variable is continuous), F tests. If the respective test for a given pair of predictor categories is not statistically significant as defined by an alpha-to-merge value, then it will merge the respective predictor categories and repeat this step (i.e., find the next pair of categories, which now may include previously merged categories). If the statistical significance of the respective pair of predictor categories is significant (less than the respective alpha-to-merge value), then (optionally) it will compute a Bonferroni adjusted p-value of the set of categories for the respective predictor.

Selecting the Split Variable: The next step is to choose the split the predictor variable with the smallest adjusted p-value, i.e., the predictor variable that will yield the most significant split; if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits have performed, and the respective node is a terminal node. Continue this process until no further splits can be performed [12].

Classification and Regression Tree (CART): Classification and Regression Trees has based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. Unlike ID3 and C4.5 algorithms, CART produces binary splits. The Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost, complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy. To measure the degree of impurity are Gini Index that are defined as

$$\text{Gini(T)} = 1 - \sum_{j=1}^{n} P_j^2$$

The Gini Index of a pure table consists of a single class is zero because the probability is 1 and $1\text{-}1^2 = 0$. Similar to Entropy, Gini Index also reaches maximum value when all classes in the table have equal probability. To work out the information gain for A relative to S, first it needs to calculate the Gini Index of S.

$$\text{Gini Index (S)} = 1 - P_{\text{first}}\log_2(P_{\text{first}}) - P_{\text{second}}\log_2(P_{\text{second}}) - P_{\text{third}}\log_2(P_{\text{third}})$$

To determine the best attribute for a particular node information gain is calculated. The information gain has defined as,

$$P_{\text{fail}}\log_2(P_{\text{fail}})$$

Gini Index and information gain have calculated for all the nodes. As the result of the calculation, the attribute has used to expand the tree. Then delete the attribute of the samples in these sub-nodes and compute the Gini Index and the Information Gain to expand the tree using the attribute with highest gain value [13]. Repeat this process until the Entropy of the node equals null. At that moment, the nodes have unexpanded anymore because the samples in this node belong to the same class.

Iterative Dichotomiser (ID): Iterative Dichotomiser (ID3) is a simple decision tree-learning algorithm developed by Ross Quinlan [14]. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node, in order to select the attribute which is most useful for classifying a given set. A statistical property called information gain has defined to measure the worth of the attribute [15].

Given a data table that contains attributes and the class of the attributes, we can measure homogeneity (or heterogeneity) of the table based on the classes. If a table is pure or homogenous, it contains only a single class. If a data table contains several classes, then it says that the table is impure or heterogeneous [16]. To measure the degree of impurity or entropy,

$$\text{Entropy} = \sum -P_j \log_2 P_j$$

The entropy of a pure table (consist of single class) is zero because the probability is 1 and the log (1) = 0. Entropy reaches maximum value when all classes in the table have equal probability.

To work out the information gain for A relative to S, it first need to calculate the entropy of S [17]. To determine the best attribute for a particular node in the tree, information gain is applied. The information gain, gain (S, A) of an attribute A, relative to the collection of examples S, information gain is calculated to all the attributes [18].

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{V \in \text{Values}(A)} \frac{|S_V|}{|S|} \text{Entropy}(S_V)$$

Of the result of the calculations, the root attribute has used to expand the tree. Then delete the attribute of the samples in these sub-nodes and compute the Entropy and the Information Gain to expand the tree using the attribute with the highest gain value [19].

Repeat this process until the Entropy of the node equals null. At that moment, the nodes have been not expandable anymore because the samples in this node belong to the same class.

C4.5: C4.5 algorithm [20] is a successor of ID3 that uses gain ratio as a splitting criterion to partition the data set. The algorithm applies a kind of normalization to information gain using a "split information" value. To determine the best attributes for a particular node in the tree it uses the measure called information gain. The information gain, gain (S, A) of an attribute A, relative to a collection of examples S, is defined as

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{V \in \text{Values}(A)} \frac{|S_V|}{|S|} \text{Entropy}(S_V)$$

where values (A) is the set of all possible values for attribute A, and Sv is the subset of S for which attribute A has value v (i.e., Sv = {s ∈ S | A(s) = v}). The first term in the equation for Gain is just the entropy of the original collection S and the second term is the expected value of the entropy after S is partitioned using attribute A. The expected entropy described by this second term is simply the sum of the entropies of each subset weighted by the fraction of examples | S | / | S | that belong to Gain (S, A) is therefore the expected reduction in entropy caused by knowing the value of attribute A [19].

$$\text{Split Information}(S, A) = -\sum_{i=1}^{n} \frac{|S_t|}{|S|} \log_2 \frac{|S_t|}{|S|}$$

and

$$\text{Gain Ratio}(S, A) = \frac{\text{Gain Ratio}(S,A)}{\text{Split Information}(S,A)}$$

The process of selecting a new attribute and partitioning the training examples has now repeated for each non-terminal descendant node. Attributes that had incorporated higher in the tree have excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either of two conditions has met:

- Every attribute has already been included along this path through the tree,
- The training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

Gain ratio has used for attribute selection, before calculating Gain ratio Split Information calculated [21]. Take the original samples as the root of the decision tree. As the result of the calculation, the attribute has used to expand the tree. Then delete the attribute of the samples in these sub-nodes and compute split information to split the tree using the attribute with highest gain ratio value. This process continues, until all data have classified perfectly or run out of attributes. Repeat this process until the Entropy of the node equals null [22]. At that moment, expanding the node has not been possible anymore because the samples in this node belong to the same class.

Experimental Design

Bio Software: To conduct this experiment we have selected UGENE 1.4 bio software. UGENE is free, open-source Bioinformatics software that helps biologists to analyze various biological data, such as sequences, annotations, multiple alignments, phylogenetic trees, NGS assemblies, and others. The data can be stored both locally (on a personal computer) and on a shared storage. UGENE integrates dozens of well-known biological tools and algorithms, as well as original tools in context of genomics, evolutionary biology, virology and other branches of life science. UGENE provides a graphical interface for the pre-built tools so biologists.

Data Set: We have collected qualitative data such as genetic disorder affected and non-affected DNA from the repository of the World Protein Data Bank (PDB). From the collected data, 100 samples have taken for this experiment and data are available in PDB format, which has supported and processes by UGENE 1.4 tool.

RESULTS AND DISCUSSION

Table 1, describes the classification accuracy of ID3, C4.5, CHAID and CART algorithms when we applied on the collected DNA data sets.

When ID3 algorithm is applied, 67 instances are correctly classified and 31 instances are misclassified (i.e. for which an incorrect prediction has made) and two instances are unclassified. Since 67 instances are

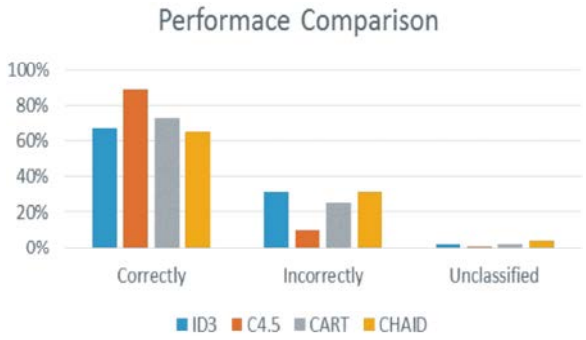

Fig. 1: Results of the comparison between the four different classifiers

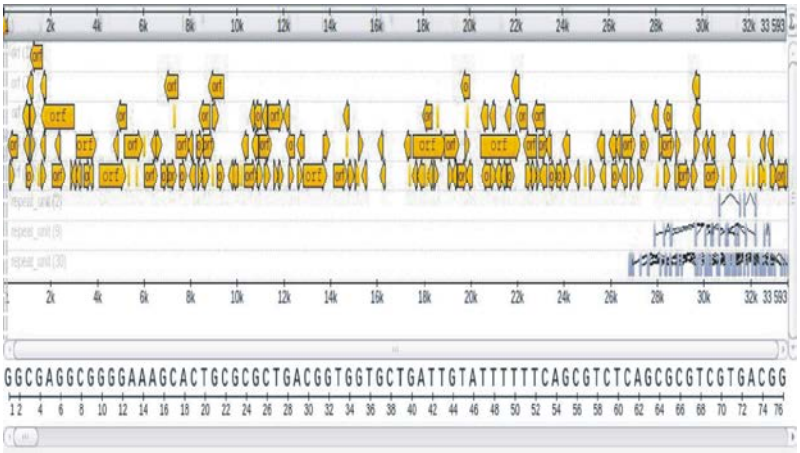

Fig. 2: Visualization of generating decision tree

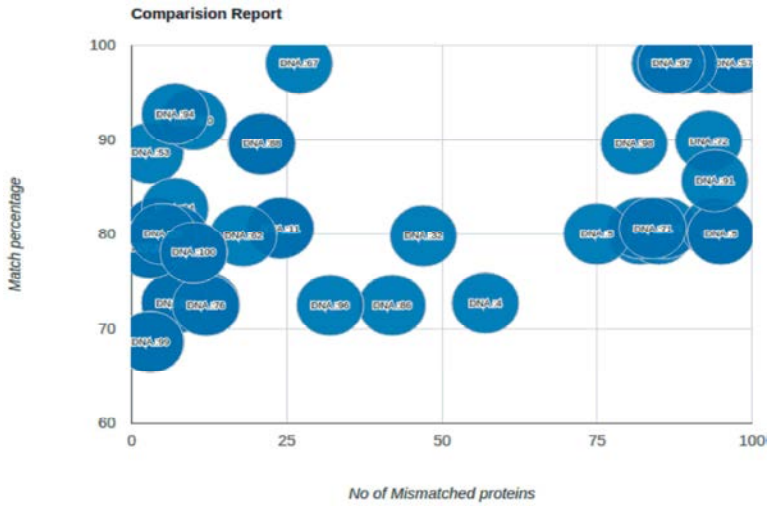

Fig. 3:Visualization of diagonized DNA dataset

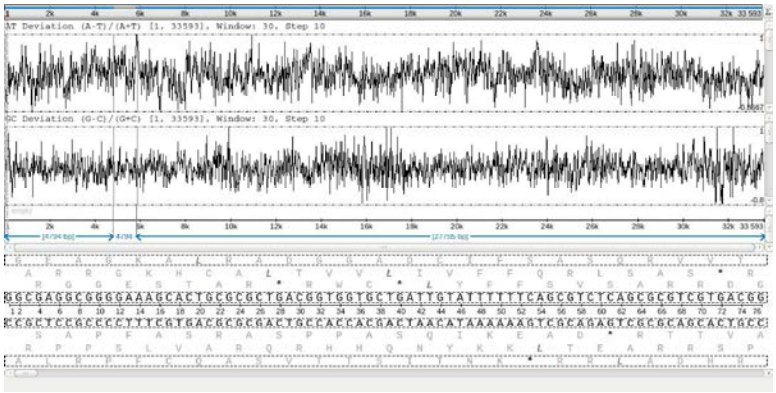

Fig. 4: Visualization of sequence analysis using UGENE tool

Table 1: Classification accuracy

Algorithms / Classified Instances	ID3	C4.5	CART	CHAID
Correctly	67%	89%	73%	65%
Incorrectly	31%	10%	25%	31%
Unclassified	2%	1%	2%	4%

misclassified, two instances are unclassified the ID3 algorithm does not obtain higher accuracy. The graph, which has depicted in Fig. 1, shows the results of the comparison between the four different algorithms interms of their classification accuracy.

From Table 1, it has found that the C4.5 yields the highest accuracy of 89% compared to ID3, CART and CHAID. However, the CART algorithm yields the highest accuracy of 73% when compared with other two algorithms.

The visualization of generating decision tree has depicted in Fig. 2 and the diagnosed DNA dataset has visualized in Fig. 3. The analysis of DNA sequence has illustrated in Fig. 4. The knowledge represented by decision tree has extracted and represented in the form of IF-THEN rules. From the above set of rules, it has found that the protein patterns have significantly related to sample genetic disorder affected DNA from the result has obtained.

CONCLUSION

This research work compares the performance of ID3, C4.5, CHAID and CART algorithms. The experimental result show that the C 4.5 has the best classification accuracy. This experimentation significance also concludes that CHAID was missing 4% of the data set and 31% of incorrect in genetic disorder identification. ID3 has also had the same amount mismatched identification with the CHAID. Overall, the 89% is better and near most one when compare to another ID3, CHAID and CART algorithms.

In future, the same experiment would have administered to the distinctive sort of infection like the diabetes, heart sicknesses and numerous different infections.

REFERENCES

1. Guishu Ji, Peiling Chen and Hang Song, 2007. Study the survey into the decision tree classification algorithms rule. Science Mosaic.
2. Yu, Q. and D.A. Clausi, 2007. SAR sea-ice image analysis based on iterative region growing using semantics. IEEE Trans. Geosci. Remote Sens., 45(12): 3919 -3931.
3. Nithya, R. and B. Santhi, 2014. A Data Mining Techniques for Diagnosis of Breast Cancer Disease. World Applied Sciences Journal, 29(29) (Data Mining and Soft Computing Techniques), pp: 18-23.
4. Moore, A.W. and D. Zuev, 2005. Internet traffic classification using Bayesian analysis techniques. ACM SIGMETRICS, New York: ACM Press, pp: 50-60.
5. Shukla, P., I. Basu, D. Graupe, D. Tuninetti and K.V. Slavin, 2012. A neural network-based design of an on-off adaptive control for Deep Brain Stimulation in movement disorders. Engineering in Medicine and Biology Society (EMBC), Annual International Conference of the IEEE, 4140-4143: 28.
6. Abe, H., H. Yokoi, M. Ohsaki and T. Yamaguchi, 2007. Developing an Integrated Time-Series Da ta Mining Environment for Medical Data Mining. Seventh IEEE International Conference on Data Mining- Workshops, pp: 127-132.
7. Zhou Xusheng, 2008. A P2P Traffic Classification Method Based on SVM. International Symposium on Computer Science and Comutational Technology, 2: 53-57.
8. Baraldi, 2009. Impact of radiometric calibration and specifications of spaceborne optical imaging sensors on the development of operational automatic remote sensing image understanding systems. IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., 2(2): 104-134.
9. Sjahputera, O., C.H. Davis, B. Claywell, N.J. Hudson, J.M. Keller, M.G. Vincent, Y. Li, M. Klaric and C.R. Shyu, 2008. GeoCDX: An automated change detection and exploitation system for high resolution satellite imagery. Proc. IGARSS, pp.V-467 -V-470.
10. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz and P. Blonda, 2010. Automatic spectral rule-based preliminary classification of radiometrically calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/ QuickBird/ OrbView/ GeoEye and DMC/SPOT-1/-2 imagery—Part I: System design and implementation. IEEE Trans. Geosci. Remote Sens., 48(3): 1299-1325.
11. Laha, N.R. Pal and J. Das, 2006. Land cover classification using fuzzy rules and aggregation of contextual information through evidence theory. IEEE Trans. Geosci. Remote Sens., 44(6): 1633-1641.
12. Nguyen, T.T.T and G. Armitage, 2008. A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials, 10(4).
13. Santaniello, S., G. Fiengo, L. Glielmo and W.M. Grill, 2011. Closed-loop control of deep brain stimulation: A simulation study. IEEE Transactions on Neural Systems and Rehabilation Engineering, 19: 15-24.
14. Dainotti, W. de Donato, A. Pescape and P.S. Rossi, 2008. Classification of Network Traffic via Packet-Level Hidden Markov Models. IEEE GLOBECOM, pp: 1-5.
15. Okun, O. and H. Priisalu, 2007. Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues. IbPRIA.
16. Hsia, T., A. Shie and L. Chen, 2008. Course planning of extension education to meet market demand by using data mining techniques-an example of chinkuo technology university in Taiwan. Expert Systems with Applications, 34(1).
17. Elenbogen, B.S. and N. Seliya, 2007. Detecting outsourced student programming assignments. Journal of Computing Sciences in Colleges, pp: 50-57. ACM.
18. YANG Zhe, L.I. Ling-zhi, J.I. Qi-jin and Z.H.U. Yan-qin, 2012. Network traffic classification using decision tree based on minimum partition distance. Journal of Communications, 33(3): 90-102.
19. Alireza, M., F. Mohammad and E. Artemis, 2008. Assessment of nasal volume and cross-sectional area by acoustic rhinometry in a sample of normal adult Iranians. Archives of Iranian Medicine, 11: 555-558.
20. Hatami, N. and R. Ebrahimpour, 2007. Combining multiple classifiers: diversity with boosting and combining by stacking. International Journal of Computer Science and Network Security, 7(1): 127-131.
21. Waxman, J.A., D. Graupe and D.W. Carley, 2010. Automated prediction of Apnea and Hypopnea, using a LAMSTAR artificial neural network. American journal of respiratory and Critical Care Medicine, 181(7): 727-733.
22. Worachartcheewan, Nantasenamat, Isarankura-Na-Ayudhya C, Pidetcha P and Prachayasittikul, V., 2010. Identification of metabolic syndrome using decision tree analysis. Diabetes Research and Clinical Practice, 90: e15-e18.

23. Ayodele A. Alaiya, Halima A. Alsini, Mohammed O. Gad El-Rab and Syed M. Hasnain, 2014. Protein Profiles of Indigenous and Commercial Extracts of Amaranthus Pollen for the Diagnosis of Allergy and Asthma Patients. World Applied Sciences Journal, 32(12): 2354-2361.

24. Krishna Anand, S., Manyam Pratyusha Chinni and Mounica K. Vineesha, 2014. Design and Analysis of a Back Propagation Neural Network in Estimating Risk of Coronary Artery Disease. American-Eurasian Journal of Scientific Research, 9(1): 16-25.