

A New Modified Decision Tree Algorithm based on ID3

Vibha Maduskar¹, Y.Kelkar²

tomailvibha@gmail.com,

Abstract: Decision Tree is an important method of Classification in data mining, it is used for prediction and forecasting from historical data . ID3 is one of the popular decision tree algorithm. In this work a new modified decision tree algorithm is proposed which combines the concept of similarity measure and decision tree to overcome the problem with conventional ID3 algorithm is that “to choose the attribute with many values. The experimental result shows that proposed algorithm perform well than conventional ID3 algorithm. The results is evaluate using N-cross validation technique. The new proposed algorithm is compute better accuracy then conventional ID3 algorithm.

Keywords: Decision Tree , Data Mining, Classification , ID3 algorithm ,accuracy ,instance based learning , N-cross validation.

1.Introduction:

A formal definition of data mining (DM), also known – historically – as data fishing, data dredging (1960-), knowledge discovery in databases (1990-), or – depending on the domain, as business intelligence, information discovery, information harvesting or data pattern processing – is data mining [3]

Definition: Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[1]

Classification is the task of learning a function that maps data objects to one or several classes in a predefined class set. To learn this function, classification methods need a training set, containing data objects that are already mapped to the class they belong to. After analyzing the training set, classification methods can map new unknown objects to the classes.“The objective of this work, “A New Modified Decision Tree Algorithm based on ID3”,

is to Modify and Develop ID3 algorithm and compare the Performance with conventional ID3 algorithm”.

The major challenge in field of data mining is to find hidden and useful pattern from large dataset. ID3 is popular decision tree algorithm for classification in data mining, but this method having the shortcoming .the main purpose of do this work is to overcome the that problem. many researcher proposed different methods for improving ID3, but there is No work done in ID3 using the distance metric . so this work motivated from different methods used for improve ID3 , main motivation for proposed work is improve the performance of classification Model.

2.Related Work:

2.Related Work: Due to the resource constraints of the research the main method used in this work is the ID3, improved ID3 Algorithm, which is used for the classification in data mining. This work is mainly focused on the classification which is used to divide the user defined categories. Numerous works in literature related with id-3 and improved id3 . id 3 used in different applications such as medical diagnosis, score analysis. Some of the work discuss following:

2.1 An improved id-3 decision tree algorithm: it represented improved id-3 algorithm[5] overcome the deficiency of general id3 which tend to the take attributes with many values. It introduces the entropy with the association function then calculate the highly information attribute by choosing highest information gain . Association function is not only overcome the attribute with many values but also represent the relations between all elements and their attributes.

2.2 An improved decision tree classification algorithm based on id3 used in score analysis [6]of students. in this proposed algorithm this combines the principle of Taylors formula with information entropy .it overcome the difficulty in id-3 algorithm is choosing attribute with many values. This work is implement on java platform using eclipse. this algorithm compares the id3 with the node no.

2.3 A novel tree based classification using id3 is used[7] . it user CAIM based online discretization and attributes selection process. It shows that modified id-3 using CAIM is work better then id-3. It gives better accuracy then id3.

2.4 An algorithm for better decision tree [8] used the sampling technique before apply id3 algorithm after find the samples it will proceed for calculating information gain and gini to select attribute.

Disease Diagnosis:

ID3 algorithm is also used in field of diagnosis of diseases.

For common disease diagnosis [9]on the basis of 12 input attributes firstly it perform k clustering algorithm to make cluster then perform id3 to find common disease . it also compares with the neural network and find that id3 is performs well then neural network.

For heart disease diagnosis : by using efficient supernova kernel for disease diagnosis. Three supervised learning algorithm is used to analyzing the dataset i.e. naïve bays ,KNN and decision tree. Data is evaluated using 10 fold cross validation . training is performed on 3000 instances with 14 different attributes. naïve bays gives better results.

Student performance:

Data mining technique can be used in educational field to enhance understanding of learning process to identifying, extracting and evaluating variables related to the learning process of students . mining in educational environment is called educational data mining.

A prediction for performance improvement of engineering students using classification [10] classification methods decision tree , Bayesian network can be applied on the educational data for predicting the student performance in examination.

III Improved ID3

To overcome the shortcoming of ID3 algorithm and improve the performance of classification model, the proposed work introduces the similarity measure is used for clustering and ID3 is used for construct decision tree. This work uses the principle of distance metric is within a data set will generally exist close proximity to other instances that have similar properties

.if instances are matched found that place in to same cluster otherwise in other cluster.

2 Proposed Decision Tree algorithm:

Input: data set as input to the algorithm

Output: Decision Tree

Steps of algorithm:

Begin:

1. Load input data set for training .
2. If attribute is uniquely identify in data set , remove from it.
3. On the basis of distance metric divide the given training data in to subsets.
4. Calculate the distance for (1.....N)each instance in available dataset

$$DE_{ud}(x,Y) = \sqrt{(\sum_{i=1}^n (X_i - Y_i)^2)}$$

Where X is selected instance
and Y is comparing instance.

5. if ($DE_{ud} > 55\%$ && $DE_{ud} < 70\%$)
 then instance is belong to same group and ,add in to new set and remove from original data set.
 otherwise
 do nothing .
6. Repeat the step 4 and 5 for each instance until all matched is not found
7. On each subset apply ID3 algorithm recursively.
 - If at target attribute all example are positive then return single tree root with label positive.
 - If at target attribute all example are negative then return single tree root with label negative.
 - If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
 - Otherwise

- Calculate the entropy of decision node ,if entropy is not equal to zero than calculate information gain for each attribute.
- For spitting each choose that value whose information gain is maximum.
- Apply algorithm is recursively until entropy is not reaches to the zero of each attribute.

End

IV Performance Measures

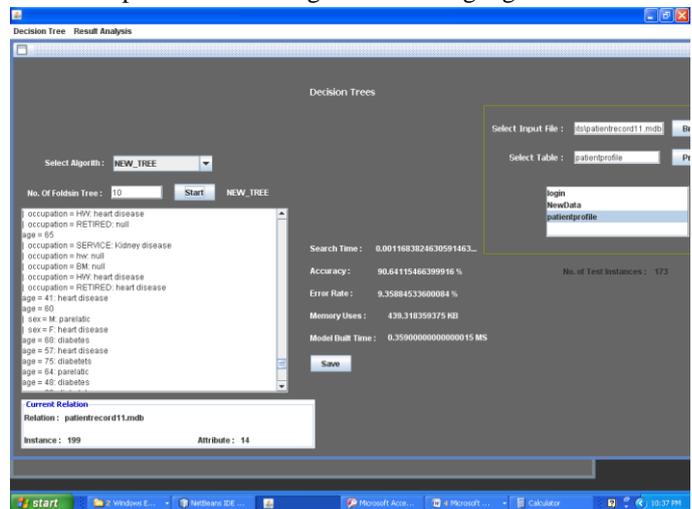
In order to minimize the bias associated with the random sampling of the training and test data samples k-Fold Cross Validation was adopted. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” D1, D2, : : , Dk, each of approximately equal size. Training and testing is performed k times (Han and Kamber, 2006). As Witten and Frank (2005) stated, extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up. Although these arguments are by no means conclusive, and debate continues to rage in machine learning and data mining circles about what is the best scheme for evaluation, 10-fold cross-validation has become the standard method in practical terms. Tests have also shown that the use of stratification improves results slightly. Thus the standard evaluation technique in situations where only limited data is available is stratified 10-fold cross-validation.

V User Navigation:

Below fig 4.1 shows the implementation window using netbeans .

To implement on patient profile data set which is in mdb format .select the decision tree on mdb format , then click on browse button select file using J file chooser than select the table and process the data click on process button and connection with data base is successfully completed .implementation Screen of the System proposed selected algorithm is named as New tree .There is a text box where input file is selected through the button named

as browse. input dataset is processed with start is contains the processing of selected algorithm in dropdown box. The result and performance parameter are provided in the right side of the screen. A text box given to input the number of folds which is used during the performance analysis of the algorithm. The left side screen provides the tree generated using algorithm.



Implemented Screen using New Tree

VI EXPERIMENTAL EVALUATION

6.1 Experimental Setup:

Experiment are perform on patient profile data set which containing 14 attributes and 200 training instances. N cross validation is used for random sampling of training and test set.

6.2 Results Evaluation:

For the processing of results N cross fold validation process is used to evaluate the results. In this process produce the training data in a random sequence and evaluate the correct outputs and incorrect classified outputs.

6.3 Result Analysis: N cross validation technique is used to evaluate results the below given table shows the comparative results in term of accuracy, build time ,error rate .

Table 6.1: Comparison ID-3 and Proposed algorithm

Algorithm used	Dataset	No of folds	Model search time ms	Accuracy %	Error rate %	Model built time ms
ID-3	Patient profile	2	0.0075	90.477	9.522	0.047
		4	0.011	90.61	9.38	0.063
		10	0.01	90.49	9.501	0.046
Modified ID-3		2	0.0023	90.53	9.56	0.362
		4	4.1	91.46	8.5	0.20
		10	0.001	90.64	9.35	0.35

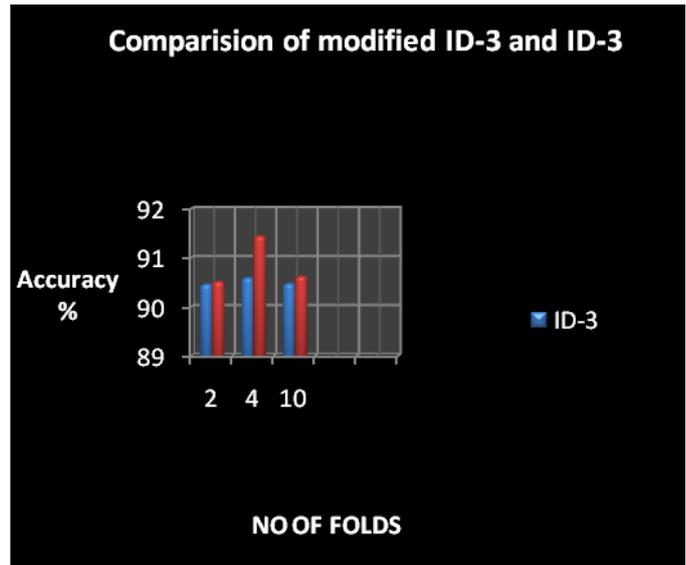


Fig 6.1: Graph for comparison of accuracy in ID-3 and modified ID-3

Above table shows that result on patient profile database ,evaluation is done by N-cross validation where different no of folds . experiment is done on total 200 instances.

When no of folds is 2 than find the 90.477% accurate result with traditional ID3 algorithm. At the same iteration is performed using modified ID3 algorithm on patient profile dataset it compute 90.53% accurate result, Modified ID3 is compute 0.5% more accurate than traditional ID3 algorithm.

6.4 Graph Results:

Comparison of ID3 and Modified ID3 algorithm is shown in different graph results.

- **Accuracy:** Graph for comparison between id-3 and modified id3.

The above graph shows the accuracy of ID3 and modified ID3 found that most of the time it provides the higher performance results thus new proposed algorithm perform well than Conventional algorithm.

The above graph shows the memory used by ID3 and modified ID3 and found that most of the time ID3 takes less memory then modified ID3.

VII CONCLUSION AND FUTURE WORK

These are following facts about this work.

1. Accuracy of Modified ID3 is better than ID-3.
2. Memory uses of modified ID3 is higher than ID3
3. Error rate of ID3 is more than Modified ID3

Memory used by ID3 is more than ID3 algorithm because it divides into clusters each cluster uses the memory .for improvement of result another algorithm is develop to improving result. memory improvement by adding some parameter is perform.

International Journal of Computer Architecture and Mobility (ISSN 2319-9229) Volume 1-Issue 9, July 2013

REFERENCES

- [1] Jiawei Han and Micheline Kamber, “*Data Mining: Concepts and Techniques*”, Second Edition
- [2]. D. Jiang, “*Information Theory and Coding [M]*”, Science and Technology of China University Press, 2001.
- [3] Huang Ming¹, NiuWenyong¹, Liang Xu , “*An improved decision tree classification algorithm based on ID3 and the application in score analysis*”, 2009 Chinese Control and Decision Conference (CCDC 2009).
- [4] Chen Jin, Luo De-lin, Mu Fen-xiang, “*An Improved ID3 Decision Tree Algorithm*”, Proceedings of 4th International Conference on Computer Science & Education 2009
- [5] Aashoo Bais, Kavita Deshmukh, Manish Shrivastava, *Implementation of Decision Tree*, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-2, December 2012
- [6]. M.R.Lad, R.G.Mehta, D.P.Rana , “*A Noval Tree Based Classification*”, [IJESAT] International Journal of Engineering and Advanced Technology Volume-2, Issue-3, 581 – 586 may 2012.
- [7] T.Jyothirmayi, Suresh Reddy , “*An Algorithm for Better Decision Tree*”, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 09, 2827-2830,2010
- [8] Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal ,”*The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree*” ,International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012.
- [9].] Rahul A. Patil, Prashant G. Ahire, Pramod. D. Patil, Avinash L. Golande ,”*A Modified Approach to Construct Decision Tree in Data Mining Classification*” , International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 1, July 2012
- [10].Brijesh Kumar Baradwaj, Saurabh Pal , “*Mining Educational Data to Analyze Students Performance*”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011
- [11] L.Sathish Kumar, Mrs.A.Padmapriya ,”*ID3 Algorithm Performance of Diagnosis For Common Disease*”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5, May 2012 .
- [12] Jun-Hui Liu, “*Optimized ID3 algorithm based on attribute importance and convex function*”, IT in Medicine and Education (ITME), 2011 International Symposium on (Volume:2)
- [13] Kilian Q.Weinberger, Lawrence K. Saul, “*Distance Metric Learning for Large Margin Nearest Neighbor Classification*”, Journal of Machine Learning Research 10 (2009) 207-244 Submitted 12/07; Revised 9/08; Published 2/09
- [14] Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K. C. Wong, Fellow, IEEE, and Yang Wang, Member, IEEE ,”*Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data*”, Manuscript received Sep. 15, 2004; revised Dec. 1, 2004; accepted March 1, 2005. The work by W.-H. Au and K. C. C. Chan was supported in part by The Hong Kong Polytechnic University under Grants A-P209 and G-V958.