ISSN: 2250–3676

[IJESAT] INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY Volume-2, Issue-3, 581 – 586

A NOVEL TREE BASED CLASSIFICATION

M.R.Lad[1], R.G.Mehta[2], D.P.Rana[3]

[1]*M.Tech, Computer Dept., SVNIT, Surat, India,* ***p10co953@coed.svnit.ac.in***
[2]*Associate Professor, Computer Dept., SVNIT, Surat, India,* ***rgm@coed.svnit.ac.in***
[3]*Assistant Professor, Computer Dept., SVNIT, Surat, India,* ***dpr@coed.svnit.ac.in***

Abstract

Classification is a data mining (DM) technique used to predict or forecast the unknown information using the historical data. There are many classification techniques. ID3 is a very popular tree based classification algorithm for a categorical data which does not support continuous data. Attribute selection process plays major role in building a classification tree model. Attribute Selection in ID3 is based on the information gain. For a continuous attribute there is an extended version of ID3, called C4.5, which uses information gain ratio for discretization. Class-attribute interdependency Relation (CAIR) and Class-attribute interdependency maximization (CAIM) are different discretization schemes. CAIR gives better relationship between attribute and class then information gain and CAIM is a better discretization scheme then the one used in C4.5. In this paper A Novel Tree Based Classification Algorithm is proposed. It generates classification tree model in the similar manner as ID3. It uses CAIM based online discretization and attributes selection process. The work is extended with CAIR based attribute selection process. The results are tested on different datasets, which have shown the improved results.

Index Terms: *Classification, CAIR, CAIM, ID3, C4.5.*

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

1. INTRODUCTION

Classification is a data mining (DM) technique used to predict or forecast the unknown information using the historical data. Many classification algorithms have been developed such as Decision tree [1], Classification and regression tree [2], Bayesian classification [3], Neural networks [4] and K nearest neighbor classification [5] etc. Among them, the decision tree has become more popular algorithm as it has several advantages over others [6].

Decision tree based classification is a most widely used classification method in the field of data mining, where core point is to choose a splitting attribute. ID3 algorithm uses the information theory to choose the attribute. Whichever the attribute having the highest information gain is selected as a splitting attribute in each step while building a classification tree. Recursively it will generate a decision tree until some stopping criteria is reached. ID3 algorithm is widely used in various applications [7][8][9], but it contains a major limitations as it works only for nominal attributes, doesn't supports continuous attributes. Attribute selection process plays major role in building a classification tree model, ID3 uses information gain for the same.

Discretization is used to transform a continuous attribute into the finite number intervals so that attribute is treated as a categorical attribute. There are many discretization algorithms, proposed by the researchers [10][11][12][13]. But CAIM [14] has shown better result compare to other algorithms. C4.5 algorithm, which is an extended version of ID3, is developed to short out the limitation with the ID3. It discretizes the continuous attribute using Information gain ratio (or entropy). In this paper novel tree based classification algorithm is proposed which uses CAIM based online discretization and attribute selection process. The work is extended with CAIR [15] based attribute selection process. The results are tested on different datasets, which have shown the improved results.

In section 2 the literature review for tree based classification and discretization is covered. Tree based classification is discussed in section 3 and CAIM discretization is discussed in section 4. Section 5 shows the proposed work which is modified C4.5 with CAIR and CAIM with CAIM as a discretization scheme. Then in section 6, comparative analysis of the results of the all these methods are given.

2. RELATED WORK

Extensive research is done for classification algorithms and ID3 is the most popular algorithm for tree base classification. To improve ID3, researchers have proposed many methods, such as, use weighting instead of information gain [16], user's interestingness [17] and attribute similarity [18] to information gain as weight. Chun Guan and x. Zeng also have proposed improved ID3 based on weighted modified information Gain called ωID3 [19].

Ross Quinlan has developed tree based classification algorithm known as C4.5, an extension to the ID3 algorithm [20]. It uses the Information gain ratio during the tree building process. The C4.5 deals with continuous attributes which was not supported by ID3. It divides the values of a continuous attribute in a two subsets. He also proposed method of pruning, which deals with the removal of unwanted branches which are generated by noise or too small size of training data [21].

Discretization algorithms are mainly categorized as supervised and unsupervised algorithms. Popular unsupervised top-down algorithms are Equal Width, Equal Frequency [10] and standard deviation. While the supervised top-down algorithms are maximum entropy [11], Paterson-Niblett [13] which uses dynamic discretization, Information Entropy Maximization(IEM) [12] and class attribute interdependence Maximization (CAIM) [14]. Kurgan and Cios have shown the outperforming results of CAIM discretization algorithm compared to the other algorithms. As CAIM considers the highest interdependence between class and attribute it improves classification accuracy. Unlike other discretization algorithm CAIM automatically generate the intervals and interval boundaries for the given data without any user input.
In the following section, C4.5 a tree based classification is discussed.

3. TREE BASED CLASSIFICATION

Classification is a data mining (DM) technique used to predict or forecast the unknown information using the historical data. There are many tree based classification algorithms like ID3, C4.5 etc. But among them, C4.5 is the most popular algorithm which is an extended version of ID3. Figure 1 [22] shows the flowchart for tree based classification.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the information gain ratio. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. It calculates the information gain for all the attributes. Than the attribute with the highest information gain is chosen to make the decision. Then on the bases on that attribute, divide the given training set into a subsets. Then recursively apply the algorithm on each subset until the set contains instances of the same class. If the set contains instances of the same class, then return that class.

Model tree cannot be created for continuous valued attributes as it will create over fitting problem in the tree. Continuous attributes are discretizes in different intervals, known as a process of discretization. Next section explores discretization techniques.

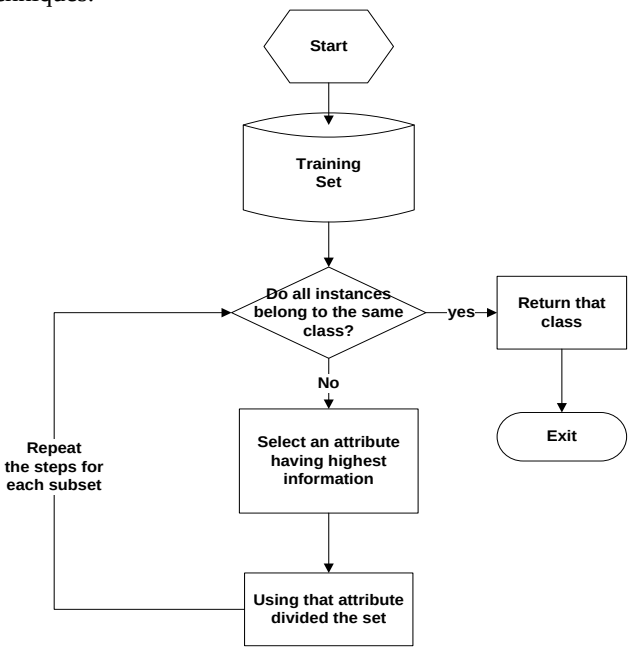

Fig-1: General Tree Based Classification

4. DISCRETIZATION

Classification algorithm like ID3 does not support continuous attribute. So, Discretization used to transform a continuous attribute's values into a finite number of intervals. C4.5 uses information gain or gain ratio to select the range of continuous attribute. CAIM discretization algorithm, proposed by L.A. Kurgan and K.J. Cios [14], automatically selects a number of discrete intervals and, at the same time, finds the width of every interval based on the interdependency between class and attribute values.

The CAIM algorithm discretizes an attribute into a smallest number of intervals and maximizes the class-attribute interdependency. They compared this algorithm with six well-known discretization algorithm and shown the outperforming results of CAIM over other discretization techniques.
Let's have a training data set of M examples, S classes, F indicates the continuous attribute. Discretization scheme D on F with n discrete intervals:

$$D: \{[d_0, d_1], [d_1, d_2], \ldots., [d_{n-1}, d_n]\}$$

Where d_0 is the minimum value and d_n is the maximum value of attribute F and the values in D are arranged in ascending order. The class variable and the discretization variable of attribute F are treated as two random variables defining a two-

dimensional frequency matrix (called quanta matrix) that is shown in Table 1 [14].

Class	Interval					Class Total
	$[d_0, d_1]$	.....	$[d_{r-1}, d_r]$	.....	$[d_{n-1}, d_n]$	
C_1	q_{11}	.....	q_{1r}	.....	q_{1n}	M_{1+}
:	:	.....	:	.....	:	:
C_i	q_{i1}	.....	q_{ir}	.....	q_{in}	M_{i+}
:	:	.....	:	.....	:	:
C_s	q_{s1}	.....	q_{sr}	.....	q_{sn}	M_{s+}
Interval Total	M_{+1}	.....	M_{+r}	.....	M_{+n}	M

Tab 1: Quanta Matrix for Attribute F and Discretization D

In Table 1, q_{ir} is the total number of continuous values belonging to the i^{th} class and interval $[d_{r-1}, d_r]$. M_{+r} is the total number of values within the interval $[d_{r-1}, d_r]$, where r=1,2,...,n and M_{i+} is the total number of values of the i^{th} class, where i=1,2,...,S.

Equation for CAIM is defined as:

$$CAIM(C, D|F) = \frac{\sum_{r=1}^{n} \frac{max_r^2}{M_{+r}}}{n} \quad (1)$$

Where max_r is the maximum value among all q_{ir}.
The algorithm starts with a single interval that covers all possible values of a continuous attribute and divides it iteratively. It chooses the highest value of the CAIM criterion.

The joint probability of the occurrence of values within the interval D_r and belongs to class C_i is given as:

$$p_{ir} = p(C_i, D_r|F) = \frac{q_{ir}}{M} \quad (2)$$

The class marginal probability for the values of attribute F belongs to class C_i, p_{i+}, and the interval marginal probability for the values are within the interval D_r, p_{+r} are as follows:

$$p_{i+} = p(C_i) = \frac{M_{i+}}{M} \quad (3)$$

$$p_{+r} = p(D_r|F) = \frac{M_{+r}}{M} \quad (4)$$

The Class-Attribute Mutual Information (CAMI) [15] between the class variable C and the feature interval boundaries of F_j, from the associated quanta matrix Q_j using equation (3) and (4), is defined as:

$$I(C, D|F) = \sum_{i=1}^{S} \sum_{r=1}^{n} p_{ir} \log_2 \frac{p_{ir}}{p_{i+} p_{+r}} \quad (5)$$

The Shannon's entropy of the quanta matrix measures the randomness of the distribution of data points with respect to class variable, and interval variable, and is defined as:

$$H(C, D|F) = \sum_{i=1}^{S} \sum_{r=1}^{n} p_{ir} \log_2 \frac{1}{p_{ir}} \quad (6)$$

The Class-Attribute Interdependence Redundancy (CAIR) was introduced by Wong and Liu (1975). It is the CAMI normalized by entropy H. Using equation (5) and (6) CAIR measure is defined as [15]:

$$CAIR(C, D|F) = \frac{I(C, D|F)}{H(C, D|F)} \quad (7)$$

CAIR can also generate a better discretization scheme, but due to its high-computational cost, it is inapplicable for discretization of continuous attribute having very large number of unique values. While CAIM perform the discretization task at reasonable computation cost so that it can be applied to the continuous attribute with large number of unique values[14].

With some modification in ID3 and using CAIM as a discretization scheme, this paper proposes a new technique for classification which is given in following section.

5. PROPOSED CLASSIFICATION ALGORITHM

The purpose of this paper is to present a modified ID3, tree based classification algorithm, which uses CAIR/CAIM criterion for an attribute selection instead of information gain in ID3 and information gain ratio in C4.5 algorithm. Main idea is to reduce the computational complexity and improve the classification accuracy.

Information gain or gain ration is calculated by considering all the class values while CAIM takes only the maximum count, so it reduces the computational complexity and gives better result than information gain or gain ratio. So in the proposed algorithm CAIM is used as online discretization during tree based classification model preparation.

CAMI gives better relationship between class and attribute than information gain which is proven by Kendall Giles, Kweku-Muata Bryson and Qin Weng[23]. So, for the betterment in the classification accuracy, CAIR is used for the attribute selection stage on model preparation. Following algorithm explains the process:

Algorithm:

Step 1. Select an attribute except that attribute whose value has to be predicted.

Step 2. If attribute is continuous, discretize it using CAIM.

Step 3. Calculate CAIR/CAIM value for that attribute. (Equation (1) and (7))

Step 4. Repeat steps 1 and 2 for each attribute.

Step 5. Then select an attribute for which CAIR/CAIM is maximum

Step 6. Make node containing that attribute.

Step 7. Then on the basis of that attribute, divide the given training set in to subsets.

Step 8. Then recursively apply the algorithm on each subset until the set contains instances of the same class. If the set contains instances of the same class, then return that class.

In this proposed algorithm, if attribute is continuous than it discretizes it using CAIM. Then calculate the values of CAIR/CAIM for each attribute. Then it selects the attribute having highest value of CAIR/CAIM. That attribute will be the value of the current node of the tree. Then using the different values of that attribute, divide the given training set into subsets. Then recursively apply this algorithm on each subset. If the set contains instances of the same class, then return that class.

It is clear that the number of calculation performed in a CAIR/CAIM is less compared to the gain ratio calculation in C4.5 algorithm. So the execution time of proposed method will be less compared to the existing C4.5 algorithm and the output of the decision tree will be more or less similar to the one generated by C4.5. But the accuracy is always more in CAIR. Results of the proposed method compare to the existing, is discussed in the next section.

6. RESULTS

In this section, the results of the proposed algorithm with CAIR and CAIM as a attribute selection method on some datasets with some continuous attributes are presented. For the comparative analysis, there is a result for an existing algorithm.

6.1 Experimental Setup

To perform the proposed method, we have used java in MyEclipse. Datasets are obtained from the UCI repository [24]. Pre-processing task like missing value replacement with mean value is done using Weka tool, before applying to these algorithms. Existing ID3 algorithm is performed in Weka, to compare with proposed work. Table 2 shows the description of the data sets which are used in the experiments.

DataSet	No. of Attributes	No. of continuous Attributes	No. of training Records	No. of Test Records
Ozone	**73**	**72**	**1000**	**2534**
Adult	**15**	**3**	**5000**	**32561**
Car	**7**	**0**	**1000**	**1728**

Tab 2: Data Set Description

After preparing data sets using weka, training data set is applied to the implemented tool. It builds the classification tree using training data and then test data is applied to evaluate the model. Tool shows us a classification tree, number of instances classified, number of instances not classified and the accuracy of the proposed method with CAIR and CAIM.

6.2 Results Analysis

Table 3 shows the results for the proposed as well as existing classification C4.5 algorithm with three different datasets.

Data Set	Method	Accuracy(%)
Ozone	Info. Gain	93.37
	CAIR	94.27
	CAIM	94.23
Adult	Info. Gain	75.82
	CAIR	79.71
	CAIM	75.62
Car	Info. Gain	85.35
	CAIR	85.35
	CAIM	84.54

Table 3: Comparison of three Attribute Selection Methods

From the table 3 it is clearly seen that CAIR always gives better accuracy compare to information gain and CAIM. It is observed that CAIR gives almost similar result as information gain when it is applied to the purely categorical data. But it gives a different and better result with continuous data. It is clearly seen form the table 3 that it gives same accuracy with CAIR and information gain for the Car data because it has no

continuous attribute. CAIR improves the accuracy for adult data and ozone data which having continuous attributes. Without using tree pruning, the proposed algorithm shows the outperforming result that shows the efficiency of the algorithm. Still some works are remaining to be done which is discussed in next section.

7. CONCLUSION

After analyzing and comparing the proposed algorithm with the existing algorithm, CAIR gives the outperforming result. It improves the accuracy for the classification. CAIR gives similar results with the data sets having all categorical attributes. But it is more accurate then information gain or gain ratio when using dataset with continuous attributes.

8. FUTURE WORK

Tree pruning will be used to prune the less used branches or less densed branches of the tree. The research scope is wide open to implement the tree process with batter data structure for fast access of the data and fast pruning from the classification tree to reduce the complexity of the overall process. To increase the classification accuracy, unclassified records are to be considered. Incremental classification technique will help in modifying the model for the unclassified data.

REFERENCES

[1]. S. Cohen, L. Rokach, O. Maimon," Decision-tree instance-space decomposition with grouped gain-ratio," *Information Sciences*, pp. 3592–3612, 2007.

[2]. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.,"Classification and Regression Trees," Monterey, CA: *Wadsworth & Brooks/Cole Advanced Books & Software* , 1984.

[3]. David HeckerMann, "A Tutorial On Learning With Bayesian Networks," March 1995 (Revised November 1996)

[4]. Raul Rojas," Neural Networks - A Systematic Introduction," *Springer-Verlag*, 1996.

[5]. Cover, T., Hart, P.," Nearest neighbour pattern classification," *IEEE Trans. on Information Theory*, vol.13, no.1,pp. 21–7, 1967.

[6]. R. Rastogi, K. Shim, "a decision tree classifier that integrates building and pruning," *Proc. of the twentyforth Int'l Conf. on Very Large Databases*, pp. 404–415, 1998.

[7]. Isao Hayashi, "An Application Fuzzy ID3 to Wireless Lan Access Point Optimal Location Problem", *Proceeding of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou*, 18-25, August 2005.

[8]. Jianna Zhao and Zhipeng Chang, "Neuro-Fuzzy Decision Tree by Fuzzy ID3 Algorithm and Its Application to Anti-Dumping Early-Warning System", *Proceedings of the 2006 IEEE International Conference on Information Acquisition*, August 20 - 23, 2006

[9]. Yan Ke-wu, Zhu Jin-fu and Sun Qiang, "The Application of ID3 algorithm in Aviation Marketing" *Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services*, November 18-20, 2007

[10]. Chiu D., Wong A. & Cheung B., "Information Discovery through Hierarchical Maximum Entropy Discretization and Synthesis," *Piatesky-Shapiro G., Frowley W.J. (Eds.) Knowledge Discovery in Databases*, MIT Press, 1991

[11]. A.K.C. Wong and D.K.Y. Chiu, "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, pp. 796-805, 1987.

[12]. Fayyad, U.M., and Irani, K.B. ,"Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. of the Thirteenth Int'l Joint Conf. on Artificial Intelligence*, San Francisco, CA, Morgan Kaufmann, pp.1022-1027, 1993

[13]. A. Paterson and T.B. Niblett, ACLS Manual. *Edinburgh: Intelligent Terminals*, Ltd, 1987.

[14]. Kurgan, L., & Cios, K.J., "CAIM Discretization Algorithm," *IEEE Transactions of Knowledge and Data Engineering*, Vol.16, No.2, February 2004.

[15]. Y. Chaoqun, L. Jianping, and D. Enming, "A Discretization Algorithm Based on Clustering and CAIR Criterion," *Seventh International Conference on Natural Computation* pp. 1424-1429, 2011.

[16]. Chen Jin, Luo De-lin ,Mu Fen-xiang. "An Improved ID3 Decision Tree Algorithm," *Proceedings of 2009 4th International Conference on Computer Science & Education* , pp.127-130, 2009.

[17]. Miao Wang,Ruimin Chai. "Improved Classification Attribute Selection Scheme for Decision for Decision Tree," *Computer Engineering and Application*, pp. 127-129. , 2010.

[18]. Rong Wang. "New Decision Tree Algorithm," *Science Technology and Engineering*, Vol.9, No.5, 2009.

[19]. Chun Guan and Xiaoqin Zeng, "An Improved ID3 Based on Weighted Modified Information Gain", *Seventh International Conference on Computational Intelligence and Security*, 2011

[20]. Quinlan J R, "C4.5 program for machine learning," *San Marteo Morgan Kaufmann Publisher* -s,pp.21-30,1993.

[21]. J. R. Quinlan. "Improved use of continuous attributes in c4.5."*Journal of Artificial Intelligence Research*, 4:77-90, 1996.

[22]. D. Gupta, D. S. Kohli, and R. Jindal, "Taxonomy of tree based classification algorithm," *2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, pp. 33-40, Sep. 2011.

[23]. Kendall Giles, Kweku-Muata Bryson, Qin Weng, "Comparison of Two Families of Entropy-Based Classification Measures with and without Feature Selection", *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001

[24]. UCI Repository, *"archive.ics.uci.edu/ml/datasets.html"*

BIOGRAPHIES

Mehul R. Lad
M.Tech Scholar,
Computer Department
SVNIT,
Surat,
India.

Rupa G. Mehta
Associate Professor,
Computer Department,
SVNIT,
Surat,
India.

Dipti P. Rana
Assistant Professor,
Computer Department,
SVNIT,
Surat,
India.