# Enhance Efficiency of Classification by Improving Iterative Dichotomiser 3

Ms.Khyati B. Naik
PG- Student Computer Engineering
Parul Institute Of Engineering and Technology
Vadodara, India

Mr. Arpit M. Rana
Assistant Professor Computer Science & Engineering
Parul Institute Of Engineering and Technology
Vadodara, India

***Abstract*— Data Mining (the analysis step of the Knowledge Discovery in Databases process or KDD) is the process of discovering or extracting new patterns from large data sets involving methods from statistics and artificial intelligence. Classification and prediction are the techniques used to make out important data classes and predict probable trend .The Decision Tree is an important classification method in data mining classification. It is commonly used in marketing, surveillance, fraud detection, scientific discovery. Iterative Dichotomiser 3algorithm is the most widely used algorithm in the decision tree so far. Aiming at deficiency of Iterative Dichotomiser 3 algorism, a new improved classification algorism is proposed in this paper.**

***Keywords*— Iterative Dichotomiser 3, classification, Decision tree***

## I. Introduction

Data Mining is the search for useful information in large volumes of data. Data mining is the process of extracting or mining knowledge from large amount of data. Data mining is the process of automatic classification of data tuples obtained from a dataset . Data Mining includes techniques from multiple disciplines such as database and data warehouse technologies, statistics, machine learning, pattern recognition, neural networks and data visualization.[9] A number of Algorithms have been developed and implemented to extract information and knowledge patterns that may be constructive for decision support. Once these patterns are extracted they can be used for automatic classification of data tuples. In other words, Data mining is the efficient discovery of valuable, non-obvious information from a large collection of data. It extracts hidden analytical information from large databases. It is a powerful new technology with great potential to help in analysis of data and for decision making. Data mining functionalities are used to specify the kind of patterns to be found in general data mining tasks.

There are many data mining techniques, like classification, clustering etc. Under classification, the cases are placed in differing groups. The procedures behind this methodology create rules as per training and testing individual cases. A number of algorithms have been developed for classification based data mining. Some of them include decision tree, k-Nearest Neighbor, Bayesian and Neural-Net based classifiers.

At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure. Quinlan in 1979 put forward a well-known Iterative Dichotomiser 3 algorithm, which is the most widely used algorithm in decision tree. But that algorithm has a defect of tending to select attributes with many values. It has also problem of over classification which leads to have less accuracy. Aiming at the shortcomings of the Iterative Dichotomiser 3 algorithm, in this paper, a Relation Function is introduced to improve Iterative Dichotomiser 3 algorithm. It reduces time complexity and increases accuracy.

## II. Iterative Dichotomiser 3 Algorithm

The basic principle of Iterative Dichotomiser 3 algorithm is as follows:

Supposes D=D1×D2×…×Dn is n-dimensional finite vector space, the Dj is finite discrete symbol set, the element d=<v1,v2,..vn>, is called the example[4,5] and vjϵDj, j=1,2,...,n. Supposing PT(Positive Tuples) and NT(Negative Tuples) is the two example sets, it supposes the sizes of PT and NT respectively are p and q, Iterative Dichotomiser 3 algorithm based on the following two suppositions:

- The class probability that a correct decision tree classify to random example set is consistent with the probability of positive-tuples and negative-tuples in vector space D.
- The information entropy needed for a decision can make correct judgment to an example take A as the root is:

$$E(A) = \sum_{i}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i) \quad . \qquad (1)$$

In (1):

$$\mathrm{I}\,(\mathrm{p},\,\mathrm{n}) = \left( \frac{p}{p+n}\log_2\frac{p}{p+n} + \frac{n}{p+n}\log_2\frac{n}{p+n} \right) \quad (2)$$

- The information gained refers to the effective decrement of information entropy, information gain namely take A as the root:

$$\mathrm{Gain(A)} = \mathrm{I(\,p,\,n)} - \mathrm{E(A)} \quad (3)$$

Iterative Dichotomiser 3 algorithm chooses the attribute with maximum Gain (A) as the root node, which means the attributes with the minimum E (A).

## III. Optimized Iterative Dichotomiser 3 Using Taylor Series [3]

According to the basic theory and the improved underlying principle of Iterative Dichotomiser 3 algorithm, we may change the information gain formula, thus seek a new standard of choosing attribute.

From the (3),

$$\mathrm{Gain(A)} = \mathrm{I(\,p,\,n)} - \mathrm{E(A)}$$

The I (p, n) is a quota to each node, so selects the value E(A) of the A attribute as the standard between the nodes .now:

$$E(A) = \sum_{i}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

In this equality

$$\mathrm{I}\,(\mathrm{p},\,\mathrm{n}) = \left( \frac{p}{p+n}\log_2\frac{p}{p+n} + \frac{n}{p+n}\log_2\frac{n}{p+n} \right)$$

Put the E (A) in the above equation and obtain under formula:

$$E(A) = \sum_{i}^{v} \frac{1}{(p+n)\ln 2} \left(-p_i \ln\frac{p_i}{\mathrm{p_i} + n_i}, -n_i \ln\frac{n_i}{\mathrm{p_i} + n_i}\right)$$

Because (P+N)ln2 is a constant, we can suppose the function e(A) satisfy the following formula:

$$\mathrm{e\,(A)} = \sum_{i}^{v} \left(-p_i \ln\frac{p_i}{\mathrm{p_i} + n_i}, -n_i \ln\frac{n_i}{\mathrm{p_i} + n_i}\right) \quad (4)$$

From Taylor series we can say that ln(1+x) = x, we may simplify function e(A):

$$\ln\frac{n_i}{\mathrm{p_i} + n_i} = \ln\left(1 - \frac{p_i}{\mathrm{p_i} + n_i}\right) \approx -\frac{p_i}{\mathrm{p_i} + n_i}$$

Put above two formula in (4):

$$\mathrm{e\,(A)} = \sum_{i}^{v} \left(p_i \frac{n_i}{\mathrm{p_i} + n_i} + n_i \frac{p_i}{\mathrm{p_i} + n_i}\right) = \sum_{i}^{v} \frac{2p_i n_i}{p_i + n_i}$$

Supposed the values of each attribute are N, which is multiplied by function e(A) simplified, than obtain the improved formula:

$$\mathrm{e(A)} = \left(\sum_{i}^{v} \frac{2p_i n_i}{p_i + n_i}\right)\mathrm{N} \quad (4)$$

It's obvious that the operation time of the improved E (A) with addition, subtraction, multiplication, division, is shorter than E (A) with logarithmic

Using this formula, time complexity is reduced but accuracy problem remains same. It selects attributes with many values leading to the problem of over classification.

## IV. Proposed Work

Suppose A is an attribute of data set D', and C is the category attribute of D'. the association degree function between A and C can be expressed as follows:

$$AF\ (A) = \frac{\sum_{i=1}^{n} \left| x_{i1} - x_{i2} \right|}{n} \quad (5)$$

Where x ij ( j = 1, 2 represents two kinds of cases) indicates that attribute A of D' takes the ith value and category attribute C takes the sample number of the jth value, n is the number of values attribute A takes. Then, the normalization of association degree function value is followed.

Suppose that there are m attributes and each attribute relation degree function value are AF(1), AF(2),…AF(m), respectively. Thus, there is

$$V(k)=\frac{AF(k)}{AF(1)+AF(2)+\cdots+AF(m)} \qquad (6)$$

Which 0< k< m . Then, (4) can be modified as

$$\mathrm{e(A)}=(\sum_{i}^{v}\frac{2p_i n_i}{p_i+n_i})\mathrm{N}\ \ x\ V(k) \qquad (7)$$

e(A) can be used as a new standard for attribute selection to construct decision tree according to the procedures of Iterative Dichotomiser 3 algorithm. Namely, decision tree can be constructed by selecting the attribute with the largest e (A) value as test attribute. By this way, the shortcomings of using Iterative Dichotomiser 3 can be overcome. It construct the decision tree, this tree structure will be able to effectively overcome the inherent drawbacks of Iterative Dichotomiser 3 Algorithm.

## V. Investigational Results

The classification accuracy of Iterative Dichotomiser 3 and our new implemented Iterative Dichotomiser 3(our proposed decision tree algorithm) were compared with the data samples. The below table shows the results

TABLE I. Evluation of Iterative Dichotomiser 3 and New Iterative Dichotomiser 3

| **No. Of Tuples** | **Precision (Percentage)** | | **Time (ms)** | |
|---|---|---|---|---|
| | ***Iterative Dichoto-miser 3*** | ***New Iterative Dichoto-miser 3*** | ***Iterative Dichoto-miser 3*** | ***New Iterative Dichoto-miser*** |
| | | | | |
| 520 | 69.7 | 72.1 | 152 | 120 |
| 730 | 72.5 | 76 | 250 | 200 |
| 900 | 80.4 | 83.2 | 350 | 260 |
| 1100 | 85 | 88 | 419 | 309 |

The end result of experimentation shows that the consequence of New Iterative Dichotomiser 3 is better than that of Iterative Dichotomiser 3 in the four aspects. With more quantity, more attributes, and the advantages is more obvious. As growing data volumes, the time is linear increasing of improved Iterative Dichotomiser 3, the little upward trend but steady corresponds to principle of Taylor, thus proved usability of the New Iterative Dichotomiser 3 algorithm.

## VI. Conclusion

The significant task of classification process is to classify new and unseen sample correctly. In this paper we’ve tried to overcome the deficiency of algorithm. By the changes done in algorithm, the classification accuracy is improved and time complexity is reduced. In experimental work we’ve shown this. This makes the algorithm more effective. This will reduce the problem of over classification.

## VII. Future Work

As the further capacity, the predictive accuracy of algorithm may still be improved by investigating other kinds of methods. An algorithm based on the input parameter combination can also be investigated for better results. It cal also be apply to various machine learning applications.

## References

[1] Aman Kumar Sharma, Suruchi Sahni,”A Comparative Study of Classification Algorithms for Spam Email Data Analysis”, International Journal on Computer Science and Engineering,, Vol. 3 No. 5 May 2011, pp. 1890-1896

[2] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, “Decision trees: an overview and their use in medicine”, Journal of Medical Systems, Kluwer Academic/Plenum Press,Vol. 26, Num. 5, pp. 445-463, October 2002

[3] Huang Ming, NiuWenying, Liang Xu,”An improved decision tree classification algorithm based on ID3 and the application in score analysis”, Chinese Control and Decision Conference,pp 1876-1880, 2009

[4] Linna Li,Xuemin Zhang“Study of Data Mining Algorithm Based on Decision Tree” International Confetence on Computer Design And Applications, Vol. 1, pp 155-159,2010

[5] Guanggun Zhai, Chunyan Liu “Research and Improvement on ID3 Algorithm in Intrusion Detection System” ,Sixth International Conference on Natural Computation, pp 3217–3220,2010

[6] ChengWenwei.Data warehousing and data mining, Beijing,Tsinghua University Press:2006 Jiawei Han, Micheline Kamber.Data mining:concepts and techniques.Morgan Kaufmann.2006.58-61

[7] S. Anumitha, S. Diana, D. Suganya, S. Shanthi,” Improvisation of ID3 Algorithm Explored On Wisconsin Breast Cancer Dataset”, International Conference on Computing and Control Engineering, 12 & 13 April, 2012

[8] Mrs.P.Nancy, Dr.R.Geetha Ramani,” A Comparison on Performance of” Data Mining Algorithms in Classification of Social Network Data International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011, pp. 47- 54

[9] Jiawei Han, Michline Kamber,”data Mining –Concepts and Technique”,Second Edition,2006

Athors Profile:

**Khyati** has received B.E. in Computer Engineering from Gujarat University, India 2010 and is currently pursuing masters in Computer Engineering in PIET. Her area of interest is DATA MINING.

**Arpit Rana** has received M.E. in Computer Science Engineering from PIET, India 2012. The author is currently working as Assistant Professor Senior in PIET.

.