ISSN: 2321-7782 (Online)
Volume 2, Issue 5, May 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining

Brijain R Patel[1]
Department of Computer Engineering
G.E.C Modasa – Gujarat
India

Kaushik K Rana[2]
Prof
Department of Computer Engineering
G.E.C Modasa – Gujarat
India

Abstract: ***ID3 (Iterative Dichotomiser) algorithm is largely used algorithm for generation of decision tree [1] which is developed by Ross Quinlan [2]. ID3 algorithm generates Entropy which is use for calculation of information Gain. ID3 algorithm finds entropy using Shannon Entropy base calculation which has some limitation. This Paper introduces new methods for calculating entropy based on Renyi Entropy instead of Shannon Entropy. By applying new calculation method we build the most optimized tree. This decision tree helps to take the decision for better analysis of data.***

Keywords: Data Mining, Decision tree, ID3, Shannon entropy, Renyi entropy, Entropy change method.

I. INTRODUCTION

Data mining is the process of fetching information from large data set [1]. In recent years, data mining technology has been widely used in security, web, stocks, real estate, healthy care, education and other fields. It is useful to people to obtain valuable and needed information more easily and flexibly. Classification and prediction are the two techniques used to make out important data classes and predict probable trend. Decision tree is one of the most useful tools for people to do data mining. Compared with other classification ways, decision tree is simple, faster and more accurate [2, 3]. Besides the mode [4] generated from decision tree can be understood more easily. So ID3 is playing more and more important role in data mining field and has an irreplaceable status.

In this Paper we propose a new entropy calculation method. We use Renyi Entropy calculation method for finding Entropy of the attribute behalf of Shannon Entropy, using new generated entropy according we calculate Information gain and generate tree with highest information gain.

II. ID3 ALGORITHM

ID3 *(Iterative Dichotomiser)* is a simple decision tree algorithm developed by Ross Quinlan (1983) [2]. ID3 algorithm is create a decision tree of given data set, by using top-down greedy approch to check each attribute at every tree node. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. So, entropy of each attribute is calculated first and accordingly information Gain is calculated. Attribute which has maximum information gain set at a root node of the tree and accordingly it generates sub tree with another node and accordingly set as child node [5].

Assumed that the data set D have m multiple attributes and for m different attributes the Entropy calculation is given by:

$$Info(D) = -\sum_{i=1}^{m} (P_i)\log_2(p_i) \qquad (1)$$

In this formula, the value Pi is the probability of I number Attribute.

Information gain is defined as the difference between the original information requirement and the new requirement which is obtain by partitioned on particular A, That is:

$$Gain(A) = Info(D) - Info_A(D) \quad (2)$$

In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, is chosen as the splitting attribute at node N.

III. Applications Shortcoming and its important on ID3 Algorithm

A. *Application of ID3 Algorithm in E-Commerce*

Decision trees are use in visualization of probabilistic business models. It also use in customer relationship management and uses for credit scoring for credit card users [6]. Now a day ID3 algorithm is widely used in e-commerce. It is use for generated online catalog which is essence for the success of an e-commerce web site.

Through generation of the tree easily gets idea about customer's area of interest for the products. It helps to improve a selling and increases business growth. It also helps to get relation between different criteria like shipping charges, delivery time, and item weight.

B. *Shortcomings of ID3 Algorithm*

Shannon Entropy finds its use and application in many areas. Here, Shannon Entropy has been used in ID3 algorithm to calculate the Information Gain contained by data, which helps to make Decision Tree for E-commerce business.

However, the results obtained from Shannon Entropy are rather complex, have more numbers of node and leaf node and in appropriate Decision Rules. Thus it makes the decision making process time consuming.

Therefore, to handle these problems, a new algorithm has been proposed by modifying ID3 algorithm using Renyi Entropy instead of Shannon Entropy.

C. *Proposed ID3 Algorithm*

The measure of tree component is one of the most important problems of ID3. Such problems occur when we have to take decision for who will be the root of the tree. We use Renyi Entropy calculation method instead Shannon Entropy for select appropriate root node.

Let P = (p_1, p_2…, p_n) be a probability distribution, p denotes the probability and q is a constant inherent parameter. Then Renyi [7, 8] gave the entropy measure by formula shown under:

$$H_q(P) = \frac{1}{1-q} \ln \sum_{i=1}^{n} p_i^q \quad (3)$$

This formula calculates Entropy. To avoid deduced solution in decision tree making process, Renyi entropy based ID3 algorithm is proposed which gives good solution in reasonable time. Such algorithm can give short and fast decision for customer interest in product and future demand of product by customers.

Algorithm:

Step 1: [select data set, find number of attribute and total number of instance of dataset]

// Where n is total number of instance. And A is selected attribute

Step 2: [Calculate entropy of each attribute using renyi entropy calculation method]:

$$H_q(P) = \frac{1}{1-q} \ln \sum_{i=1}^{n} p_i^q$$

Where H(p) is calculated entropy and q is inherent parameter.

Step 3: [Calculate the information Gains of all attributes.]

Gain (A) = Info (D) – $Info_A(D)$

//where Gain (A) tells us how much would be gained by branching on A attribute

Step 4: [split tree which has maximum value of info gain]

// set child node according the max info gain values

Step 5: For each child of the root Node, apply algorithm recursively until reach node that has entropy=0 or reach leaf node.

Step 6: Display generated final Decision tree.

IV. Implementation and Analysis

The summarized data of customer dispatch information in a section period (one month) from an information system database of a 3PL, which including 19 items in this sample data set shown in a table I [12]. In this example all sample data is divided by Customer login into two classes, which are Sign Customer (S) and Not Signed Customer (N) respectively, and has four properties: freight fee, Item Price, Item Weight, and Delivery Time. On the one hand, the summarizing data is integrated data from different sections and different consignment nodes. The values of these four properties are: freight fee (<50, 50-100, >100); Item Price (<500, 500-5000, >5000); Item weight (<1 kg, 1 kg - 5 kg, >5 kg); Delivery Times (days) (<2, 2-6, >6). The meanings of these properties are: The Delivery Charge is paid by customer for the transport cost; the Item Price is Transportation Company brings the money of the goods from the receiver to dispatcher; the Item Weight is measured by kilogram; the Delivery time is the period of time to deliver product to customer at customers address.

We can calculate needed information by taking probability of customer type here S class have 8 items and N has 11 items. Therefore, needed information gain by taken sample by putting q = 0.25, 0.5, 0.75, 2, 3, 4 etc.

Sr no	freight _fee	Item Price (Rs)	Item weight (kg)	Delivery Time (days)	Customer Login
1	50-100	<500	1-5	<2	S
2	50-100	<500	>5	2-6	S
3	50-100	<500	1-5	>6	S
4	50-100	500-5000	1-5	>6	S
5	50-100	>5000	<1	2-6	N
6	>100	<500	>5	>6	S
7	>100	<500	>5	2-6	S
8	>100	<500	1-5	<2	N
9	<50	>5000	1-5	2-6	N
10	<50	<500	1-5	2-6	N
11	<50	500-5000	1-5	>6	N
12	<50	<500	<1	2-6	N
13	<50	500-5000	<1	<2	N
14	<50	<500	<1	<2	N
15	<50	>5000	<1	2-6	S

16	<50	<500	<1	<2	N
17	<50	<500	<1	2-6	S
18	<50	500-5000	<1	<2	N
19	<50	<500	<1	<2	N

Table I: Information of Customer Dispatch goods

Assuming q=1.5

Parameter	ID3	Improve ID3
Correctly classified instance	78.94%	84.21%
Incorrectly classified instance	21.05%	15.78%
Kappa statistic	0.5682	0.6816
Root Node	weight	Freight fee

Table II: Comparison

Above, table shows comparison of Original Id3 and generated new Id3 algorithm. Table shows that new algorithms number of correctly classified instance are more than original ID3 and number of incorrectly classified instance are less compare to ID3.It also Kappa static value of generated ID3 is higher than ID3 algorithm. Also selection of root node is Freight Fee attribute.

The generated tree is:

```
freight_fee = <50
| Delivery_time = <2: N
| Delivery_time = 2-6
| | weight = <100
| | | payment = <500: S
| | | payment = 500-5000: null
| | | payment = >5000: S
| | weight = 100-700: N
| | weight = >700: null
| Delivery_time = >6: N
freight_fee = 50-100
| payment = <500: S
| payment = 500-5000: S
| payment = >5000: N
freight_fee = >100
| weight = <100: null
| weight = 100-700: N
| weight = >700: S
```

Fig I: Decision Tree

Figure I show a generated Decision tree from new ID3 algorithm, in which base root node is Freight fee. Below table shows comparison of Different dataset in which Data type of data set, attribute type, Number of attribute, Number of instance, classified by ID3 algorithm and Classified by new generated ID3 algorithm results are shown.

V. Conclusion

According to our observations, the performances of the algorithms are strongly depends on the entropy, information gain and the features of the data sets. There are various work has been done using the Decision tree Algorithm, but they all are like Static in Nature. In this thesis Improve ID3 algorithm is used, Instead of using Shannon Entropy, Renyi Entropy has been used to find the information of different properties which is used as the node of decision tree. This modification help find detail information of data, which will help to understand customer characteristics. We have also tried number of Data set and shows that improve algorithm help to generate efficient decision tree which is help to gain large number of information for customer or user of the Dataset.

No.	Name of Data base	Data Type	Attribute Type	Number of attribute	Number of instance	ID3 Algorithm Correct classified	Generated ID3 Algorithm
1	Tic - Tack – Toe Game	Multivariate	Classification	9	958	82.29%	83.61%
2	Congressional Voting Records Data Set	Multivariate	Classification	16	435	81.37%	82.45%
3	Lenses	Multivariate	Classification	4	24	72.12%	70.83%
4	BLOGGER Data	Multivariate	Classification	6	100	83.25%	81.79%

VI. Scope and Future Work

In this work, the value of alpha (α) has been taking as 1.5 fix value, but for further research we can take varying value of alpha (α) which may give different trees. Also, Instead of using Renyi Entropy, Different Entropy can also be used for further research like Arimoto, Taneja, Sharma-Taneja, Ferreri, Havrda and Charvat, Sharma-Mittal, Sant'anna—Taneja, Picard, Aczel-Daróczy.

References

1. Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining: concepts and techniques" Morgan kaufmann, 2006.
2. Quinlan J. R. (1986). "Induction of decision trees. Machine Learning," Vol.1-1, pp. 81-106.
3. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, Inc., 1993.
4. Ding Xiang-wu and Wang Bin, "An Improved Pre-pruning Algorithm Based on ID3," Jisuanji Yuxiandaihua, Vol. 9, pp. 47,2008.
5. Ming Fan, Xiaofeng Meng translated, "Data mining techniques and concepts", Machinary Industry Press, Beijing, pp. 136-145, Feb., 2004.
6. N R Srinivasa Raghavan, "Data mining in e-commerce: A survey," Sadhana Vol. 30, Parts 2 & 3, April/June 2005, pp. 275–289.
7. A Renyi. "On measure of Entropy and information", In Proc Forth Berkeley Symp, Math Stat. Prob., 1960, volume 1, page 547 , Berkeley ,196. University of California Press.
8. Bromiley P A N Thacker and E Bouhova-Thacker."Shannon entropy Renyi entropy and information" Statistic and Inf Series (2004-04).
9. Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
10. Fisher, Douglas H. "Knowledge acquisition via incremental conceptual clustering." Machine learning 2.2 (1987): 139-172.
11. Soleimanian Gharehchopogh, Farhad, and Seyyed Reza Khaze. "Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study."International journal of computer applications 47 (2012).
12. Q. Wang, Y. Wu, J. Xiao, and G. Pan, "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics. Automation and Logistics," presented at 2007 IEEE International Conference on 08 October 2007, Jinan.

