Fuzzy Classification to Classify the Income Category Based On Entropy

Srinivasan vaiyapuri
Department of MCA, Velalar College of Engineering and Technology, Erode, Tamil Nadu, INDIA- 638 012, Email: newsrini@rediffmail.com
Rajenderan Govind
School of Science & Humanities, Kongu Engineering College, Erode, Tamil Nadu, INDIA – 638 052. E-mail id: rajendranjv@gmail.com
Vandar Kuzhali Jagannathan
Department of MCA, Velalar College of Engineering and Technology, Erode, Tamil Nadu, INDIA- 638 012, E-mail id: vandarkuzhali@yahoo.com
Aruna Murugesan
Department of MCA, Velalar College of Engineering and Technology, Erode, Tamil Nadu, INDIA- 638 012, Email: newsrini@rediffmail.com

ABSTRACT

The classification problem is one of the main issues in data mining because it aims to extract a classifier which can be used to predict the classes of objects whose class table are unknown. This paper deals with classifying the income database with the entropy based method for analyzing the income is high or low. This method incorporates two mathematical techniques Entropy and Information Gain (IG) with Interactive Dichotomize 3 Algorithm (ID3). Subsets are calculated through Entropy. We fix the threshold point based on the fuzzy approach and the factors are identified using IG. The ID3 algorithm is used to derive a decision tree which classifies the income. This method also helps to extract logical rules that could be used in classifying high or low based on income with various attributed.

Keywords: Classification, Entropy, Information Gain, ID3, Decision Tree, fuzzy.

1. INTRODUCTION

Classification can loosely be defined as the process of identifying commonalities in a data set sufficient for discriminating among a finite number of groups. Classification accuracy is a key factor. Other important factors include compactness and expressiveness. Compactness refers to the length of the learned model, whereas expressiveness relates to the understandability of the knowledge represented by the learned model [1]. As one of the most fundamental data mining tasks, classification has been extensively studied

and various types of classification algorithms have been proposed.

It is common that any organization, state, or country may want to classify their salaried and business people into low or high categories. We take the original database and find out which attribute gives more information, taking these into consideration and find out which will play the highest and lowest role in classifying this problem. This can be done with the entropy to find out the various attribute which give more information with the data based on the information gain from the entropy [2, 3].

2. IMPLEMENTATION

2.1 Entropy

Entropy is a measure of variability in a random variable. It will measure how the particular attribute divides the training examples in to the number of result classes. Information which is more useful for classification is selected. For defining gain Entropy is obtained from Information Theory. Entropy is used to calculate the amount of information in an attribute. This is calculated as

$$\text{Entropy}(S) = -\Sigma P(x_i) \log_b P(x_i) \quad (1)$$

S- Collection of Samples.
x_i - Set of outcomes.
$P(x_i)$ - Proportion of S to the class x_i

This Entropy was developed by J. Ross Quinlan who has used Entropy in ID3 algorithm. This algorithm is based on the Concept Learning System (CLS) [4, 5]. Here Entropy is taken as 0 if all members of S belong to the same class. Entropy ranges from 0 to 1 which is a fuzzy value.

2.2 Information Gain (IG)

The information gain is based on the decrease in entropy after a dataset is split on an attribute. First the attribute that creates the most homogeneous branches are identified.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \Sigma((|S_v|/|S|) * \text{Entropy}(S_v)) \quad (2)$$

2.3 Proposed Method

From the proposed method one can identify with the gain value. Based on the gain value which scores more for the information gain can place the attribute to the root of the node and further attributes can be placed below and finally classify with high and low category. The sample data set includes 100 of records taken from the UCI repository [6], the original data is converted into high and low and the partial data is shown in Table I. This data is reduced based on the information gain, threshold point is set and based on the threshold point we take only the attributes with greater or equal to threshold point, other attribute are reduced to make effective calculation which are shown in the Table II. This is done for the calculation efficiency.

Table I – The Sample Data Set of Partial Record is Shown

S. N	AGE	WC	OCC	NC	EDU	H/P
1	L	H	L	H	L	H
2	H	H	H	H	L	L
3	L	L	L	H	L	H
4	H	L	L	H	L	L
5	L	H	H	L	L	H
6	L	H	H	H	H	H
7	H	L	L	L	L	L
8	H	H	H	H	L	H
9	L	H	H	H	H	H
10	H	H	H	H	L	L
11	H	H	H	H	L	H
12	L	H	H	L	L	H
13	L	L	L	H	L	L
14	L	L	L	H	H	H
15	H	L	L	L	L	L

The above table has six attributes, for each attributes the entropy and the IG is calculated. We fix the threshold point and consider the attribute which has greater than or equal to the threshold point [7]. We take into consideration that only these attribute gives more information when compared to the other attributes, so we reduce the other attributes which does not give much information for our fuzzy problem [8, 9, and 10]. The table II gives the attributes which has more information based on the information gain.

Table II – Effective Attribute after the Threshold Point is Fixed

S. No	AGE	WC	OCC	EDU
1	L	H	L	L
2	H	H	H	L
3	L	L	L	L
4	H	L	L	L
5	L	H	H	L
6	L	H	H	H
7	H	L	L	L
8	H	H	H	L
9	L	H	H	H
10	H	H	H	L
11	H	H	H	L
12	L	H	H	L
13	L	L	L	L
14	L	L	L	H
15	H	L	L	L

Initially Entropy is calculated for the total set. The data set may contain different attributes. Each attribute is treated as a subset. IG is calculated for each subset. Threshold value is set. Attributes which are above or equal to the threshold value are considered for decision tree generation, subsequently for the rule creation [11].

The purpose of including entropy is to define the IG more clearly. This Entropy measures the purity of sample datasets. IG is used to find the highest income attribute to recognize the high factor. By considering the highest factor as the root node, the Decision Tree is formed [12]. Decision trees can produces understandable rules and Classification can be done without much

computation. Decision trees provide a clear indication of which fields are most important for classification.

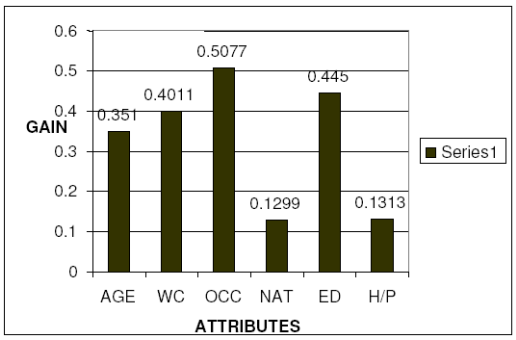
Steps involved to Calculate IG.

1. Calculate Entropy of the total dataset.
2. The entropy for each attribute is calculated.
3. The subset Entropy is subtracted from the total dataset Entropy.
4. The result is the IG; Find the Entropy for all the subsets.
5. The attribute that yields the largest IG is chosen as the root node.

3. RESULTS

Entropy and Information Gain Calculation: First, Entropy for the total data set is calculated. Entropy(S) = 0.9569. Second, IG for each attribute is calculated. Gain (AGE) = 0.351, Gain (WC) = 0.4011, Gain (OCC) = 0.5077, Gain (NAT) = 0.1299. Gain (ED) = 0.4450, Gain (H/P) = 0.1313. Here, the Threshold value is set as 0.3. This is shown in the Figure – I, Third, the IG values greater than the Threshold values are taken to create the Decision Tree.

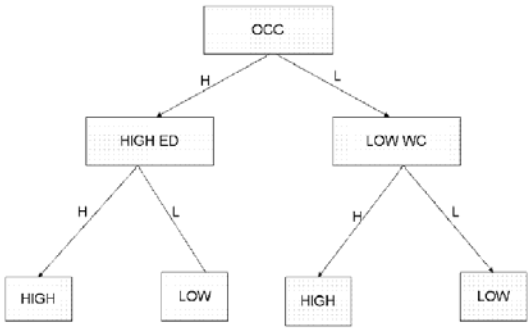
Figure I - Results of Gain Calculation



The figure II shows that the attribute OCC has the highest information when compared to all other attributes. Next highest information is WC so when these two attribute has High value than our output for this result will be high. If both of these value are low than our output for this result will be low.

Decision Tree: The attribute which is having the highest value is the root node for the decision tree. Attribute OCC is having the highest gain. So it is used as the root node. The attribute second highest gain is taken and inserted in the tree. This will be done until all data is classified completely [13] Decision Tree is shown below (Figure II).

Figure II – The Decision Tree



Rules from Decision Tree:

The decision tree can be expressed in the rule format.

IF OCC = HIGH AND ED = HIGH THEN
INCOME=HIGH
IF OCC = HIGH AND ED = LOW THEN
INCOME = LOW
IF OCC = LOW AND LOW WC = HIGH

THEN INCOME = HIGH
IF OCC =LOW AND LOW WC = LOW
THEN INCOME = LOW

4. CONCLUSION

This result is concerned with Entropy that identifies the attribute which plays highest role in classifying with more information. We present a Decision Tree model using ID3 Algorithm. First the sample data set is converted for calculation efficiency and the threshold point is fixed for fuzzy approach. Second this paper offers to use Entropy and Information Gain for identifying the high or low income category. Finally, ID3 is applied to derive a decision tree in order to mine the rules. Our result shows that the attribute OCC, Occupation and Low Education are the major factor that will drive to income classification. Through the Entropy, IG and ID3, it is possible to identify the attribute which plays highest role in classifying the income.

6. REFERENCES

[1]. J. and Kamber M., Data Mining Concepts and Techniques, Morgan Kaufmann Publishers,2000.

[2]. K. Alsabti, S. Ranka, and V. Singh, CLOUDS: A Decision Tree Classifier for Large Datasets, Proc. Fourth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '98), 1998.

[3]. X. Yin and J. Han, CPAR: Classification based on Predictive Association Rules, Proc. Third SIAM Int'l Conf. Data Mining, 2003.

[4]. J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publisher Inc., 1993.

[5]. J.C. Fodor, On Fuzzy Implication, Fuzzy sets and systems, vol. 42, pp. 293-300, 1991.

[6]. C.J. Mertz and P.M. Murphy, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/pub/machinelearining-databases, 2008

[7]. L. Breslow and D. Aha, Simplifying Decision Trees, Knowledge Eng. Rev., vol. 12, no. 1, pp. 1-40, 1997.

[8]. L.X. Wang. Adaptive Fuzzy Systems and Control. PTR Prentice Hall, 1994.

[9]. D. Dubois and H. Prade, Rough Fuzzy Sets and Fuzzy Rough Sets, Int. J. General Systems, vol. 17, nos. 2-3, pp. 191-209,1990

[10]. S. Greco, M. Inuiguchi, and R. Showinski, Fuzzy Rough Sets and Multiple-premise Gradual Decision Rules, Inr.J. Approximate Reasoning, vol. , no. , pp. 179-211, 2006.

[11]. Andrew Colin, Building Decision Trees with ID3 Algorithm. Dr. Dobbs Journal, June 1996.

[12]. Z. Pawlak, Rough Sets, Decision Algorithms and Bayes Theorem, European J. Operational Research, vol.136, pp.-2002

[13]. H. Wang and C. Zaniolo, CMP: A Fast Decision Tree Classifier Using Multivariate Predications, Proc. 17th Int'l Conf. Data Eng. (ICDE '01), 2001.

Received: Sept. 2011. Accepted: Sep. 2011.