2012 International Conference on Information and Computer Applications (ICICA 2012)
IPCSIT vol. 24 (2012) © (2012) IACSIT Press, Singapore

Classification Techniques of Datamining to Identify Class of the Text with Fuzzy Logic

Renuka D. Suryawanshi [1] and PROF. D. M. Thakore[2]

[1]M.Tech computer IInd Year Student (Bharti Vidyapeeth, Pune)
[2]Bharti Vidyapeeth

Abstract. Decision tree (DT) is a very practical and popular approach in the machine learning domain for solving classification problems in data mining . Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. In the past, ID3 was the most used algorithm in this area . This algorithm is introduced by Quinlan, using information theory to determine the most informative attribute. A disadvantage of decision tree is its instability. Decision tree is recognized as highly unstable classifier. The structure of the decision tree may be entirely different if some things change in the dataset. To overcome this problem, some scholars have suggested Fuzzy Decision Tree (e.g. FuzzyID3) by utilizing the fuzzy set theory to describe the connected degree of attribute values, which can precisely distinguish the deference of subordinate relations between different examples and every attribute values. After some years PFID3 was also introduced which was called as probabilistic fuzzy ID3. In this paper, a comparative study on ID3, FID3 and PFID3 is done.

Keywords: ID3 (Iterative Dichotomiser 3), FID3 (Fuzzy Iterative Dichotomiser 3), PFID3 Proba bilistic Fuzzy Iterative Dichotomiser 3), CLS (Concept Learning System), KDD (knowledge Discovery in Database)

1. Introduction

We often meet decision-making problems in our daily life or working environment.

Sometimes it is very difficult for us to make good decision. In practice, we usually use our past experiences to make a decision. We can see these past experiences as a form of performing experiments to come to a correct decision. However, executing experiments costs time and money. Fortunately, the developments of computer technologies and automatic learning techniques can make this easier and more efficient. In the domain of machine learning where it always lets computers decide or come up with suggestions for the right decision, there exist many approaches of decision making techniques, such as decision trees, artificial neural networks and Bayesian learning. This paper focuses on the decision tree approach with fuzzy logic to solve decision making problems.

2. What is decision tree?

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision [18].

Decision trees are commonly used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take actions.From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

3. What is decision learning algorithm?

Decision tree learning is a common method used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

For inductive learning, decision tree learning is attractive for 3 reasons [1]:

1. If the instances are described in terms of features that are correlated with the target concept Decision tree is a good generalization for unobserved instance.

2. The methods are efficient in computation that is proportional to the number of observed training instances.

3. The resulting decision tree provides a representation of the concept those appeals to human because it renders the classification process self-evident.

In data mining, trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.Data comes in records of the form:**(x,y)=(x1, x2, x3…xk, Y)…….1.2.1**

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **x** is composed of the input variables, x1, x2,x3 etc. that are used for that task.

4. Id3 (Iterative Dichotomiser 3)

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node [18].

ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The example has several attributes and belongs to a class (like yes or no). The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute.

ID3 uses information gain to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from an expert.

The sample data used by ID3 has certain requirements, which are:

1. Attribute-value description - the same attributes must describe each example and have a fixed number of values.

2. Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3.

3. Discrete classes - classes must be sharply delineated. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect.

4. Sufficient examples - since inductive generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences.

ID3 decide which attribute is the best? A statistical property, called **information gain**, is used for this purpose.

Gain measures how well a given attribute separates training examples into targeted classes. The one with proportion of positive examples (information being the most useful for classification) is selected.

In order to define gain, we first borrow an idea from information theory called entropy. **Entropy** measures the amount of information in an attribute.

For example, if S is (0.5+, 0.5-) then Entropy(S) is 1, if S is (0.67+, 0.33-) then Entropy(S) is 0.92, if P is (1+, 0 -) then Entropy(S) is 0. Note that the more uniform is the probability distribution, the greater is its information.

Given a collection S of c outcomes [18]

Entropy(S) = - P (positive) $\log_2$ P (positive) - P (negative) $\log_2$ P (negative) ……1.2.2

P (positive): proportion of positive examples in S

P (negative): proportion of negative examples in S

For example, if S is (0.5+, 0.5-) then Entropy(S) is 1, if S is (0.67+, 0.33-) then Entropy(S) is 0.92, if P is (1+, 0-) then Entropy(S) is 0. Note that the more uniform is the probability distribution, the greater is its information.

5. Fid3 (Fuzzy Id3)

Fuzzy decision tree is an extension of classical decision tree and an effective method to extract knowledge in uncertain classification problems. It applies the fuzzy set theory to represent the data set and combines tree growing and pruning to determine the structure of the tree [21].

1. Create a *Root* node that has a set of fuzzy data with membership value 1

2. If a node t with a fuzzy set of data *D* satisfies the following conditions, then it is a leaf node and assigned by the class name

The proportion of a class C*k* is greater than or equal to θr

$$\frac{|Dci|}{|D|} \geq \theta n$$

- The number of a data set is less than θn.
- There are no attributes for more classifications

3. If a node *D* does no satisfy the above conditions, then it is not a leaf-node. And a new subnode is generated as follows:

For Ai's (*i* =1…*L*) Calculate the information gain G (2.8), and select the test attribute A_{max} that maximizes them.

Divide *D* into fuzzy subset *D*1... *Dm* according to Amax , where the membership value of the data in D_j is the product of the membership value in *D* and the value of Fmax , *j* of the value of A max in *D* .

Generate new nodes t1 , …, t_m for fuzzy subsets *D* 1 , ... , D_m and label the fuzzy sets F max, *j* to edges that connect between the nodes t*j* and t

Replace *D* by *D j* (j=1, 2… m) and repeat from 2 recursively.

6. Algorithm Probabilistic Fuzzy Id3

1) Create a *Root* node that has a set of fuzzy data with membership value that fits the condition of *well-defined sample space* .

2) Execute the Fuzzy ID3 algorithm from step 2 to end.

7. Comparing The Algorithms Among Id3, Fid3 and Pfid3

Because FID3 and PFID3 are based on ID3, these three methodologies have similar algorithms. However, there also exist some differences.

a. Data representation: The data representation of ID3 is crisp while for FID3 and PFID3, they are fuzzy, with continuous attributes. Moreover, the membership functions of PFID3must satisfy the condition of *well-defined sample space* . The sum of all the membership values for all data value *i x* must be equal to 1.

b. **Termination criteria**: ID3: if all the samples in a node belong to one class or in other words, if the entropy equals to null, the tree is terminated. Sometimes, people stop learning when the proportion of a class at the node is greater than or equal to a predefined threshold. This is called pruning. The pruned ID3 tree tops early because the redundant branches have been pruned. FID3 & PID3: there are three criteria's.

1) If the proportion of the dataset of a class is greater than or equal to a threshold θr

2) If the number of a data set is less than another threshold θn

3) If there are no more attributes at the node to be classified If one of these three criteria's is fulfilled, the learning is terminated.

8. Conclusion and Future Research

Here summarization of the average performances of FID3 and PFID3 and the best one from ID3. In general, PFID3 performs the best, follows by ID3 and finally FID3. The best hit-rates are 94.7%, 94.9% and 95% under condition that θn =0.1, θn=0.2 and θn =0.4 respectively.

First, we compare the results of FID3 and ID3, the performance of PFID3 is always better.θn=0.1, ID3 is 0.025 worse than PFID3, while for θn=0.2, it is 0.027 and for θn =0.4, it is 0.009. following shows the percentage of how much ID3 is worse than PFID3. We conclude that applying the *well-defined sample space* to the fuzzy partition have a positive effect on the performance. We only compare PFID3 and FID3. PFID3 performs much better than FID3 under all conditions. The main difference between the learning PFID3 and FID3 is the *well-defined sample space* . The weight of the each data point of PFID3 is equal to one. Therefore, the data reacts on the learning with the same weight; each data has the same contribution to reasoning. On the contrary, the data point of FID3 can be overweight or underweight. Thus, the learning is inaccurate due to the imbalanced weight of the data. In other words, the origin of the better accuracy of PFID3 is the weight consistency of the data.We consider this phenomenon as the *effect of well-defined sample space* . But we need more evidences to support this viewpoint. In the further research, more experiments will be executed and evaluated. The leaf decision threshold is very important to the performance of the learning. In general, the performance increases along with the increasing of leaf decision threshold. This happens because when the leaf decision threshold increases, the learning has pruned the redundant branches. However, if the threshold increases too much, it causes *underfitting* . Finding the best leaf decision threshold will be done in the future research. The partitions of the fuzzy data do significantly have effect on the performance of the learning. In this thesis, we use the cluster centers to do the fuzzy partition. We find that the number of the clusters determines the number of the membership functions, which then affects the fuzzy partition. How to define the number of the clusters is not discussed in this thesis, it is left for the future research.

9. References

[1] An Implementa tion of ID3 Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, Schoo l o f Computer Science & Engineering,Sydney, NSW 2032, Australia weipengtig er@hotmail.com

[2]http://www.roselladb .com

[3] J. Ha n and M. Kamber. Data Mining: Concepts and Techniq ues. Morgan Kaufmann, 2000

[4] Top 10 algorithms in data mining Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007Published online: 4 December 2007 © Springer-Verlag London Limited 2007

[5]Ahmed S, Coenen F, Leng PH (2006) Tree-ba sed partitioning of date for association rule mining. KnowlInf Syst 10(3):315 –33 1

[6] Bloch DA, Olshen RA, Wa lker MG (2002) Risk estima tion fo r classificatio n trees. J Comput Graph Sta t11:263–288

[7] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. adsworth,Belmont

[8] Breiman L (19 68) Probability theory. Addison-Wesley, Reading . Republished (1991) in Classics of mathematics.SIAM, Philad elphia.

[9] O. Cord ó n, F. Gomide, F. Herrera, F. Hoffmann, a nd L. Magdalena. Ten years of genetic fuzzy systems:Current framework a nd new trends. Fuzzy Sets and Systems141: 5–31. 2001.

[10] W. W. Cohen. Fast effective rule inductio n. In Machine Lea rning: Proceedings ofthe 12th Interna tional Conference, Morga n Kaufmann, pp. 115–1 23. 199 5.

[11] R. B. Bhatt and M. Gopal. FRID: Fuzzyrough interactive dichotomizers. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'0 4). Piscataway, NJ: IEEE Press, p p. 1 337–1342. 2004

[12] 1.Knowledge Discovery using Text Mining: A Programmab le Implementation on Information Extraction and Categorization by Atika Mustafa, Ali Akbar, and Ahmer Sulta n Na tiona l University of Computer and Emerging Sciences-FAST, International Journal of Multimedia and Ub iquito us Engineering Vol. 4, No. 2

[13].Differentiating Data and Text-Mining Termino logy JAN H. KROEZE, MACHDEL C. MATTHEE AND THEO J.D. BOTHMAUniversity of Pretoria.

[14] Automated Learning of Decisio n Rules for Text Categorizatio n C. Apte, F. n Damerau, and S.M. Weiss ACM Transactions on Info rmation Systems, 1994

[15] Text Classification in Information Retrieval using Winnow P.P.T.M. van Mun Department of Computing Science, Catholic University of Nijmegen Toernooiveld 1, NL-6 525 ED, Nijmegen, The Netherla nds

[16] Von Altrock, Constantin (1995). Fuzzy logic and Neuro Fuzzy applications expla ined. Upper Sad dle River, NJ: Prentice Hall PTR. ISBN 0-13-368465-2 . Zimmerma nn, H. (2001). Fuzzy set theory and its applications. Boston: Kluwer Acad emic Pub lishers. ISBN 0-7923-743 5-5 .

[17] "Building Decision Trees with the ID3 lgorithm", by: Andrew Colin, Dr. Dobbs Journal, June 19 96

[18] HAN, J. AND KAMBER, M. 2001. Data mining: concepts and techniques. Morgan Kaufmann, San Francisco, CA.

[19] Von Altrock, Constantin (1995). Fuzzy logic and NeuroFuzzy applications exp lained. Upper Saddle River,NJ: Prentice Hall PTR. ISBN 0-13-368465-2 .

[20] Zimmermann, H. (2001). Fuzzy set theory and its applications. Boston: Kluwer Academic Publishers. ISBN 0-7923-7435-5.