The Study On A Decision Tree Based On The Classification Preference Ratio

Jing LIN
School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan,China
whpu_linjing@163.com

Keywords: decision tree; hierarchical granularity; classification preference ratio; condition attribute

Abstract. The decision tree is an important data mining method for classification. Granular computing has been applied to the decision tree, then a new decision tree based on the classification preference ratio of attribute is proposed. This decision tree is a hierarchical granularity tree. In information systems, each condition attribute divides the domain into several parts of a granular space. In the granular space, the classification preference ratio is used to describe the condition attribute. The classification preference ratio of every condition attribute is computed, and then the maximum attribute is chosen to divide the domain. According to different values of the attribute, the sample set is divided into several subsets. Each subset is a node or branch of the decision tree. If all the objects in a node are the same class, this node is a leaf node without further division. Otherwise, the node is not a consistent node. Above process of division will be repeated for all inconsistent nodes until all nodes become leaf nodes. Now the decision tree is finished. An example is given, which shows that the algorithm is feasible.

1.Introduction

In knowledge discovery, the concept of granularity is borrowed from physics. As the measurement of the data information and knowledge, the granularity is used to analyze and process information both macroscopically and microcosmically. As a new concept of information processing, granular computing covers theories, methods, technologies and tools of the granularity. We can describe relationships of knowledge via the size, thickness and classification of granularity. In this way, massive, rough, fuzzy and uncertain information will be processed. In recent years, the granular computing has become one of the most important branches in the artificial intelligence field. The application areas of granularity are wider and wider.

Data classification is one of the main tasks in data mining and knowledge discovery. According to a set of samples, we can create the model to classify and forecast unresolved problems. The design of algorithm is critical for the model. Currently the decision tree algorithm is popular, e.g. ID3 algorithm and C4.5 algorithm. Most of traditional decision tree algorithms are based on the information entropy. The strategy for selection of attributes is partly optimal. Each time an attribute is selected to decompose inconsistent nodes, only the current node may be considered. So different attributes will appear at the same level of the tree. And the same attribute may also appear at different levels. It reduces the effectiveness and efficiency of classification. On the basis of the granular computing theory, this paper presents a new decision tree algorithm. Different from those traditional algorithms, this new algorithm takes the classification preference ratio as the selection criteria of attribute. This decision tree is a hierarchical granularity tree. An instance proves the method, and analyzes its effectiveness and advantages.

2. Granularity and granular computing

In quotient space, the granularity can be considered as a class, a cluster or a subset of the domain. In the domain, objects with the same granularity are indistinguishable, equivalent or similar. Under the granularity, the domain is divided into several subsets.

Definition 1: S=<U,A>. S is an information system. U is a domain. The domain U is a collection of objects. A is a collection of attributes. $A=C\cup D$. C is a collection of condition attributes and D is a collection of decision attributes. R(R∈A) is an equivalence relation on the domain. A division U/R determined by R is a granular space about R in S, denoted by GS_R. The element in GS_R is called granularity about R in U, denoted by G.

Table 1 describes an information system. In the system, there are fourteen samples. The data set contains four condition attributes (Outlook, Temperature, Humidity, Windy) and a decision attribute (PlayTennis).

Table 1 A Sample Set

U	Outlook	Temperature	Humidity	Windy	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

As an equivalence relation, attribute Outlook determines a division in a domain U:

Outlook={Sunny，Overcast，Rain}

This division is a granular space of U. In the granular space, an element is an elementary particle, such as Sunny.

3. A decision tree method based on the classification preference ratio of attribute

3.1 Overview of decision tree algorithms

The decision tree is a classification algorithm in data mining. In the decision set, the samples are classified to create the decision tree. Then classification rules are concluded from the tree. The creation of a decision tree is a top-down process. Starting from the root node, for each non-leaf node, we test the sample set by an attribute. According to the results of the test, we will divide the sample set into several subsets. Each subset is a new node. Above process of division is repeated on the new nodes until all training samples can be classified correctly. In the complete decision tree, every leaf node is a class. The key of creating the decision tree is which attribute to select and how to divide the sample set.

At present, ID3 algorithm and C4.5 algorithm is relatively common. ID3 algorithm is based on the information entropy theory. The attribute with maximum information gain value is selected to divide the sample set. C4.5 algorithm improves ID3 algorithm via information gain ratio to select an attribute. These traditional algorithms are partly optimal. At the same level, each inconsistent node may select a different attribute. This will cause some attributes is repeatedly tested in a path of the decision tree. This is the disadvantage of traditional algorithms.

3.2 A new decision tree algorithm

This paper introduces the concept of granularity to the decision tree. The paper improves the selection criteria of the test attribute and presents a new decision tree. In information systems, every condition attribute divides the domain. In a granular space, the classification preference ratio is used to described attributes. In the equivalence relation about a condition attribute, if most objects belong to the same decision class, the attribute is more likely to divide a collection of objects into this class. Compared with the information entropy, the classification preference ratio of condition attribute can better reflect the classification effect. So in the new decision tree algorithm, the classification preference ratio is used as the selection criteria of the attribute. We calculate each classification preference ratio of the condition attribute and choose the maximum attribute to divide the sample set. For each value of this attribute, there is a sub sample set, which is a branch or a node of the tree. If the classes of objects in this node are different, the node needs to be divided further until all the objects in each subset are the same class. Finally, the decision tree is finished. Each time when selecting attributes, we have to consider all inconsistent nodes at the same level. Based on the hierarchical granularity, the decision tree is top-down and depth-first.

Definition 2 S=<U,A>. S is an information system. U is a domain. Domain U is a collection of objects. A is a collection of attributes. A=C∪D. C is a collection of condition attributes and D is a decision attribute. The condition attribute q(q∈C) divides domain U: q={x_1,x_2,…,x_m}(m>0). The classification preference ratio of q, denoted by CW_q , is:

$$CW_q = \frac{1}{m} \times \sum_{i=1}^{m} \frac{\max\{CU(d_j \cap x_i)\}}{CU(x_i)} \tag{1}$$

d_j is a class of a decision attribute.

CU(x_i) is the number of the samples, and in these samples the value of the condition attribute q is x_i.

CU($d_j \cap x_i$) is the number of the samples, and in these samples the value of the condition attribute q is x_i , and the value of the decision attribute d is d_j.

T is a set of samples that will be divided. SA is a set of condition attributes that has been chosen. UA is a set of condition attributes that has not been chosen. The algorithm steps are as follows:

(1) T=U, SA= Φ ,UA=C;

(2) The classification preference ratio of each condition attribute q(q∈UA) is calculated, and the attribute with the maximum classification preference ratio is chosen as the division attribute, denoted by q_{max}. If the classification preference ratios of several attributes are equal and maximum, the condition attribute that can divide the sample set into the most numerous subsets is chosen.

(3) According to different values of q_{max}, the set T is divided to subsets of T_1， T_2， …， T_j. Every subset is a branch or a node of the decision tree. For T_i(1⩽i⩽j), if the objects in this subset are the same class, this node is a leaf node and does not need to be divided. Otherwise, this node is an inconsistent node and needs to be divided.

(4) SA=SA∪q_{max}, UA =UA-q_{max} ,T=T-T_i;

(5) Repeat steps1-3 until all nodes are leaf nodes. The decision tree is completed.

From above, this decision tree is a hierarchical granularity tree. At each level of the tree, there is only an attribute. The values of the condition attribute form branches of the tree. Firstly, an optimal attribute becomes the node of the first level. For each value of the attribute, if the objects belong to the same class, the value of this attribute can completely classify data, and the objects of this attribute are removed from the sample set. Secondly, we choose the optimal in the remaining attributes for the rest of samples, and form the next level of the decision tree. The process is repeated until objects in each branch of the tree are the same class. At each level, the standard of selecting the optimal attribute is the classification preference ratio. The classification preference ratio describes the ability of attribute dividing domain objects. The larger the value of classification preference ratio is, the more likely the attribute intends to divide the domain objects into a category. This attribute is preferred as the division attribute.

3.3 Extraction of classification rules

Similar to the C4.5 algorithm, in order to make decision tree more readable, we can transform each path in the decision tree to a classification rule. The classification rule is the IF-THEN statement. The IF statement is the intersection of each condition attribute value in the path. The THEN statement is the final class of the leaf node. IF-THEN statement is denoted as follows:

If (q_i=x_m) and (q_j=x_n) and …… Then d_j

In the above formula, q is the condition attribute(q∈C) and d_j is a class of a decision attribute.

4. An example

An example is given to prove the decision tree algorithm. There is a decision system in table 2. U is a set of objects. C is a set of condition attributes. D is a decision attribute.

U={1.2，……，14}

C={Sex,Age,Income,Education}

D={Purchase}

Table 2 A Decision System

	Sex	Age	Income	Education	Purchase
1	m	young	i2	e2	Yes
2	m	young	i2	e1	Yes
3	f	young	i1	e2	No
4	m	young	i3	e2	No
5	f	young	i3	e3	Yes
6	f	old	i3	e2	No
7	m	young	i4	e3	Yes
8	f	young	i4	e2	Yes
9	m	old	i4	e2	No
10	f	old	i3	e1	No
11	m	young	i1	e1	No
12	m	young	i1	e2	Yes
13	f	old	i4	e1	No
14	m	young	i3	e3	No

The condition attribute Sex, Age, Income and Education can respectively divide the domain U.

Sex={m,f}

Age={young,old}

Income={i1,i2,i3,i4}

Education={e1,e2,e3}

The classification preference ratio of each condition attribute is calculated as follows:

$$CW_{sex}=\frac{1}{2}\times(\frac{4}{8}+\frac{4}{6})=0.5835$$

$$CW_{age}=\frac{1}{2}\times(\frac{6}{10}+\frac{4}{4})=0.8$$

$$CW_{income}=\frac{1}{4}\times(\frac{2}{3}+\frac{2}{2}+\frac{4}{5}+\frac{2}{4})=0.7418$$

$$CW_{education}=\frac{1}{3}\times(\frac{3}{4}+\frac{4}{7}+\frac{2}{3})=0.663$$

From above, the attribute with the maximum classification preference ratio is the Age. So the attribute Age is selected as the root of the decision tree to divide the sample set. The result is as follows:

t1={1,2,3,4,5,7,8,11,12,14}   (Age=young)
t2={6,9,10,13}   (Age=old)

For all subjects in subset t2, the value of the decision attribute Purchase is "no". So these subjects belongs to the same class. This branch is does not need to be divided. Next, we can remove the condition attribute Age and subject 6,9,10,13, recalculate the classification preference ratio of the attribute Sex,Income and Education.

$$CW_{sex}= \frac{1}{2}\times(\frac{4}{7}+\frac{2}{3}) = 0.619$$

$$CW_{income}= \frac{1}{4}\times(\frac{2}{3}+\frac{2}{2}+\frac{2}{3}+\frac{2}{2}) = 0.8325$$

$$CW_{education}= \frac{1}{3}\times(\frac{1}{2}+\frac{3}{5}+\frac{2}{3}) = 0.589$$

Similar to the first attribute selection, the attribute Income with the maximum classification preference ratio is selected to divide the rest of the sample set into subsets of {1,2}， {3,11,12}， {4,5,14}， {7,8}. Both subset {1,2} and subset {7,8} are leaf nodes. The other two subsets need to continue the division. Repeat this process, until all subsets are leaf nodes. The final decision tree is as follows:

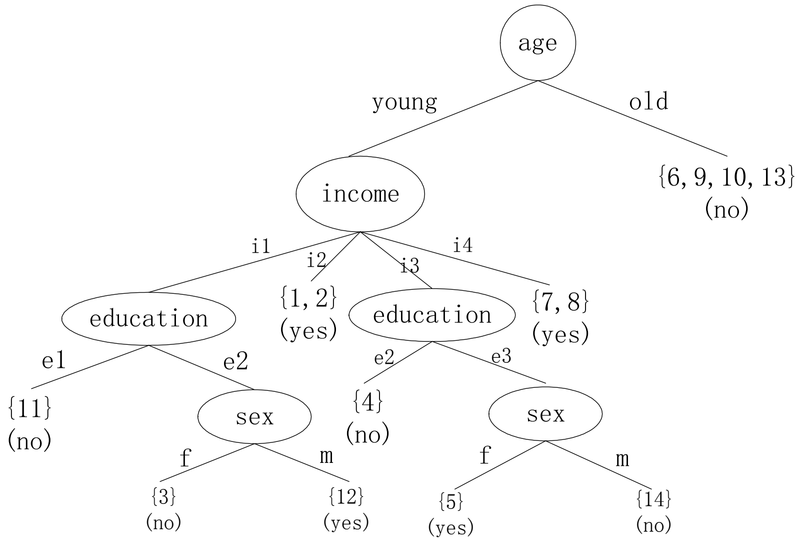

Fig. 1 Decision Tree

From above decision tree, the classification rules are as follows:

If (age=young) and (income=i1) and (education=e1) Then purchase=no
If (age=young) and (income=i1) and (education=e2) and (sex=f) Then purchase=no
If (age=young) and (income=i1) and (education=e2) and (sex=m) Then purchase=yes
If (age=young) and (income=i2) Then purchase=yes
If (age=young) and (income=i3) and (education=e2) Then purchase=no
If (age=young) and (income=i3) and (education=e3) and (sex=f) Then purchase=yes
If (age=young) and (income=i3) and (education=e3) and (sex=m) Then purchase=no
If (age=young) and (income=i4) Then purchase=yes
If (age=old) Then purchase=no

5. Algorithm analysis

Compared with traditional decision tree algorithms, the new decision tree has several advantages.

Firstly, in traditional decision tree algorithms such as ID3 or C4.5, the attribute selection strategy is the importance of attribute. In the algorithm of this paper, the classification preference ratio of attribute is calculated as the criteria of selecting attribute. Compared with the importance of attribute, the classification preference ratio describes the ability of attribute dividing domain objects. The domain is better classified with the classification preference ratio.

Secondly, in traditional decision tree algorithms, each time an attribute is selected to decompose inconsistent nodes, only one node may be considered. So different attributes will appear at the same level of the tree. And the same attribute may also appear at different levels. In the new decision tree algorithm, all inconsistent nodes can be considered when selecting an attribute. So only one attribute will appear at each level in the new decision tree. This greatly improves the classification results.

Because there is only one attribute at a level of the tree, the tree may become very big when the number of attributes in decision system increases. Similarly, many domain objects can lead to the decline of classification accuracy. And it takes more time to build the decision tree. This means that the effect of the algorithm needs to be improved when processing large amounts of data. The solution is that the decision system is divided into several sub systems. Or the attribute reduction is implemented. This will be the way of future research.

6. Conclusion

This paper introduces the granularity to the decision tree and presents a new decision tree algorithm based on the classification preference ratio. In a granular space, the classification preference ratio is the criteria to select the attribute. We choose the attribute with the maximum classification preference ratio to divide the sample set. Each value of the attribute is a branch. If the samples in a branch are different classes, the division will continue until all samples can be accurately classified.

The new algorithm improves the traditional decision tree algorithm. An example proves it. The algorithm is easy to program. It has strong feasibility. Some issues, such as the decision tree optimization, or the processing of large amounts of data, will be studied in the future.

References

[1] ZhouJun,LinQing, Quotient GD based algorithm for design decision tree，Computer Engineering and Design,2009,30(16)

[2]Zhai Junhai I,Wang Xizhao,Zhang Sufang, Information granularity，information entropy and decision tree, Computer Engineering and Applications,2009,45(12)

[3] ChenTing, LuoJingqing, Recognition Model of Radar Signal Based on Rough Set with Granulation and Decision Tree,Microelectronic and Computer, 2008,12

[4] GAOPingan, MengZuQiang, CaiZixin, Data Classification Modeling Based on Granular Computing, Application Research of Computers,2007.3

[5] DunYijie,ZhangXiaofeng,SunHao,ZhaoLi, A rules mining method based on granularity, Journal of Lanzhou University of Technology, 2006,2

[6] Xu Jiucheng,Shen Junyi,An Qiusheng,Li Naiqia, Study on Decision Subdivision Based on Information Granularity an d Rough Sets, Journal of Xi’an Jiaotong University,2005,4

[7] Yao J.T. and Yao Y.Y,Induction of classification roles by granular computing，Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing, LNAI 2475,331-338,2002