ELSEVIER

Expert Systems with Applications 28 (2005) 799–812

Expert Systems with Applications

www.elsevier.com/locate/eswa

MMDT: a multi-valued and multi-labeled decision tree classifier for data mining

Shihchieh Chou[a], Chang-Ling Hsu[b,*]

[a]Department of Information Management, National Central University, Taiwan, ROC
[b]Department of Computer Science and Information Management, Hungkuang University of Technology, 34 Chung-Chie Road, Shalu, Taichung County, 433 Taiwan, ROC

Abstract

We have proposed a decision tree classifier named *MMC* (multi-valued and multi-labeled classifier) before. MMC is known as its capability of classifying a large multi-valued and multi-labeled data. Aiming to improve the accuracy of MMC, this paper has developed another classifier named *MMDT* (multi-valued and multi-labeled decision tree). MMDT differs from MMC mainly in attribute selection. MMC attempts to split a node into child nodes whose records approach the same multiple labels. It basically measures the average similarity of labels of each child node to determine the goodness of each splitting attribute. MMDT, in contrast, uses another measuring strategy which considers not only the average similarity of labels of each child node but also the average appropriateness of labels of each child node. The new measuring strategy takes scoring approach to have a look-ahead measure of accuracy contribution of each attribute's splitting. The experimental results show that MMDT has improved the accuracy of MMC.

Keywords: Multi-valued attribute; Multiple labels; Classification; Decision tree; Data mining

1. Introduction

The purpose of the decision tree classifier is to classify instances based on values of ordinary attributes and class label attribute. Traditionally, the data set is single-valued and single-labeled. In this data set, each record has many single-valued attributes and a given single-labeled attribute (i.e. class label attribute), and the class labels that can have two or more than two types are exclusive to each other or one another. Prior art decision tree classifiers, such as *ID3* (Quinlan, 1979, 1986), *Distance-based method* (Mantaras, 1991), *IC* (Agrawal, Ghosh, Imielinski, Iyer, & Swami, 1992), *C4.5* (Quinlan, 1993), *Fuzzy ID3* (Umano et al., 1994), *CART* (Steinberg & Colla, 1995), *SLIQ* (Mehta, Agrawal, & Rissanen, 1996), *SPRINT* (Shafer, Agrawal, & Mehta, 1996), *Rainforest* (Gehrke, Ramakrishnan, & Ganti, 1998) and *PUBLIC* (Rastogi & Shim, 1998), all focus on this single-valued and single-labeled data set.

However, there is multi-valued and multi-labeled data in the real world as shown in Table 1. Multi-valued data means that a record can have multiple values for an ordinary attribute. Multi-labeled data means that a record can belong to multiple class labels, and the class labels are not exclusive to each other or one another. Readers might have difficulties to distinguish multi-labeled data from two-classed or multi-classed data mentioned in some related works. To clarify this confusion, we discuss the exclusiveness among classes, number of class and representation of the class label attribute in the related works as follows:

1. *Exclusiveness:* Each data can only belong to a single class. Classes are exclusive to one another. ID3, Distance-based Method, IC, C4.5, Fuzzy ID3, CART, SLIQ, SPRINT, Rainforest and PUBLIC are such examples.
2. *Number of class:* Data with classes classified into two types in the class label attribute is called two-classed data. ID3 and C4.5 are such examples. Data with classes

* Corresponding author. Tel.: +886 4 2631 8652x5400; fax: +886 4 2652 1921.
E-mail addresses: scchou@mgt.ncu.edu.tw (S. Chou), johnny@sunrise.hk.edu.tw (C.-L. Hsu).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2004.12.035

Table 1
A training data set for 15 products

Id	Maker	Price	Performance	Color	Class label
p1	A	$100	Not good	Yellow	A, B, C
p2	B	$880	Good	Yellow	B, C
p3	A	$370	Not good	Yellow, green	A
p4	C	$1230	Good	Blue	B
p5	B	$910	Good	Yellow, blue	B, C
p6	B	$770	Not good	Yellow	A, B, C
p7	B	$590	Not good	Yellow, green	A, B
p8	C	$1350	Good	Green	A, B, C
p9	C	$1250	Good	Yellow, green	A, B, C
p10	B	$1140	Good	Yellow, green	A
p11	A	$340	Not good	Yellow, blue	A, C
p12	C	$1300	Good	Yellow	A, B
p13	B	$1090	Good	Blue	C
p14	B	$810	Good	Green	A
p15	B	$520	Not good	Yellow, blue, green	C

classified into more than two types in the class label attribute is called multi-classed data. IC, CART and Fuzzy ID3 are such examples.

3. *Label representation:* Data with a single value for the class label attribute is called single-labeled data. ID3, Distance-based Method, IC, C4.5, Fuzzy ID3, CART, SLIQ, SPRINT, Rainforest and PUBLIC are such examples.

According to the discussion above, a multi-valued and multi-labeled data as we defined here can be regarded as a non-exclusive, multi-classed and multi-labeled data.

In our previous work (Chen, Hsu, & Chou, 2003), we have explained why the traditional classifiers are not capable of handling this multi-valued and multi-labeled data. To solve this multi-valued and multi-labeled classification problem, we have designed a decision tree classifier named *MMC* (Chen et al., 2003) before. MMC differs from the traditional ones in some major functions including growing a decision tree, assigning labels to represent a leaf and making a prediction for a new data. In the process of growing a tree, MMC proposes a new measure named *weighted similarity* for selecting multi-valued attribute to partition a node into child nodes to approach perfect grouping. To assign labels, MMC picks the ones with numbers large enough to represent a leaf. To make a prediction for a new data, MMC traverses the tree as usual, and as the traversing reaches several leaf nodes for the record with multi-valued attribute, MMC would union all the labels of the leaf nodes as the prediction result. Experimental results show that MMC can get an average predicting accuracy of 62.56%.

Having a decision classifier developed for the multi-valued and multi-labeled data, this research steps further to improve the classifier's accuracy. Considering the following over-fitting problems (Han & Kamber, 2001; Russell & Norving, 1995) of MMC, improvement on its predicting accuracy seems possible. First, MMC neglects to avoid the situation when the data set is too small. Therefore, it may choose some attributes irrelevant to the class labels. Second, MMC appears to prefer the attribute which splits into child nodes with larger similarity among multiple labels. Therefore, MMC exists inductive bias (Gordon & Desjardins, 1995).

Trying to minimize the over-fitting problems above, this paper proposes solutions as: (1) Set a constraint of size for the data set in each node to avoid the data set being too small. (2) Consider not only the average similarity of labels of each child node but also the average appropriateness of labels of each child node to decrease the bias problem of MMC.

Based on the propositions above, we have designed a new decision tree classifier to improve the accuracy of MMC. The decision tree classifier, named *MMDT* (multi-valued and multi-labeled decision tree), can construct a multi-valued and multi-labeled decision tree as Fig. 1 shows.

The rest of the paper is organized as follows. In Section 2, the symbols will be introduced first. In Section 3, the tree construction and data prediction algorithms are described. In Section 4, the experiments are presented. And, finally, Section 5 makes summaries and conclusions.

2. Notation

The symbols for the multi-valued and multi-labeled classification problem are formally stated as follows:

(a) Given that D is a training set, $|D|$ denotes the number of records.
(b) C denotes a set of class labels, $C=\{C_i|C_i$ is a class label, $i=1,\ldots,k\}$. The number of class labels in C is known in advance. $|C|$ denotes the number of class labels k.
(c) A denotes a set of attributes, $A=\{A_i|A_i$ is any ordinary attribute of D, $i=1,\ldots,n\}$. $|A|$ denotes the number of attributes n.