Expert Systems with Applications 41 (2014) 4625–4637

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data

Carlos J. Mantas, Joaquín Abellán *

Department of Computer Science & Artificial Intelligence, University of Granada, ETSI Informática, c/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain

ARTICLE INFO

Keywords:
Imprecise probabilities
Imprecise Dirichlet Model
Uncertainty measures
Credal decision trees
C4.5 algorithm
Noisy data

ABSTRACT

In the area of classification, C4.5 is a known algorithm widely used to design decision trees. In this algorithm, a pruning process is carried out to solve the problem of the over-fitting. A modification of C4.5, called *Credal-C4.5*, is presented in this paper. This new procedure uses a mathematical theory based on imprecise probabilities, and uncertainty measures. In this way, Credal-C4.5 estimates the probabilities of the features and the class variable by using imprecise probabilities. Besides it uses a new split criterion, called Imprecise Information Gain Ratio, applying uncertainty measures on convex sets of probability distributions (credal sets). In this manner, Credal-C4.5 builds trees for solving classification problems assuming that the training set is not fully reliable. We carried out several experimental studies comparing this new procedure with other ones and we obtain the following principal conclusion: in domains of class noise, Credal-C4.5 obtains smaller trees and better performance than classic C4.5.

1. Introduction and justification

A decision tree (DT) is a very useful tool for classification. Its structure is simple and easy to interpret. Moreover, to build the classification model normally requires a short time. When a DT is used for classification, a key question is the adjustment degree of the model to the training set. If the algorithm to build a DT employs a tight stopping criteria, then it tends to create small and underfitted DTs. On the other hand, if the algorithm uses a loose stopping criteria, then it tends to generate large DTs that over-fit the data of the training set. Pruning methods were developed for solving this dilemma. According to this methodology, a loosely stopping criterion is used, letting the DT to over-fit the training set. Then the over-fitted tree is cut back into a smaller tree by removing subbranches that are not contributing to the generalization accuracy (Rokach & Maimon, 2010). It has been shown in various studies that employing pruning methods can improve the general performance of a DT, especially in noisy domains.

The ID3 algorithm (Quinlan, 1986) and its extension C4.5 (Quinlan, 1993) are widely used for designing decision trees. C4.5 improves to ID3 algorithm with several characteristics: handling of continuous attributes, dealing training data with missing attribute values and a process for pruning a built tree.

There are different post-pruning processes for DTs (see Rokach & Maimon (2010) for a revision). They are based on estimating the generalization error and then removing useless sub-branches according this information. Usually, the basic idea of this estimation is that the ratio of error, calculated by using the training set, is not quite reliable. The training error is corrected in order to obtain a more realistic measure.

On the other hand, C4.5 algorithm uses a measure of information gain ratio for selecting an input variable in each node (split criterion). This variable selection process is based on the precise probabilities calculated from the training set. Therefore, C4.5 considers that the training set is reliable when the variable selection process is carried out, and it considers that the training set is not reliable when the pruning process is made. This situation can be unsuitable, specially when noisy data are classified. Let us see an example of this situation.

Example 1. Let us suppose a noisy data set composed by 15 instances, 9 instances of class A and 6 instances of class B. We consider that there are two binary feature variables X_1 and X_2. According with the values of these variables, the instances are organized in the following way:

$$X_1 = 0 \rightarrow (3 \text{ of class } A, 6 \text{ of class } B)$$
$$X_1 = 1 \rightarrow (6 \text{ of class } A, 0 \text{ of class } B)$$
$$X_2 = 0 \rightarrow (1 \text{ of class } A, 5 \text{ of class } B)$$
$$X_2 = 1 \rightarrow (8 \text{ of class } A, 1 \text{ of class } B)$$

* Corresponding author. Tel.: +34 958 242376.
E-mail addresses: cmantas@decsai.ugr.es (C.J. Mantas), jabellan@decsai.ugr.es (J. Abellán).

http://dx.doi.org/10.1016/j.eswa.2014.01.017
0957-4174/

If this data set appears in the node of a tree, then the C4.5 algorithm chooses the variable X_1 for splitting the node (see Fig. 1).

We can suppose that the data set is noisy because it has an outlier point when $X_2 = 1$ and class is B. In this way, the clean distribution is composed by 10 instances of class A and 5 instances of class B, that are organized as follows:

$X_1 = 0 \rightarrow (4 \text{ of class } A, 5 \text{ of class } B)$
$X_1 = 1 \rightarrow (6 \text{ of class } A, 0 \text{ of class } B)$
$X_2 = 0 \rightarrow (1 \text{ of class } A, 5 \text{ of class } B)$
$X_2 = 1 \rightarrow (9 \text{ of class } A, 0 \text{ of class } B)$

If this data set is found in the node of a tree, then the C4.5 algorithm chooses the variable X_2 for splitting the node (see Fig. 2).

We can observe that C4.5 algorithm generates an incorrect subtree when noisy data are processed, because it considers that the data set is reliable. Later, the pruning process considers that the data set is not reliable in order to solve this problem. However, the pruning process can only delete the generated incorrect subtree. It can not make a detailed adjustment of the correct subtree illustrated in Fig. 2. The ideal situation is to carry out the branching shown in Fig. 2 and then to make the pruning process. This situation is achieved by using decision trees based on imprecise probabilities as it will be shown later.

In the last years, several formal theories for manipulation of imprecise probabilities have been developed (Walley, 1996; Wang, 2010; Weichselberger, 2000). By using the theory of imprecise probabilities presented in Walley (1996), known as the Imprecise Dirichlet Model (IDM), Abellán and Moral (2003) have developed an algorithm for designing decision trees, called *credal decision trees* (CDTs). The variable selection process for this algorithm (split criterion) is based on imprecise probabilities and uncertainty measures on credal sets, i.e. closed and convex sets of probability distributions. In particular, the CDT algorithm extends the measure of information gain used by ID3. The split criterion is called the Imprecise Info-Gain (IIG).

Recently, in Mantas and Abellán (2014), credal decision trees are built by using an extension of the IIG criterion. In this work, the probability values of the class variable and features are estimated via imprecise probabilities. The CDT algorithm obtains good experimental results (Abellán & Moral, 2005; Abellán & Masegosa, 2009). Besides, its use with bagging ensemble (Abellán & Masegosa, 2009, 2012; Abellán & Mantas, 2014) and its above mentioned extension (Mantas & Abellán, 2014) are especially suitable when noisy data are classified. A complete and recent revision of machine learning methods to manipulate label noise can be found in Frenay and Verleysen (in press). Here, the credal decision tree procedure is included as a *label noise-robust method*.

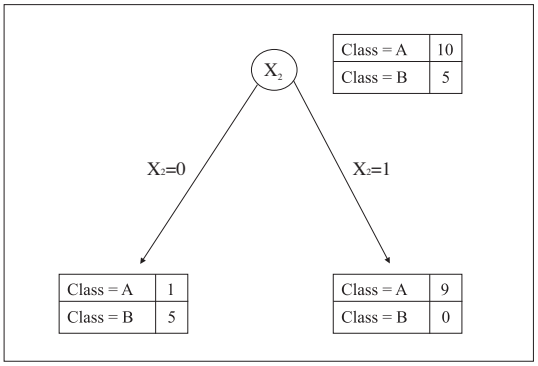

Fig. 2. Branching of a node with clean data produced by C4.5 algorithm.

According to the previous paragraphs, the CDT algorithm and its extensions consider that the training set is not reliable when the variable selection process is carried out. Hence, the problem shown in Example 1 is solved. If the noisy data set appears in the node of a credal tree, then the variable X_2 is chosen for splitting it (see Fig. 3).

Therefore, if we design a new credal tree algorithm inspired on C4.5 (with its improvements and advantages), then we can obtain an algorithm that considers the training set as unreliable when the processes of pruning and variable selection are made. This algorithm will be especially suitable for designing DTs in noisy domains.

Hence, C4.5 algorithm is redefined in this paper by using imprecise probabilities (Credal-C4.5). A new measure called Imprecise Information Gain Ratio (IIGR) is presented as split criterion. IIGR estimates the probability values of the class variable and features with imprecise probabilities as it is presented in Mantas and Abellán (2014). Besides, all the improvements of C4.5 are available: handling of continuous attributes, dealing of missing values, post-pruning process and so on. Credal-C4.5 and classic C4.5 are compared when they classify noisy data. It will be shown that Credal-C4.5 obtains smaller trees and better accuracy results than classic C4.5 with significant statistical difference.

Section 2 briefly describes the necessary previous knowledge about decision trees, C4.5 and credal decision trees. Section 3 presents Credal-C4.5 algorithm. Section 4 analyzes the differences between Credal-C4.5 and classic C4.5. Section 5 compares the action of Credal-C4.5 with the one performed by pessimistic pruning. In Section 6, we describe the experimentation carried out on a wide range of data sets and comments on the results. Finally, Section 7 is devoted to the conclusions and future works.

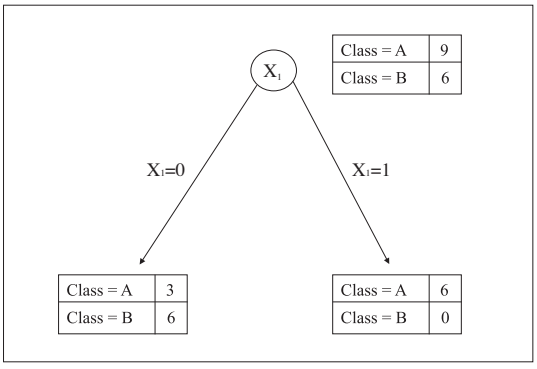

Fig. 1. Branching of a node with noisy data produced by C4.5 algorithm.

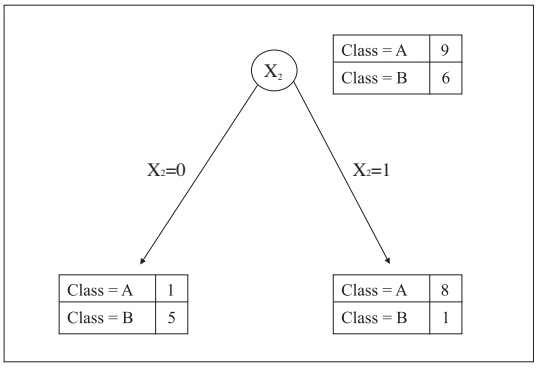

Fig. 3. Branching of a node with noisy data by a credal tree.