Expert Systems with Applications 41 (2014) 2514–2525

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Analysis and extension of decision trees based on imprecise probabilities: Application on noisy data

Carlos J. Mantas, Joaquín Abellán *

Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

ARTICLE INFO

Keywords:
Imprecise probabilities
Imprecise Dirichlet model
Uncertainty measures
Credal Decision Trees
Noisy data

ABSTRACT

An analysis of a procedure to build decision trees based on imprecise probabilities and uncertainty measures, called CDT, is presented. We compare this procedure with the classic ones based on the Shannon's entropy for precise probabilities. We found that the handling of the imprecision is a key part of obtaining improvements in the method's performance, as it has been showed for class noise problems in classification. We present a new procedure for building decision trees extending the imprecision in the CDT's procedure for processing all the input variables. We show, via an experimental study on data set with general noise (noise in all the input variables), that this new procedure builds smaller trees and gives better results than the original CDT and the classic decision trees.

1. Introduction

In the area of machine learning, supervised classification learning can be considered an important tool for decision support. Classification can be defined as a machine learning technique used to predict group membership for data instances. It can be applied to decision support in medicine, character recognition, astronomy, banking and other fields. A classifier may be represented using a Bayesian network, a neural network, a decision tree, etc.

A Decision Tree (DT) is a very useful tool for classification. Its structure is simple and easy to interpret. Moreover, the normal time required to build classification model based on a DT is low.

The ID3 algorithm (Quinlan, 1986) and its extension C4.5 (Quinlan, 1993) for designing decision trees are widely used. These algorithms use the *Divide and Conquer* technique. The splits are carried out in terms of the input variable values. The variable selection process in each node is based on the probabilities calculated from training examples. The probabilities are used via an uncertainty measure: the Shannon's entropy (Shannon, 1948).

In this manner, the variable selection process for a node is determined by the arrangement of the examples in the descendant nodes. However, this process does not depend on the training set size in each node. Because overspecialization relies on the number of examples in the training set, the trees designed by ID3 and C4.5 have this problem.

On the other hand, recently several formal theories for imprecise probabilities manipulation have been developed (Walley, 1996; Wang, 2010; Weichselberger, 2000). The use of imprecise probabilities instead of precise ones implies some advantages:

- The manipulation of total ignorance is coherently solved.
- Indeterminacy and inconsistency are adequately represented.

The different manipulation of two training sets with distinct sizes is possible thanks to the second advantage. The probabilities of the larger training set will be more reliable than those of the smaller set.

By using the theory of imprecise probabilities presented in Walley (1996), Abellán and Moral (2003) have developed an algorithm for designing decision trees. The variable selection process for this algorithm is determined from operations based on imprecise probabilities and uncertainty measures. This method obtains good experimental results, as shown in Abellán and Moral (2005), Abellán and Masegosa (2009), Abellán and Masegosa (2009).

An analysis of this kind of trees, called Credal Decision Trees (CDTs), is exposed in this paper. CDTs complement the ID3 and C4.5 algorithms because they use a total uncertainty measure that adds good properties to Shannon entropy (H). This new measure is the *Maximum of Entropy of a Probability Convex Set* (H^*), which has been previously defined and analyzed (Abellán & Moral, 2003; Abellán, Klir, & Moral, 2006; Abellán & Masegosa, 2008).

It has been demonstrated in Abellán et al. (2006) H^* is a disaggregated measure of information that combines two elements:

(a) A randomness measure that indicates the arrangement of the samples of each class in the training set. This measure corresponds to the entropy of the probabilities in the convex set.

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

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.050

(b) A non-specificity measure that shows the uncertainty derived from the training set size. This measure corresponds with the size of the convex set.

Two objectives are achieved by using the total uncertainty measure for the variable selection process. It serves to:

(1) Consider the arrangement of the samples in a node when the variable is selected. This fact is similar to the process followed by the ID3 and C4.5 algorithms.
(2) Take into account the training set size of a node and the available example numbers for each class. This property is different to the process carried out by the ID3 and C4.5 algorithms.

Differences between CDT methodology and the ID3 and C4.5 algorithms will be analyzed later on in this paper.

On the other hand, recent data mining literature (Khoshgoftaar & Van Hulse, 2009; Nettleton, Orriols-Puig, & Fornells, 2010; Van Hulse, Khoshgoftaar, & Huang, 2007; Zhu & Wu, 2004) is paying more attention to classification problems with attribute noise. As mentioned in Zhu and Wu (2004), class information in real-world data is usually much cleaner than attributes information. A medical data set can be considered as an example. In this case, when new data (for example, a patient's data) are entered, medical staff pay more attention to the diagnosis and to the accuracy of the prediction done.

However, research on handling attribute noise has not made much progress. Research on noise classification has focused on class noise. This is the case for CDTs, where only the probability distribution for the class variable is considered imprecise.

Let us see the following example in order to show the importance of considering attribute noise when a classification model is designed.

Example 1. Let us suppose a medical binary classification problem from patient's data (*type_A* versus *type_B*), and the following instance from a data set with possible incorrect values:

$$(Age = 23, \quad Weight = 100, \quad Class = type\_A)$$

It is possible that the error of this instance is produced by the measure on the class variable, *type_B* instead of *type_A* (class noise). However, other error source can be the measure on the input attributes. For instance, let us suppose that we know the exact classification rule, and it is:

$$\begin{array}{ll} \text{If } Age = 25 \text{ and } Weight = 100 \text{ then} & Class = type\_A \\ \qquad\qquad Otherwise & Class = type\_B \end{array}$$

In this case, the error of the instance can be derived from an imprecision on the input variable *Age* (value 23 instead of 25). Hence, it can be interesting to design classification models where attribute noise is taken into account.

With this motivation, an extension of the CDT's procedure is presented in this paper. This new classification model considers that both the probabilities for the class variable and the ones for the attributes (or features) are imprecise.

The new model, called Complete Credal Decision Trees (CCDTs), is coherently defined by using the principle of maximum uncertainty (see Klir (2006)), analyzed and experimentally compared with other models. The conclusions will be that CCDTs are more adequate for general noisy data and build smaller decision trees.

Section 2 briefly describes the necessary previous knowledge on decision trees and split criteria. Section 3 analyses the performance of the Credal Decision Tree (CDT). Section 4 presents an extension of the CDT procedure where the imprecision for manipulating all input variables is considered. In Section 5 we describe the experimentation carried out on a wide range of data sets and comments on the results. Here, we apply different percentages of noise on all the variables in the experiments. Finally, Section 6 explores the conclusions.

2. Previous knowledge

2.1. Decision Trees

Decision Trees (DTs), also known as Classification Trees or hierarchical classifiers, started to play an important role in machine learning with the publication of Quinlan's ID3 (Iterative Dichotomiser 3) (Quinlan, 1986). Subsequently, Quinlan also presented the C4.5 algorithm (Classifier 4.5) (Quinlan, 1993), which is an advanced version of ID3. Since then, C4.5 has been considered a standard model in supervised classification. It has also been widely applied as a data analysis tool to very different fields, such as astronomy, biology, medicine, etc.

Decision trees are models based on a recursive partition method, the aim of which is to divide the data set using a single variable at each level. This variable is selected with a given criterion. Ideally, they define a set of cases in which all the cases belong to the same class.

Their knowledge representation has a simple tree structure. It can be interpreted as a compact set of rules in which each tree node is labelled with an attribute variable that produces branches for each value. The leaf nodes are labelled with a class label.

The process for inferring a decision tree is mainly determined by the followings aspects:

(i) The criteria used to select the attribute to insert in a node and branching (split criteria).
(ii) The criteria to stop the tree from branching.
(iii) The method for assigning a class label or a probability distribution at the leaf nodes.
(iv) The post-pruning process used to simplify the tree structure.

Many different approaches for inferring decision trees, which depend upon the aforementioned factors, have been published. Quinlan's ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993) stand out among all of these.

Decision trees are built using a set of data referred to as the training data set. A different set, called the test data set, is used to check the model. When we obtain a new sample or instance of the test data set, we can make a decision or prediction on the state of the class variable by following the path in the tree from the root to a leaf node, using the sample values and tree structure.

2.1.1. Split criteria

Let us suppose a classification problem. Let C be the class variable, $\{X_1, \ldots ,X_n\}$ the set of features, and X a feature. We can find the following split criteria to build a DT.

2.1.1.1. Info-Gain. This metric was introduced by Quinlan as the basis for his ID3 model (Quinlan, 1986). The model has the following main features: it was defined to obtain decision trees with discrete variables, it does not work with missing values, a pruning process is not carried out and it is based on Shannon‘s entropy H (Shannon, 1948).

The split criterion of this model is *Info-Gain* (IG) which is defined as:

$$IG(C,X) = H(C) - \sum_i P(X = x_i)H(C|X = x_i).$$