Computers and Electrical Engineering 000 (2017) 1–14

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

An improved Id3 algorithm for medical data classification☆

Shuo Yang, Jing-Zhi Guo*, Jun-Wei Jin

Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macao

ARTICLE INFO

Article history:
Received 15 November 2016
Revised 8 August 2017
Accepted 8 August 2017
Available online xxx

Keywords:
Id3 algorithm
Decision tree
Balance function
Numeric attribute discretization
Rule representation of classifier model

ABSTRACT

Data mining techniques play an important role in clinical decision making, which provides physicians with accurate, reliable and quick predictions through building different models. This paper presents an improved classification approach for the prediction of diseases based on the classical Iterative Dichotomiser 3 (Id3) algorithm. The improved Id3 algorithm overcomes multi-value bias problem when selecting test/split attributes, solves the issue of numeric attribute discretization and stores the classifier model in the form of rules by using a heuristic strategy for easy understanding and memory savings. Experiment results show that the improved Id3 algorithm is superior to the current four classification algorithms (J48, Decision Stump, Random Tree and classical Id3) in terms of accuracy, stability and minor error rate.

1. Introduction

In the medical and health fields, researchers have been trying different data mining techniques in an attempt to improve the precision of medical diagnosis. Methods with better precision and reliability would provide more supportive data for the identification of prospective patients via accurate disease prediction. Until now, techniques such as support vector machines [1–3], neural networks [4–6], logistic regression [7,8] and clustering algorithms [9,10] have been applied in medical field. Experiment results show that these approaches give high accuracy in prediction [11,12]. These techniques support physicians to make accurate diagnostic decisions in clinical diagnosis. However, among them, the decision tree as a kind of classification method is still preferred in engineering applications, since tree models as classifiers can be easily applied. Moreover, their usages in bio-informatics have been reported in numerous research papers [2,5,7]. For example, based on historical patient data, decision tree classifiers are used to predict whether a patient has a kind of disease (e.g., heart diseases, lung cancers or breast cancers). The internal nodes of a decision tree are test attributes including properties such as symptoms and signs of patients. Based on the decision tree, an illness can be predicted by passing a new example down from the root to a leaf, testing the appropriate attribute at each internal node and following the branch corresponding to the attribute's value. If the accurate rate is high, then the patient can be suggested to receive proper treatments as early as possible.

Iterative Dichotomiser 3 (Id3) is an important classification algorithm which was proposed by Quinlan [13] in 1986 and has been applied in fields such as economics, medicine and science. In a situation where large data sets are used and the information for classification is complex, Id3 provides a useful solution. However, it has problems including multi-value bias, only handling nominal attributes as well as loosing relations among attributes in the classifier model. To resolve these problems, many approaches have been proposed, such as Gain ratio [13–15] and Gini index [16–18] for multi-value bias,

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. Sioutas.

* Corresponding author.
E-mail addresses: yb37416@umac.mo (S. Yang), jzguo@umac.mo (J.-Z. Guo), jinjunwei24@163.com (J.-W. Jin).

http://dx.doi.org/10.1016/j.compeleceng.2017.08.005
0045-7906/

partition number [19] for numerical attribute discretization. However, these approaches have both pros and cons. Gain ratio tends to prefer unbalanced splits in which one partition is much small than the other. Gini index is biased towards multi-value attributes. It also has difficulties when the number of classes is large and tends to favor tests that result in equal-sized partitions and purity in both partitions. For numeric attribute discretization, pre-assigning partition number is a common way [19]. However, its drawback is that it may be difficult to choose an optimal partition number for any numeric attribute or the threshold for an interval partition. More often, when the number of numeric attributes is large, it is time-consuming to set an optimal partition number for each of them.

This paper proposes an improved Id3 algorithm. The new algorithm uses a balance function to offset the increase of the information gain when an attribute has many different values. Besides, it is capable to handle numeric attributes by using a novel discretization method. The final classifier model is represented in the form of rules rather than a tree structure for the sake of easy interpretation and memory saving. Through the experiments with five data sets from UCI [20], the improved algorithm achieves better accuracy and reliability compared to other four decision-tree classification algorithms. Therefore, it can be used to improve the performance of disease prediction. The main contributions of this paper are:

- This paper proposes an improved Id3 algorithm for disease prediction with a novel balance function for test attribute selection, a novel numeric attribute discretization strategy as well as a new rule-based heuristic method for classifier representation.
- This paper mathematically proves that the classical Id3 algorithm is biased to multi-variated attributes and how the proposed balance function can address this issue.

The rest of the paper is organized as follows: Section 2 discusses related works. In Section 3, the fundamental theory of Id3 algorithm and its problems are discussed. Section 4 proposes improvement strategies. Section 5 shows experimental results based on five benchmark data sets. Finally, a conclusion is given with summary, contributions and future work.

2. Related work

Decision trees (DT) are becoming popular with the growth of data mining in the field of bioinformatics. DT-based classification algorithms have tree structures consisting of internal nodes, branches and leaves. The tree structure is equivalent to a set of decision rules. Each internal node in the tree represents a decision on a property (or an attribute), and the branch of each internal node represents the output of a decision result. The purpose of DT is to search for a set of decision rules to predict an outcome for a set of input instances. A leaf node of the tree represents a sample group or a sample classification. The number of nodes of the tree has an important influence on the accuracy of the classification and the size of the tree. Common decision tree construction methods include classification and regression trees [21,22], chi-square automatic interaction detector (CHAID [23]), Random Forest and Random Tree [24]. Compared to Boosted Decision Tree (BDT) methods and Decision Tree Forest (DTF) techniques, Id3 as a single decision tree classification algorithm provides a rapid and effective method to categorize data sets.

To resolve the issue of multi-variated attribute and select the most discriminatory attribute as splitting node, several methods have been proposed. Gain ratio [13–15] and Gini index [16–18] are two famous indices designed to address this problem. By computing these indices, the fitness of an attribute can be determined, and the attribute with the best fitness is chosen as the test attribute for the current node when building a decision tree. Specifically, Gain ratio normalizes information gain through dividing information gain by splitting information amount that represents the potential information generated by splitting the training data set D into v partitions. The strategy of Gini index chooses the attribute whose Gini Index is minimum after splitting. Besides, some researches [19,25] focus on multi-valued and multi-labeled data where the traditional decision tree algorithms have been proved to be not applicable. However, the issue of multi-value attribute is a special case of multi-variated attribute, since multiple values of a certain attribute in a sample can be seen as a special attribute value. Since multi-value attributes have different combination forms of values, this kind of attribute has high information gain and thus it is easily chosen to be test/split attribute.

To address the issue of numeric attribute discretization, some studies (e.g., C4.5 [15]) assign attribute A with two possible outcomes: $A \leq t$ and $A > t$, where t is called threshold. The threshold is calculated by means of a linear search in the whole training set. Thus, it needs to compute information gain many times with the change of t. Some researchers propose improved methods to update the procedure of numeric attribute discretization. For example, in [19], a partition number needs to be pre-assigned, and then the sorted data records are segmented proportionally to the partition number. The problem is that it may be difficult for users to set an optimal partition number for any numeric attribute and the threshold for an interval partition.

3. Classical Id3 algorithm and its problems

The core part of the Id3 algorithm is that it computes the information gain as a selection criteria of test attributes for hierarchical levels of non-leaf nodes in a decision tree. The aim of the algorithm is to acquire the largest class information about the sub-dataset when making a decision on an internal node. Specifically, its first step is to compute the information gain for each attribute and select the one that holds the largest information gain as the root node of the decision tree. Based on the different values of the attribute, the node generates different branches. Then, for the sub-dataset assigned to each

Please cite this article as: S. Yang et al., An improved Id3 algorithm for medical data classification, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.08.005