Knowledge-Based Systems 92 (2016) 35–42

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Classification with test costs and background knowledge

Tomasz Łukaszewski*, Szymon Wilk

Computer Science Department, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

ARTICLE INFO

Article history:
Received 30 April 2015
Revised 30 September 2015
Accepted 7 October 2015
Available online 22 October 2015

Keywords:
Test costs
Levels of abstraction
Naïve Bayes classifier

ABSTRACT

We propose a novel approach to the problem of the classification with test costs understood as costs of obtaining attribute values of classified examples. Many existing approaches construct classifiers in order to control the tradeoff between test costs and the prediction accuracy (or misclassification costs). The aim of the proposed method is to reduce test costs while maintaining of the prediction accuracy of a classifier. We assume that attribute values are represented at different levels of abstraction and model domain background knowledge. Our approach sequentially explores these levels during classification – in each iteration it selects and conducts a test that precises the representation of a classified example (i.e., acquires an attribute value), invokes a naïve Bayes classifier for this new representation and checks the classifier's outcome to decide whether this iterative process can be stopped. The selection of the test in each iteration takes into account the possible improvement of the prediction accuracy and the cost of this test. We show that the prediction accuracy obtained for classified examples represented precisely (i.e., when all the tests have been conducted and all specific attribute values have been acquired) can be achieved for a much smaller number of tests (i.e., when not all specific attribute values have been acquired). Moreover, we show that without levels of abstraction and with uniform test costs our method can be used for selecting features and it is competitive to popular feature selection schemes: filter and wrapper.

1. Introduction

One of the main tasks of machine learning is to build classifiers from available data. Constructed classifiers, after their evaluation, are applied in many real-world applications, in medical diagnosis, automated testing, robotics, industrial production processes and many other areas. The most commonly used evaluation criterion of a classifier is its predictive accuracy. The measure of the prediction accuracy of a classifier is often replaced by the measure of the misclassification costs of a classifier, because different errors may have different costs. On the other hand, more and more attention is paid to test costs, that is the cost of obtaining attribute values (features) of classified examples. The cost associated to a feature can be related to different concepts: expenses, risks or computational costs [1]. In order to decrease the total cost of these tests we may reduce their number allowing for missing attribute values in the representation of classified examples. However, missing values of relevant attributes in the representation of classified examples usually degrade the predictive accuracy of a classifier (or increases the misclassification costs of a classifier) [2,3]. Therefore, we have to decide which tests should be carried out in order to control the *tradeoff* between the cost of these tests and the accuracy of a classifier (or the tradeoff between test costs and misclassification costs of a classifier) [4,5]. However, in many real applications it is very difficult to evaluate misclassification costs. For example, in medical diagnosis, how much money you should assign for a misclassification cost, when a misclassification hurts a patient's life? In such cases, we should concentrate on the tradeoff between test costs and the accuracy of a classifier. The appropriate approach may be to reduce test costs while maintaining the prediction accuracy of a classifier. This goal may be achieved by cost-based feature selection methods [1].

Let us notice that standard feature selection methods were designed to handle plain data without any type of generalization of attribute values. However, there are areas where attribute values are represented at different levels of abstraction. These levels model domain background knowledge and have usually a form of a tree-like hierarchy. In such a tree the root represents a missing value, leaves represent specific attribute values, and the remaining nodes represent abstract attribute values (e.g., sets of specific values). Importantly, such hierarchies entail the existence of tests that replace a more abstract value by a less abstract value. Moreover, this replacement may be taken in several stages for a given attribute, going from the root of a hierarchy towards less abstract values. We assume that for some classified data a decision may be taken based on (less or more) abstract attribute values. Without levels of abstraction, precise attribute values had to be acquired in such a case. Assuming that the

* Corresponding author. Tel.: +48 616652920.
E-mail address: tlukaszewski@cs.put.poznan.pl (T. Łukaszewski).

http://dx.doi.org/10.1016/j.knosys.2015.10.008
0950-7051/

cost of obtaining an abstract value is less than the cost of obtaining a precise value that is a refinement of this abstract value, we see that introducing levels of abstraction should allow to further reduce test costs. Moreover, the exploration of these levels of abstraction in the context of a classified example should result in lower test costs than their exploitation during learning or even earlier, during the data preprocessing. Unfortunately, the approaches proposed so far that take into account levels of abstraction or more general ontologies are intended to obtain only models that are simpler and their classification accuracy is preserved or improved (test costs are not considered) (e.g., [6–9]).

In this paper we present a novel approach to the problem of a classification with test costs. Our approach sequentially explores levels of abstraction during classification – in each iteration it selects and conducts a test that precises the representation of a classified example (i.e., acquires an attribute value), invokes a naïve Bayes classifier for this new representation and checks the classifier's outcome to decide whether this iterative process can be stopped. The selection of the test in each iteration takes into account the possible improvement of the prediction accuracy and the cost of this test. We show that the prediction accuracy obtained for classified examples represented precisely (i.e., when all the tests have been conducted and all specific attribute values have been acquired) can be achieved for a much smaller number of tests (i.e., when not all specific attribute values have been acquired). Moreover, we show that without levels of abstraction and with uniform test costs our method can be used for selecting features and it is competitive to popular feature selection schemes: filter and wrapper.

The novelty of the paper is twofold. First, the stopping criterion of this sequential process explores the classifier outcome for the current and previous stages of the sequential process. Second, our approach allows for representing attribute values at different levels of abstraction in order to model domain background knowledge. These two elements allow to achieve the aim of our research.

The method presented in the paper is based on the results of our earlier works. In [3] we showed that missing values of attributes with small value of information gain does not reduce prediction accuracy. In [10] we showed that the prediction accuracy of the sequential classification process, that is applied in the paper converges very quickly to the prediction accuracy achieved for the examples represented precisely. The method presented in the paper adds the stopping criterion to this sequential classification and presents the experimental evaluation of the proposed approach.

The paper is organized as follows. Section 2 recalls the existing approaches to the problem of classification with test costs. Section 3 presents the idea of representing background knowledge (attribute values and tests) by levels of abstraction. It also describes a naïve Bayes classifier generalized to these levels of abstraction. Section 4 describes the concept of sequential classification and the stopping criterion for this strategy. Section 5 shows the results of the experimental evaluation of the proposed method. Section 6 concludes the paper.

2. Related works

The detailed review of algorithms that take into account test costs and/or misclassification costs are presented in [11,12]. However, not all the algorithms consider the aforementioned tradeoff. Thus, we intend to indicate these approaches, where this tradeoff is considered.

The problem of the tradeoff between the cost of tests and the accuracy of a classifier was considered in [13] (IDX), [14] (EG2), [15] (CS-ID3) and [16] (Clarify). All these approaches combine information gain and test costs in order to construct decision trees.

The problem of a tradeoff between the cost of tests and the misclassification cost of a classifier also was extensively analyzed. In [4] a system called ICET, which uses a genetic algorithm to build a decision tree to minimize the cost of tests and misclassifications was presented. In [17] the theoretical aspects of active learning with test costs using a PAC learning framework were studied. It is a theoretical work on a dynamic programming algorithm searching for best diagnostic policies measuring at most a constant number of attributes. The obtained result is not applicable in practice, because it requires a predefined number of training data in order to obtain suboptimal policies. In [18] an algorithm based on formulating the classification process as a Markov Decision Process (MDP), whose optimal policy gives the optimal diagnostic procedure was presented. While related to other work, this approach may incur very high computational cost to conduct the search. In [19] a tree-building strategy was proposed that uses minimum cost of tests and misclassifications as the attribute split criterion. In [20] a naïve Bayesian based cost-sensitive learning algorithm, called csNB was proposed in order to minimize the sum of test costs and misclassification costs. In [21] tree-building strategies were proposed: sequential test strategy, single batch strategy and multiple batch strategy. The comparison of these strategies showed that the total cost of the sequential test strategy is the lowest. In [22] a framework based on game theory was employed in order to build a cost-sensitive decision tree. The empirical evaluation of the proposed algorithm showed that it is possible to induce decision trees that maintain prediction accuracy but also minimize test and misclassification costs. However, there are a number of parameters which can be set in order to change the behavior of the algorithm in response to the differing test costs and misclassification costs. In [23–26] the problem of cost-sensitive classification with multiple cost scales was considered. The empirical comparison of cost-sensitive decision tree induction algorithms was presented in [27]. This comparison took into account 30 algorithms, which can be organized into 10 categories. The lowest cost is produced by a system ICET. It was indicated that high accuracy rates do not always mean low classification costs. Moreover, having an inexpensive decision tree does not automatically mean that it is an accurate decision tree.

The problem of reducing test costs while maintaining the prediction accuracy of a classifier is also considered in the context of cost-based feature selection. Methods that can deal with large-scale and real-time applications are urgently needed since costs must be budgeted and accounted for [1]. In [28] a genetic algorithm was used to perform feature selection where the fitness function combined two criteria: the accuracy of the classification realized by the neural network and the cost of performing the classification. In [29] a similar approach was presented, where a genetic algorithm is used in feature selection and parameters optimization for a support vector machine. The fitness function aggregated classification accuracy, the number of selected features and the feature cost. However, the above mentioned methods have the disadvantage of being computationally expensive. Therefore, a modification of a filter model, which is known to have a low computational cost was proposed in [30]. The presented modification adds to the features evaluation function a term to take into account the cost of the features. In [31] two main components of test-time CPU cost were examined (i.e., classifier evaluation costs and feature extraction costs) and it was shown how to balance these costs with classification accuracy.

3. Representing background knowledge by levels of abstraction

Let us notice that there are areas where attribute values are represented at different levels of abstraction. These levels model domain background knowledge and have usually a form of a tree-like hierarchy [6]. In such a tree the root represents a missing value, leaves represent specific attribute values, and the remaining nodes represent abstract attribute values (e.g., sets of specific values). Importantly, such hierarchies entail the existence of tests that replace a more abstract value by a less abstract value. Moreover, this replacement may