Knowledge-Based Systems 95 (2016) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features

Qifeng Zhou [a], Hao Zhou [a], Tao Li [b,c,*]

[a] School of Aerospace Engineering, Automation Department, Xiamen University, Xiamen, 361005, China
[b] School of Computing and Information Sciences, Florida International University, Miami, FL, 33199, United States
[c] School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210046, China

ARTICLE INFO

Article history:
Received 31 March 2015
Revised 9 September 2015
Accepted 11 November 2015
Available online 11 December 2015

Keywords:
Cost sensitive
Feature selection
Random forest

ABSTRACT

Feature selection aims to select a small subset of informative features that contain most of the information related to a given task. Existing feature selection methods often assume that all the features have the same cost. However, in many real world applications, different features may have different costs (e.g., different tests a patient might take in medical diagnosis). Ignoring the feature cost may produce good feature subsets in theory but they can not be used in practice. In this paper, we propose a random forest-based feature selection algorithm that incorporates the feature cost into the base decision tree construction process to produce low-cost feature subsets. In particular, when constructing a base tree, a feature is randomly selected with a probability inversely proportional to its associated cost. We evaluate the proposed method on a number of UCI datasets and apply it to a medical diagnosis problem where the real feature costs are estimated by experts. The experimental results demonstrate that our feature-cost-sensitive random forest (FCS-RF) is able to select a low-cost subset of informative features and achieves better performance than other state-of-art feature selection methods in real-world problems.

1. Introduction

The feature selection (FS) problem has been studied by the statistics and machine learning communities for many years. Its main theme is to select a small subset of informative features that best discriminate the data objects of different classes [1]. In many data analysis tasks, feature selection is an important and frequently used dimensionality reduction technique and is often considered as a critical step in data pre-processing. In addition, feature selection can significantly improve the understandability of the machine learning models and often help build models with better generalization [2,32]. As a result, in many situations, finding a good subset of features is an important problem in its own right.

Feature cost, literally meaning the cost consumed in acquiring a features value, is a special case of various cost types in machine learning and data mining [9–11]. It may involve different factors such as money, time, and implementation difficulty. In many real world applications, however, different features may have different costs and the difference may be sufficiently large to influence the result of feature selection. Take medical diagnosis as an example, we collected a dataset regarding hepatitis from a local hospital, the costs of most features from this dataset are estimated by experts, in terms of price, time, implementation difficulty, and side effect (i.e., medical tests, their detailed descriptions are given in Section 5). As seen in Table 1, different features may vary considerably across the costs. If a classifier is constructed with many expensive features, to predict new samples, it needs to pay a high cost and the classifier can be thought as lack of practicability. In such cases, it would be better to use a feature subset with an acceptable classification performance but a much lower cost. This kind of feature selection considering the cost is called cost-sensitive feature selection whose aim is to select both low-cost and informative features.

Existing feature selection techniques can be approximately divided into three categories: embedded, filter, and wrapper methods [1,3–8,31]. However, all these three categories of FS methods usually only focus on the features' distinguishing ability or their contribution to classification, and ignore their cost. One simple case of the filter methods Relief-F, for example, is to seek the subsets of features whose pairwise correlation is low. Support vector machine recursive feature elimination (SVM-RFE) [5], a more sophisticated embedded feature selection method, operate by eliminating the least weighted feature in the construction of the SVM model recursively. These traditional methods, in effect, assume all the features have the same cost.

* Corresponding author at: School of Computing and Information Sciences, Florida International University, United States. Tel.: +1 305 348 6036; fax: +1 305 348 3549.
E-mail address: taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.knosys.2015.11.010
0950-7051/

Table 1
Some medical tests and their costs in hepatitis treatment.

Cost / test	G	S	ALT	AST	HBV-DNA	Gene-type	MTCT
Price(¥)	100–200	100–200	89	89	140	390	–
Time(day)	3–7	3–7	1	1	2-3	4-5	–
Difficulty(0–10)	7	7	3	3	4	5	2
Side effect(0–10)	3–4	3–4	1	1	1	1	0

A branch specialized to handle cost-related problems, in data mining, is cost-sensitive learning [6,13]. The taxonomy of different types of costs is summarized by Turney [12] and the most common types of cost are the misclassification cost and the feature cost [14,15]. Both of them often arise in practical applications. For example, in medical diagnosis, the feature cost (or the test cost), is based on the fact that the medical tests, whose results the physician will refer to in diagnosing a patient, vary in the running time, the expense, and the side effect etc. The physician needs to estimate the tradeoff between the test effect and its associated cost before deciding which tests the patient should take. The test cost is actually one special case of the feature cost (i.e., the cost of acquiring one feature's value). In the following, we use these two terms (feature cost and test cost) interchangeably and name the process of acquiring a feature's values as having or taking a test.

Compared with the misclassification cost, the feature cost has been studied much less. In reality, feature cost is not only difficult to quantify, but also can have more complicated scenarios, e.g., some feature cost is variable, and different features costs may be connected. These scenarios are summarized by Min and Liu [13] who construct a hierarchical model covering six possible test-cost-sensitive decision systems.

Basically, there are two strategies to reduce the feature cost for a data mining task. The first one is to work out some principles or rules about how to utilize the features for a new test instance. This strategy is especially suitable for a small amount of test instances with many missing values: for every new instance, whose feature values are partially unknown, the strategy decides which feature should be used or tested (if its value is unknown). Note that different instances may vary on unknown features. To predict a new instance, a tailored classification model are needed (usually unsophisticated, like trees). This strategy, which effectively focuses on the classification process rather than the regular feature selection, finds its way in many applications, e.g., the decision tree-based methods, including ICET [12], minimal cost tree [14,15]; and other applications including Markov Decision Process [17,18], and some general test tactics [14,16]. The second strategy for reducing feature cost is to search for both informative and low-cost feature subsets. Models trained with such a feature subset can retain its structure in the test stage, and therefore, is usually fast and applicable for a great amount of test instances. This strategy, in essence, is an improvement on the ordinary feature selection. However, it is a more complex global optimization process, as it takes cost into consideration. Compared against the first strategy, the second strategy as well as its potential benefit has been rarely studied to date.

In order to obtain low-cost subsets of informative features, one straightforward solution is to take a two-step approach: first performing feature selection in regular way, then further analyzing the generated feature subsets or rank based on the costs. However, since these two steps are conducted separately, the interaction between the features discriminative ability and the costs will be neglected. As a result, the ultimate outcome will depend primarily on the regular feature selection, and the goal of "cheap" often cannot be satisfied. In addition, since the second analysis step is conducted manually based on experts experience, there is much uncertainty for the results. Another solution to seeking good and cheap feature subsets is making use of SVM [19], as it can assign weights to the features (this characteristic is already used in the classical feature selection method SVM-RFE [8]). By incorporating the cost factor into the SVM optimization processing, it is possible to approximately re-calibrate the weights for the features. The expanded SVM model can be viewed in [20]. The main limitation of this method is that the outcomes are very sensitive to the model parameters (for themselves, they are also difficult to set) thus weakening its practicability.

To overcome the aforementioned limitations, in this work, we propose a random forest-based cost-sensitive feature selection method named feature-cost-sensitive random forest (FCS-RF). FCS-RF can sort the features based on their comprehensive performance both on the distinguishing ability and costs. The top ranked features will be selected into the final feature subset first. Specifically, the FCS-RF incorporates the feature cost information into the base decision tree construction process. When constructing a base tree, a feature is selected with a probability inversely proportional to its associated cost, instead of being selected randomly. By means of the underlying mechanism of random forest, the importance of all features is calculated and a feature rank can be obtained considering both the feature cost and the distinguishing ability.

The contributions of our work are summarized as follows:

(1) We propose a cost-sensitive feature selection method FCS-RF, which overcomes the limitation of two-step cost-sensitive feature selection methods. FCS-RF incorporates both the distinguishing ability (or quality) of features and their costs as criteria into one optimization process. Therefore, FCS-RF is an approximate global optimization method which can consider the correlations among the features;

(2) We perform a series of empirical evaluations on benchmark datasets to demonstrate the effectiveness of FCS-RF. Compared with the commonly used feature selection approaches, FCS-RF can reduce the feature costs while maintaining a comparable classification performance;

(3) We apply FCS-RF to a real-world medical diagnosis to find cheap and good feature subset for interferon-α (IFN-α) treatment of hepatitis B virus(HBV) . The evaluating results on more than 300 real cases of patients demonstrate the effectiveness of FCS-RF.

The rest of the paper is organized as follows. Section 2 gives the formulation of our problem. Section 3 presents the feature-cost-sensitive random forest algorithm. Section 4 describes how to produce a cost-sensitive feature rank using the improved random forest. Section 5 shows the experiments and the results analysis. Finally, Section 6 makes conclusions and discusses the future work.

2. Problem formulation

A basic feature-cost-sensitive decision system can be summarized according to [15] as

$$S = (U, F, D, V_a | a \in F \cup D, I_a | a \in F \cup D, c^*), \tag{1}$$

where U is a finite set of objects called the universe, F is the set of features, D is the set of class variables (decisions), V_a is the set of values for each $a \in F \cup D$, I_a: $U \rightarrow V_a$ is an information function for each $a \in F \cup D$, $c^* : F \rightarrow R^+ \cup 0$ is the feature cost function. Considering some records of patients $\{x_1\ x_2\ \ldots\ x_r\}$ with each record containing the information of age, gender and gene type, then $U = \{x_1\ x_2\ \ldots\ x_r\}$, $F = \{age, gender, genetype\}$, $D = \{response, noresponse\}$. It is noticed that if c^* is empty, the system will ignore the feature cost (i.e.,