Expert Systems with Applications 40 (2013) 3918–3925

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A maximum-margin genetic algorithm for misclassification cost minimizing feature selection problem

Parag C. Pendharkar *

Information Systems, School of Business Administration Pennsylvania State University at Harrisburg, 777 West Harrisburg Pike, Middletown, PA 17057, United States

ARTICLE INFO

Keywords:
Feature selection
Cost sensitive classification
Support vector machines
Classification
Data mining

ABSTRACT

We consider a feature selection problem where the decision-making objective is to minimize overall misclassification cost by selecting relevant features from a training dataset. We propose a two-stage solution approach for solving misclassification cost minimizing feature selection (MCMFS) problem. Additionally, we propose a maximum-margin genetic algorithm (MMGA) that maximizes margin of separation between classes by taking into account all examples as opposed to maximizing margin of separation using a few support vectors. Feature selection is carried out by either an exhaustive or a heuristic simulated annealing approach in the first stage and a cost sensitive classification using either MMGA or cost sensitive support vector machines (SVM) in the second stage. Using simulated and real-world data sets and different misclassification cost matrices, we test our two-stage approach for solving the MCMFS problem. Our results indicate that feature selection plays an important role when misclassification cost asymmetries increase and the MMGA shows equal or better performance than the SVM.

1. Introduction

Feature selection (FS) problem, also known as dimensionality reduction problem, has been extensively studied in statistics, pattern recognition (Zavaschi, Britto, Oliveira, & Koerich, 2013), finance (Won, Kim, & Bae, 2012), medicine (Gorunescu, Belciug, Gorunescu, & Badea, 2012) and data mining literature (Pinheiro, Cavalcanti, Correa, & Ren, 2012; Saeys, Inza, & Larranaga, 2007; Zhuang, Widschwendter, & Teschendorff, 2012). Among the advantages of FS are better generalization due to lower model overfitting; and efficient and cost effective data mining models (Saeys et al., 2007; Stracuzzi & Utgoff, 2004). There are several statistical and machine learning methods that are available for solving the FS problem. Guyon and Elisseeff (2003) provide a good introduction to some of these methods. While FS methods are well established, for datasets containing large number of decision-making attributes, selecting features from training data is not always easy and is often computationally intensive (Saeys et al., 2007).

For datasets containing a large number of input decision-making attributes, an automatic approach is required for identifying low cardinality subset of relevant input decision-making attributes (Stracuzzi & Utgoff, 2004). Reducing cardinality of input decision-making attributes reduces organizational information acquisition cost and improves generalizability of a classifier. Since several real-world classification problems require FS (Fahlman & Lebiere, 1990; Kivinen & Warmuth, 1997), a variety of FS approaches are proposed in data mining literature. These FS approaches from the literature can be divided into two main categories: *filter* methods and *wrapper* methods. The primary difference between filter and wrapper methods is that the filter methods select input decision-making attributes independent of the learning algorithm, whereas the wrapper methods make learner-dependent selection of input decision-making attributes.

Filter methods use statistical or information theoretical measures to reduce dimensionality. Among the techniques used in filter methods are cross entropy, correlations, principal component analysis, χ^2 test and distance metrics (Dunteman, 1989; Hall, 1999; Kira & Rendell, 1992; Koller & Sahami, 1996; Liu & Setino, 1997). Wrapper methods, on the other hand, use search based methods for selecting input decision-making attributes. These search based methods include greedy search (Caruana & Freitag, 1994), backward elimination beam search (Aha & Bankert, 1994), nearest neighborhood search (Langley & Sage, 1994), backward best first search (Kohavi & John, 1997), randomized search (Stracuzzi & Utgoff, 2004), and heuristic genetic algorithms (Vafaie & Jong, 1995). A general consensus in the literature is that the wrapper methods outperform the filter methods (Stracuzzi & Utgoff, 2004).

To our knowledge, most studies on the FS problem have assumed symmetric misclassification error costs. The FS problem under asymmetric misclassification error cost did not receive a whole lot of attention in the literature. It is possible that the

* Tel.: +1 (717) 948 6028; fax: +1 (717) 948 6456.
E-mail address: pxp19@psu.edu
URL: http://www.personal.psu.edu/pxp19/

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.12.086

misclassification cost asymmetry and the number of selected features may interact with each other making certain features more relevant when misclassification costs are symmetric and certain other features more relevant when misclassification costs are asymmetric. Additionally, the benefit of using FS when misclassification cost asymmetries increase is not clearly established. That is, it is not clear if FS is more beneficial, less beneficial or about the same when misclassification cost asymmetries increase. In our research, we implemented a misclassification cost minimizing feature selection (MCMFS) problem to gain better understanding of the FS problem under varying symmetric and asymmetric misclassification costs.

When solving the MCMFS, we are essentially solving two problems: the FS problem and the misclassification cost minimization problem. In our research, we solved these two problems separately in two different stages. In the first stage, we selected decision-making features and in the second stage, we learned a misclassification cost minimizing discriminant function based on features selected in the first stage.

We propose a two-stage wrapper method procedure for solving the MCMFS. Additionally, we propose two search based methods for solving the FS problem in the first stage. The first FS method uses exhaustive backtracking search and considers all possible features of cardinality greater than or equal to two decision-making attributes. The major drawback of exhaustive backtracking search is increased computational requirements making it unrealistic for a large number of input decision-making attributes. As a result, we propose a second FS method that uses simulated annealing (SA) heuristic to identify features of cardinality greater than one. The SA heuristic procedure does not guarantee an optimal solution but uses realistic computational time to provide a heuristic solution.

For learning a discriminant function that minimizes misclassification cost in the second stage, we use two different methods as well. Our first method is the misclassification cost minimizing support vector machine (SVM) proposed by Masnadi-Shirazi and Vasconcelos (2010). We design our second method by considering some principles of the SVM (Bradley, Fayyad, & Mangasarian, 1999) where a classification function that minimizes misclassification cost also minimizes an upper bound on generalization error. While there are several classification algorithms that allow decision-makers to minimize misclassification costs, Pendharkar and Nanda (2006) showed that genetic algorithm (GA) based misclassification cost minimizing classifiers generally perform the best and are more flexible. Thus, we use a GA based linear classifier that minimizes misclassification cost and provides a separating plane that maximizes the margin of separation between classes. We call our misclassification cost minimizing classifier a max-margin genetic algorithm (MMGA).

The rest of the paper is organized as follows. In Section 2, we formally present the MCMFS and propose a framework for two-stage solution approach. In Section 3, we describe the SVM and the MMGA classifier for solving classification problems involving asymmetric misclassification costs. In Section 4, using the proposed framework of two stage solution approach, we describe three hybrid approaches – backtracking-MMGA (BT-MMGA), BT-SVM and simulated annealing MMGA (SA-MMGA) – for solving the MCMFS. In Section 5, we describe our simulated and real-world data; and experiments and results. In Section 6, we conclude our paper with the summary of our findings.

2. The MCMFS problem and a two stage solution approach

In order to mathematically describe the MCMFS problem, we assume that $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ is a vector of $n > 2$ input attributes that are available for use by a machine learning classification algorithm that minimizes misclassification cost. Let $\boldsymbol{u} = [u_1, u_2, \ldots, u_n]^T$, where $u_i \in \{0,1\}\ \forall\ i \in \{1,\ldots,n\}$ be a binary n-dimensional vector that indicates the input attributes that are considered, when $u_i = 1$, or discarded, when $u_i = 0$. Further, assume a vector $\boldsymbol{z}$ defined as $\boldsymbol{z} = \{x_i | \forall i,\ u_i \neq 0\}$. If $A = \{a\}$ is an output decision with $a \in \{0,1\}$ then the MCMFS problem can be represented as $f(\boldsymbol{z}) \rightarrow A$. The MCMFS problem consists of finding a combination of misclassification cost minimizing classification function $f()$ and a vector $\boldsymbol{z}$ such that $f(\boldsymbol{z})$ has the lowest misclassification cost among all possible values of vector $\boldsymbol{z}$. Since $\boldsymbol{z}$ depends on $\boldsymbol{u}$, which is a binary vector of cardinality n, there are a total of $2^n - 1$ non-null unique values for vector $\boldsymbol{z}$ that have to be considered.

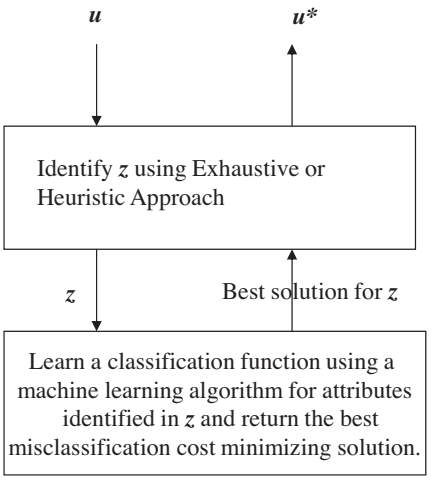

Fig. 1. A general framework for solving the MMCA problem.

In our research, to avoid solutions with only one input attribute[1], we imposed a constraint $|\boldsymbol{z}| > 1$, where $|\boldsymbol{z}| \leqslant n$ denotes cardinality of vector $\boldsymbol{z}$. Our constraint led to a total of $2^n - (n+1)$ possible solutions search space for finding a set of input attributes that provided the lowest misclassification cost. Since from the Binomial theorem, we have $2^n = \sum_{g=0}^{n} \binom{n}{g}$, the search space complexity of the MCMFS problem is exponential.

We used a two-stage solution procedure to solve the MMCA problem. Our solution procedure consisted of systematically identifying all values of $\boldsymbol{u}$, learning $f(\boldsymbol{z})$ for each value of $\boldsymbol{u}$ using two different misclassification cost minimizing algorithms (described in next section), and selecting the optimal value $\boldsymbol{u}^*$ that provided the lowest misclassification cost (optimal solution) on the training dataset. Given that the search space complexity is exponential, for small values of n, all possible values of $\boldsymbol{u}$ could be identified and $f(\boldsymbol{z})$ was solved for all values of $\boldsymbol{u}$. However, for large values of n, it was not practical to solve $f(\boldsymbol{z})$ for all possible values of $\boldsymbol{u}$. An SA heuristic approach was used for large values of n, and sub-optimal heuristic solutions were obtained for the MCMFS problem in reasonable time.

Fig. 1 illustrates the two-stage general framework used for solving the MMCA problem. In the first stage at the top of the Fig. 1, we identified different values of $\boldsymbol{u}$ and related $\boldsymbol{z}$ using either an exhaustive search or a heuristic technique. In the second stage, we solved a classification problem using the attributes identified in the first stage and a misclassification cost minimizing procedure. The first and second stages worked iteratively and sequentially

[1] We impose this constraint to create a classifier with more than one variable. We assume that the classification problem is sufficiently complex to require the use of two or more variables.