Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 35 – 42

Procedia Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

An individualized preprocessing for medical data classification

Sarab AlMuhaideb[a], Mohamed El Bachir Menai[b]

[a]*Department of Computer Science, Prince Sultan University, 66833 Riyadh 11586, SA*
[b]*Department of Computer Science, King Saud University, 51178, Riyadh 11543, SA*

Abstract

Data preprocessing has a profound effect on the performance of the learner. Before attempting medical data classification, characteristics of medical datasets, including noise, incompleteness, and the existence of multiple and possibly irrelevant features, need to be addressed. In this paper, we show that selecting the right combination of preprocessing methods has a considerable impact on the classification potential of a dataset. The preprocessing operations considered include the discretization of numeric attributes, the selection of attribute subset(s), and the handling of missing values. The classification is performed by an ant colony optimization algorithm as a case study. Experimental results on 25 real-world medical datasets show that a significant relative improvement in predictive accuracy, exceeding 60% in some cases, is obtained.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: classification; ant colony optimization; medical data classification; preprocessing; feature subset selection; discretization.

1. Introduction

Medical data classification (MDC) refers to learning classification models from medical datasets and aims to improve the quality of health care[1]. Medical data classification can be used for diagnosis and prognosis purposes. Medical data exhibit unique features including noise resulting from human as well as systematic errors, missing values and even sparseness[2]. The quality of data has a large implication for the quality of the mining results. It is necessary to perform preprocessing steps in order to remove or at least alleviate some of the problems associated with medical data. However, each dataset is different, and there is no preprocessing method that is best across all datasets. Deciding the best combination of preprocessing methods for a specific dataset is not possible without trial and comparisons. The advent of various open-source libraries, like Weka[3] and KEEL[4], hosting an extensive set of off-the-shelf preprocessing methods, combined with the leisure of standard formats like the attribute-relation file format (ARFF) [1] and advances in computer hardware technology, encourages integration of automatic tuning for preprocessing operations into the data mining task for each dataset on an individual basis. In this research, we investigate the influence of individualized preprocessing on the classification of medical datasets, including the removal of missing values and a variety of

* Corresponding author. Tel.: +0-966-494-8360 ; fax: +0-966-454-8317.
E-mail address: smuhaideb@psu.edu.sa

[1] `http://weka.wikispaces.com/ARFF`

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.006

discretization and attribute selection methods. The rest of the paper is organized as follows. Section 2 highlights related work in the area. Next, Section 3 describes the individualized tuning procedure. Experimental results are presented in Section 4 and discussed in Section 5. The paper is concluded in Section 6.

2. Related Work

Metaheuristic methods stand as interesting techniques for classification model learning, because of their good performance and low computational requirements. Metaheuristics require little or no background knowledge of the problem at hand. In ant colony optimization algorithms[5,6], artificial ants use pheromone trails and heuristic information to guide solution construction for finding the shortest path from food sources to their nest. AntMiner[7] is the first ACO algorithm for classification tasks. Among the different variants of AntMiner, AntMiner+[8] has been chosen as the classification algorithm in this research. AntMiner+ is based on the MAX–MIN ant system[6], which is recognized as one of the best-performing algorithms in the ant colony optimization family. The classification model is constructed using the sequential covering strategy. The results reported show that AntMiner+, on average, obtained the highest rank among state-of-the-art rule-based classifiers included[8].

Although the problems associated with medical data have been documented since the nineties, not much research has been done to address the complete preprocessing task of medical data. Tanwani and Farooq[2] performed an extensive study to present the challenges associated with biomedical data and approximate the classification potential of a biomedical dataset using a qualitative measure of this complexity. The study concludes that the classification accuracy is found to be dependent on the complexity of the biomedical dataset, not on the classifier choice. The number and type of attributes have no noticeable effect on the classification accuracy, as compared to the quality of the attributes. It is shown that biomedical datasets are noisy and that noise is the dominant factor that affects the resulting classification accuracy. Lin and Haug[9] use heuristic rules that utilize that utilizes information from the medical data, metadata and sources of medical knowledge. As far as we are concerned, the individualized preprocessing of medical data has not been addressed before.

3. An Individualized Preprocessing Procedure

The AntMiner+ is based on a sequential-covering strategy and a default rule related to the majority class. In effect, rule induction focuses on classes other than the majority class. This particular strategy is advantageous in MDC because the majority of class instances are normally the negative cases of which we care less. The sequential-covering strategy helps in handling large-sized datasets; due to the removal of instances already covered by induced rules, the progressive reduction of the training set size is thus achieved. AntMiner+ algorithm cannot handle instances containing missing values. Thus, these instances are removed from the dataset in the first step. To reduce the size of the solution space, the number of attributes is limited to no more than a default value of 10. If the dataset contains a larger number of attributes, then attribute selection takes place prior to induction. Various attribute types can be handled by the AntMiner+ algorithm. These include nominal and ordinal values, as well as numeric values, including integer and continuous attributes that are discretized. In effect, numeric values are encoded as discrete intervals defined by $[lower\_bound - upper\_bound]$. The order of preprocessing steps in the concerned AntMiner+ implementation is as follows: removal of instances with missing values, discretization, then attribute selection.

3.1. Timing of Removing Instances Having Missing Values

In the context of the AntMiner+ algorithm, all instances having missing values are removed in the first step of preprocessing. The next steps in the preprocessing consist of the application of the discretization algorithm and attribute selection algorithm (if necessary). This procedure might not be the best in some cases. For example, consider datasets with large number of predictive attributes. If the removal of instances having missing values is delayed after the attribute selection step, then this would allow more instances to be available for training and testing subsets, thus perhaps improving the results. Otherwise, some instances would be removed because they include missing values in attributes that will be next removed by the attribute selection step. Thus, the removal of these instances is no longer rationalized. We hypothesize that if the removal of instances with missing values were delayed until after the attribute selection step, then better results would be obtained.