Data & Knowledge Engineering 81–82 (2012) 67–103

Contents lists available at SciVerse ScienceDirect

Data & Knowledge Engineering

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

DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets

Mina Alibeigi, Sattar Hashemi *, Ali Hamzeh

CSE and IT Dept., Engineering Campus Number 2, Mollasadra Ave., Shiraz, Iran

ARTICLE INFO

Article history:
Received 31 October 2010
Received in revised form 5 August 2012
Accepted 6 August 2012
Available online 18 August 2012

Keywords:
Feature selection
Imbalanced data set
Probability density function (PDF)

ABSTRACT

Nowadays, imbalanced data sets are pervasive in real world human practices, and hence, become a very interesting research area within machine learning communities. Imbalanced data sets introduce a significant reduction in performance of standard classifiers when they are invoked to learn data underlying concepts. The problem becomes even more sever when imbalanced data sets are involved with high dimensions.

This paper presents a novel feature ranking approach based on the probability density estimation to cope with these issues. The idea behind our approach, named Density Based Feature Selection (DBFS), is that features' distributions over classes can bring significant benefits to feature selection algorithms. In other words, to explore the contribution of each attribute and assign it an appropriate rank, DBFS takes into account features' corresponding distributions over all classes along with their correlations.

To show the effectiveness of the presented approach, well-known feature ranking methods are implemented and compared with our approach across varieties of small sample size and high dimensional data sets from microarray, mass spectrometry and text mining domains. Our theoretical analysis and experimental observations reveal that our approach is the method of choice by offering a simple yet effective feature ranking method based on well-known statistical evaluation measures.

1. Introduction

The class imbalance problem refers to the issue that occurs when one or more classes of a data set have significantly more number of instances than other classes [1]. Nowadays, imbalanced data sets, also known as skew data sets, have received a great deal of attention by researchers due to their importance in many real world human practices such as biological data analysis [2], text classification [3–5], image classification [6] and fraud detection [7,8] among many others.

Despite the prevalence of imbalanced data sets, the performance of many classification algorithms like Naïve Bayes [9], Nearest Neighbor [9], Support Vector Machines [9] and C4.5 [10] degrades significantly when they are applied on these types of data sets [11–13]. This poor performance can be attributed to the fact that almost all classifiers return a simple yet accurate hypothesis to avoid over-fitting the data. These standard algorithms assume or expect balanced class distributions or equal miss classification costs for different classes [1,3,14,66,67]. It is not surprising to see that when presented with complex imbalanced data sets, such hypothesis simply returns the majority class as the result of classification to satisfy the simplicity and accuracy trade off [1,3]. Nonetheless, when dealing with imbalanced data sets, we would prefer that classifiers perform well on the minority class, even at the expense of misclassifying instances of the majority class due to the importance of the minority class [14]. Also, it is worth mentioning that imbalanced data sets usually tend to suffer from class overlapping, lack of representative

* Corresponding author at: P.O. Box: 71348 51154. Tel.: +98 711 613 3544; fax: +98 711 647 4605.
E-mail addresses: alibeigi@cse.shirazu.ac.ir (M. Alibeigi), s_hashemi@shirazu.ac.ir (S. Hashemi), ali@cse.shirazu.ac.ir (A. Hamzeh).

0169-023X/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2012.08.001