Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets

Mina Alibeigi alibeigi@cse.shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Sattar Hashemi s_hashemi@shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Ali Hamzeh hamzeh@shirazu.ac.ir
Computer Science and Engineering Department
Shiraz University
Shiraz, 71348-51154, Iran

Abstract

Since dealing with high dimensional data is computationally complex and sometimes even intractable, recently several feature reduction methods have been developed to reduce the dimensionality of the data in order to simplify the calculation analysis in various applications such as text categorization, signal processing, image retrieval and gene expressions among many others. Among feature reduction techniques, feature selection is one of the most popular methods due to the preservation of the original meaning of features. However, most of the current feature selection methods do not have a good performance when fed on imbalanced data sets which are pervasive in real world applications.

In this paper, we propose a new unsupervised feature selection method attributed to imbalanced data sets, which will remove redundant features from the original feature space based on the distribution of features. To show the effectiveness of the proposed method, popular feature selection methods have been implemented and compared. Experimental results on the several imbalanced data sets, derived from UCI repository database, illustrate the effectiveness of the proposed method in comparison with other rival methods in terms of both AUC and F1 performance measures of 1-Nearest Neighbor and Naïve Bayes classifiers and the percent of the selected features.

Keywords: Feature, Feature Selection, Filter Approach, Imbalanced Data Sets.

1. INTRODUCTION

Since data mining is capable of finding new useful information from data sets, it has been widely applied in various domains such as pattern recognition, decision support systems, signal processing, financial forecasts and etc [1]. However by the appearance of the internet, data sets are getting larger and larger which may lead to traditional data mining and machine learning algorithms to do slowly and not efficiently. One of the key solutions to solve this problem is to reduce the amount of data by sampling methods [2], [3]. But in many applications, the number of instances in the data set is not too large, whereas the number of features in these data sets is more than one thousands or even more. In this case, sampling is not a good choice. Theoretically, having more features, the discrimination power will be higher in classification. However, this theory is not always true in reality since some features may be unimportant to predict the class labels or even be irrelevant [4], [5]. Since many factors such as the quality of the