Unsupervised Feature Selection Using Feature Density Functions

Mina Alibeigi, Sattar Hashemi, Ali Hamzeh

***Abstract*—**Since dealing with high dimensional data is computationally complex and sometimes even intractable, recently several feature reductions 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, gene expressions and etc. Among feature reduction techniques, feature selection is one the most popular methods due to the preservation of the original features.

In this paper, we propose a new unsupervised feature selection method which will remove redundant features from the original feature space by the use of probability density functions of various features. To show the effectiveness of the proposed method, popular feature selection methods have been implemented and compared. Experimental results on the several datasets derived from UCI repository database, illustrate the effectiveness of our proposed methods in comparison with the other compared methods in terms of both classification accuracy and the number of selected features.

***Keywords*—**Feature, Feature Selection, Filter, Probability Density Function

I. Introduction

SINCE data mining is capable of finding new useful information from datasets, it has been widely applied in various domains such as pattern recognition, decision support, signal processing, financial forecasts and etc [1]. However by the appearance of the internet, datasets 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 dataset is not too large, whereas the number of features in these datasets 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 data, are responsible in the success of a learning algorithm, in order to extract information more efficiently, the dataset should not contains irrelevant, noisy or redundant features [6]. Furthermore, high dimensionality of data may cause the “curse of dimensionality” problem [7]. Feature reduction (dimensionality reduction) methods are one of the key solutions to all these problems.

Feature reduction refers to the problem of reducing the dimension by which the data is described [8]. The general purpose of these methods is to represent data with fewer features to reduce the computational complexity whereas preserving or even improving the discriminative capability [8]. Since feature reduction can brings a lot of advantages to learning algorithms, such as avoiding over-fitting and robustness in the presence of noise as well as higher accuracy, it has attracted a lot of attention in the three last decades. Therefore, vast variety of feature reduction methods was suggested which are totally divided into two major categories including feature extraction and feature subset selection. Feature extraction techniques projects data into a new reduced subspace in which the initial meaning of the features are not kept any more. Some of the well-known state-of-the-art feature extraction methods are principal component analysis (PCA) [5], non-linear PCA [12] and linear discriminant analysis (LDA) [12]. In comparison, feature selection methods preserve the primary information and meaning of features in the selected subset. The purpose of these schemas is to remove noisy and redundant features from the original feature subspace [12]. Therefore, due to preserving the initial meaning of features, feature selection approaches are in more of interest [8], [9].

Feature selection methods can be broadly divided into two categories: filter and wrapper approaches [9]. Filter approaches choose features from the original feature space according to pre-specified evaluation criterions, which are independent of specified learning algorithms.

Conversely, wrapper approaches selects features with higher prediction performances estimated according to specified learning algorithms. Thus wrappers can achieve better performance than filters. However, wrapper approaches are less common than filter ones because they need higher computational resources and are often intractable for large scale problems [9]. Due to its computational efficiency and independency to any specified learning algorithm, filter approaches are more popular and common for high

M. Alibeigi is with IT, Computer Science and Engineering Department, Shiraz University, Iran (corresponding author to provide phone: +98-711-6133544; fax: +98-711-6474605; e-mail: alibeigi@cse.shirazu.ac.ir).

S. Hashemi is with IT, Computer Science and Engineering Department, Shiraz University, Iran (e-mail: s_hashemi@shirazu.ac.ir).

A. Hamzeh is with IT, Computer Science and Engineering Department, Shiraz University, Iran (e-mail: hamzeh@shirazu.ac.ir).