Chapter 12

Simultaneous Feature Selection and Tuple Selection for Efficient Classification

Manoranjan Dash
Nanyang Technological University, Singapore

Vivekanand Gopalkrishnan
Nanyang Technological University, Singapore

ABSTRACT

Feature selection and tuple selection help the classifier to focus to achieve similar (or even better) accuracy as compared to the classification without feature selection and tuple selection. Although feature selection and tuple selection have been studied earlier in various research areas such as machine learning, data mining, and so on, they have rarely been studied together. The contribution of this chapter is that the authors propose a novel distance measure to select the most representative features and tuples. Their experiments are conducted over some microarray gene expression datasets, UCI machine learning and KDD datasets. Results show that the proposed method outperforms the existing methods quite significantly.

INTRODUCTION

It is no longer news that data are increasing very rapidly day-by-day. Particularly with Internet becoming so prevalent everywhere, the sources of data have become numerous. Data are increasing in both ways: dimensions or features and instances or examples or tuples, not all the data are relevant though. While gathering the data on any particular aspect, usually one tends to gather as much information as will be required for various tasks. One may not explicitly have any particular task, for example classification, in mind. So, it behooves for a data mining expert to remove the noisy, irrelevant and redundant data before proceeding with classification because many traditional algorithms fail in the presence of such noisy and irrelevant data (Blum and Langley 1997). As an example, consider microarray gene expression data where there are thousands of features (or genes) and only 10s of tuples (or sample tests). For example, Leukemia cancer data (Alon, Barkai et al. 1999) has 7129 genes and 72 sample tests. It has been shown that even with very few genes one can achieve the same or even better prediction ac-

DOI: 10.4018/978-1-60566-748-5.ch012
Simultaneous Feature Selection and Tuple Selection for Efficient Classification

curacy than whole data. Similarly not all sample tests are necessary for the classification task. In addition to this advantage, removing such noisy features and tuples improves the understanding of the user by bringing focus.

Thus, as a solution, in this chapter we study how to remove the noisy and irrelevant features and tuples while improving or at least not significantly reducing the prediction accuracy. This task is called “focusing” in this chapter, i.e., select the relevant features and tuples in order to improve the overall learning. In the literature many feature selection and tuple selection methods have been proposed (Dash and Liu 1997; Liu and Motoda 1998). Feature selection algorithms can be broadly grouped as wrapper and filter (Blum and Langley 1997; Kohavi and John 1997; Liu and Motoda 1998; Inza, Larranaga et al. 2004).

In the wrapper method the classifier is used as an evaluation function to evaluate and compare the candidate feature subsets. In filter method evaluation of feature subsets is usually independent of the final classifier. In tuple selection, a broad category of methods let the learning algorithm to select the relevant tuples (Blum and Langley 1997; Liu and Motoda 2001; Fragoudis, Meretakis et al. 2002).

Contribution of this chapter is that we propose a distance measure to select representative features and tuples. Experimental results over several microarray datasets and UCI machine learning repository datasets (Asuncion and Newman 2007) and UCI KDD archive (Hettich and Bay 1999) show that the proposed TwoWayFocused algorithm performs better than the existing feature selection methods.

The rest of the chapter is organized as follows. In the next section we briefly discuss the related work in feature selection and tuple selection for classification. In Section 3 we describe the distance measure, and then describe our proposed feature selection method for classification using the distance measure. In Section 4 we describe our proposed tuple selection method for classification using the distance measure. In Section 5 we describe how to handle different types of data such as binary, nominal, discrete, and continuous. Section 6 describes the empirical study where we show the comparison results between the proposed algorithm and some existing methods. The chapter concludes in Section 7.

RELATED WORK

We study the related work in feature selection and tuple selection. We will mainly draw our references from the work that has been done in the area of machine learning.

Feature Selection

Feature selection is defined by many authors by looking at it from different angles. A common definition is: Definition 1. The aim of feature selection is to choose a subset of features for improving prediction accuracy or decreasing the number of features without significantly decreasing prediction accuracy of the classifier built using only the selected features (John, Kohavi et al. 1994).

Ideally feature selection methods search through the subsets of features and try to find out the best one among the competing $2^m$ subsets according to some evaluation criterion. Here $m$ is the number of features. But this procedure is exhaustive and impractical even for moderate size $m$. Other methods are based on heuristic or random search. These methods need a stopping criterion to prevent an exhaustive search. So, a typical feature selection process consists of (a) search and generate procedure, (b) an evaluation function, (c) a stopping criterion, and (d) a validation function to validate the selected features (Dash and Liu 1997).

The evaluation function measures the goodness of a feature subset. It can either be wrapper (Kohavi and John 1997; Kim, Street et al. 2000) or filter (Liu and Setiono 1996; Hall 2000; Yu and...