Chapter 5
An Optimal Categorization of Feature Selection Methods for Knowledge Discovery

Harleen Kaur
Hamdard University, India

Ritu Chauhan
Hamdard University, India

M. Afshar Alam
Hamdard University, India

ABSTRACT

With the continuous availability of massive experimental medical data has given impetus to a large effort in developing mathematical, statistical and computational intelligent techniques to infer models from medical databases. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. However, there have been relatively few studies on preprocessing data used as input for data mining systems in medical data. In this chapter, the authors focus on several feature selection methods as to their effectiveness in preprocessing input medical data. They evaluate several feature selection algorithms such as Mutual Information Feature Selection (MIFS), Fast Correlation-Based Filter (FCBF) and Stepwise Discriminant Analysis (STEPDISC) with machine learning algorithm naive Bayesian and Linear Discriminant analysis techniques. The experimental analysis of feature selection technique in medical databases has enable the authors to find small number of informative features leading to potential improvement in medical diagnosis by reducing the size of data set, eliminating irrelevant features, and decreasing the processing time.

DOI: 10.4018/978-1-4666-2455-9.ch005

INTRODUCTION

Data mining is the task of discovering previously unknown, valid patterns and relationships in large datasets. Generally, each data mining task differs in the kind of knowledge it extracts and the kind of data representation it uses to convey the discovered knowledge. Data mining techniques has been applied to a variety of medical domains to improve medical decision making. The sheer number of data mining techniques has the ability to handle large associated medical datasets, which consist of hundreds or thousands of features. The large amount of features present in such datasets often causes problems for data miners because some of the features may be irrelevant to the data mining techniques used. To deal with irrelevant features data reduction techniques can be applied in many ways, by feature (or attribute) selection, by discretizing continuous feature-values, and by selecting instances. There are several benefits associated with removing irrelevant features, some of which include reducing the amount of data (i.e., features). The reduced factors are easier to handle while performing data mining, and is capable to analyze the important factors within the data.

However, the feature selection has been an active and fruitful field of research and development for decades in statistical pattern recognition (Mitra, Murthy, & Pal, 2002), machine learning (Liu, Motoda, & Yu, 2002; Robnik-Sikonja & Kononenko, 2003) and statistics (Hastie, Tibshirani, & Friedman, 2001; Miller, 2002). It plays a major role in data selection and preparation for data mining. Feature selection is the process of identifying and removing irrelevant and redundant information as much as possible. The irrelevant features can harm the quality of the results obtained from data mining techniques; it has proven that inclusion of irrelevant, redundant, and noisy attributes in the model building process can result in poor predictive performance as well as increased computation.

Moreover, the feature selection is widely used for selecting the most relevant subset of features from datasets according to some predefined criterion. The subset of variables is chosen from input variables by eliminating features with little or no predictive information. It is a preprocessing step to data mining which has proved effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility in medical databases. Many methods have shown effective results to some extent in removing both irrelevant features and redundant features. Therefore, the removal of features should be done in a way that does not adversely impact the classification accuracy. The main issues in developing feature selection techniques are choosing a small feature set in order to reduce the cost and running time of a given system, as well as achieving an acceptably high recognition rate. The computational complexity of categorization increases rapidly with increasing numbers of objects in the training set, with increasing number of features, and increasing number of classes. For multi-class problems, a substantially sized training set and a substantial number of features is typically employed to provide sufficient information from which to differentiate amongst the multiple classes. Thus, multi-class problems are by nature generally computationally intensive. By reducing the number of features, advantages such as faster learning prediction, easier interpretation, and generalization are typically obtained.

Feature selection is a problem that has to be addressed in many areas, especially in data mining, artificial intelligence and machine learning. Machine learning has been one of the methods used in most of these data mining applications. It is widely acknowledged that about 80% of the resources in a majority of data mining applications are spent on cleaning and preprocessing the data. However, there have been relatively few studies on preprocessing data used as input in these data mining systems.