Chapter 3
A Unified Feature Selection Model for High Dimensional Clinical Data Using Mutated Binary Particle Swarm Optimization and Genetic Algorithm

Thendral Puyalnithi
Vellore Institute of Technology(VIT), Vellore, India

Madhuviswanatham Vankadara
Vellore Institute of Technology(VIT), Vellore, India

ABSTRACT

This article contends that feature selection is an important pre-processing step in case the data set is huge in size with many features. Once there are many features, then the probability of existence of noisy features is high which might bring down the efficiency of classifiers created out of that. Since the clinical data sets naturally having very large number of features, the necessity of reducing the features is imminent to get good classifier accuracy. Nowadays, there has been an increase in the use of evolutionary algorithms in optimization in feature selection methods due to the high success rate. A hybrid algorithm which uses a modified binary particle swarm optimization called mutated binary particle swarm optimization and binary genetic algorithm is proposed in this article which enhanced the exploration and exploitation capability and it has been a verified with proposed parameter called trade off factor through which the proposed method is compared with other methods and the result shows the improved efficiency of the proposed method over other methods.

DOI: 10.4018/978-1-7998-8048-6.ch003

INTRODUCTION

Feature reduction is an important process in the domain of data preprocessing. There are many reasons to perform feature reduction, including the need to decrease the size of a dataset in classification and clustering algorithms. In some cases, a redundant attribute can influence the decision attribute. Moreover, a redundant attribute can decrease the performance of classification algorithms. It is important to reduce the features of high-dimensional datasets while maintaining significant attributes. Feature reduction is performed on the datasets of variety of applications and it can be predominantly found the applications corresponding to image processing and clinical data analysis.

Clinical datasets usually have very large number of attributes few numbers of tuples, so that the reduction of attributes becomes immensely important, since the presence of many attributes increases the probability of the existence of noisy attributes and due to the presence of unimportant attributes the efficiency of classifiers that are created out of that will be degraded. So, it has become a very common practice to just keep the important attributes in clinical datasets and there are lot of researches has been done over this topic of feature reduction to support analysis of clinical data sets. The electronic medical records are clinical data can be categorized into many types. Clinical datasets include Patients' symptoms data, Patient's medical history data such as treatment data, demographic data, diagnostics data, laboratory test data, physiology data, pharmacy data, radiology image and report, hospital admission, transfer and discharge information and discharge summary. For clinical decision support system to work efficiently the clinical data would tend to have these varieties of clinical data and due to this aspect only, the clinical data sets need to be trimmed to get better classifier out of that.

Moreover, due to its social advantage too, feature reduction is often used to reduce high-dimensional clinical datasets. The reason for using feature selection, for example, if a patient must undergo clinical tests prior to a diagnosis, knowledge of the tests' significant features will identify parameters. Thereby, the patient will be both physically and economically comfortable. The attribute reduction should not degrade the performance of classification. There should be a trade-off between the feature reduction and performance of the models.

There are three ways to perform feature reduction:

1. **Feature Subset Selection:** Generates a subset of the features.
2. **Feature Reduction Through Transformation:** Updates the values of the features.
3. **Feature Generation:** Generates new features from existing features. The generated feature replaces at least one existing feature, which reduces the size of the dataset.

The proposed method focuses on the feature subset selection and enhances the classifier's accuracy.

Figure 1. Classification of feature subset selection methods