Chapter 3
Data Compaction Techniques

R. Raj Kumar
RGMCET, India

P. Viswanath
IIITS Chittoor, India

C. Shoba Bindu
JNTUA, India

ABSTRACT

A large dataset is not preferable as it increases computational burden on the methods operating over it. Given the Large dataset, it is always interesting that whether one can generate smaller dataset which is a subset or a set (cardinality should be less when compare to original dataset) of extracted patterns from that large dataset. The patterns in the subset are representatives of the patterns in the original dataset. The subset (set) of representing patterns forms the Prototype set. Forming Prototype set is broadly categorized into two types. 1) Prototype set which is a proper subset of original dataset. 2) Prototype set which contains patterns extracted by using the patterns in the original dataset. This process of reducing the training set can also be done with the features of the training set. The authors discuss the reduction of the datasets in the both directions. These methods are well known as Data Compaction Techniques.

INTRODUCTION

The large datasets are always increases computational burden on the methods (algorithms) operating over them. In pattern recognition and its allied fields it is always interesting to generate smaller dataset which is a representative of the original

DOI: 10.4018/978-1-5225-2805-0.ch003

set. A dataset can be represented using a set of attributes or features and patterns or objects. The reduction of the original set can be achieved in both directions *i.e.* reducing the number of patterns and reducing the number of features. Reducing the number of patterns is called Prototype selection or Prototype generation. Reducing the number of features is called Feature Selection or Feature Extraction.

Forming Prototype set is basically categorized into two types.

- Prototype set which is a proper subset of original dataset.
- Prototype set which contains patterns extracted by using the patterns in the original training set.

Given the large dataset in which patterns are represented by large number of features, it is efficient to select a set of features which can best describe the dataset. Like prototype set discussed above, selection of feature set also basically divided into two categories.

- Feature set which is a proper subset of set of features in the original dataset.
- Feature set which contains features extracted from the features of the original dataset.

Both the methods attract the researchers of Big Data, Pattern Recognition and its allied fields.

This chapter is organized as follows. In the section Data Compression using Prototype selection methods the novel methods that are used for Prototype selection were presented. Nearest Neighbour rule is used for computing the Prototype set. The section Data Compaction using Feature selection, some of the important methods which are useful for reducing the data using based on feature selection were discussed. The next section presents how we can combine the methods present in the previous two sections so as to reduce the data both vertically and horizontally. Each section is strengthened by giving suitable examples and experimental results over datasets which are widely used in Machine Learning and its allied fields. The section conclusion and future enhancement gives the future scope for the researchers related to this area.

The last section presents brief summary of the chapter.