Chapter 5
Minimum Database Determination and Preprocessing for Machine Learning

Angel Fernando Kuri-Morales
ITAM, Mexico

ABSTRACT

The exploitation of large databases implies the investment of expensive resources both in terms of the storage and processing time. The correct assessment of the data implies that pre-processing steps be taken before its analysis. The transformation of categorical data by adequately encoding every instance of categorical variables is needed. Encoding must be implemented that preserves the actual patterns while avoiding the introduction of non-existing ones. The authors discuss CESAMO, an algorithm which allows us to statistically identify the pattern preserving codes. The resulting database is more economical and may encompass mixed databases. Thus, they obtain an optimal transformed representation that is considerably more compact without impairing its informational content. For the equivalence of the original (FD) and reduced data set (RD), they apply an algorithm that relies on a multivariate regression algorithm (AA). Through the combined application of CESAMO and AA, the equivalent behavior of both FD and RD may be guaranteed with a high degree of statistical certainty.

DOI: 10.4018/978-1-5225-7268-8.ch005

INTRODUCTION

Nowadays, commercial enterprises are importantly oriented to continuously improving customer-business (CRM) relationship. With the increasing influence of CRM Systems, such companies dedicate more time and effort to maintain better customer-business relationships. The effort implied in getting to better know the customer involves the accumulation of very large data bases where the largest possible quantity of data regarding the customer is stored.

Data warehouses offer a way to access detailed information about the customer's history, business facts and other aspects of the customer's behavior. The databases constitute the information backbone for any well established company. However, from each step and every new attempted link of the company to its customers the need to store increasing volumes of data arises. Hence databases and data warehouses are always growing up in terms of number of registers and tables which will allow the company to improve the general vision of the customer.

Data warehouses are difficult to characterize when trying to analyze the customers from company's standpoint. This problem is generally approached through the use of data mining techniques (Palpanas, T., 2000; Silva, D. R., 2002; Han, J., Pei, J., & Kamber, M. 2011; Tan, P. N. 2006, Chaudhuri, S., & Dayal, U. (1997). To attempt direct clustering over a data base of several terabytes with millions of registers results in a costly and not always fruitful effort. There have been many attempts to solve this problem. For instance one may use parallel computation, optimization of clustering algorithms, alternative distributed and grid computing and so on. But still the more efficient methods are unwieldy when attacking the clustering problem for databases as considered above. In this work we present a methodology derived from the practical solution of an automated clustering process over large database from a real large sized (over 20 million customers) company. We emphasize the way we used statistical methods to reduce the search space of the problem as well as the treatment given to the customer's information stored in multiple tables of multiple databases.

Because of confidentiality issues the name of the company and the actual final results of the customer characterization are withheld.

CHAPTER OUTLINE

The outline of the chapter is as follows. First, we give an overview of the analysis of large databases; next we give an overview of the methodology we applied. We emphasize the problem of adequately pre-processing non-numerical attributes so that numerical algorithms are applicable, in general. We describe two possible methods