Chapter 7
Classification with Incomplete Data

Pedro J. García-Laencina
Universidad Politécnica de Cartagena, Spain

Juan Morales-Sánchez
Universidad Politécnica de Cartagena, Spain

Rafael Verdú-Monedero
Universidad Politécnica de Cartagena, Spain

Jorge Larrey-Ruiz
Universidad Politécnica de Cartagena, Spain

José-Luis Sancho-Gómez
Universidad Politécnica de Cartagena, Spain

Aníbal R. Figueiras-Vidal
Universidad Carlos III de Madrid, Spain

ABSTRACT

Many real-word classification scenarios suffer a common drawback: missing, or incomplete, data. The ability of missing data handling has become a fundamental requirement for pattern classification because the absence of certain values for relevant data attributes can seriously affect the accuracy of classification results. This chapter focuses on incomplete pattern classification. The research works on this topic currently grows wider and it is well known how useful and efficient are most of the solutions based on machine learning. This chapter analyzes the most popular and proper missing data techniques based on machine learning for solving pattern classification tasks, trying to highlight their advantages and disadvantages.

DOI: 10.4018/978-1-60566-766-9.ch007

INTRODUCTION

Pattern classification is the discipline of building machines to classify data (patterns or input vectors) based on either a priori knowledge or on statistical information extracted from the patterns (Bishop, 1995; Duda *et al.*, 2000; Jain et al., 2000; Ripley, 1996). This research field was developed starting from the 1960's, and it has progressed to a great extend in parallel with the growth of research on knowledge-based systems and artificial neural networks. Pattern classification has been successfully applied in several scientific areas, such as computer science, engineering, statistics, biology, and medicine, among others. These applications include biometrics (personal identification based on several physical attributes as fingerprints and iris), medical diagnosis (CAD, computer aided diagnosis), financial index prediction, and industrial automation (fault detection in industrial process). Many of these real-word applications suffer a common drawback, missing or unknown data (incomplete feature vector). For example, in an industrial experiment some results can be missing because of mechanical/electronic failures during the data acquisition process (Lakshminarayan *et al.*, 2004; Nguyen *et al.*, 2003). In medical diagnosis some tests are not possible to be done because both the hospital lacks the necessary medical equipment or some medical tests may not be appropriate for certain patients (Jerez *et al.*, 2006; Liu *et al.*, 2005; Markey & Patel, 2004; Proschan *et al.*, 2001). In the same context, another example could be an examination by a doctor, who performs several different kinds of tests; some test results may be available instantly, and some may take several days to complete. Anyway, it might be necessary to reach a preliminary diagnosis instantly, using only test results that are available. Missing data is a subject which has been treated extensively in the literature of statistical analysis (Allison, 2001; Little & Rubin, 2002; Schaffer, 1997), and also, but with less effort, in the pattern recognition literature. The unavailability of the data hinders the decision making processes due to the dependencies of decisions on information. Most scientific, business and economic decisions are somehow related to the information available at the time of making such decisions. As an example, most business evaluations and decisions are highly dependent on the availability of sales and other information, whereas advances in research are based on discovery of knowledge from various experiments and measured parameters. The ability of handling missing data has become a fundamental requirement for pattern classification because inappropriate treatment of missing data may cause large errors or false results on classification. In addition, it is being a more common problem in real-world data. Another clear example of the importance of handling missing data is that 45% of data sets in the UCI repository have missing values, what is one of most used collection of data sets for benchmarking machine learning procedures.

In general, pattern classification with missing data concerns two different problems, handling missing values and pattern classification. Most of the approaches in the literature can be grouped in four different types of approaches depending on how both problems are solved. Figure 1 resumes the different approaches in pattern classification with missing data.

Intuitively the easiest way to deal with missing values is simply deleting the incomplete data. In a multivariate environment missing values may occur on one or more attributes and missing components are often a significant portion of the whole data set, and so, the deletion of these incomplete items may cause a substantial loss of information.

Another approach to handle missing data is to try to estimate the missing data. The process of estimating the missing data components is referred to as imputation. This chapter distinguishes between two different types of imputation methods,