A Survey on Evolutionary Instance Selection and Generation

Joaquín Derrac, University of Granada, Spain

Salvador García, University of Jaén, Spain

Francisco Herrera, University of Granada, Spain

ABSTRACT

The use of Evolutionary Algorithms to perform data reduction tasks has become an effective approach to improve the performance of data mining algorithms. Many proposals in the literature have shown that Evolutionary Algorithms obtain excellent results in their application as Instance Selection and Instance Generation procedures. The purpose of this article is to present a survey on the application of Evolutionary Algorithms to Instance Selection and Generation process. It will cover approaches applied to the enhancement of the nearest neighbor rule, as well as other approaches focused on the improvement of the models extracted by some well-known data mining algorithms. Furthermore, some proposals developed to tackle two emerging problems in data mining, Scaling Up and Imbalance Data Sets, also are reviewed.

Keywords: *Data Reduction, Evolutionary Algorithms, Imbalanced Data, Instance Generation, Instance Selection, Prototype Generation, Prototype Selection, Scaling Up, Training Set Selection*

1. INTRODUCTION

Data reduction (Pyle, 1999) is one of the data preprocessing tasks which can be applied in a data mining process. The main objective in data reduction is to reduce the original data by selecting its most representative information. This way, it is possible to avoid excessive storage and time complexity, improving the results obtained by any data mining application, ranging from predictive processes (classification, regression) to descriptive processes (clustering, extraction of association rules, subgroup discovery).

Data reduction processes can be performed in many ways, some of the more remarkable being:

- Selecting features (Liu & Motoda, 2007), reducing the number of columns in a data set. This process is known as Feature Selection.
- Making the feature values discrete (Liu et al., 2002), reducing the number of possible values of features. This process is known as attribute Discretization.

DOI: 10.4018/jamc.2010102604

- Generating new features (Guyon et al., 2006) which describe the data in a more suitable way. This process is known as Feature Extraction.
- Selecting instances (Liu & Motoda, 2001; Liu & Motoda, 2002), reducing the number of rows in a data set. This process is known as Instance Selection (IS)
- Generating new instances (Bezdek & Kuncheva, 2001; Lozano et al., 2006), which describes the initial data set by generating artificial examples. This process is known as Instance Generation (IG).

This article discusses a wide number of IS and IG proposals. They can be divided into two types of techniques depending on the goal followed by the reduction. If the set of selected or replaced instances will be used as the reference data to instance-based classification, then we refer to Prototype Selection (PS) and Prototype Generation (PG). On the other hand, if the set of instances obtained will be used as input or training set of any data mining algorithm for building a model, then we refer to Training Set Selection (TSS).

In spite of the differences between PS and PG (the first one finds suitable prototypes, while the second one generates them), both have been mainly employed to improve the same classifier, the Nearest Neighbor rule (Cover & Hart, 1967; see also Papadopoulos & Manolopoulos, 2004; Shakhnarovich et al., 2006). This predicts the class of a new prototype by computing a similarity measure (Cunningham, in press) between it and all prototypes from the training set. In the k-Nearest Neighbors classifier, k nearest prototypes vote to decide the class of the new instance to classify. This algorithm is the baseline of the instance based learning field (Aha et al., 1991).

On the other hand, TSS consists of the selection of reduced training sets to improve the efficiency and the results obtained by any data mining algorithm. It has been mainly applied to improve the performance of decision trees, neural networks and subgroup discovery techniques. Although there exists a wide number of TSS approaches, no IG work on TSS has been reported yet, until our knowledge.

In recent years, the data mining community has identified some challenging problems in the area (Yang & Wu, 2006). Two of these are the *Scaling Up Problem* and the *Imbalance Data Sets Problem*. They are closely related to the data reduction field.

The *Scaling Up Problem* (Provost & Kolluri, 1999; Domingo et al., 2002) appears when an overwhelming amount of data must be processed, overcoming the capabilities of the traditional data mining algorithms. The *Imbalance Data Sets Problem* (Chawla et al., 2004; Batista et al., 2004) appears when the distribution of the class in the training data is not balanced, thus the number of instances of some classes is too low. This distribution can cause several problems in the classification of examples which belong to the minority classes.

Evolutionary Algorithms (Eiben & Smith, 2003) are general-purpose search algorithms that use principles inspired by natural genetic populations to evolve solutions to problems. The basic idea is to maintain a population of chromosomes which represent plausible solutions to the problem and evolve over time through a process of competition and controlled variation.

Evolutionary Algorithms have been successfully used in different data mining (Freitas, 2002; Ghosh & Jain, 2005; Abraham et al., 2006) and data reduction (Cano et al., 2003; Oh et al., 2004) problems. Given that the IS problem can be defined as a combinatorial problem, Evolutionary Algorithms have been used to solve it with promising results (Ho et al., 2002; García et al., 2008); these applications of Evolutionary Algorithms to tackle IS problems are usually called EIS (Evolutionary Instance Selection) methods. Furthermore, Evolutionary Algorithms have shown interesting behavior in their application to IG due to it can be defined as a parameter optimization problem (Fernández & Isasi, 2004; Nanni & Lumini, 2008).