Chapter 3.4
Artificial Intelligence Techniques for Unbalanced Datasets in Real World Classification Tasks

Marco Vannucci
Scuola Superiore Sant'Anna, Italy

Valentina Colla
Scuola Superiore Sant'Anna, Italy

Silvia Cateni
Scuola Superiore Sant'Anna, Italy

Mirko Sgarbi
Scuola Superiore Sant'Anna, Italy

ABSTRACT

In this chapter a survey on the problem of classification tasks in unbalanced datasets is presented. The effect of the imbalance of the distribution of target classes in databases is analyzed with respect to the performance of standard classifiers such as decision trees and support vector machines, and the main approaches to improve the generally not satisfactory results obtained by such methods are described. Finally, two typical applications coming from real world frameworks are introduced, and the uses of the techniques employed for the related classification tasks are shown in practice.

INTRODUCTION

When dealing with real world classification tasks it often happens to face problems related to unbalanced datasets. Although there is no prearranged rule for the definition of such datasets, they are characterized by a not uniform distribution of the samples in terms of the *class* variable which is also the one to be predicted by the classifier.

DOI: 10.4018/978-1-60960-818-7.ch3.4

The effect of the class unbalance is, in most cases, very detrimental for the predictive performances of any classifier, in facts most of them, such as decision trees, neural networks and SVM, are designed to obtain optimal performances in terms of global errors (Estabrooks, 2000) thus, as a result, when coping with this kind of datasets they achieve good performance when classifying the most represented patterns while the others are practically ignored. In these cases the classification abilities of the predictors are compromised by several interacting factors. A part from rare cases where patterns belonging to different classes are clearly discernible and samples in the input space are easily separable, the little number of samples corresponding to infrequent events prejudices their correct characterization and makes the separation of the classes difficult for the classifier. Moreover in many real world problems the presence of noise in the data plays as well a detrimental role for the classifiers as it introduces further uncertainties.

Unbalanced datasets concern many real world problems. In the industrial framework malfunction detection databases are often unbalanced as when monitoring industrial processes most observations are related to the normal situations while the number of abnormal ones is represented only by a little percentage. In the same framework, in quality control tasks, the quantity of defective products is much lower than the number of those which have been produced without defects. A similar thing is observed in certain classification tasks in the medical field such as for instance in the diagnosis of breast cancer from the analysis of biopsy images: also in this case the dataset is unbalanced in favor of negative tests. Furthermore in the financial framework the fraud detection belongs to the same set of problems, in facts among the transactions constituting the database to be analyzed for the characterization of frauds a very high percentage of them corresponds to normal situations.

Another aspect to be considered when dealing with these kind of problems is that in certain fields, as those just cited, the rare events correspond to critical situations which should be identified as the different kinds of misclassification errors do not have the same relevance. In facts it is very important to detect a machinery malfunctioning in order to restore a normal situation in the production line by avoiding possible losses in terms of time and money; on the other hand it is not a big problem if a normal situation is misclassified as a malfunctioning, as a *false alarm* is generated, which would only lead to supplementary controls on the machinery without any substantial drawback. Similarly in the medical field the missed detection of a disease could bring to dreadful consequences while a false alarm simply to further medical exams.

Unfortunately most of these critical situations would not be identified by standard classifiers for the previously mentioned reasons, thus, in order to overcome this problem, many methods have been developed. Two different methodological approaches can be distinguished for dealing with unbalanced datasets: the *external* and *internal* ones. Internal approaches are based on the creation of new algorithms expressly designed for facing uneven datasets while the external ones exploit traditional algorithms but with suitably re-sampled databases in order to reduce the detrimental effect of unbalance.

Within this chapter the effect of unbalanced datasets in classification tasks will be described and analyzed; afterwards the main internal and external methods for coping with this problem will be presented and discussed together with some practical examples. Subsequently some case studies taken from real world applications will be described and finally conclusive remarks will be drawn.

CLASSIFICATION TASKS WITH UNBALANCED DATASETS

The detrimental effect of dataset imbalance on the predictive performances of standard classifiers can be observed in most of the datasets affected by such drawback. The performance reduction