Chapter 3[1]

Predictive Classification with Imbalanced Enterprise Data

Sophia Daskalaki
Dept. of Engineering Sciences, University of Patras, Greece, sdask@upatras.gr
Ioannis Kopanas
OTE S.A, Hellenic Telecommunications Organization, Patras, Greece, ikopanas@ote.gr
Nikolaos M. Avouris
Dept. of Electr. and Computer Engin., University of Patras, Greece, avouris@upatras.gr

Abstract: Enterprise data present several difficulties when are used in data mining projects. Apart from being heterogeneous, noisy and disparate, they may also be characterized by major imbalances between the different classes. Predictive classification using imbalanced data necessitates methodologies that are adequate for such data, and particularly for the training of algorithms and evaluation of the resulting classifiers. This chapter suggests to experiment with several class distributions in the training sets and a variety of performance measures, especially those that are known to better expose the strengths and weaknesses of classification models. By combining classifiers into schemes that are suitable for the specific business domain, may improve predictions. However, the final evaluation of the classifiers must always be based on the impact of the results to the enterprise, which can take the form of a cost model that reflects requirements of existing knowledge. Taking a telecommunications company as an example, we provide a framework for handling enterprise data during the initial phases of the project, as well as for generating and evaluating predictive classifiers. We also provide the design of a decision support system, which embodies the above process with the daily routine of such company.

***Key Words*:** Predictive classification, Knowledge discovery from data, Imbalanced datasets, Performance measures, Voting schemes.

[1] Liao, T.W. and E. Triantaphyllou, (Eds.), **Recent Advances in Data Mining of Enterprise Data**, *World Scientific*, Singapore, pp. 147-188, 2007.