Prof. George A. Tsihrintzis

Department of Informatics

University of Piraeus

Piraeus 185 34, Greece

TITLE: Machine Learning-enhanced Software

Abstract:

Classification is a very common supervised machine learning and data analytics task, in which a piece of data needs to be assigned by the learning algorithm to one of a given number of potential classes of origin. More specifically, in classification, the machine is given a set of training samples for each of which the class of origin is known. The machine is then required to learn inductively from the given samples and generalize into a rule for assigning data into classes of origin that allows it to classify samples other than the ones used for training. It is the usual assumption of the binary classification problem that the number of training samples available from one class is comparable to the number of training samples available from the other class. However, it is not uncommon in certain applications for the number of training samples from one class to be significantly higher than the number of training samples from the other class. For example, users of recommender systems are very willing to provide examples (samples) of items they like, but are reluctant to provide samples of items they do not like. Similarly, in a protected system, the number of samples of intruders may be relatively limited, while the number of available samples of allowed/legal users may be quite high. Classification problems with class imbalance arise in nature as well. For example, the immune system in vertebrate organisms needs to be able to discriminate between self cells and other antigens, so as to respond accordingly. A high number of samples from the class of self cells are available to train the immune system. On the other hand, the class of antigens is very broad, including cancer cells, cells from other organisms, molecules and other intruding substances, viruses, bacteria, and parasitic worms. The number of available training samples from the class of antigens is very limited when compared to the size and diversity of this class.

The imbalance in the number of samples from each class affects the performance of traditional binary classifiers. Indeed, in probabilistic terms, classification problems in which training samples from one class are significantly higher in number than training samples from the other class result in significantly uneven prior probabilities of the two classes. The class from which a higher number of samples is available (target class) will have higher prior probability, while the class from which only a limited number of samples is available (outlier class) will have much lower prior probability. In turn, this affects the posterior probabilities of a sample coming from one or the other class. As a result, a binary classifier will erroneously tend to decide more often that an unknown sample comes from the target class than from the outlier class. In recommender system applications, this would mean that the system would tend to recommend items that the user might not like. Similarly, in a protected system, intruders and other threats might not be recognized.

In this presentation, we will discuss machine learning and analytics with extremely-imbalanced data and investigate the applicability of these methodologies in the design of recommender systems that support software systems.