Analyzing Imbalanced Data in Life Sciences: Methods and Alternatives

Professor: Dr. A. Fazel Famili

ffamili@uottawa.ca

Over the last 10-20 years, extensive developments of tools and software systems to design experiments, automatically monitor, collect and warehouse large amounts of data, from applications such as life sciences and industrial processes, have been the prime motivation for an evolving data mining paradigm. This has created several challenges among which are situations where the data is imbalanced. The class imbalance problem corresponds to cases where majority of samples belong to one class and a small minority belongs to the other, which in many cases is equally or even more important. The goal is still to learn from this data and discover models that are useful and unknown to the producers of this data. Most machine learning algorithms are overwhelmed by the majority class and ignore the minority class since the traditional classifiers focus more on minimizing the overall error rate instead of paying special attention to the minority class. This could result in classifying all the data into the majority class in order to achieve higher accuracy. As an example, decision trees tend to over-generalize the class that is represented by most of the examples in the data. This obviously creates a major problem.

To deal with this problem a number of approaches have been studied in the past and have been applied to many domains. In this talk we will provide an overview of some existing methods. We will then introduce a novel approach that is different than almost all existing ones and is based on identifying the inherent characteristics of one class vs the other. We present the results of our studies focusing on real data from life science applications.