Data Mining from Extreme Data Sets: Very Large and/or Very Skewed Data Sets

Lawrence O. Hall
Department of Computer Science & Eng, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, Fl 33620
hall@csee.usf.edu

Abstract

This talk describes an approach to the construction of classiers from imbalanced data sets. A data set is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of *normal* examples with only a small percentage of *abnormal* or *interesting* examples. It is also the case that the cost of misclassifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classier to the minority class. We will discuss a combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than only under-sampling the majority class. Our method of over-sampling the minority class involves creating synthetic minority class examples. Performance is measured using the area under the Receiver Operating Characteristic curve. It is shown that generally a more diverse set of operating points can be found with the combination of over and undersampling of an imbalanced data set. Usually, the best of the true positives with minimal false negatives is found when compared with loss ratios, different classification costs, etc. Details will be provided.

For very large data sets with approximately balanced classes recent work has shown that breaking the data into disjoint subsets, creating classifiers from the subsets and then uniformly voting the classifiers can have surprisingly good results. One can compare this approach on large data sets with bagging on small data sets. It is possible to get bagging like performance in a parallel implementation. This can be contrasted with creating a subset of the random data, learning from the subset, then choosing a new subset to learn on based on the performance of the first subset. This can perform well on small data sets and also medium-size data sets (it is related to boosting), but may be quite expensive for very large data sets. Other possibilities are parallelizing learning algorithms (scaling is an issue), choosing a single strict subset, utilizing multiple classifier types or cooperative distributed learning. A discussion of trade-offs and potentials of the different approaches will be given. Results will be given which show that distributed learning can be effectively applied to very large data sets. In particular, secondary structure prediction of amino acids in proteins can be improved from 78% to 84% accuracy with distributed machine learning.

0-7803-7087-2/01/$10.00 © 2001 IEEE