Selecting Relevant Features from Data Sets with many Variables.

A Proposal for Independent Study in Computer Science, Yale University

Submitted by: Eleanor Wieschaus
Course: CPSC 490
Advisor: Prof. Martin Schultz
Date: 26 September 2003
To be completed: 12 December 2003

Abstract:

The problem of classification has been well examined when the number of variables is small, say, fifty or fewer. Less work has been done as the number of variables increases. Many algorithms shown to be accurate and efficient with few features become noisy and imprecise as the number grows towards the thousands. This has ramifications in the application of many classification problems. Imagine a blood sample taken from a patient that is to be used to determine if he or she has cancer. If the diagnostician does not know which the relevant attributes are, he or she can test for thousands. The results of these tests can then be cross-referenced with a database of other patient’s variable levels and their ultimate diagnoses. The question that arises is the most efficient way to organize the data such that the diagnoses can be made quickly. Machine learning techniques like decision trees, neural networks, genetic algorithms and nearest neighbor can be employed to organize data such that classification is efficient when the number of features is small. The purpose of this project is to study methods for classification when the number of variables is quite large. Generally, the goal is to select the most relevant features from the thousands so that one of the previously mentioned methods can be employed.

Topic:

Machine Learning involves computer programs that automatically improve their performance through experience. The improvement comes either with the addition of new information or adaptation of behavior. Machine Learning methods have been applied to problems such as learning to drive a vehicle, learning to recognize human speech, learning to detect credit card fraud, and learning strategies for game playing. Classification is one of the fundamental problems in machine learning theory. Given $n$ classes of objects and faced with a new, previously unseen object, how does one assign it to one of the classes? The solution to a classification problem is a rule to categorize an object according the values of its features. The question this project seeks to answer is how to determine that rule in a manner that is both accurate and efficient.

As the number of features to be considered increases, the likelihood that some are irrelevant also grows, as does the likelihood of noise. Additionally, the time it takes to run increases exponentially. Thus, a multi-tiered approach to classification involving the elimination of irrelevant features is ideal. Using several steps, one is able to decompose the problem into simpler sub-problems which can be more efficiently solved. The final
step involves utilizing one of many well-established machine learning techniques: decision trees, neural networks, etc. In order to get to that point, several feature-reduction techniques can be involved.

There are three types of feature selection algorithms: exhaustive search, heuristic search and randomized search. Exhaustive search involves considering every possible combination of features and testing its accuracy. It is the only type of algorithm to guarantee the optimal subset as a result. However, with large sets of features, exhaustive search becomes impractical in terms of both space and time. Thus, heuristic search and randomized search are more likely candidates for a practical application. Heuristic search involves finding a greedy algorithm to either add the most relevant attributes or get rid of the irrelevant ones, while randomized search involves testing random subsets of data and using those results to determine a solution in the larger set.

The purpose of this project is to research, implement and test examples of these two types of algorithms. It is, in part, an extension of work begun by Manfred Lau and Professor Martin Schultz in 2002. In his project, Manfred worked to implement a randomized search algorithm he called Random Bagging. Random Bagging is similar to a more common algorithm called Random Forests. In its implementation, a set of decision trees is created using random subsets of the data. Those features that appear more often in the nodes of the trees are more likely to be relevant to classification and are then more likely to be selected. Random Bagging differs from Random Forests in the way it selects the subsets of data to be considered. One of the initial goals of the project will be to test the Random Bagging method and eventually to do so against other implemented algorithms.

The practical applications of classification for large data sets is endless. Besides the diagnosis of medical patients, one could imagine using it to classify and perhaps even remedy environmental issues, predict basketball outcomes, and determine actions to better-satisfy customers in a corporate database. Thus, this project is not about the solution to the specific data-sets involved, but about finding a general method for solving all classification problems where the number of variables is large.

To aid in testing the above, a bell-labs software product, Mirage and Weka, will be utilized. Both are Java-based tool for analysis and visualization of multi-dimensional numerical data. Currently, Mirage the tool shows projected images of points, point classes, or proximity structures in one, two, or higher dimensional subspaces, in views of tables, histograms, scatter plots, parallel coordinate plots, graphs, and trees. It also provides some facilities manual and automatic classification. If its extensibility proves adequate, Mirage will be modified for use in the implementation various methods of classification and feature reduction. Weka is a well-respected data-mining tool that will be extended to implement several classification algorithms.

Goals:

1. To test previously implemented methods for feature-reduction in a Java environment using various data sets.
2. To implement the decision forest method for feature reduction.
3. To compare these methods for efficiency and accuracy with a Java-based software tool, Weka to test methods and compare results.

Meetings:

Meetings will be held regularly with Professor Schultz, and he will also point to relevant research and readings as he sees fit.

Readings:

Readings will, for the most part concern machine learning, feature selection, pattern recognition, and optimization.

Deliverables:

1. 250-to-300 word abstract, describing the project, sent to the DUS.
2. Conference-style paper documenting achievements and conclusions.
3. Set of web pages containing information on the progress of the project and detailing the theory behind, implementation, and outcome the project.
4. Reading list containing all sources used throughout the course of the project.
5. Any code, executables, and resulting data from the project implementation.