

Lecture 14: Biological Data Mining

Paper reference

E.S. Correa and A.A. Freitas and C.G. Johnson. A new discrete particle swarm algorithm applied to attribute selection in a bioinformatics data set. Proc. Genetic and Evolutionary Computation Conference (GECCO-2006). Pages 35-42.

Abstract

This paper addresses the problem of *Feature Selection* in data classification. A classifier models the class of an instance as depending on multiple attributes or features. Noisy or unimportant attributes tend to derail classifiers and hurt classification accuracy. In addition, these attributes also increase computational running time. It is therefore important to select an appropriate subset of *good* features for classification. Most of the existing literature dealt with particle swarm algorithms for attribute selection, which dealt only with continuous variables. Particle swarm algorithms are a family of methods to heuristically search the space of all subsets of attributes to be selected to obtain the *best* set. The methods involve initializing with a certain number of subsets of attributes which are called particles. Each particle is associated with a velocity which is updated depending on its score of *goodness*. The new position of the particle is a stochastic function of the old position and velocity.

This paper extended the idea to discrete variables and proposes a discrete Particle Swarm Optimization (PSO) or DPSO algorithm.

Discussion

The authors compare the performance of DPSO to Binary PSO and find that they manage to get a slight improvement in accuracy over Binary PSO but with a much smaller set of attributes. Both these methods perform much better than the approach of classifying with all attributes. The Naïve Bayes classifier is used to evaluate classification accuracy using the selected set of features and thereby evaluate the *fitness* of the feature set. Naïve Bayes classifiers assuming conditional independence of attributes given the class label, and consequently their performance is adversely affected when this assumption is violated. Specifically, the Naïve Bayes classifier tends to perform poorly with correlated or noisy attributes, and feature selection is known to be very useful for improving its classification accuracy

The method proposed in the paper is reasonably generic and is applied to a postsynaptic dataset which has information from sensors in neurons and the task is to classify a protein into whether or not it exhibits postsynaptic activity.

This method is arguably not too different from the PSO class of algorithms but still has some useful features that enable it to improve over Binary PSO. DPSO differs from PSO since it forces the particles to have a constant number of attributes across iterations. Also the DPSO algorithm does not use a vector of velocities for each particle but instead uses proportional likelihoods for each particle. The particle is updated stochastically based on its current position, the attributes of the local optimum particle in the neighborhood, and the attributes of the globally optimal particle. Using the different treatment of velocities as proportional likelihoods, the updates performed to the current particle, the optimal particle in the neighborhood and the globally optimal particle is done exactly identical to the Binary PSO algorithm.

A disadvantage of this paper is that it does not evaluate the classification accuracy with a discriminative state-of-the-art classifier such as Support Vector Machines (SVMs) or Logistic Regression. Discriminative classifiers are relatively robust to noisy or less important attributes, and comparing accuracy with feature selection using these classifiers would have been a stricter test and also more insightful. The main advantage of this approach is that it helps prune the set of attributes very effectively. Though this might not result in any significant improvement in accuracy with discriminative classifiers, this would certainly help in enhancing the computational efficiency of the classifier.