Perspective

Swarm Tools for Optimization and Feature Selection

Ahmed Husain*

Department of Computer Sciences and Informatics, University of Allahabad, Allahabad, India

DESCRIPTION

Feature selection is a major step in the classification system and is the procedure for selecting a subset of the original features. Feature selection is one of the biggest challenges in text classification. The high dimensionality of the feature space plays an important role in the text classification process, which adds to the complexity of the text classification process. This paper presents a new feature selection method based on particle swarm optimization to improve text classification performance. Particle swarm optimization inspired by social behavior of fish schooling or bird flocking. The complexity of the proposed method is very low due to the use of a simple classifier.

Feature Selection (FS) is used in many areas as a tool to eliminate irrelevant and redundant features. Feature selection simplifies a data set by reducing its dimensionality and identifying relevant features without decreasing the prediction accuracy. The dimensionality of data set is often very large, since learning algorithm might not work as well before removing these irrelevant features. Reducing the number of irrelevant features significantly reduces the running time of a learning algorithm. Feature selection is having many applications, including Text Categorization (TC), data mining, pattern recognition and signal processing. The goal of TC is to automatically assign predefined categories to text documents. This goal is of great practical importance given the sheer volume of online text available through websites, email, and digital libraries. The main issue of TC is the high dimensionality of the feature space. The original functional space contains many unique terms found in text, and even for medium-sized text collections, the number of terms can reach hundreds of thousands. This is very costly for many mining methods. Therefore, it is highly desirable to reduce the feature space without degrading the classification accuracy.

Feature selection approaches

Feature selection is the process of selecting a subset from a feature set. The optimality of a subset of features is evaluated using criteria. Typical feature selection procedures include subset generation, subset evaluation, outage criteria, and result validation. The subset selection process implements a search procedure that selects a feature subset to evaluate based on a particular search method. These search methods include forward selection, backward removal, and forward/backward of the combinations. The process of subgroup selection and evaluation is repeated until certain termination conditions are met. The selected subset of optimal features usually needs to be validated using a different test dataset. Subset feature selection methods can be categorized into filters, wrappers, and embedded approaches. The filter model separates feature selection from classifier learning and selects feature subsets that are independent of each learning algorithm. The wrapper method uses a scoring function to select a functional subset. The scoring function is based on the same learning algorithm used to learn later. In this method, the scoring function calculates the suitability of the functional subset generated by the subset generation procedure, compares it to the previous best candidate, and replaces it if it turns out to be better. Wrappers can generate better solutions, but because they use learning algorithms when evaluating subsets, they are complex to execute and can be decomposed into a large number of functions. When feature selection and learning algorithms are nested, the feature selection process is a kind of embedding process.

Particle swarm optimization for feature selection

Particle swarm optimization is a computational approach that optimizes problems in a continuous multidimensional search space. Particle swarm optimization begins with a random swarm of particles. Each particle is assigned a velocity. The velocity of the particles is adjusted to match the historical behavior of each particle and its neighbors as they fly through the search space. Therefore, the particles tend to move in the direction of a better exploration space.

Particle swarm optimization was originally considered for searching multidimensional continuous spaces. In this paper, it applies to the discrete feature selection problem. Each subset of features can be thought of as a point in the functional space. Sweet spots are the shortest length and the most accurate subset.

The first swarm is randomly distributed throughout the search space, with each particle occupying a position. The goal of particles is to fly to the best position. By passing the time, their position is changed by communicating with each other, and they search around the local best and global best position. Finally, they should converge on good, possibly optimal, positions since they have exploration ability that equips them to perform feature selection and discover optimal subsets.