FEATURE SELECTION USING CLUSTERING ALGORITHMS

Sashikala Parimi, ICFAI Business School
Padmanabha Aital, Narsee Monjee Institute of Management Studies

Feature selection is one of the important aspects of Data mining which is most useful in pattern recognition. Once the data which is in millions and trillions of tuples obtained from the user, the feature selection stage comes. It is used in the removal of redundant data from the highly dimensional data sets leading to reducing the dimensions of the complex data.. In this the features are extracted based upon the various attributes which are selected for the object or class of objects based upon the classifier generated from the trial data (Jianyu et al, 2016). There are two types of algorithms which are used such as supervised and unsupervised feature selection algorithms.

Supervised approach needs the labelled instances to train the classifier. Decision tree, Support Vector Machine, Naïve Bayes etc. are supervised algorithms. Supervised approach has limitations such as labelled instances are rare and difficult to obtain and it forces mapping of instances to one of the known class without detecting new ones.

Unsupervised learning is one of the Machine learning paradigms which is learnt using experience. The main task of this is to recognize the similar patterns in a data set and group them together to form clusters. This task is involved in extracting features from a data sets and classify them according to the similarity of the feature sets. In this the features of the data set cannot be decided in the beginning as the data considered is raw. It is a class where unlabeled instances are used and based on the inner similarity between instances, clusters are formed. K-Means, DBSCAN, Birch etc. are unsupervised algorithms. The clusters formed are homogeneous within the same group and are as much as possible heterogeneous outside the group (Jennifer et al, 2004).

Unsupervised feature selection algorithms can be divided as Filter approaches and wrapper approaches. Filter approaches discover relevant and important features by analyzing the correlation and dependence among features without any clustering algorithms. Wrapper approaches aim to identify a feature subset where the clustering algorithm trained on this feature subset can achieve the optimal value of the predefined.

Due to the outgrowth of the genetic algorithms, the evolutionary algorithms (EA) are developed called the Estimation of Distribution Algorithms (EDAs). This algorithm is used by replacing the crossover and mutation operators with the learning and sampling technique. In this the probability distribution of each of the criterion is determined and the best criterion is selected among the population of data sets at each iteration.

In EDAs, the correlations between different variables are explicitly expressed through the joint probability distribution associated with the individuals selected at each iteration. Hence EDAs are promising methods for capturing the structure of variable interactions, identifying and manipulating crucial building blocks. Since it has certain limitations on non-availability of theory to ensure that the solution is optimum, another technique based incremental learning was developed called Population-Based Incremental Learning (PBIL).

It assumes that all variables considered are independent of each other. It proved to be very successful in solving a variety of real-world problems. It's combined with the mechanisms of a generational genetic algorithm which is far simpler than a GA, and out-performs a GA on large set of optimization problems in terms of both speed and accuracy (Liu & Yu (2005).

REFERNCES

Jianyu., M.L.N (2016) A Survey on Feature Selection, Procedia Computer Science 91:919-926.

Jennifer, G. Dy, Carla E. Brodley (2004) Feature Selection for Unsupervised Learning, *Journal of Machine Learning Research 5*: 845–889.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering, 17*(4), 491-502.