ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2 Issue Special 1 Jan-March 2014
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Feature Subset Selection Algorithm for Large Volumes of Data Based on Clustering

[I]P.Manimaran, [II]M.Selvan
[I]Assistant Professor, Dept. of CSE, K.S.Rangasamy College of Technology, India
[II]PG student, Department of CSE, K.S.Rangasamy College of Technology, India

Abstract
Clustering which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different type of clusters. In the generative clustering model, the form of parametric data generation is assumed, and the main goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the data given the model. The FAST algorithm works in two steps. The first step of the algorithm is, features are divided into clusters by using graph-theoretic clustering methods. The second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. The Features in the different clusters are relatively independent the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, To assume the efficient minimum-spanning tree (MST) clustering method. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between without any labeled training data and with completely labeled training data. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features.

Keywords
Feature subset selection, filter method, feature clustering, supervised learning, semi-supervised learning, graph-theoretic clustering

I. Introduction

The main aim of choosing a subset of good features with respect to the target concepts, the feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy.

Many of the feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features.

II. Preprocess Feature Selection

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. The commonly used as a preliminary data mining practice, the data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

III. Fast Process

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. The efficiency concerns the time required to find a subset of the features, the effectiveness is related to the quality of the subset of features. Based on these criteria, fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this module.

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because 1)The irrelevant features do not contribute to the predictive accuracy and 2) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

IV. Irrelevant Based Feature Selection

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. The efficiency concerns the time required to find a subset of the features, the effectiveness is related to the quality of the subset of the features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated.

Many feature subset selection algorithms, some can effectively eliminate the irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. The proposed FAST algorithm falls into the second group. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the later removes redundant features from relevant ones via choosing representatives from different feature clusters, and it can produces the final subset.

V. Redundant Based Feature Selection

The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. The hybrid method mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. Redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

VI. Graph Based Cluster

An algorithm to systematically add instance-level constraints to the graph based clustering algorithm. Unlike other algorithms which use a given static modeling parameters to find clusters, Graph based cluster algorithm finds clusters by dynamic modeling. Graph based cluster algorithm uses both closeness and interconnectivity while identifying the most similar pair of clusters to be merged. Graph based cluster algorithm works in two phases. The first phase, it finds the k-nearest neighbors based on the similarity between the data points. Then, using an efficient multi-level graph partitioning algorithm sub-clusters are created in such a way that similar data points are merged together. In the second phase, these sub-

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

clusters are combined by using a novel agglomerative hierarchical algorithm. Clusters are merged using Relative Interconnectivity RI and Relative Closeness RC metrics are defined below. Let X, Y are two clusters.

VII. Semi Supervised

Semi supervised learning is closely related to the problem of density estimation in statistics. However semi supervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. The various types of methods employed in unsupervised learning are based on data mining methods used to preprocess data. Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

VIII. Affinity Propagation Algorithm

Affinity Propagation (AP) algorithm can take as input also general non metric similarities. For example, in the domain of image clustering, AP has been used to solve a wide range of clustering problems, such as image processing tasks gene detection tasks, and individual preferences predictions. Clustering algorithm that works by finding a set of prototypes in the data and assigning other data points using a supervised learning and semi supervised process. Input: pair-wise similarities (negative squared error), data point preferences (larger = more likely to be an exemplar) Approximate maximization of the sum of similarities to exemplars Some limited amounts of side information All points sharing the same label should be in the same cluster. Points with different labels should not be in the same cluster.

IX. Conclusion

Clustering algorithm that works by finding a set of prototypes in the data and assigning other data points using a supervised learning and semi supervised process. However semi supervised learning also encompasses many other techniques that seek to summarize and explain key features of the data. Affinity Propagation (AP) algorithm can take as input also general non metric similarities. For example, in the domain of image clustering, AP has been used to solve a wide range of clustering problems, such as image processing tasks gene detection tasks, and individual preferences predictions.

References

[1] *J. Biesiada and W. Duch (2008), "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249.*

[2] *R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici(2005), "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584.*

[3] *S. Das(2001), "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74- 81.*

[4] *M. Dash and H.Liu(2003), "Consistency-Based Search in Feature Selection," Artificial Intelligence, vol. 151, nos. 1/2, pp. 155-176.*

[5] *L. Yu and H. Liu(2004), "Redundancy Based Feature Selection for Microarray Data," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 737-742.*