IMPLEMENTATION OF FEATURE SUBSET SELECTION USING GRAPH BASED CLUSTERING

1PUSHPA GHONGE, 2NEHA KHATRI-VALMIK

1Student, CSE Department, at Dr. Seema Quadri Institute of Tech, Aurangabad, Maharashtra, India
2Assistant Professor, Department of Computer, Dr. Seema Quadri Institute of Tech, Aurangabad, India
E-mail: pushpaghonge@yahoo.in, nehavalmi@gmail.com

Abstract- Feature selection contains identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. A fast clustering-based feature selection algorithm (FAST) is proposed. Features in 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, we adopt the efficient minimum-spanning tree (MST) clustering method.

Keywords- Feature Subset Selection, Feature Clustering, Graph-Based Clustering

I. INTRODUCTION

In this paper selecting a good features with respect to target class, this concept is an effective way for reducing size, removing irrelevant data, increasing correctness and improving the result. We use minimum spanning tree based clustering algorithm. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. Subset selection evaluates a subset of features as a group for suitability. Subset selection algorithms can be broken up into Wrappers, Filters and Embedded. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset.

Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model. We propose a Fast Clustering-based feature Selection algorithm (FAST). The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic Clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Data mining tasks are specified by its functionalities that tasks are classified into two forms:
1. Descriptive mining tasks: Portray the general properties of the data.
2. Predictive mining tasks: Perform the implication on the current data order to craft prediction.

Data mining Functionalities are:
- Characterization and Discrimination
- Mining Frequent Patterns
- Association and Correlations
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis.

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

Research on feature selection has been done for last several decades and is still in focus. Reviews and books on feature selection can be found in. Recent papers such as address some of the existing issues of feature selection. Feature subset selection is an effectual way for dimensionality reduction, elimination of inappropriate data, rising learning accurateness, and recovering result unambiguousness.
Numerous feature subset selection methods have been planned and considered for machine learning applications. They can be separated into four major categories such as: the Wrapper, Embedded, and Filter and Hybrid methods. In particular, we accept the minimum spanning tree based clustering algorithms, for the reason that they do not imagine that data points are clustered around centers or separated by means of a normal geometric curve and have been extensively used in tradition.

II. RELATED WORK

A. Lei Yu, Huan Liu in” Efficient Feature Selection via Analysis of Relevance and Redundancy”- we show that feature relevance alone is insufficient for efficient feature selection of high-dimensional data. We define feature redundancy and propose to perform explicit redundancy analysis in feature selection. A new framework is introduced that decouples relevance analysis and redundancy analysis. We develop a correlation-based method for relevance and redundancy analysis, and conduct an empirical study of its efficiency and effectiveness comparing with representative methods. Yanxia Zhang, Ali Luo, and Yongheng Zhao in” An automated classification algorithm for multi-wavelength data” we applied a kind of filter approach named Relief to select features from the multi-wavelength data.

B. Then we put forward the naive Bayes classifier to classify the objects with the feature subsets and compare the results with and without feature selection, and those with and without adding weights to features.

C. The result shows that the naive Bayes classifier based on Relief algorithms is robust and efficient to preselect AGN candidates. N.Deepika, R.Saravana Kumar in” A Fast Clustering Based Flexible and Accurate Motif Detector Technique for High Dimensional Data” present an algorithm that uses FLAME as a building block and can mine combinations of simple approximate motifs under relaxed constraints.

The approach we take in FLAME explores the space of all possible models. In order to carry out this exploration in an efficient way, we first construct two suffix trees: a suffix tree on the actual data set that contains counts in each node (called the data suffix tree), and a suffix tree on the set of all possible model strings (called the model suffix tree). To get effective and accurate motif detection.

III. EXISTING SYSTEM

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality.

Disadvantages
1. The generality of the selected features is limited and the computational complexity is large.
2. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

IV. PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible.

This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many 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.

Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weights each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

Advantages:
1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

V. FEATURE SUBSET SELECTION ALGORITHM

CONCLUSION

This paper explains about the data mining functionalities and also about the feature subset selection. In this we have explained different methods proposed for feature subset selection. The proposed method is used to extract the features based on clustering. This also provides the implementation details of the proposed algorithm. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

REFERENCES


ABOUT AUTHOR:

Add your details here

Student Details:
Neha Khatri Valmik is currently working as Assistant Professor in the Department of Computer, Dr. Seema Quadri Institute of Tech, Aurangabad, India She has 5 years and 4 months of teaching experience..Her research area include Security, image processing ,Data Structure, Android.

BIOGRAPHY

Student Name -I am pursuing degree in computer Science & engineering from Dr. Seema Quadri Institute of Tech, Aurangabad,My area of interest is java, ASP.net.

***