Improving the Efficiency of Fast Using Semantic Similarity Algorithm

D.KARTHIKA[1], S. DIVAKAR[2]

Final year M.E. [1], Assistant Professor [2]
Department of Computer Science and Engineering
Mail id: karthikadharuman@gmail.com, reachdivakar@gmail.com
Arunai Engineering College, Tiruvannamalai.

Abstract- A Feature selection for the high dimensional data clustering is a difficult problem because the ground truth class labels that can guide the selection are unavailable in clustering. Besides, the data may have a broad number of features and the irrelevant ones can run the clustering. A novel feature weighting scheme is proposed, in which the weight for each feature is a measure of its contribution to the clustering task. A well defined objective function is given, which can be explicitly solved in an iterative way. A fast clustering-based feature selection algorithm (FAST) works in two steps. In the first step, graph-theoretic clustering methods are used to divide the features into clusters. In second step, the most nearest feature that is completely related to destination is chosen from each cluster to design a subset of features. Features in each cluster are independent. Efficiency of FAST is measured by using Minimum Spanning Tree (MST). Accuracy of image comparison is efficient by using semantic similarity algorithm.

Index Terms- Feature subset selection, filter method, feature clustering, graph-based clustering

I. INTRODUCTION

With the aim of selecting a subset of good features is based on target concept. Feature subset selection is an effective measure to reduce dimensionality, eliminate irrelevant data, increasing learning efficiency and improving result accuracy. Four main categories of feature selection algorithms: the Embedded, Filter, Wrapper and Hybrid approaches.

The Embedded methods are a catch-all group of techniques which perform feature selection as part of model construction. It is more dynamic than the other categories. For example tradition machine learning algorithms like decision tree or artificial neural networks.

The Wrapper methods use a predictive model to score feature subset. The goodness of the chosen subsets is determined by the predictive accuracy of predetermined learning algorithms. The generality of the chosen features is limited and large computationally intensive, but it should provide the best performing feature subset. The Filter methods use a sproxy measure instead of error rate to score a feature subset. These methods are independent of learning algorithms, with better generality, Filters are less computationally intensive, but the accuracy of the learning algorithms in not assured. The hybrid methods are a combination of filter and wrapper methods. The wrapper methods are computationally large and applicable for only small training sets. The filter methods are better choice when number of features is huge.

In cluster analysis, graph-theoretic methods are used in many applications. Sometime their outcomes have best agreement with human performance. The graph theoretic clustering is used to evaluate a neighborhood graph of instances. Fast clustering based feature selection algorithm (FAST) works based on MST method. The proposed algorithm improves the performance of different classifiers and reduces the number of features.

II. BACKGROUND AND RELATED WORK

Feature subset selection is a process of discovering and eliminating as many irrelevant and redundant features as much as possible. Reason for this is 1) irrelevant features do not alliteration to the predictive accuracy, and 2) redundant features provide most of the information which is already present in the other features, so it does not redound to getting good predictor. FAST algorithm taking care of both irrelevant and redundant features.

Traditionally, feature subset selection is used to find out the relevant features. A well known example is Relief. Relief is not dependent on heuristics, it requires only linear time of number given features and training instances. For two predictive Relief is highly correlated features but it is not efficient for redundant feature elimination. Proposed some updates in the algorithm which is called as Relief-F. It work with incomplete data set and generalizing it to multi-class problems, but still it cannot find out redundant features. Accuracy and speed of learning algorithm can be affected.

A Correlation Feature Selection (CFS) evaluates subsets of features on the basis of observations that a good feature subset contains features highly correlated with the destination yet uncorrelated with each other.

Fast Correlation Based Filter (FCBF) identifies both relevant and redundant features without pair wise correlation analysis. Different from these algorithms, FAST algorithm works

based on clustering-based method to select features. Hierarchical clustering is a process of word selection method in the context of text classification. Strategies for hierarchical clustering fall into two types: agglomerative and divisive. Agglomerative hierarchical clustering used to remove redundant features. FAST algorithms cluster the features by using Minimum Spanning Tree method.

III. FEATURE SUBSET SELECTION ALGORITHM

A. FEATURE SUBSET SELECTION

Feature subset selection algorithm chooses a best subset from available features. To
Obtain a good feature subset; a novel algorithm can effectively and efficiently deal with redundant and irrelevant features.

Proposed FAST algorithm involves 1) minimum spanning tree is constructed from a weighted complete graph; 2) The partitioning of the MST is a cluster example forest with each tree; 3) the select the features from clusters.

B.FEATURE EXTRACTION

Transfer the input data into the set of feature is known as feature extraction. Feature extraction are chosen, feature set should extract the relevant information from input data set to achieve the desire goal. Large set of data can be described accurately by using simplified amount of resource. It materializes new features from functions of the pristine features, whereas feature selection returns a subset of the features.

C. CONTENT BASED IMAGE RETRIEVAL

Content based image retrieval is a technique which uses visual contents to search images from large scale image databases. The term content refers to context in colors, textures, shapes, or any other information that should be derived from image itself.

- HISTOGRAM: It is a measure used to describe the image and distribution of color brightness across the image.

- REGION BASED: Histogram measure is taken only for local images that is different image regions but not for globally.

Content based image retrieval has two techniques, Query techniques and Semantic retrieval.

- QUERY TECHNIQUES: It is query by example technique which involves providing the CBIR system with an example image then it starts its search. Difficulties will arise when trying to describe images with words; it can be removed by query technique.

- SEMANTIC RETRIEVAL: A user point of view the ideal CBIR system involves where the user makes a query.

CONTENT COMPARISON USING IMAGE DATABASE

An image distance measure is common method for two images comparison in content based image retrieval. Comparison of two similar images in various dimensions such as color, shape, texture measured by image distance measure method.

- COLOR: Color similarity is done by finding a color histogram for every image that identifies the proportion of pixels within an image holding values. Color search involves color histogram comparison.

- COLOR SPACE: Image pixels are represented as point in a 3D color space. RGB, HSV are used as color space for image retrieval. Uniformity is one of the characteristics of color space for image retrieval.

- TEXTURE: Texture measures viewed as visual patterns in image that are spatially defined. It computes brightness of pair of pixels such as degree of contrast, regularity, coarseness and directionality.

- SHAPE: Shape does not refer to the shape of an image but it refers the shape of particular region. Shapes are used to measure the segmentation or edge detection to an image.

SIMILARITY/DISTANCE MEASURES

Instead of exact matching, content-based image retrieval calculates visual similarities between a query image and images in a database. Accordingly, the retrieval result is not a single image but a list of images ranked by their similarities with the query image.

INDEXING SCHEME

CBIR indexing is effective indexing and also fast in searching of images based on visual features.

SEMANTIC SIMILARITY

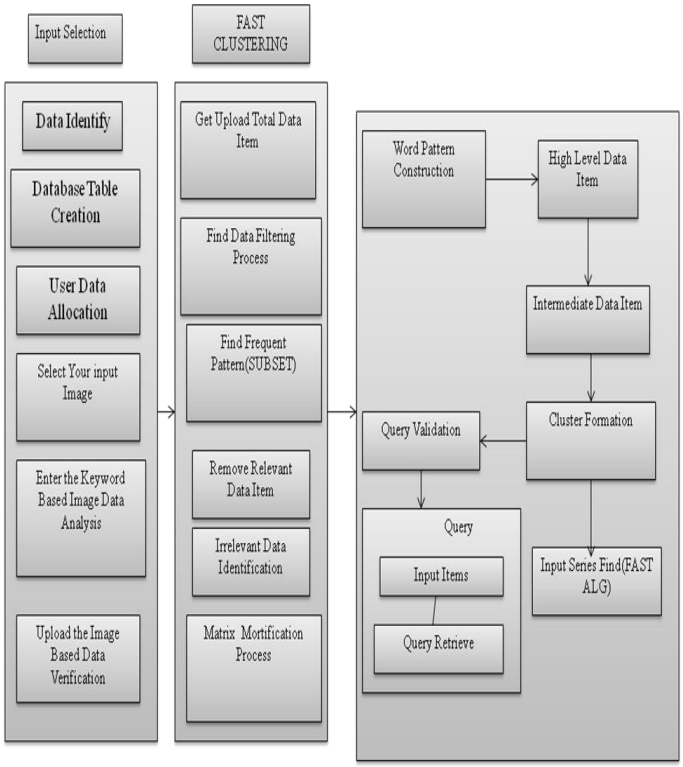
It is a concept whereby terms within the term list or a set of documents are assigned based on their semantic content or meaning. This can be achieved by concentrate on topological similarity. Accuracy can be increased efficiently.

V. SYSTEM DESIGN

A. SYSTEM ARCHITECTURE

The input selection initially identifies the data using which it creates the database table. The database table has allocated with user data memory. The user selected image is the input image. The details of the selected image are saved in the database table created. The keyword is given as a selection criterion the keyword based image data analysis takes place. In this, the images related to the keyword are analyzed for its properties. Uploading the total data item in the database in fast clustering. Data filtering process takes place and the subset patterns are found. Then the redundant and irrelevant data items are identified and removed. Matrix mortification process takes place. The word pattern construction provides the high level data items which produce the intermediate data items for cluster formation and the query is validated. From the query the input

items are analyzed and the relevant output of the query is retrieved.



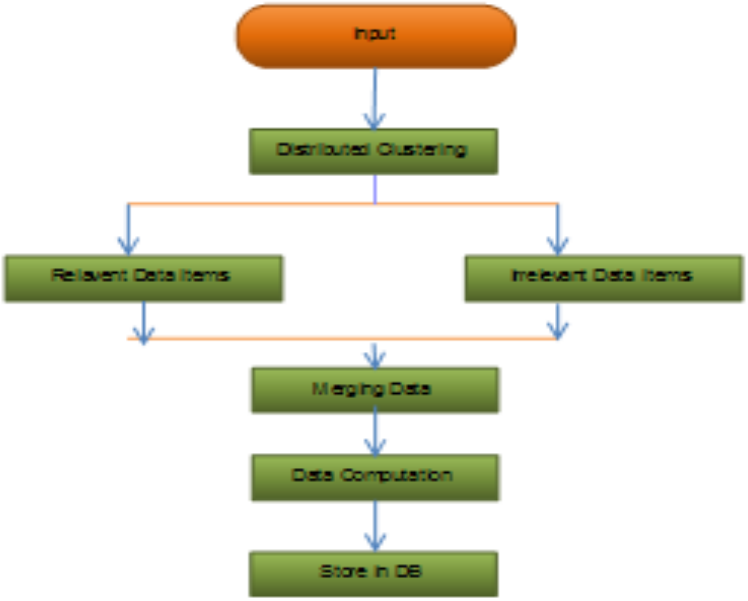
B. MODULE DESIGN

- Distributed clustering
- Subset Selection Algorithm
- Graph Partitioning
- Data Resource
- Irrelevant Feature Removal

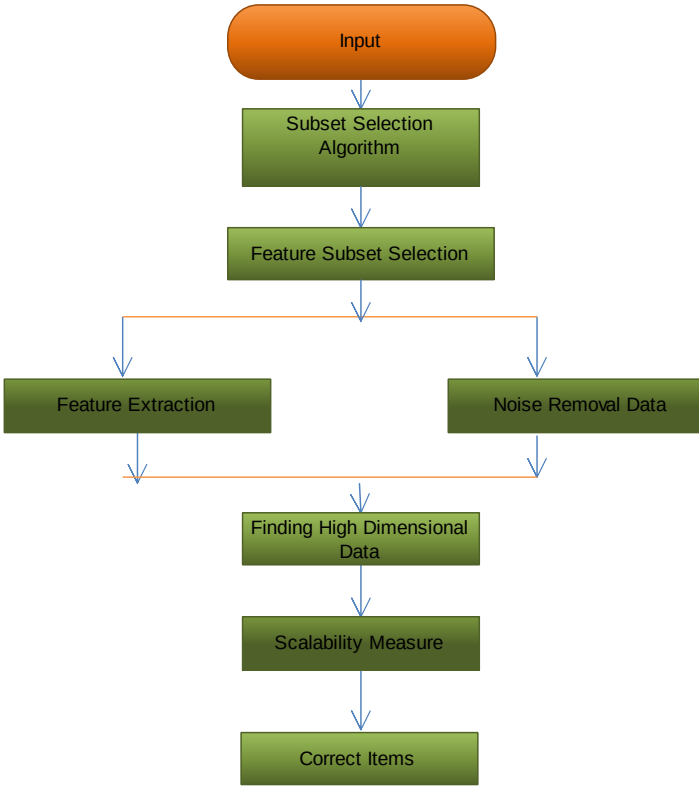
Module Description:

Distributed clustering

Distributed clustering algorithms embrace this trend of merging computations with communication and explore all the facts of the distributed computing environments. Thus a distributed algorithm must take under consideration that the data may be distributed inherently. Distributed version of the algorithm manages to distinguish the number of clusters present in a dataset with satisfactory accuracy in a distributed clustering; the dataset is spread over a number of different types. Thus, let us assume that the entire dataset X is distributed among m, each one string Xi for i = 1; m, so X = S i=1;m Xi . Furthermore let us assume that there is a central one O that will hold the final clustering results. So, the distributed clustering algorithm will form the number of partitions into groups, which helps us to find relevant and irrelevant data.



Subset Selection Algorithm



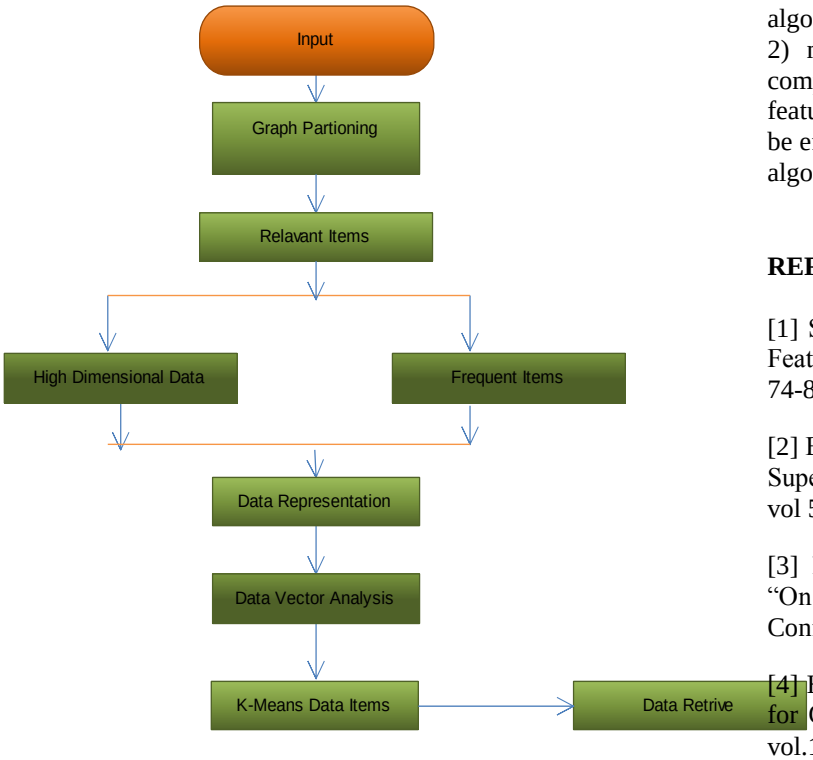
Subset selection is an important pre-processing tool in data mining. It can be an active field in research and development for past three decades. As the datasets are getting bigger both in terms of instances and feature count in the fields of research, customer relationship management, this enormity causes scalability and performance issues in learning algorithms.

Subset selection solves the scalability issue and increases the performance of classification models by eliminating redundant, irrelevant or noisy features from high dimensional datasets. Subset selection is a process of selecting a subset of relevant features by applying certain evaluation criteria.

In general, feature selection process consists of three phases. It starts with selecting a subset of original features and evaluating each feature's worth in the subset. Secondly, using this evaluation, some features in the subset may be eliminated or enumerated to the existing subset. Thirdly, it checks whether the final subset is good enough using certain evaluation criterion. Subset selection can be classified into feature subset selection and feature ranking. Feature ranking calculates the score of each attribute and then sorts them according to their scores. Feature subset selection selects a subset of attributes which collectively increases the performance of the model.

Graph Partitioning

Documents are usually represented as high-dimensional vectors in term space. This method of using graph partitioning for initializing the feature map. For the binary datasets, we can find the frequent associations between features and build the association graph of the features. Then we use the graph partitioning algorithm to partition the features into several parts, i.e., to induce the initial feature map.



Data Resource

Document clustering methods can be mainly categorized into two types: partitioning and Fast clustering. Partitioning method decompose a collection of documents into a given number disjoint clusters which are optimal in terms of some predefined criteria functions. For example, the traditional K-means method tries to minimize the sum-of-squared-errors criterion function. The criteria functions of adaptive K-means approaches usually take more factors into consideration and hence more complicated. The numbers of features of the 35 data sets vary from 37 to 49, 52 with a mean of 7,874. The dimensionalities of the 54.3 percent data sets exceed 5,000, of which 28.6 percent data sets have more than 10,000 features. The 35 data sets cover a range of application domains such as text, image and bio microarray data classification.

Irrelevant Feature Removal

The irrelevant feature removal is a technique if the right relevance measure is defined or selected, while the redundant feature elimination is easy. 1) Minimum spanning tree is constructed from a weighted complete graph; 2) The partitioning of the MST is a cluster example forest with each tree; 3) the select the features from clusters.

Relevant features should be strong correlation with destination. So are always necessary for a best output, while redundant features are not because their values are completely correlated with each other.

V. CONCLUSION

A novel clustering based algorithm can effectively and efficiently deal with redundant and irrelevant features. This algorithm involves 1) irrelevant features are removed efficiently; 2) minimum spanning tree is constructed from a weighted complete graph, 3) partitioning the MST and the select the features from clusters. The result set shows that the accuracy can be efficiently increased in the proposed system than conventional algorithm.

REFERENCES

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

[2] Battiti.R, Using "Mutual Information for selecting Features in Supervised Neural Net Learning" Proc.IEEE Neural Networks, vol 5, 2002

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

[4] Huan Liu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE. Knowledge and Data, vol.17, 2005

[5] S. Garcia and F. Herrera, "An Extension on Statistical Comparisons of Classifiers over Multiple Data Sets for All Pairwise Comparisons," J. Machine Learning Res., vol. 9, pp. 2677-2694, 2008.

[6] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc.IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

[7] J.Demsar,"Statistical Comparison of Classifiers over Multiple Data Sets," J. Machine Learning Res., vol. 7, pp. 1-30, 2006.

[8] Z. Zhao and H. Liu, "Searching for Interacting Features in Subset Selection," J. Intelligent Data Analysis, vol. 13, no. 2, pp. 207-228,2009.

[9] H. Park and H. Kwon, "Extended Relief Algorithms in Instance-Based Feature Filtering," Proc. Sixth Int'l Conf. Advanced Language Processing and Web Information Technology (ALPIT '07), pp. 123-128,2007.

AUTHORS

First Author - Karthika D, M.E CSE, Arunai engineering college, and karthikadharuman@gmail.com .

Second Author – Divakar S, Asst Prof (M.E), Arunai engineering college , and reachdivakar@gmail.com.

Correspondence Author - Karthika D, M.E CSE, Arunai engineering college, and karthikadharuman@gmail.com .