ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A Fast Clustering Based Feature Subset Selection Using Affinity Propagation Algorithm

Mr. M. Senthil Kumar[1], Ms. V. Latha Jothi M.E., [2]

Department of CSE, Velalar College of Engineering & Technology, Erode[1]

Assistant Prof, Department of CSE, Velalar College of Engineering & Technology, Erode[2]

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 clusters, under a particular similarity metric. In the generative clustering model, a parametric form of data generation is assumed, and the goal in the maximum likelihood formulation is to find the parameters that maximize the probability (likelihood) of generation of the data. In the most general formulation, the number of clusters k is also considered to be an unknown parameter. Such a clustering formulation is called a "model selection" framework, since it has to choose the best value of k under which the clustering model fits the data. 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 unsupervised learning (without any labeled training data) and supervised learning (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. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Traditional approaches for clustering data are based on metric similarities, i.e., nonnegative, symmetric, and satisfying the triangle inequality measures using graph-based algorithm to replace this process a more recent approaches, like Affinity Propagation (AP) algorithm can be selected and also take input as general non metric similarities.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

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

Cluster Analysis

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

Data Preprocessing

Data preprocessing is used to improve the efficiency and ease of the mining process. To extract data from the data warehouse that data may be incomplete, inconsistent or contain noisy because data warehouse collect and store the data from various external resources.

Data preprocessing Techniques are:

- Data cleaning: Attempt to fill in missing values, smooth out noise, correct inconsistencies in the data. Cleaning techniques are binning, regression etc.
- Data Integration and Transformation: Merging of data from multiple data sources, these sources may include multiple database, data cubes or flat files. Transformation is the process of consolidate the data into another form, it includes aggregation, generalization, normalization and attribute construction.
- Data Reduction: Techniques can be applied to obtain a reduced version of the data set that is much smaller in quantity but maintains the integrity of original data, which contains following strategies: data cube aggregation, attribute subset selection dimensionality reduction etc.,
- Concept hierarchy Generation: Concept hierarchies can be used to condense the data by collection and replacing low level concept with high level concept.

Feature Selection

Feature selection is similar to data preprocessing technique. It is an approach of identifying subset of features that are mostly related to target model. The main aim is to remove irrelevant and redundant features, it is also known as attribute subset selection. The purpose of feature selection is to increase the level of accuracy, condense dimensionality; shorter training time and enhances generalization by reducing over fitting. 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).

II. RELATED WORK

Minimum-Spanning Tree

A minimum spanning tree (MST) is an undirected, connected, weighted graph is a spanning tree of minimum weight. A tree is an acyclic graph. The idea is to start with an empty graph and try to add edges one at a time, the resulting graph is a subset of some minimum Spanning tree. Each graph has several spanning trees. This method is mainly used to make the appropriate feature subset clustering but it take time to construct the cluster.

Graph Clustering

Graph-theoretic clustering methods have been used in many applications. The general graph-theoretic clustering is simple to compute a neighborhood graph of instances, and then delete any edge in the graph that is much longer/shorter (based on some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. The complete graph G reflects the correlations among all the target-relevant features.

Mutual Information

Consecutive features are grouped into clusters, and replaces into single feature. The clustering process based on the nature of data. This paper shows the information about feature grouping, feature clustering and functional modeling. Select the features by relevance estimation which is calculated using Mutual Information of two variables A and B, it is defined as the uncertainty reduction of B when A is known. In this method, first collect the possible candidate subset from the original data set, and then applies the forward search is used to choose the most relevant features. The process is then iterated until discover the optimal feature subset.

Hierarchical clustering

Hierarchical clustering is a procedure of grouping data objects into a tree of clusters. It has two types: 1) Agglomerative approach is a bottom up approach; the clustering processes starts with each object forming a separate group and then merge these atomic group into larger clusters or group, until all the objects are in a single cluster. 2) Divisive approach is reverse process of agglomerative; it is top down approach, starts with all of objects in the same cluster. In each iteration, a clusters split up into smaller clusters, until a termination condition holds.

Feature selection methods

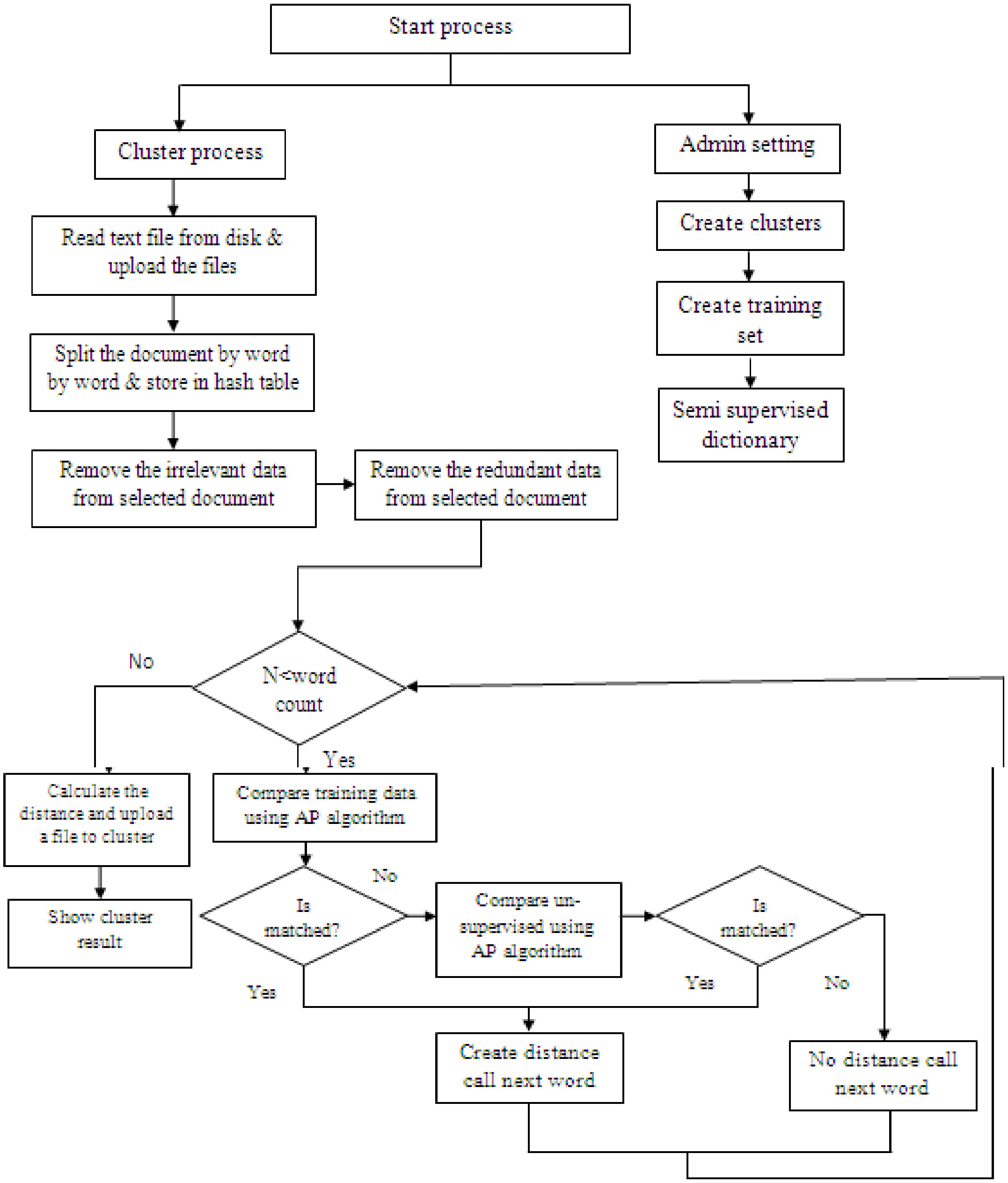
Evaluation functions are used to measure the goodness of the subset. Feature subset selection method is categorized into four types: Embedded, Filter, Wrapper, and Hybrid. Wrapper method is used to calculate the integrity of the selected subset features by using predictive accuracy of machine learning algorithm which provides greatest accuracy of learning algorithms but it has more expensive. Filter is significant choice when the selected feature is very large. It is the independent of learning algorithm and has the low computational complexity. Hybrid method is the integration of filter and wrapper method is worn better performance of learning algorithms.

III. AFFINITY PROPAGATION

Traditional approaches for clustering data are based on metric similarities, i.e., nonnegative, symmetric, and satisfying the triangle inequality measures. More recent approaches, like Affinity Propagation (AP) algorithm can also take input as general non metric similarities. AP can use as input metric selected segments of images' pairs. Accordingly, AP has been used to solve a wide range of clustering problems, such as image processing tasks gene detection tasks, and individual preferences predictions. Affinity Propagation is derived as an application of the max-sum algorithm in a factor graph, i.e., it searches for the minima of an energy function on the basis of message passing between data points.

The proposed system implements, semi supervised learning has captured a great deal of attentions. The system retrieves the data from training data or labeled data and extracts the feature of the data and compare with labeled data and unlabeled data to. 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 unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

IV. SYSTEM FLOW DIAGRAM


Start process
Cluster process
Admin setting
Read text file from disk & upload the files
Create clusters
Split the document by word by word & store in hash table
Create training set
Semi supervised dictionary
Remove the irrelevant data from selected document
Remove the redundant data from selected document
No
N<word count
Yes
Calculate the distance and upload a file to cluster
Compare training data using AP algorithm
Show cluster result
Is matched?
No
Compare un-supervised using AP algorithm
Is matched?
Yes
Yes
No
Create distance call next word
No distance call next word

MODULES

PREPROCESS FEATURE SELECTION

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

IRRELEVANT BASED FEATURE SELECTION

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

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. But FAST algorithm falls into the second group. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

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. They 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.

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. In 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-clusters are combined by using a novel agglomerative hierarchical algorithm. Clusters are merged using Relative Interconnectivity RI and Relative Closeness RC metrics defined.

AFFINITYPROPAGATION ALGORITHM

- Clustering algorithm that works by finding a set of exemplars (prototypes) in the data and assigning other data points to the exemplars.
- 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.

V. CONCLUSION

A novel clustering-based Feature Subset Selection Algorithm for high dimensional data are designed for removing irrelevant features, constructing a minimum spanning tree from relative ones, and partitioning the MST and selecting representative features. A cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

REFERENCES

[1] H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[2] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.

[3] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[4] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.

[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[6] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

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

[8] 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.

[9] C. Cardie, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.

[10] 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.