International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Determining and Exploring Dimensions in Subspace Clustering for Value Decomposition

Saranya Sagambari Devi .S[1]

[1]Sri Jayendra Saraswathy Maha Vidyalaya, College of Arts and Science, Coimbatore-641005, India

Abstract: ***Clustering a large sparse and large scale data is a open research in the data mining. To discover the significant information through clustering algorithm stands inadequate as most of the data finds to be non actionable. Existing clustering technique is not feasible to time varying data in high dimensional space. Hence Subspace clustering will be answerable to problems in the clustering through incorporation of domain knowledge and parameter sensitive prediction. Sensitiveness of the data is also predicted through thresholding mechanism. The problems of usability and usefulness in 3D subspace clustering are very important issue in subspace clustering. Also determining the correct dimension is inconsistent and challenging issue in subspace clustering .In this thesis, we propose Centroid based Subspace Forecasting Framework by constraints is proposed, i.e. must link and must not link with domain knowledge. Unsupervised Subspace clustering algorithm with inbuilt process like inconsistent constraints correlating to dimensions has been resolved through singular value decomposition. Principle component analysis is been used in which condition has been explored to estimate the strength of actionable to be particular attributes and utilizing the domain knowledge to refinement and validating the optimal centroids dynamically. An experimental result proves that proposed framework outperforms other competition subspace clustering technique in terms of efficiency, F Measure, parameter insensitiveness and accuracy.***

Keywords: Information Retrieval, Sub space clustering, Numerical optimization, Data mining

1. Introduction

Clustering aims to find groups of similar objects and due to its usefulness, it is popular in a large variety of domains, such as geology, marketing, etc. Over the years, the increasingly effective data gathering has produced many high-dimensional data sets in these domains. As a consequence, the distance (difference) between any two objects becomes similar in high dimensional data, thus diluting the meaning of cluster [1]. A way to handle this issue is by clustering in subspaces of the data, so that objects in a group need only to be similar on a subset of attributes (subspace), instead of being similar across the entire set of attributes (full space) [2]. The high-dimensional data sets in these domains also potentially change over time. We define such data sets as three-dimensional (3D) data sets, which can be generally expressed in the form of object-attribute-time, e.g., the stock-ratio-year data. In Existing CATSs, which are used to clusters of objects that suggest profits or benefits to users and users are allowed to incorporate their domain knowledge, by selecting their preferred objects as centroids of the clusters. . We propose algorithm Forecasting Technique, which uses a hybrid of SVD, optimization algorithm, and 3D frequent itemset mining algorithm, Principle component analysis to mine Forecasted data which is actionable data in the cluster in an efficient and parameter insensitive way. . We conduct a comprehensive list of experiments to verify the effectiveness of Forecasting technique and to demonstrate its strengths over existing approaches:

The rest of the paper is organized as follows: Sections 2 describes the related work of state of art methods about Subspace clustering and extraction with alignment technique, section 3 describes the overall framework with Methods and solution to achieve the sub space clustering using PCA for determining a actionable data. Section 4 describes the experimental results of our method and performance measures with state-of-the-art methods. Section 5 concludes the paper and outlines possible future work.

2. Related Works

2.1 Data Collaboration based extraction and Content based Prediction

In Big data Analysis, Collaborative filtering approaches are the most popular prediction methods and are widely adopted in Data collaboration based extraction [4]. User-based approaches predict the ratings of active users based on the ratings of their similar users, and item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. Query refinement is another closely related notion, since the objective of query refinement is interactively recommending new queries related to a particular query [5].

2.2 Concept based mining and Click through Data Analysis

Concept based mining model [12][13] has also been utilized in big data community that analyzes terms on the sentence, document, function, dependency level and corpus levels is introduced. The proposed dependency measure takes full advantage of measures on the sentence, document, and corpus and function levels in calculating the dependency range between documents by the importance of dependency discovery, a method for discovering XML functional dependencies. Functional and inclusion dependency discovery is important to knowledge discovery, database semantics analysis and data quality

assessment. In Click through data analysis, the most common usage is for optimizing Web search results or rankings [10], Web search logs are utilized to effectively organize the clusters of search results by learning "interesting aspects" of a topic and generating more meaningful cluster labels.. Besides ranking, click through data is also well studied in the query clustering problem [11]. Query clustering is a process used to discover frequently asked questions or most popular topics on a search engine.

3. Proposed Methodology

Establishing the Indexing of Actionable data for Evolution of Data from Different clusters

A fundamental tool in construction of text classification is a list of 'stop' words (stop word list) that is used to identify frequent words that are unlikely to assist in classification and hence are deleted during pre-processing. Currently, we only remove English stop words (e.g., and, into, or will) as source code is almost exclusively written with English acronyms and comments. Till now, many stop word lists have been developed for English language. Then we use the cleaning filter to remove unnecessary punctuation characters like commas or semicolons at the start or end of the token that might have been inserted at formulas or (for example, name= or 'rech' from an expression like int name='rech' are changed to name and rech). Special characters that represent multiplications, equals, additions, subtractions, or divisions from formulas should have been eliminated in this process (e.g., from a =b * c +d only a, b, c, and d should get through).

Problem Formulation

The main objective of the proposed problem is to predict the user specific Query results state through an optimized clustering for the big data analysis. The linear clustering Suffix tree separates the data, but it maximizes the distance between the given data point to the nearest data point of each class.

The training data set is given by

$$D = \{(x_1, y_1)\Lambda, (x_i, y_i)\Lambda, (x_l, y_l)\}, \quad x \in R^n, y \in \{-1,1\} \tag{1}$$

Where, l – number of training data,

X_i – Training data,
y_i – class label as 1 or -1 for x_i for large data with drifting

A nonlinear function is adopted to map the original input space R^n into N-dimensional feature space of the large dataset.

$$\psi(x) = \varphi_1(x), \varphi_2(x), ..., \varphi_N(x) \tag{2}$$

The separating hyper plane is developed in this N-dimensional feature space. Then the clustering function represented as,

$$y(x) = \text{sgn}(\omega.\psi(x) + b) \tag{3}$$

Where ω - weight vector and b- scalar.

In order to obtain the optimal clustering through ensemble classifier $\|\omega\|$ should be minimized subject to the following constraints

$$y_i[\varphi(x_i).\omega + b] \geq 1 - \xi_i, \text{ i= 1, 2… l.} \tag{4}$$

The variable ξ_i is the positive slack variables, necessary for misclassification of data in different cluster.

Determining a feature Evolution and Feature selection for data classification using ensemble classification

One of the most assumptions of ancient data processing is that knowledge is generated from one, static and hidden perform from the data evolving in the data streams. However, it is hard to be true for data stream learning, where unpredictable changes are likely to eventually happen. Concept drift is said to occur once the underlying function that generates instances changes over time. The Suffix tree clustering is known to be efficient in clustering large datasets. This clustering is one in all the best and also the best far-famed unsupervised learning algorithms that solve the well-known clustering problem in terms large data through the steps of big data community.

The objective function is given in Eq. (5),

$$\min J(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \tag{5}$$

We have,

$$y_i[\varphi(x_i) * \omega + b] \geq 1 - \xi_i \tag{6}$$

where, C – margin parameter, ω - weight vector, x_i - training data, y_i - class label (1 or -1)

for x_i, ξ_i - positive slack variables; $\xi_i \geq 0, i = 1,....,l$, b – scalar, l – number of training data.

Objective function obeys the principle of structural risk minimization in order to obtain the optimal solution with less false positive rate for the data clustered. The objective function in Eqn (5) can be re-modified by following Legrangian principle for the data segmentation and prediction as,

$$L(\omega,b,\xi,a,\gamma)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}a_i(y_i[\varphi(x_i)*\omega+b]-1+\xi_i)-\sum_{i=1}^{l}\gamma_i\xi_i \quad (7)$$

Below equation explains the similarity assignment as follows

$$\text{Where, } a_i \geq 0, \gamma_i \geq 0 (i=1,2,.....,l),\ a_i,\gamma_i - \quad (8)$$

On substituting Eq. (8) in Eq. (7), the dual problem becomes,

$$\max W(a)=-\frac{1}{2}\sum_{i,j=1}^{l}a_i a_j y_i y_j(\varphi(x_i),\varphi(x_j))+\sum_{i=1}^{l}a_i$$

$$\max W(a)=-\frac{1}{2}\sum_{i,j=1}^{l}a_i a_j y_i y_j k(x_i,x_j)+\sum_{i=1}^{l}a_i$$

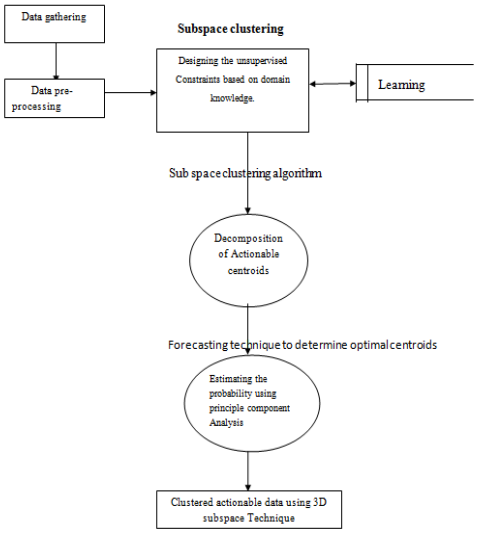

Figure 1: Architecture Diagram of Centroids based Subspace forecasting Clustering

The suffix algorithm aims to partition a group of objects supported their attributes/features, into no. of feature clusters, wherever x may be a predefined or user-defined constant into x clusters

Prediction based on the query preferences and query frequency suggestions

The prediction of the query relevance is calculated based on the query preferences and query frequency of the user or community to the particular type of data. Frequency suggestion is employed through prediction and equivalence of the system in the data evolution and concept drifting in the data streaming in the network to the server.

Determining temporal probability and temporal pattern relevance of data to the query

Temporal probability is carried out the density based clustering technique and its cluster employed through the ranking of the document, temporal pattern relevance is also estimated from the cluster in terms of entropy and Euclidean calculation.

Ranking Based on the Integration Values through following process

Pair wise alignment through Similarity estimation.

Pair wise alignment is carried through the ranking based on the analysis and pairwise alignment is carried out through the algorithm is based on the observation that the data values belonging to the same attribute usually have the same data type and may contain similar strings, especially since results records of the query for the user query.

Nested structure Alignment through user specific clustering

Holistic data value alignment constrains a data value in a Result set to be aligned to at most one data value from another Result set. If a Result set contains a nested structure such that an attribute has multiple values, then some of the values may not be aligned to any other values. Therefore, nested structure processing identifies the data values of a Result set that are generated by nested structures.

4. Experimental Results

In this section, Experimental Results for query based prediction from big data with data evolution and feature evolution were carried out using web data and results were performed with performance system configurations to perform the data scaling and extracting into the proper clusters through suffix tree clustering. Initially extracting the framework has been utilized by training, validation and testing data for classification of results using historical prediction models identify the results set estimation efficiently and effectively in large dataset.

4.1Query frequency estimation and temporal probability estimation

The temporal prediction states observed from the large data set are as follows: supervised data, unsupervised data and semi-supervised data.

- **Feature extraction through user query modeling**

Feature Extraction is employed in large dataset with data drifting and information retrieval with estimating various factors in the query analysis to the large dataset

Feature extraction:

(1) The data in the big data is evolved with several feature classification with novel features estimation in each sample such as, y1, y2, y3, y4 and y5, are extracted by the equation as follows:

$$y_k = \frac{c^k}{\max_{i=1}^{5}(c^i)} \quad (9)$$

where k=1, 2,. . ., 5,
c^k – Absolute feature data per one sample.

(2) The absolute information is calculated for different samples given by,

$$Y_6=\log_{10}\left(\max_{m=1}^{5} c^m\right) \quad (11)$$

Table 1: Parameters of classification and Prediction of data classification

Parameters	*Notations used*	*Values*
Learning rate	Λ	0.01
Scaling factor	Σ	1

Table 2: performance Parameters to compute Data Extraction mechanism

Parameters	*Notations used*	*Values*
Number of iteration	I	15000
Order of the polynomial	Order	3
Scaling factor	Σ	1

4.3 Result Analysis

The proposed framework is implemented and tested using different types of datasets using user specific cluster modeling and multi correlation estimation. An extensive experimental study was conducted to evaluate the efficiency and effectiveness of the proposed methodology on various parameters of benchmark instances and the prediction states are obtained in the graph.

User Specific Clustering has been utilized by the training the data through the analysing the user behaviour in the personalization methods in the literatures. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

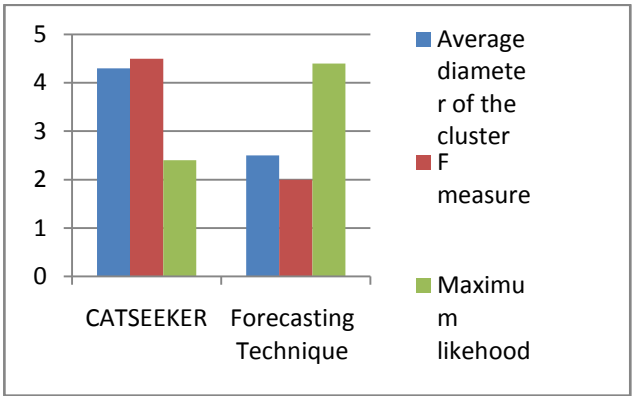

Figure 2: Estimation of the proposed framework against concept based mining

The following parameters are utilized to estimate the performance of the big data classification and prediction of data for user queries.

Minimize average diameter of clusters

This factor estimates the performance of proposed framework in the classifying the data with concept drift .Proposed framework by suffix tree clustering proves the accuracy results set with precision and recall in the cluster achieved.

Maximum likelihood:

It is a method of estimating the parameter of a statistical model. When applied to a data set and given a statistical model, maximum - likelihood estimation provides estimates for the model's parameters.

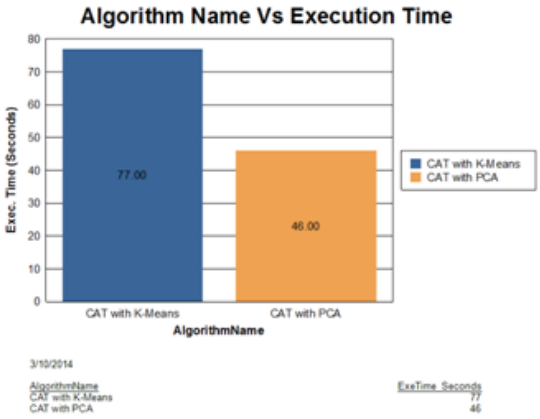

Figure 3: Performance evaluation of the proposed system

We have proved the performance of system in clustering the query based on the several factors included in the framework and experiment to determine the performance factors with better results.

5. Conclusion

We have developed E**xploring dimension in subspace clustering for value decomposition to** Mining actionable 3 D subspace clusters from continuous valued 3D (object-attribute-time) data is useful in domains ranging from finance to biology. But this problem is nontrivial as it requires input of users' domain knowledge, clusters in 3D subspaces, and parameter insensitive and efficient algorithm. We developed and utilized a novel algorithm CAT Seeker to mine CATS, which concurrently handles the multifacets of this problem. In our experiments, we verified the effectiveness of CATSeeker in synthetic data. In financial application, we show that CATSeeker is 70-80 percent better than the next best competitor in the return/risk (maximizing profits over risk) ratio. Hence we conclude that system performs better clustering in terms of precision, recall and f measures of performance factors.

References

[1] A. Bonnaccorsi, "On the Relationship between Firm Size and Export Intensity," Journal of International

Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)

[2] R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)

[3] M. Clerc, “The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization,” In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)

[4] H.H. Crokell, “Specialization and International Competitiveness,” in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)

[5] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, “A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II,” KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)

[6] J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)