International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Principle Component Analysis Based on Optimal Centroid Selection Model for SubSpace Clustering Model

G. Raj Kamal[1], A. Deepika[1], D. Pavithra[1], J. Mohammed Nadeem[1], V. Prasath Kumar[2]

[1]UG Student, [2]Associate Professor,
[1,2]Department of Information Technology,
[1,2]Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Clustering a large sparse and large scale data is an 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. . The Solutions is highly helpful benefit for police departments and law enforcement organisations to better understand stock issues and provide insights that will enable them to track activities, predict the likelihood. 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, Fmeasure, parameter insensitiveness and accuracy.

How to cite this paper: G. Raj Kamal | A. Deepika | D. Pavithra | J. Mohammed Nadeem | V. Prasath Kumar "Principle Component Analysis Based on Optimal Centroid Selection Model for SubSpace Clustering Model" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.379-382, URL: www.ijtsrd.com/papers/ijtsrd31374.pdf

IJTSRD31374

KEYWORDS: *Clustering, Unsupervised Learning, Subspace, Principle Component Analysis, Singular value Decomposition*

1. INTRODUCTION

Streaming data identify the need to mine actionable data through subspace clustering, which are 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. Forecasting algorithm, which uses a hybrid of SVD, optimization algorithm, and 3D frequent itemset mining algorithm to mine actionable data in the subspace of the cluster in an efficient and parameter insensitive way. Clustering can also help marketers discover distinct groups.

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 centroid dynamically. In the continuous iteration, a cluster is split up into smaller clusters[1]. It is down until each object in one cluster or the termination condition holds. The performance of the approach is evaluated with high dimensional datasets. Subspace clustering is the task of detecting all clusters in all subspaces[2]. This means that a point might be a member of multiple clusters, each existing in a different subspace. Subspaces can either be axis-parallel or affine. The term is often used synonymous with general clustering in high-dimensional data[3].

The rest of the paper is organised as follows: section II describes the related work on centroid analysis and machine learning methods; In Section III, design of proposed model using unsupervised learning methods are described and in section IV experimental results among the methods are done. The section V provides the conclusion of the work.

2. Related Work

In this section, various works on subspace clustering has been defined as follows

2.1. Pattern-based subspace clustering

Subspace clusters satisfy some distance or similarity based functions, and these functions normally require some thresholds to set. However, setting the correct thresholds to obtain significant subspace clusters from real-world data is generally a guessing game, and these subspace clusters are usually sensitive to these thresholds. Clustering also requires a global density threshold which is generally hard to set. Subspace clustering is the task of detecting all clusters in all subspaces. This means that a point might be a member of multiple clusters, each existing in a different subspace. Subspaces can either be axis-parallel or affine[4].

2.2. Hierarchical Spatio-Temporal Pattern Discovery and Predictive Modelling

CCRBoost, to identify the hierarchical structure of spatio-temporal patterns at different resolution levels and subsequently construct a predictive model based on the identified structure[4]. To accomplish this, we first obtain indicators within different spatio-temporal spaces from the raw data.

A distributed spatio-temporal pattern (DSTP) is extracted from a distribution, which consists of the locations with similar indicators from the same time period, generated by multi-clustering. Next, we use a greedy searching and pruning algorithm to combine the DSTPs in order to form an ensemble spatio-temporal pattern (ESTP). An ESTP can represent the spatio-temporal pattern of various regularities or a non-stationary pattern to determine the trend of the crime data growing in the specific region on year and state wise visualization of the dataset.

2.3. Prophet model for Predictive Modelling and Visualization

The Prophet model is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and/or daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and cover several seasons of historical data. The Prophet model is robust to missing data and shifts in the trend, and typically it handles outliers well. The Prophet model is designed to handle complex features in time series; it also designed to have intuitive parameters that can be adjusted without knowing the details of the underlying model [5].

3. Proposed methodology

3.1. Data Preprocessing

The high dimensional data considered in form as synthesis dataset as its contains more information with several attributes along huge records in difference time factors to analyse for providing accurate predictions in future cases[5]. It undergo data normalization, missing value prediction and data reduction.

3.2. Singular Value Decomposition

Actionable 3D Subspace clustering undergoes dimensionality reduction due to the high-dimensional and continuous-valued tensor data for difficult and time-consuming process. Hence, it is vital to first remove regions that do not contain CATs. A simple solution is by removing values that are less than a threshold, but it is impossible to know the right threshold[6].

On Data Acquisition, series of preprocessing step has been carried which is as follows

- Time based data is discretized into a couple of columns to allow for time series forecasting for the overall trend within the acquired data to generate effective feature space.
- Missing Value Prediction using random value samples has been carried out the data acquired with missing values.
- Feature Classification is carried out the dataset into categories with deduced attributes.
- Attribute Reduction has been carried out to eliminate irrelevant attributes of the dataset.

Hence, we propose mechanism to efficiently prune tensor in a parameter-free way, by using the variance of the data to identify regions of high homogeneity values. The figure 1 represents the proposed architecture.

3.3. Principle Component analysis

Principle Component analysis is applied towards analyzing and grouping of data is required for better understanding and examination [7]. Strategy for finding the local maxima and minima of a function subject to equality constraints has been identified on basis of the feature of the dataset in the different time frame.

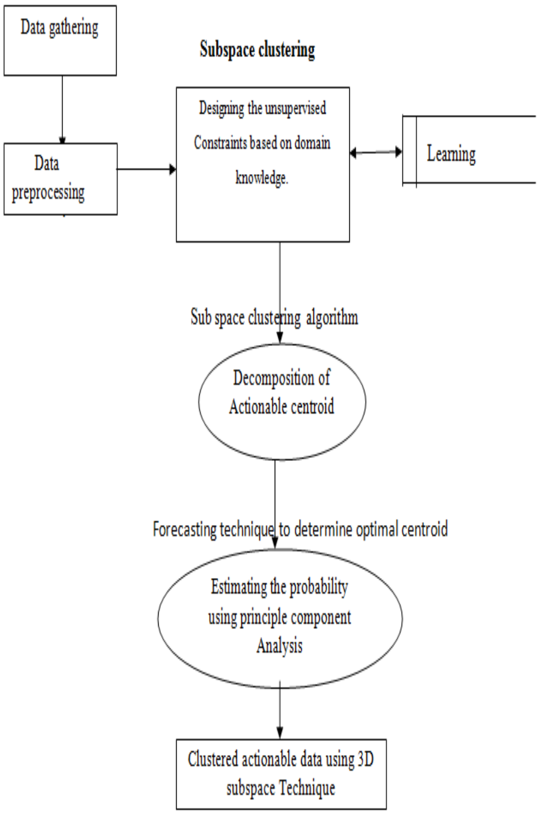
Data gathering
Subspace clustering
Designing the unsupervised Constraints based on domain knowledge.
Learning
Data preprocessing
Sub space clustering algorithm
Decomposition of Actionable centroid
Forecasting technique to determine optimal centroid
Estimating the probability using principle component Analysis
Clustered actionable data using 3D subspace Technique

Figure 1: Architecture of the proposed model

It yields a necessary condition for optimality of the cluster[8].

Algorithm: Optimal Centroid Selection

Input $|\partial| * |\beta| * |\gamma|$ is the 3D Data Cluster

Output: Actionable Data µ

Description:

Calculate the Probability of the Data based on the time constraint "T"

Condition p(T) $=$ max (∂)

Calculate the thresholds Tu and Tl

Use SVD Pruning for dummy value and high changeable data's

if ($\partial \leq$ Tu |Tl)

$\gamma = \mu$

Else

Check the ∂ to the next value

The Prophet model decomposes time series into three main components. That main component has been processed further on basis of Principle Component Analysis (PCA). It is computed using covariance and correlation matrix which is given as

Association Matrix A =

$$= \begin{pmatrix} \operatorname{cov}(a,a) & \operatorname{cov}(a,b) & \operatorname{cov}(a,c) \\ \operatorname{cov}(b,a) & \operatorname{cov}(b,b) & \operatorname{cov}(b,c) \\ \operatorname{cov}(c,a) & \operatorname{cov}(c,b) & \operatorname{cov}(c,c) \end{pmatrix}$$

Matrix value is represented as

$$A_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}\left(d1 - \overline{X}_i\right)\left(dn - \overline{X}_j\right) + Xi$$

Identifying patterns from vast amounts of data streams and identifying members of a predefined class, which is called classification, have become critical tasks Class formation is carried out through classifier which is classified into parametric and nonparametric classifiers. Significant instance are clustered and notified to police department intrinsically prominent in the data.

The class characteristics in terms of changes in the feature space covered by a class will provide the recurring classes prediction

$$F^n = 1 + \frac{Cr}{Pr} + \frac{Cr(n+1)}{Pr}$$

For L > L_m

Region of the feature space defined by the decision boundary of class c just before the class disappeared from the stream.

Calculate the Probability of the Data based on the time constraint "T"

Condition p(T) = max (∂)

```
Calculate the thresholds Tu and Tl
Use SVD Pruning for dummy value and high changeable data's
Extract ()
Feature set = {V1(x, y), V2(x, y)...}
if ($\partial \leq$ Tu|Tl)
Predict ()
$P = \frac{W \pm \sqrt{x^k a^{n-k} - 4c}}{2n}$
$\gamma = \mu$
Else
Check the $\partial$ to the next value.
```

The prediction models employed in this section has capability to learn complex functions and data structures while irrelevant variations are suppressed.hih performance has been achieved in determining the data structure of large scale data.

4. Experimental Results

Experiment analysis has been carried out on efficiency analysis. It broadly represents the different classes of subspace clustering [9]. All experiments were performed using computers with Intel Core 2 Quad 3.0 GHz CPU, 8 GB RAM. In this parameter insensitive and default parameter setting.

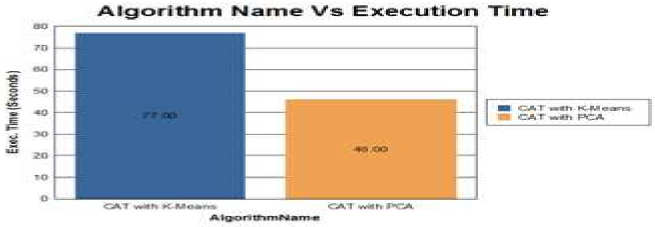

Figure 2: Performance Evaluation of proposed model against existing model

The figure 2 represents the performance of the proposed model. The proposed utility helps to improve the clustering. Centroid-based subspace clustering finds any cluster because it strictly requires a cluster to occur in every time stamp [10].

The different technique employed for Stock prediction on various trend analyses using mining of cluster through approximation with cluster indices of the data interactions.

Table 1: Performance Evaluation of the Proposed model

Technique	Precision	Execution time
Existing	82.23	56ms
Proposed	96.23	37ms

Table 1 provides the performance computation of significant clusters is intrinsically prominent in the data, and they are usually small in numbers.

The Correlation of the Attribute is calculated based on the similarity and distance function using correlation coefficient. The Coefficient measures the correlation between pairs of columns to remove one of two highly correlated data columns. Furthermore, if any earlier process reappears, the data can be handled effectively. It is complex to process the massive data. It Classifies according to their internal relevance on representative features.

The cluster has been employed to identify and realize a trade-off between precision and computation cost values. In this training clusters provides is no deviation on data association on clusters and between clusters. After computing the distance, sorting has to be made in ascending order to extract the results. It uses the Covariance matrix and correlation matrix for similarity computation.

In addition to concept and semantic of the data, features of the data tends to evolve, which can handled using ensemble model.

Conclusion

We have developed forecasting technique to Exploring 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 forecasting using PCA principles to mine subspace data, which concurrently handles the multifacets of this problem.

We also found the PCA based Prophet Model with LSTM algorithm optimal time period for the training sample to be 10 years, in order to achieve the best prediction of trends in terms of Accuracy. Optimal parameters for the Prophet and the LSTM models are also determined.

In our experiments, we verified the effectiveness of PCA in synthetic data. In financial application, we show that forecasting technique 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] K. Wang, S. Zhou, and J. Han, "Profit Mining: From Patterns to Actions," Proc. Eighth Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT), pp. 70-87, 2002.

[2] K. Wang, S. Zhou, Q. Yang, and J. M. S. Yeung, "Mining Customer Value: From Association Rules to Direct Marketing," Data Mining Knowledge Discovery, vol. 11, no. 1, pp. 57-79, 2005.

[3] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," Data Mining Knowledge Discovery, vol. 2, no. 4, pp. 311-324, 1998.

[4] J. Y. Campbell and R. J. Shiller, "Valuation Ratios and the Long Run Stock Market Outlook: An Update," Advances in Behavioral Finance II, Princeton Univ. Press, 2005.

[5] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is 'Nearest Neighbor' Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), pp. 217-235, 1999.

[6] H.-P. Kriegel et al., "Future Trends in Data Mining," Data Mining Knowledge Discovery, vol. 15, no. 1, pp. 87-97, 2007.

[7] B. Graham, The Intelligent Investor: A Book of Practical Counsel. Harper Collins Publishers, 1986.

[8] J. Y. Campbell and R. J. Shiller, "Valuation Ratios and the Long RunStock Market Outlook: An Update," Advances in Behavioral FinanceII, Princeton Univ. Press, 2005.

[9] J. F. Swain and L. M. Gierasch, "The Changing Landscape of Protein Allostery," Current Opinion in Structural Biology, vol. 16, no. 1, pp. 102-108, 2006.

[10] G. Buhrman et al., "Allosteric Modulation of Ras Positions Q61 for Direct Role in Catalysis," Proc. Nat'l Academy of Sciences USA, vol. 107, no. 11, pp. 4931-4936, 2010.

[11] R. Wang, S. Kwong, and D. D. Wang, "An analysis of ELM approximate error based on random weight matrix," International Journal of Uncertainty Fuzziness Knowledge Based System., vol. 21, pp. 1–12, 2013.

[12] Xu, J., Rahmatizadeh, R., Bölöni, L., et al.: 'Real-time prediction of taxi demand using recurrent neural networks', IEEE Trans. Intelligent Transportation System,. 2017, pp. 1–10

[13] X.-Z. Wang, Q.-Y. Shao, Q. Miao, and J.-H. Zhai, "Architecture selection for networks trained with extreme learning machine using localized generalization error model," Neurocomputing, vol. 102, pp. 3–9, 2013.

[14] S. Van Dongen, "Graph clustering via a discrete uncoupling process," Journal of Matrix Analytical. Applications, vol. 30, no. 1, pp. 121–141, 2008.

[15] Y. J. Li, Y. Li, J. H. Zhai, and S. Shiu, "RTS game strategy evaluation using extreme learning machine," Soft Computing., vol. 16, no. 9, pp. 1627–1637, 2012.