International Journal of Engineering Sciences & Emerging Technologies, Oct. 2014.
ISSN: 22316604 Volume 7, Issue 2, pp: 641-645 ©IJESET

CONSTRAINT BASED PARTITIONAL CLUSTERING – A COMPREHENSIVE STUDY AND ANALYSIS

Aparna K[1], Mydhili K Nair[2]
[1]Department of MCA, BMS Institute of Technology, Bangalore, INDIA
[2]Department of ISE, M S Ramaiah Institute of Technology, Bangalore, INDIA.

ABSTRACT

Data clustering is the concept of forming predefined number of clusters where the data points within each cluster are very similar to each other and the data points between clusters are dissimilar to each other. The concept of clustering is widely used in various domains like bioinformatics, medical data, imaging, marketing study and crime analysis. The popular types of clustering techniques are partitional, hierarchical, spectral, density-based, mixture-modelling etc. Partitional clustering is a widely used technique for most of the applications since it is computationally inexpensive. An analysis of the various research works available on partitional clustering gives an insight into the recent problems in partitional clustering domain. In this paper, nine research articles from 2005 to 2013 have been taken for survey in order to analyse the different concepts used for constrained based partitional clustering techniques. Also, a comparative analysis is carried out, to find out the importance of the different approaches that can be adopted, so that any new development in constrained based partitional data clustering can be made easier to be carried out by researchers.

KEYWORDS: *Data Mining, Clustering, K-Means Algorithm, Partitional clustering, Constraint-based partitional clustering.*

I. INTRODUCTION

Data Mining is an integral part of the process of Knowledge Discovery in Databases (KDD). KDD is the overall process of transforming the raw data into useful information. Data mining includes several important tasks such as Association Analysis, Predictive modelling, Clustering, Classification etc., before the useful information is mined from the large repository of the data. Clustering is a division of data into groups of similar objects. From the machine learning perspective, clustering can be viewed as unsupervised learning of concepts. The concept of clustering can be used in order to cluster images, patterns, shopping items, words, documents and so on. Among the different types of clustering techniques available, partitional clustering is one of the most widely used techniques. K-Means and Bisecting K-Means algorithms are the most widely used algorithms under partitional clustering. The above traditional algorithms do not scale well with high dimensional datasets. Hence the performance of the traditional algorithms can be enhanced by incorporating certain constraints. This paper focuses on the analysis and study of the possible constraints that can be applied in order to improve the performance of the traditional partitional clustering algorithms. The reset of the paper is organized as follows: Section 2 discusses the various related literature on constrained based partitional clustering. Section 3 gives a comparison of the various work carried out based on the constrained partitional clustering. The conclusion forms the section 4 of the paper.

II. LITERATURE REVIEW ON CONSTRAINT BASED CLUSTERING

Under this subcategory, nine research papers are taken for survey.

In some applications, the pairwise constraints can be collected automatically along with the unlabeled data. The pairwise constraint was incorporated with recently designed Maximum Margin Clustering (MMC). Hong Zeng and Yiu-Ming Cheung in [1] have focused on semi-supervised clustering algorithm with pairwise constraints. It uses the maximum margin principle adopted in supervised learning and also tries to find the hyper planes that partition the data into different clusters with the largest margins between them over all the possible labellings. The clustering result with basic MMC was not satisfied and the result of developed constrained MMC showed that the incorporated pairwise constraints improved the performance of the basic MMC. Also by Constrained Concave-Convex Procedure (CCCP), a sequence of convex quadratic optimization problems was solved. The experimental results of the constrained MMC algorithm showed that it is efficient, scalable, and outperforms the existing constrained MMC.

To overcome the drawbacks of clustering spatial data, V.Pattabiraman *et al* in [2] have discussed a novel spatial clustering algorithm using K-Means, K-Mediods, IKSCOS and GKSCOS etc. EDBSCOC (Edge Detection Based Spatial Clustering with Constraints) for clustering spatial images was designed with the obstacle and facilitator constraints. It was designed to cluster the spatial data with the constraints and it also compared the result with the various constraint based clustering algorithms in terms of number of clusters and its execution time. The attention for higher speed and stronger global optimum search was provided by the Edge detection based K-Mediods algorithms and also it gets down to the obstacles and facilitator constraints and practicalities of spatial clustering. The experimental result on real datasets showed that the EDBSCOC algorithm performs better than the IKSCOC (Improved Spatial Clustering with Obstacles Constraints based on K-Mediods) and GKSCOC (Spatial Clustering with Obstacles Constraints based on Genetic Algorithms and K-Mediods) in terms of execution time. Compared to other algorithms, the designed EDBSCOC algorithm calculated more number of similar objects for a given period of time.

Due to the several drawbacks of the existing clustering algorithms, Di Wang *et al*, [3] have designed an output-constrained clustering algorithm for system identification. Gaussian membership function was used to represent a nonlinear subsystem. To obtain a globally optimal system, the training data were first clustered and then each cluster was considered to be an initial fuzzy rule for a fuzzy system, which was then further adjusted and refined. The main objective of the clustering analysis was to have an appropriate and efficient subsystem and also to keep the total number of initial fuzzy rule sets small and efficient. Separability was the key concept of clusters within each output constraint. It automatically found an appropriate small and efficient number of clusters for each output constraint. The designed clustering method was unlike most existing clustering algorithms for structure identification of fuzzy systems, where the focus was on input or combined input–output clustering.

Xueping Zhang *et al*, in [4], have discussed about the spatial clustering with obstacles constraints. To the best of their knowledge, only three clustering algorithms such as, COD-CLARANS based on the Partitioning approach of CLARANS, AUTOCLUST+ based on the Graph partitioning method of AUTOCLUST, and DBCluC based on the Density-based algorithm were used before for clustering spatial data with obstacle constraints. To overcome the drawbacks of partitional based clustering methods, they designed a novel Spatial Clustering with Obstacles Constraints based on GAs and K-Medoids, called GKSCOC. It was designed to cluster spatial data with obstacle constraints. The comparison proved that their method could not only give attention to higher local constringency speed and stronger global optimum search, but also get down to the obstacles constraints and practicalities of spatial clustering. The experimental result on real data set showed that GKSCOC method was better than standard GAs and K-Medoids in clustering spatial data with obstacles constraints.

To overcome the drawbacks of Subspace clustering, Xianchao Zhang *et al*, in [5] have addressed the importance of feature correlation and distance divergence. The feature correlation can be used to reduce the search space for the relevant features and distance divergence helps to improve the accuracy of the search. By the integrated solutions of the dimension correlation and distance divergence problems, a semi-supervised subspace clustering algorithm called CDCDD (Constraint based Dimension Correlation and Distance Divergence) was designed. The CDCDD algorithm was designed from previous unsupervised subspace clustering algorithm called FINDIT (**F**ast and **I**Ntelligent subspace clustering algorithm using **D**imension voting, **I**nformation and software **T**echnology). The must-links and cannot-links were the two pair-wise constraints used here to

overcome the drawback of subspace clustering at high dimensional data. The performance of the CDCDD algorithm was demonstrated by experimenting it on both synthetic data sets and real data sets. The result showed that the designed CDCDD algorithm outperformed FINDIT in terms of accuracy and also outperformed the other constraint based algorithm called SC-MINER in terms of both accuracy and efficiency.

Balancing constraints can be viewed as a special case of size constraints where all the clusters have the same size. Balanced constraints are helpful in generating more meaningful initial clusters and avoiding outlier clusters. Shunzhi Zhu *et al* in [6] have extended the balancing constraints to size constraints, i.e., based on the prior knowledge of the distribution of the data, the size of each cluster is assigned to find a partition which satisfies the size constraints. To transform size constrained clustering problems into integer linear programming optimization problems, a heuristic algorithm was designed. They also included the Instance-level cannot-link constraints into their designed size constrained clustering. The improved clustering performance was showed by the experimental result on UCI data sets.

The limitation of non parametric kernel learning is the difficulty in linking the non parametric kernels to the input patterns. In a nonparametric setup, the kernel matrix was learned and also it was often convenient to cast the kernel learning problem into an optimization problem, which included the entire kernel matrix as a single variable. So, the dependency of kernel matrices on the input patterns of examples was difficult to be explored directly within this framework. In order to solve this issue, Steven C. H. Hoi *et al*, in [7] have designed a novel algorithm for nonparametric kernel learning. They constructed a graph called Laplacian matrix based on the pair wise similarities measured between any two examples. The Laplacian matrix was then used to regularize the nonparametric kernel learning. The Semi-Definite Programming (SDP) problem was solved in this paper with a large number of unlabeled data. The experimental result on clustering with pairwise constraints showed that the nonparametric kernel learning method was more effective than other state-of-art kernel learning techniques.

The development of scalable clustering algorithms to satisfy balancing constraints on the cluster sizes was an important problem addressed by Arindam Banerjee and Joydeep Ghosh in [8]. To overcome this problem, a general framework for scalable balanced clustering was designed. The designed method was divided into three steps which were, sampling, clustering of the sampled set and populating and refining the clusters to satisfy the balancing constraints. They showed that to obtain representative subset with high probability, a simple uniform sampling from the original data was sufficient. They focused on some popular parametric algorithm and designed algorithms to populate and refine the clusters. The experimental results showed that the sampled balanced algorithms performed competitively, and often better, compared to a given base clustering algorithm which was a run on the entire data.

Decision function, generally, was computed using Support Vector Clustering (SVC) algorithm which transformed the data into a high dimensional feature space. Dragomir Yankov *et al*, in [9] have used a Mixture of Factor Analyzers (MFA) to improve the performance of SVC in the case of Gaussian distributed noise and also to obtain better control over the number of detected clusters, they explored the density variability of the data in very small regions. The mixture model was analyzed and the points that deviated from the main trajectory of the data were detected. The information about those locally deviated points was used to determine the soft margin tradeoffs between the outliers and the accuracy of the one-class Support Vector Machine (SVM) learner. The regularization results in smoother contours were shrunk towards the dense regions in the data, instead of trying to accommodate all outliers. The subsequent clustering often allowed for easier interpretation too. Because of the local dimensionality reduction performed by MFA and the nonlinear feature map, the "locally constrained" SVC method was further demonstrated to correctly identify the topological structure of the data, when the clusters reside on a lower dimensional nonlinear manifold.

III. COMPARATIVE ANALYSIS

The Table 1below gives a comparative insight into the various approaches incorporated based on the constraint based partitional clustering. The advantages and disadvantages of each of the approaches are also discussed.

Table 3. Comparison of the papers considered for Constraint Based Clustering

Paper Details	Techniques	Advantages	Drawbacks
1. Hong Zeng and Yiu-Ming Cheung (2012)	Constrained maximum margin clustering (MMC)	Effectively improved the baseline MMC	Pairwise constraints were provided earlier
2. V.Pattabiraman, R.Parvathi, R.Nedunchezhian and S.Palaniammal (2009)	EDBSCOC (Edge Detection Based Spatial Clustering with Constraints).	Computational time was very less.	-------
3. Di Wang, Xiao-Jun Zeng and John A. Keane (2011)	Output-Constrained Clustering method	-----	Not effective when data were not evenly distributed in the output domain
4. Xueping Zhang, Jiayao Wang, Fang Wu, Zhongshan Fan and Xiaoqing Li (2006)	GKSCOC	Higher local constringency speed, stronger global optimum search and practicalities of spatial clustering	Slower clustering speed
5. Xianchao Zhang, Yao Wu and Yang Qiu (2010)	CDCDD (Constraint based Dimension Correlation and Distance Divergence)	The increasing number of constraints reduced the computational time	Difficulty in clustering high dimensional data
6. Shunzhi Zhu, Dingding Wang and Tao Li (2010)	Heuristic algorithm	-----	Various constraints were required for better result
7. Steven C. H. Hoi, Rong Jin and Michael R. Lyu, (2007)	NPK-SMO algorithm (Non-Parametric Kernel – Sequential Minimal Optimization)	It remained unchanged when the number of examples were increased	Extensive evaluation was required
8. Arindam Banerjee and Joydeep Ghosh, (2006)	Scalable clustering algorithm	-----	High dimensional data clustering was difficult
9. Dragomir Yankov, Eamonn Keogh and Kin Fai Kan, (2007)	"Locally constrained" SVC (LSVC)	Interpretation was easier in subsequent clustering	-----

IV. CONCLUSION

A detailed survey of various constrained based partitional clustering methods is presented here. It can be inferred from the above set of comparisons that the incorporation of different types of constraints enhances the performance of the various algorithms used for different applications. The incorporation of the constraints also has overcome the disadvantages of the previous methods applied for the same applications. But many of them are computationally expensive and time consuming. This can be further improved by optimizing the methods by using some traditional optimization techniques.

REFERENCES

[1] Hong Zeng, Member and Yiu-Ming Cheung, "Semi-Supervised Maximum Margin Clustering with Pairwise Constraints", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, 2012.

[2] V.Pattabiraman, R.Parvathi, R.Nedunchezhian and S.Palaniammal, "A Novel Spatial Clustering with Obstacles and Facilitators Constraint Based on Edge Deduction and K- Mediods", International Conference on Computer Technology and Development, 2009.

[3] Di Wang, Xiao-Jun Zeng and John A. Keane," An Output-Constrained Clustering Approach for the Identification of Fuzzy Systems and Fuzzy Granular Systems", IEEE Transactions on Fuzzy Systems, Vol. 19, No. 6, December 2011.

[4] Xueping Zhang, Jiayao Wang, Fang Wu, Zhongshan Fan and Xiaoqing Li, "A Novel Spatial Clustering with Obstacles Constraints Based on Genetic Algorithms and K-Medoids", Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, Vol.1, pp. 605 – 610, 2006.

[5] Xianchao Zhang, Yao Wu and Yang Qiu, "Constraint Based Dimension Correlation and Distance Divergence for Clustering High-Dimensional Data", In Proceedings of 10th International Conference on Data Mining, pp. 629-638, 2010.

[6] Shunzhi Zhu, Dingding Wang and Tao Li, "Data clustering with size constraints", Knowledge-Based Systems, Vol. 23, No. 8, pp. 883–889, 2010.

[7] Steven C. H. Hoi, Rong Jin and Michael R. Lyu, "Learning Nonparametric Kernel Matrices from Pair wise Constraints", In Proceedings of the 24th International Conference on Machine Learning, pp.361-368, 2007.

[8] Arindam Banerjee and Joydeep Ghosh, "Scalable Clustering Algorithms with Balancing Constraints", Data Mining and Knowledge Discovery, Vol.13, No. 3, pp. 365-395, 2006.

[9] Dragomir Yankov, Eamonn Keogh and Kin Fai Kan, "Locally Constrained Support Vector Clustering", In Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp. 715 – 720, 2007.

AUTHORS

Aparna K is working as Associate Professor in the Dept. of MCA at BMS Institute of Technology, Bangalore, affiliated to Visvesvaraya Technological University, Belgaum. She has received B.Sc degree in Computer Science from Bangalore University, MCA from VTU, Belgaum and M.Phil degree in Computer Science from Bharathidasan University, Tiruchirappalli and currently pursuing Ph.D in Computer Science at VTU in the area of Data Mining. She is a life member of CSI and ISTE and she has presented and published various papers in different national and international conferences. Her areas of research include Data Mining, Information Retrieval and MANET protocols.

Mydhili K Nair was a project manager in one of the major IT Company, and is now working as an Associate professor in MS Ramaiah Institute of Technology, Bangalore, India. She is an accomplished individual, with a doctorate in Philosophy to her credit in the field of Information and Communication Engineering. She has presented a paper on Intelligent Agents and Multi-Agents, IAMA 2009 in IEEE which also has been given the "Best paper award". She is currently doing her research in Cloud computing and in computer networks.