An Efficient Clustering Technique for Cluster Extraction from Unlabeled Datasets Using Nonlinear Methods

Satish Kumar Soni, Ramjeevan Singh Thakur, Anil Kumar Gupta

Abstract— Clustering is an important task in machine learning to identify the unique groups within the data, based on some similarity measures. In this paper we are trying to study the effect of Nonlinear Methods to optimize clustering results and based on the findings thereafter we proposed a clustering optimization technique to further improve the quality of clusters experimented in Educational and Iris Datasets.

Index Terms— Nonlinear Methods, Clustering, k_means, sk_means, optimization.

1 INTRODUCTION

"The Next Cold War Is Here, and It's All About Data" the article published in a website [5] shows the importance of data and data based applications for future technological development. As according to [6] the 90 percent of the data created in past two years nearly 2.5 quintillion bytes of data produced daily. Such a huge amount of data generation, handling and processing is a challenging task for the data analyst as well as researchers. As data is increasing the demand of new and improved techniques to process that data is also increasing. Unsupervised machine learning or clustering is all about to find the hidden patterns in newly generated datasets, for that there is always the strong need of optimized clustering techniques [4]. The k-means is one of the classical yet simple and robust clustering technique to find the groups on data. This technique uses distance functions to identify the similar points in a cluster. If there are linearly separable boundaries can be identified in data, then this methods convergence rate is high but if nonlinear multidimensional feature space exists in data then the convergence rate might slow down and needs optimization of parameters [1]. Nonlinear numerical methods can be used to solve the multidimensional feature space problems [8]. In this we will represent the datasets as a nonlinear function with finite intervals [9] which further approximates the parameters to find cluster centers. In this paper an optimization method using numerical methods for k-means to improve the performance and cluster quality of clustering named SK-means clustering method is proposed.

- *Satish Kumar Soni, PhD (pursuing) in Computer Sciences, Barkatullah University, Bhopal. MCA from RGPV University, Bhopal, BSc (Statistics, Mathematics, Physics) from APS University, Rewa, currently teaching MSc Students of CS&IT in Computer Science & Application Dept. B.U., Bhopal. Research areas are Data Mining, Machine Learning, Pattern Recognition and Mathematical Modeling.*
- *Dr. Ramjeevan Singh Thakur, Associate Professor(MCA), MANIT, Bhopal, Teacher, Researcher and consultant in the field of Computer Science and Information Technology, Ph.D. (Computer Science) from RGPV Bhopal, areas of interest include Data Mining, Data Warehousing, Web Mining, Text Mining, and Natural Language Processing, member of the CSI, IEEE, ACM, IAENG, ISTE, GAMS and IACSIT.*
- *Dr. Anil Kumar Gupta, Head, Computer Science & Applications Dept., Barkatullah University, Bhopal, PhD in Computer Science, Active Researcher in the areas of Data Mining, Artificial Intelligence and Machine Learning.*

2 PREVIOUS WORK

Finding the natural groups in the new data is always a challenging task among the researchers and data analysts. Several methods have been developed during the years of study with several optimization techniques but no single method can be used for every dataset. Some analytical and numerical methods have been used to optimize the existing methods to improve the quality and accuracy. In [10] the authors have presented Cluster Number Density-Exact Method and Cluster Number Density-Numerical Method to find the distance among short inter cluster distances. The functional nature of instances is presented in [11] herein the clustering method involve two phases: fitting the functional data by B-splines and subdividing the estimated model coefficients using a k-means algorithm, this study depicts the functional clustering having optimizations. The authors in [12] has given the exact and approximate methods for clustering which solves the k-center problem during the grouping of points. In [13] authors study on survey of numerical methods for grouping gives the in depth site of methods can be used. [14] shows the use of hyperbolic smoothing functions in clustering problem which apply special differentiable class functions. In [15] author gives the insight about how finite element method in clustering of time series data can be used with averaged clustering function. In past years the data type was the major issue but in last decades, dimensionality and multi-variability is also included in the key issues. Every single passing day creates huge amount and verity of data poses the big challenge for prediction and pattern recognition needs better optimization so that method works for requirements.

3 PROPOSED WORK

In clustering we separate the data points in a finite number of groups either in similarity basis or dissimilarity basis. Let us given a dataset R= {x1, x2, x3……..., xn} and the number of clusters as K= {k1, k2, k3……, kn} then we need to evaluate a function like

$$g: \mathcal{R} \rightarrow \{k1, k2, k3, \ldots\ldots kn\} \quad (1)$$

where k indexes cluster centers, g is the clustering function. There are many clustering algorithms developed by the researchers of which the k-means [3] is state of the art method

to find initial clusters. K-means got lot more variations for varying datasets and applications. If it is not viable to represent the datasets as linear functional problems, then the non-linear functional solution for the data will be explored. But if there exists a non-linearly separable structure in data, it fails to converse to the desired point. To overcome this limitation k-means with nonlinear optimization is applied. In this approach before clustering the data instances are mapped into a polynomial feature space by a nonlinear function and then, k-means clustering is performed on that feature space. In this paper we are using Bisection and Newton methods to optimize the multidimensional feature space [1]. The clustering function g defined by R is used to assign the points to the neighboring centers described as:

$$g(xi) = \arg.\min ||\emptyset(xi) - \vartheta i||^{2} \quad \textbf{(2)}$$

Where Ø(xi) is the nonlinear function and the ui is the centroids, in each iterations the value of g(xi) and ϑi is updated till the function converses to the solution which satisfies the optimization criteria.

Sloped K-means(SK-means) algorithm:

1. Define the number of clusters.
2. Find two points, say m and n where m < n and g(m) * g(n) < 0
3. Find the center interval of m and n, call it p.
4. p is the approximate value of function if g(p) = 0, else.
5. Divide the interval (m, n) to fined midpoints further. If g(p) * g(n) < 0, if m = p, else if g(p) * g(m) < 0 then say n = p these are called the bracketing methods.
6. Repeat from 2 to 5 steps until g(p) = 0 means the centroid value obtained.
7. Select initial points as cluster centers randomly.
8. Assign each instance to the closets centroids calculated by some distance function.
9. Calculate mean of each cluster instances and reassign as centroids.
10. Repeat steps 2, 3 and 4.
11. Final clusters.

For optimizing the k-means clustering the Bisection Method [2] is applied for the nonlinear function obtained from multidimensional feature space to approximately find the cluster centers with deferring intervals. Figure 1 describes the proposed SK-means method of clustering with optimization.

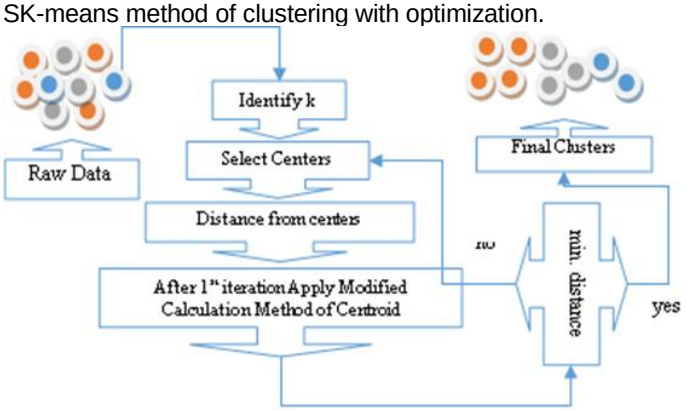

Figure 1 Process diagram of proposed SK-means clustering method.

4 Datasets & Experimental Setup

For experimentation we use Education dataset and Iris dataset downloaded from UCI web repository available for public use for research [7]. Education dataset contains 145 instances and 5 attributes of numerical types. Iris dataset contains 145 instances with 4 attributes of real type. Both of these datasets are multidimensional and nonlinear nature [7]. A glimpses of datasets are given in the table 1 and table 2. Table 1 shows the first 50 instances of education dataset, the details of the attribute type and names are given in the UCI website given in the reference list. The associated task with this dataset is to predict the knowledge level of student as very high, high, middle and low [7].

Table 1 shows the first 50 instances of education dataset

STG	SCG	STR	LPR	PEG
0	0	0	0	0
0.08	0.08	0.1	0.24	0.9
0.06	0.06	0.05	0.25	0.33
0.1	0.1	0.15	0.65	0.3
0.08	0.08	0.08	0.98	0.24
0.09	0.15	0.4	0.1	0.66
0.1	0.1	0.43	0.29	0.56
0.15	0.02	0.34	0.4	0.01
0.2	0.14	0.35	0.72	0.25
0	0	0.5	0.2	0.85
0.18	0.18	0.55	0.3	0.81
0.06	0.06	0.51	0.41	0.3
0.1	0.1	0.52	0.78	0.34
0.1	0.1	0.7	0.15	0.9
0.2	0.2	0.7	0.3	0.6
0.12	0.12	0.75	0.35	0.8
0.05	0.07	0.7	0.01	0.05
0.1	0.25	0.1	0.08	0.33
0.15	0.32	0.05	0.27	0.29
0.2	0.29	0.25	0.49	0.56
0.12	0.28	0.2	0.78	0.2
0.18	0.3	0.37	0.12	0.66
0.1	0.27	0.31	0.29	0.65
0.18	0.31	0.32	0.42	0.28
0.06	0.29	0.35	0.76	0.25
0.09	0.3	0.68	0.18	0.85
0.04	0.28	0.55	0.25	0.1
0.09	0.26	0.6	0.45	0.25
0.08	0.33	0.62	0.94	0.56
0.15	0.28	0.8	0.21	0.81
0.12	0.25	0.75	0.31	0.59
0.15	0.3	0.75	0.65	0.24
0.1	0.26	0.7	0.76	0.16
0.18	0.32	0.04	0.19	0.82
0.2	0.45	0.28	0.31	0.78
0.06	0.35	0.12	0.43	0.29
0.1	0.42	0.22	0.72	0.26
0.18	0.4	0.32	0.08	0.33
0.09	0.33	0.31	0.26	0
0.19	0.38	0.38	0.49	0.45
0.02	0.33	0.36	0.76	0.1
0.2	0.49	0.6	0.2	0.78

0.14	0.49	0.55	0.29	0.6
0.18	0.33	0.61	0.64	0.25
0.12	0.35	0.65	0.27	0.04
0.17	0.36	0.8	0.14	0.66
0.1	0.39	0.75	0.31	0.62
0.13	0.39	0.85	0.38	0.77
0.18	0.34	0.71	0.71	0.9
0.09	0.51	0.02	0.18	0.67

Table 2 depicts the 1st 50 instances of the Iris dataset in which details of three flower are given attribute names and type can be fined from the UCI website given in reference list at the end of this paper. Associated task with this dataset is to predict the class of iris plant [7].

Table 2 The 1st 50 instances of the Iris dataset in which details of three flower

S.Length	S.Width	P.Length	P.Width
4.7	3.2	1.3	0.2
5	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
4.9	3.1	1.5	0.1
5.4	3.7	1.5	0.2
4.8	3.4	1.6	0.2
4.8	3	1.4	0.1
5.8	4	1.2	0.2
5.7	4.4	1.5	0.4
5.4	3.9	1.3	0.4
5.1	3.5	1.4	0.3
5.7	3.8	1.7	0.3
5.4	3.4	1.7	0.2
5.1	3.7	1.5	0.4
4.6	3.6	1	0.2
5.1	3.3	1.7	0.5
4.8	3.4	1.9	0.2
5	3.4	1.6	0.4
5.2	3.5	1.5	0.2
5.2	3.4	1.4	0.2
4.7	3.2	1.6	0.2
5.4	3.4	1.5	0.4
5.2	4.1	1.5	0.1
5.5	4.2	1.4	0.2
4.9	3.1	1.5	0.2
5	3.2	1.2	0.2
4.9	3.6	1.4	0.1
5	3.5	1.3	0.3
4.4	3.2	1.3	0.2
5.1	3.8	1.9	0.4
4.8	3	1.4	0.3
4.6	3.2	1.4	0.2
5.3	3.7	1.5	0.2
5	3.3	1.4	0.2
5.5	2.3	4	1.3
6.5	2.8	4.6	1.5
5.7	2.8	4.5	1.3
6.3	3.3	4.7	1.6
4.9	2.4	3.3	1
6.6	2.9	4.6	1.3
5.2	2.7	3.9	1.4
5.9	3	4.2	1.5
5.6	2.9	3.6	1.3
6.7	3.1	4.4	1.4
5.6	3	4.5	1.5
6.2	2.2	4.5	1.5
5.6	2.5	3.9	1.1
6.3	2.5	4.9	1.5
6.4	2.9	4.3	1.3

The preprocessing of datasets is not required due to the standard data provided. The tools used for the analysis is weka 3.8 and anaconda 3 having windows 7 machine with 4gb RAM, core i5 processor and 500gb hard drive.

After finding the appropriate nonlinear function for the dataset we apply Bisection method to find the intervals which contains the midpoint for that interval called centroids [2]. In this way this SK-means clustering with slope interval optimization will work. The graphical representation of the comparison of functional optimization for Iris dataset is shown in the Figure 2 and Figure 3.

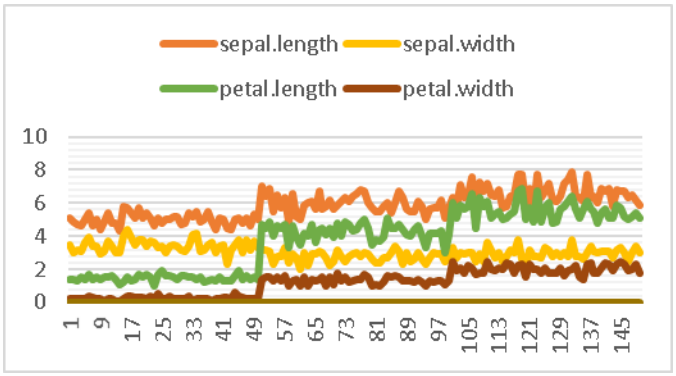

Figure 2 The data curves without optimization for Iris Dataset

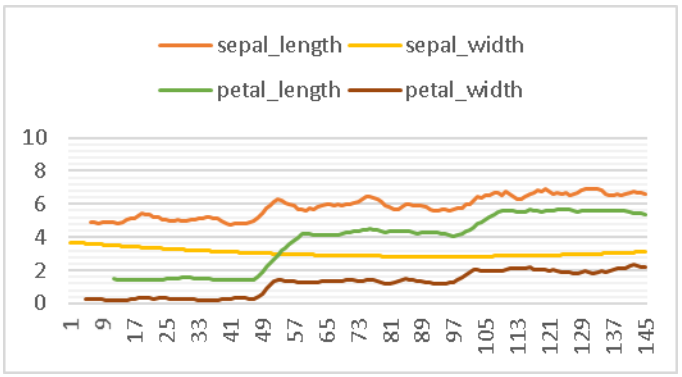

Figure 3 The data curves with functional optimization for Iris Dataset

Clearly we can see the difference between both figures 2 and figures 3.

The graphical representation of the comparison of functional optimization for Education dataset is shown in the Figure 4 and Figure 5

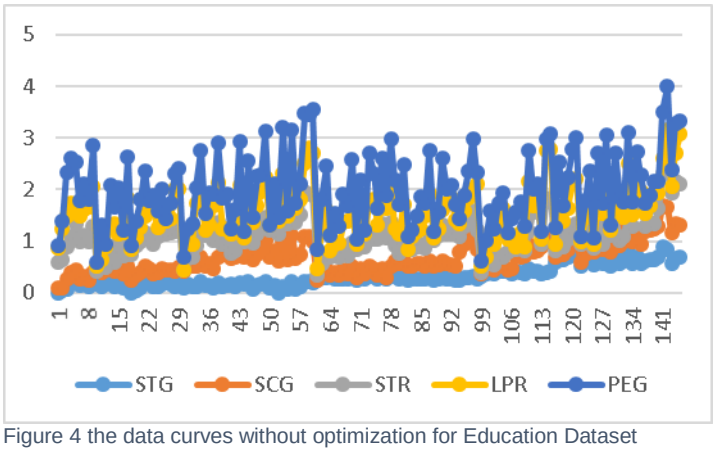

Figure 4 the data curves without optimization for Education Dataset

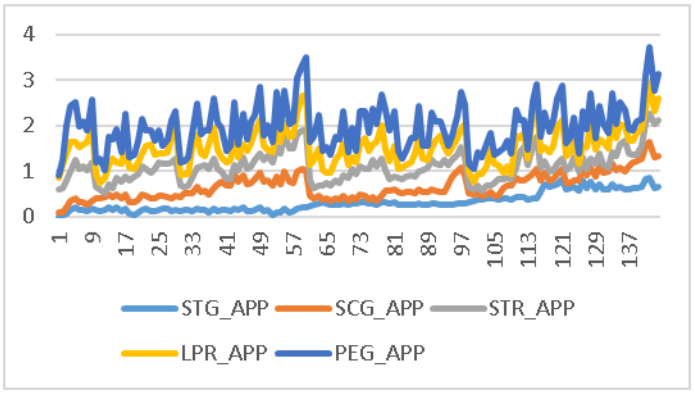

Figure 5 the data curves with functional optimization for Education Dataset

It is partially clear that the optimization significantly improves data to find the appropriate interval for the nonlinear function thereafter the centroids can be calculated from min-max functional space.

5 Results & Discussion

The comparative analysis of the results obtained for K-means and SK-means, for Education and Iris datasets are described in this section. Table 3 shows comparison of k-means clustering and sk-means with optimization, clustering results for the educational data, the parameters for comparison when the dataset is considered to be new (no class defined a-prior) are number of Iterations, Time taken to process, sum of squared errors and initial cluster selection. In our analysis the sk-means with applied optimizations perform significant improvement over simple k-means for educational data in terms of SSE with equal time and Iterations having random selection of initial cluster centers

Table 3 Comparison of k-means clustering before and after applying optimization for education dataset.

Methods	Iterations	Time	SSE	Initial Point
K-means	10	0.04	30.25	Random
SK-means with Optimization	10	0.04	27.32	Random

In figure.6 and figure.7 the graphical plot generated in Weka 3.8 of clustered instances is shown, clearly green points in the plot in figure.7 illustrates the improvements over figure.6 of simple k-means.

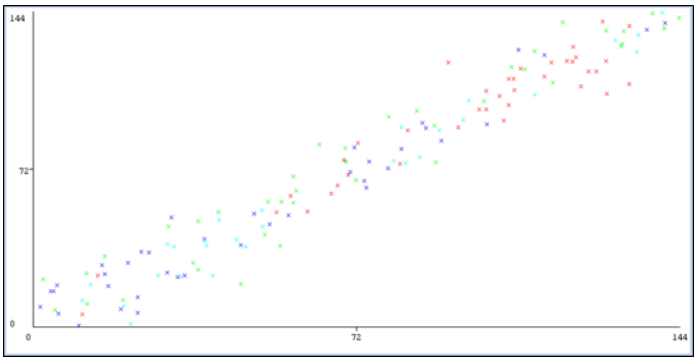

Figure 6 clustered points of education data using k-means

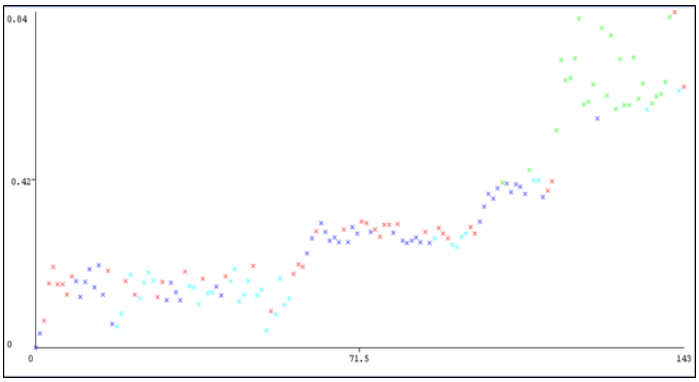

Figure 7 clustered points of education data using sk-means with optimization

In table 4 the comparison for Iris dataset is given. On applying same optimization with k-means as applied in education data above, gives more significantly comparable clusters for Iris data. As we can see that time and SSE both are improved for sk-means with optimization, while iterations and initial cluster centers are the same.

Table 4 Comparison of k-means clustering before and after applying optimization for Iris dataset

Methods	Iterations	Time	SSE	Initial Point
K-means	5	0.02	5.24	Random
SK-means with Optimization	5	0.01	4.12	Random

If we see the graphical plots of the clusters generated by both methods in figure.8 and figure.9 visibly too we can describe that after applying the proposed optimization, results have improved. As all the three red, green and blue dots are well separated from each other in figure.9.

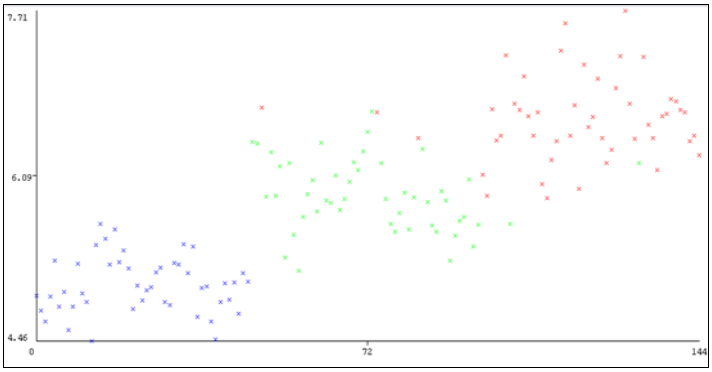

Figure 8 clustered points of Iris data using k-means

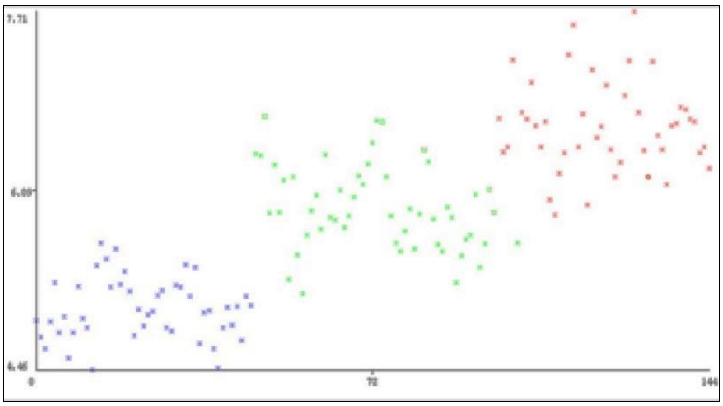

Figure 9 clustered points of Iris data using sk-means with optimization

6 Conclusion

In this paper the comparison of the simple k-means clustering and proposed sk-means clustering with proposed numerical method based optimization for given datasets has done. After analyzing results for Education and Iris datasets, we can conclude that the improvements of clustering methods with proposed optimization have comparable impacts on cluster quality. As it is not worthwhile for new dataset that the result of any clustering technique cannot be generalized for all datasets, so only hit and trial gives the feasible solution for clustering. Further we are trying other nonlinear methods for better quality clusters for new datasets.

References

[1] Lee, H. and Singh, R. (2019). Unsupervised kernel parameter estimation by constrained nonlinear optimization for clustering nonlinear biological data.

[2] Anon, (2019). [online] Available at: https://math.tutorvista.com/calculus/bisection-method.html [Accessed 3 Jul. 2019].

[3] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[4] Witten, I. and Frank, E. (2011). Data mining. 3rd ed. San Francisco, Calif.: Morgan Kaufmann.

[5] Pendergast, T. (2018). The Next Cold War Is Here, and It's All About Data. [online]. Available at: https://www.wired.com/story/opinion-new-data-cold-war/ [Accessed 3 Jul. 2019].

[6] Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. [online]. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#47a1168b60ba [Accessed 3 Jul. 2019].

[7] Marshall, M. and Fisher, A. R. (1988). UCI Machine Learning Repository: Iris Data Set. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/iris [Accessed 9 Jul. 2019].

[8] Solomon, J. (2016). Numerical Algorithms. Hoboken: Taylor and Francis.

[9] Gilat, A. (2013). Numerical methods for engineers and scientists. John Wiley & Sons.

[10] Wang, X.C. and Gao, J.H. (2016), "A Study on Numerical Calculation Method of Small Cluster Density in Percolation Model". Journal of Applied Mathematics and Physics, **4**, 1507-1512. http://dx.doi.org/10.4236/jamp.2016.48159

[11] Abraham, C., Cornillon, P., Matzner-Lober, E. and Molinari, N. (2003). Unsupervised Curve Clustering using B-Splines. Scandinavian Journal of Statistics, 30(3), pp.581-595.

[12] Agarwal, P. and Procopiuc, C. (2002). Exact and Approximation Algorithms for Clustering. Algorithmica, 33(2), pp.201-226.

[13] BERGAN, T. (1971). Survey of Numerical Techniques for Grouping. American Society for Microbiology, 35(4), pp.379-389.

[14] Xavier, A. (2010). The hyperbolic smoothing clustering method. Pattern Recognition, 43(3), pp.731-737.

[15] Horenko, I. (2010). Finite Element Approach to Clustering of Multidimensional Time Series. SIAM Journal on Scientific Computing, 32(1), pp.62-83.

[16] Tiwari,V. and Thakur, R. S.(2015), "P²MS: a phase-wise pattern management system for pattern warehouse," International Journal of Data Mining, Modelling and Management, vol. 7, no. 4, pp. 331-350.

[17] Tiwari,V. and Thakur, R. S.(2017), "Towards important issues of pattern retrieval: pattern warehouse," International Journal of Data Science, vol. 2, no. 1, pp. 1-14.

[18] Thakur, R. S. (2019), "Associative Analysis among Attribute of ILPD Medical Datasets Using ARM, IJITEE, 8(4), pp. 321-328.

[19] Thakur, R. S., Jain, R. C., Pardasani, K. R. (2008), Graph Theoretic Based Algorithm for Mining Frequent Patterns. In Proc. IEEE World Congress on Computational Intelligence, Hong Kong, pp. 629-633.

[20] Satish Kumar Soni, Dr. R. S. Thakur," Applications of Runge-Kutta methods in Data Mining", 18th Annual Cum 1st International Conference of Vijnana Parishad of India on Computational and Integrative Sciences &

International Symposium on Computational Biology, pp-116, MANIT Bhopal (M.P.), December 11-14, 2015.

[21] Satish Kumar Soni, Dr. R. S. Thakur, Dr. A.K. Gupta, “A Survey on Hyperspectral Image Segmentation Approaches with the Integration of Numerical Techniques”, In Proceedings of International Conference on Recent Advancement on Computer and Communication, vol. 34. Springer, Singapore, 2018

[22] Satish Kumar Soni, Dr. R. S. Thakur, Dr. Anil Kumar Gupta,” Clustering of Hyperspectral Images Using K-Means and Runge-Kutta Methods”, In National Mathematics Day Seminar on Mathematical Methods in Science & Engineering, TEQIP II, SATI Vidisha (M.P.), December 22, 2016.