Biomedicine and Chemical Sciences 1(4) (2022) 207-214

27-BCS-1072-84

Content list available at:
https://journals.irapa.org/index.php/BCS/issue/view/15

Biomedicine and Chemical Sciences

Journal homepage: https://journals.irapa.org/index.php/BCS

Hybrid Clustering Approach for Time Series Data

V Harsha Shastri[a*], Prathipati Ratna Kumar[b], Madhavi Kolukuluri[c], D Radha[d], Donthireddy Sudheer Reddy[e], B N Siva Rama Krishna[f]

[a] Department of Computer Systems and Engineering, Loyola Academy, Secunderabad, Telangana – India
[b, e, f] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad – India
[c] Department of Computer Science and Engineering, NSRIT, Visakhapatnam – India
[d] Department of Computer Science and Engineering, Raghu Engineering college, Visakhapatnam – India

ARTICLE INFO

Article history:
Received on: November 13, 2021
Revised on: July 20, 2022
Accepted on: July 20, 2022
Published on: October 01, 2022

Keywords:
Clustering
Data mining
Distance measure
K-means
K-means++
K-median
Time series data

ABSTRACT

The clustering of data series was already demonstrated to provide helpful information in several fields. Initial data for the period is divided into sub-clusters Recorded in the data resemblance. The grouping of data series takes 3 categories, based on which users operate in frequencies or programming interfaces on original data explicitly or implicitly with the characteristics derived from physical information or through a framework based on raw material. The bases of series data grouping are provided. The conditions for the evaluation of the outcomes of grouping are multi-purpose time constant frequently employed in dataset grouping research. A clustering method splits data into different groups so that the resemblance between organisations is better. K-means++ offers an excellent convergence rate compared to other methods. To distinguish the correlation between items the maximum distance is employed. Distance measure metrics are frequently utilized with most methods by many academics. Genetic algorithm for the resolution of cluster issues is worldwide optimization technologies in recent times. The much more prevalent partitioning strategies of large volumes of data are K-Median & K-Median methods. This analysis is focusing on the multiple distance measures, such as Euclidean, Public Square and Shebyshev, hybrid K-means++ and PSO clubs techniques. Comparison to orgorganization-basedthods reveals an excellent classification result compared to the other methods with the K++ PSO method utilizing the Chebyshev distance measure.

1. Introduction

Digitally, there was a rapid expansion of IT and a vast volume of data acquired from many sectors. The corporate expert's more hard role is to turn enormous measures of data housed in structured data into technical knowledge. This job is accomplished via Knowledge Discovery in Databases (KDD). Data mining (Aghdasi, et al., 2014) is component of the Process model. In order to find heretofore unknown, meaningful patterns and connections in huge data sets, big data refers to the application of data analytic tools. One of the big data mining operations is grouping (Sethi & Mishra, 2013).

The study of clusters is the act of aggregating a number of observations just so the sustainability describes within that group are much more comparable as well as the data sets of other groupings are distinct. Unchecked approach is done cluster, since groupings are not previously known. In a data collection (Danesh, et al., 2011) the objective of grouping is to find thick and empty areas. Clusters is utilized in various fields such as system modelling, model analysis, machine intelligence, image classification, image recognition, genomics, data recovery and the finding of data. Thus, it is a significant issue of study in several fields.

Data clustering may be widely grouped into hierarchy techniques, partial approaches, cluster analysis techniques; location is strategic techniques and modelling classification algorithms.

*Corresponding author: Prathipati Ratna Kumar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad – India

E-mail: rk30111972@klh.edu.in

How to cite:

Prathipati, R. K., Shastri, V. H., Kolukuluri, M., Dharavathu, R., Reddy, D. S., & Krishna, B. N. S. R. (2022). Hybrid Clustering Approach for Time Series Data. Biomedicine and Chemical Sciences, 1(4), 207–214.

DOI: https://doi.org/10.48112/bcs.v1i4.84

1.1. Hierarchical Methods

The hierarchy approach generates the hierarchical breakdown of datasets. They might be downwards or downwards. Traditional top - down algorithms are broken into smaller groups when each data item has been placed in a particular unit for one data set. Between each data set that forms a distinct clustering Bottom-Up procedures start. It combine the nearby data points sequentially, once all groups become one.

1.2. Partitional Methods

Divides the population through into number of nodes which was before. Consists of a collection of 'N' datasets, they try to identify 'k' groups to satisfy the minimum requisites: each database should represent a group, with each person having a minimum of one.

1.3. Fuzzy Clustering Methods

Every data set may form and over 1 group in fuzzy proposed algorithm. Each of the data points is connected with both the data points. The values range from 0 to 1.

1.4. Hard Clustering Methods and Model-Based Methods

Every piece of data could be part of just one group in difficult clustering methods. Design approaches assume that a model is the most appropriate to every group, as well as the information better placed to a theory. Depending on the design they might be potentially hierarchy or partial.

Nowadays, a number of actual performance prediction have become more prominent with hybrid techniques (Aghdasi, et al., 2014). Historically, euclidean distance measure is used in the research for various clustering techniques. In this work, the efficiency of methods was studied using additional significant measurement approach, including such City Block and Chebyshev. In this study a clustering utilising multiple measurement approach is proposed which is based on K-means ++ & PSO algorithms (K++ PSO). The K++ PSO method provides strong group results on 4 benchmark problems, for example, the assessment of teacher assistants, heart, seedlings, breast cancer, as well as synthetic cancer.

1.5. Review of Literature

Aghdasi et al. (2014) developed PSO as well as Tabu searching K-Harmonic Data Optimization Algorithm. According to Sethi and Mishra (2013), the hybrid K-means grouping & Particle Swarm Optimization (PSO) algorithms for quantitative Principle Component Analysis (PCA) were proposed (PCA-KPSO). The model includes PSO's worldwide search functions and rapid K-mean method completion.

Danesh et al. (2011) developed an effective K-Harmonic Means, Particle Swarm Optimization , also GA suitable numerical method. The hybrid method contributes to solving the global optimal issue also solves the sluggish performance constraint. Chuang et al. (2012) suggested improvement of Gaussian Chaotic Clusters Particle Swarm Optimization. You utilized the radius from the interpersonal and inter for the searching of cloud services.

According to Ran, Yong, and Na (2013), this K-means technique on Chaos particle swarm was suggested (CPSOKM). The suggested approach resolves and optimizes the group output of the K-means method.

Kishirsagar et.al (2020) to develop and develop a hybrid artificially intelligent application alongside optimizations to classify and forecast diverse dataset having good accuracy, and utilize multiple methods for analysis and regression of benchmark functions, which have been beneficial in all sectors growing. In computer security machine learning technology, the algorithms utilized in various study were beneficial to achieve more accurate findings utilizing varied quality factors.

Rai & Singh (2010) cluster is a collection used to integrate comparable data components without comprehensive techniques of grouping characteristics into homogenous groupings. Structural identification in an unmarked database is a helpful method. Cluster centre in the manner that items are as close as possible to other products within such a grouping and as few as possible comparable to things in other subgroups. A particular form of grouping is the cluster of periodicity. Temporal period are dynamical structures with date change in their characteristics. Information is available in dataset in many systems - such as banking, healthcare and commerce.

Zakaria, et al., (2012), time - series has given many academics in data analysis groups the chance to analyse data set in the past decade. As a result, several study and initiatives based on time period were conducted for diverse objectives in multiple regions: subsequent matched, intrusion detection, character recognition, indexation, grouping, categorization visualization, predictive modelling, statistical analysis, summary and prediction. In addition, several extensive research initiatives are being undertaken to enhance conventional methods. There seems to be a great deal of time - series studies and surveys and implementations.

1.6. Time Series Clustering

Static data grouping, frequency classification involves a method or technique of grouping and grouping depends also on kind or goal of both the dataset and the decision of the proposed technique as well as the nature. Whether data are discerning or re-valuation, chosen, regular or non-uniform, simple or multimodal, but whether the historical data seem to be of similar or uniform duration and as far as logistic regression information is concerned, differentiation may be established. Instead of universal clustered procedures, data collected from the sample should be transformed to uniform data. A broad variety of options may be applied to this (Kumar, Patel & Woo, 2002).

Different techniques were developed to compile distinct sorts of data from regression analysis. When their variations are set apart, it is far from being the case that while in essence they are all trying to alter current cluster methods so that moment data set may be processed or records converted to static data to make straight more use current methods for clustered data format (Niennattrakul, Srisai & Ratanamahatana, 2012). In the first technique, actual data for the period are often used straight, hence the original data method, and a key change is in the replacement of a distance metric for data structure with a data set measurement that is suitable (Ni & Jinhang, 2017). First, that last technique transforms raw data for the period into a lower-dimensional column or into certain design variables, and then uses a typical clusters technique, dubbed the

features and prototype technique, for retrieved selected features or modelling parameters (Kumar, Patel & Woo, 2002). The three techniques are outlined in Figure 1: raw, functionality and design. Note which, without requiring any other scheduling technique, a moved splinter group of the prediction model educated the framework and used input variables for grouping (Kshirsagar, Akojwar & Dhanoriya, 2017).

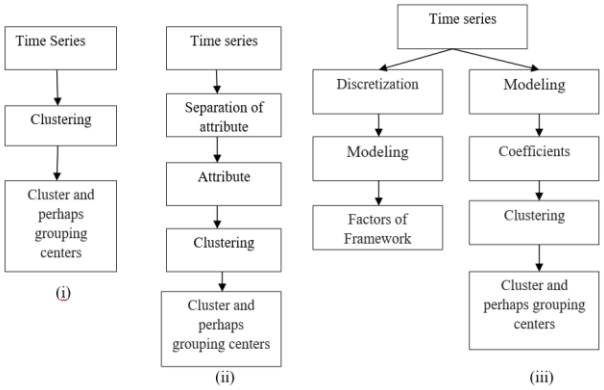

Fig. 1. Clustering methods of 3 series: (i) raw-data-based, (ii) feature-based, (iii) model-based.

1.7. Taxonomy of Time-Series Clustering

Clusters associated tasks in time series are categorized into three: grouping literally the entire time - series data; sub clustering and grouping time points as shown in Figure 2.

1.7.1. Whole Time-Series Clustering

Time series complete In relation to their resemblance, grouping is seen as a grouping of a number of potential frequency. Clustering in this case implies that the usual cluster is applied to abstract events, which are data set (Aghabozorgi & Teh, 2014).

1.7.2. Subsequence Clustering

The cluster of sub graphs implies the grouping of segment from either a big continuous serial by a collection of sub graphs of a temporal serial retrieved using a feature vector (Kshirsagar & Akojwar, 2016).

1.7.3. Time Point Clustering

The other kind of cluster is indeed the grouping of time points. It consists of the combining of life stages depending upon both spatial closeness to durations and the similitude of the related values. This method is comparable to the segment of data series. Nevertheless, the difference is that point must not be allocated to groups, that is to say, part of the data are regarded as clutter (Manoharan, et al., 2020).

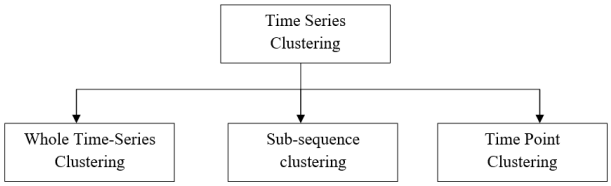

Fig. 2. Taxonomy of Time-Series Clustering

In essence, further grouping is carried out in a single string all periods, which indicates that the grouping is of little importance. Frequency cluster also is carried out in a single time series, as the goal of a frequency clustering is just to discover frequency cluster rather than time-series groupings (Madicar, et al., 2013). This study focuses on the "cluster formation of the complete time series." Table4 provides a full overview with statistical analysis grouping. A number of studies show that several approaches for the cluster of entire time series analysis have always been recommended (Lai, et al., 2010). Vast majority these, though, choose one of the following techniques to time - series data clusters:

- The current traditional cluster methods are adapted so that they really are consistent with the spirit of information from time series. Generally, its distance measurement is changed in this manner to be consistent with original information from time series.
- Transform time series information as input of standard clustering methods to specific structures.
- To use multi-stage time series resolves as an inputs to a multi-stage method. In essence, there are three alternative approaches to clusters time series, including formal, functional and model-driven, apart from this consistent theme (Zhang, et al., 2011).

Figure 3 shows a short of the methods. The form-based method combines forms of two data sets with a non-linear stretch and contraction of the time vectors as closely as feasible. This technique is often referred to as a technique based on current data since it usually functions straight with pure time - series. Pattern algorithms normally use traditional techniques of clustering that are consistent with data types while their moving object has been changed in time series (Xu, et al., 2013). The original sequence is transformed into a lower-dimensional feature vector in the functionality method. The collected extracted features are then covered by a standard cluster analysis. Typically, a different lengths vector of each time series accompanied by an incident measurement (Zakaria, Mueen & Keogh, 2012) is generated from certain technique. The input time series of models are converted onto parameter values in Figure 3. In design approaches. Then maybe an appropriate forecasting similarity and a group method are generated by applying to the modelling variables retrieved. Unfortunately, design methods generally have issues with scaling and their pressure drops if cluster are adjacent to one other (Kshirsagar, Chavan & Akojwar, 2017). Analysing previous research in the literature implies that perhaps the major parts of the cluster of time series are four: Reduce dimensions or display technique, measurement range, testing set, design and assessment of prototypes.

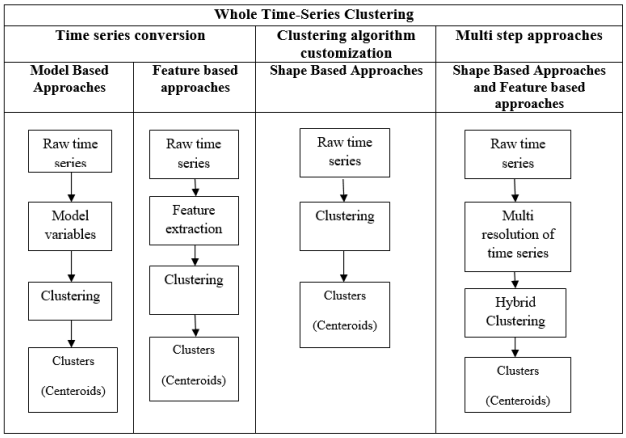

Fig. 3. The Time-Series Clustering Approaches

Figure 4 shows a summary of the elements in question. One or more of those elements rely on the difficulty in the overall procedure in the time series grouping. Data are normally represented in ways that fit within the storage that used a representations method (Seref, et al., 2014). A membership function on data is then used by a measurement of proximity. There in clustering procedure, the time series is generally synthesized by a template. The groups will finally be assessed via criterion. Inside each element, numerous relevant studies and approaches are covered there in following sub-sections (Darkins, et al., 2013).

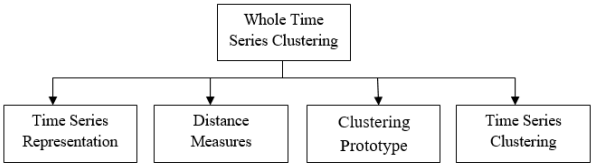

Fig. 4. A summary of four factors of the entire time series

1.8. Clustering of Time Series Analysis Method

In taxonomy of presentations four components of depictions are typically data-adaptive, data-free, design and data-driven representational techniques as described in Figure 5 (Ghassempour, Girosi & Maeder, 2014).

1.8.1. Information Adaptive

Information adjustment display techniques in larger data are done on every time series and randomly aim to reduce the intricacy of the world's largest method (Aghabozorgi, et al., 2014). This method has been adopted in a variety of methods, including extrapolation of Nonlinear algebraic expressions, stagnation of Piece - wise algebraic expressions, linear estimation of Leadership networks, estimate of the Dynamic Piecewise continuous, dissolution of dot product, language processing, innate direct quotations, estimation of the symbolic accumulation and Data adaptability can indeed be ideally suited to every serial. It is much more challenging to do many time series comparisons (Akojwar & Kshirsagar, 2016).

1.8.2. Non- Information Adaptive

Non-data alternative manner are presentations suited for time series with such a separation of the fixed dimensions as well as a real analogy of the images of different time series. Wavelet coefficients are indeed the approaches in this collective: HAAR, Randomized Mapping, Piece - wise Aggregation Estimation and Indexing capable Piece ways linearly approximate solution: HAAR, Affine, Coeiflets, Simlets, Discrete wave front Transformations, Spectrum Chebyshev polynomias, Randomized Maps (Kshirsagar, et al., 2020).

1.8.3. Model Based

Modelling methods depict a stochastically time series including such Model Parameters and the Word Embedding Models, statistical methods, ministering of the time series and the machine gradient descent. Allows users to specify the pressure ratio upon on basis of the application in hands in information adaptable, non-data adaptive and pattern recognition ways.

1.8.4. Information Dictated

In comparison, the classification accuracy is manually calculated based on basic time series, including such Tucked, with data suggested techniques.

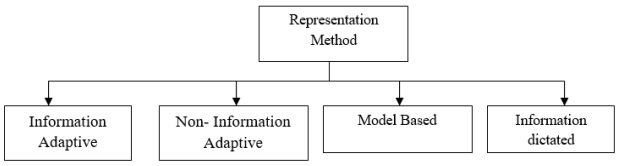

Fig. 5. Arrangement of several analysis methods in time series

1.9. Distance Measures

In time series cluster, an exact calculation is contentious. If you study the above methods as a measurement of similarity/differences, it implies that dynamic programming (DP), which have been highly expensive to operate, is by far the most reliable and timely technique. Even though some limitations are often adopted to minimize waste for these proximity and cosine similarity, careful adjustment of variables is required to be efficient and robust (Rakthanmanon, et al., 2012). Consequently, the use of these measures also should make a difference between power and agility. In another aspect the degree to which distance measurement is efficient in huge time series analysis collections is crucial –. This subject is not taken from academia, as the majority of the studies evaluated are built on very tiny sets of data (Aghabozorgi, et al., 2014).

Various problems related to distance measuring are explored in background subtraction study. The inconsistency of the distance measure only with display technique is a major difficulty (Rakthanmanon, et al., 2012). For example, several of the typical techniques for time series research is based on wavelength, and it is possible to specify similarities across episodes and to generate real worth contrasts to utilise in clusters using another environment.

The most prevalent approaches for trusting relationship in time series clusters are Feature extraction and DTW (Aghabozorgi, et al., 2011). Research has revealed that the distances to Euclid is unexpectedly aggressive in the multiclass classification; nevertheless, the ability of DTW is not to be reduced in similarity measures.

Clustering are used to measure the resemblance or discrepancy among any pair of items by using similarity metrics (Petitjean, Ketterlin & Gançarski, 2011). Range may be quantified using background subtraction among data and centroids. The main features of extracted features are as follows:

- d(a, b)≥ 0 for every a and b
- d(a, b) = 0 only if a = b
- d(a,a) = 0 for every a
- d(a, b) = d(y, x) for every a and b
- d(a, c) ≤d(a, b) + d(b, c) for every a, b and c

The productivity of K-means + + hybrid using PSO clustering techniques was examined in this work on the basis of several feature vectors such as Euclidean, City Block and Chebyshev (Oh, et al., 2013).

1.10. Euclidean Distance

It is usually used to clusters apps using this distance measure. The L2 or Pythagoras measure is sometimes termed. The following rate is calculated:

$$d(a,c) = \|a - c\| = \sqrt{\sum_{i=1}^{n}(a_i - c_i)^2} \quad (1)$$

1.11. City Block Distance

It's also referred to as L1 or Distance measure. The length of the city block from strong functionality a and centre c

$$d(a,c) = \sum_{i=1}^{n}|a_i - c_i| \quad (2)$$

1.12. Chebyshev

The radius from Chebyshev is sometimes called the greatest range number. It is a specified measure in a subspace where the distance between any two equals the largest difference here between measurements of each component.

$$d(a_i, a_j) = \max|(a_{ik} - a_{jk})| \quad (3)$$

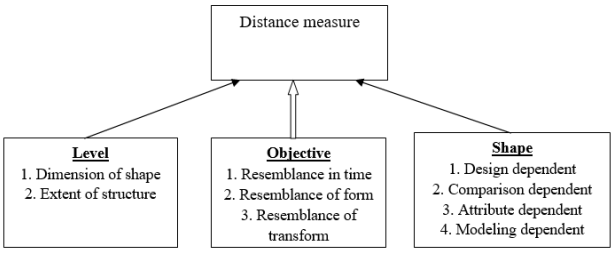

Fig. 6. Distance measure approaches

2. Materials and Methods

2.1. k-means Clustering Algorithm

K-means method seems to be a non-hierarchical approach for grouping the item to a certain cluster. The fundamental principle of the method k-means consists of calculating the size of groupings to be generated as soon as possible (Xia, Ye & Zhang, 2012). A I J(i=1,...,v;j=1,...,u) is identified in an initialization phase, where v has been the number of nodes to perform and u is really the set of attributes. The centre within each clusterd kj is selected randomly in the application of information (t = 1,..., k; j= 1,...,u). Then, using each cluster centre termed the centroid, we measures the movement among each information. The length between data-i and central k is calculated by calling ik.

Classify the information belonging to each cluster. By determining the minimum value from data that form a clustered component using Equations 4, a centre value may be obtained.

$$c_{kj} = \frac{\sum_{i=1}^{n} a_{ij}}{n} \quad (4)$$

n=Number of Cluster K Member

Cluster steps are described with the k-means method.

1. Assume data matrix A = {a ij} measuring v for u, i= 1, 2... v; j = 1, 2,..., m. 1
2. Calculate comprises components (k), set centre value randomly
3. The separation between the information and the centroid may be calculated by using equation 1
4. Use Equation 2 to perform classification into the minority and majority cluster
5. Use Equation 4 to determine the revised centre
6. Repeat step 3 to 5 until data is moved to some other group

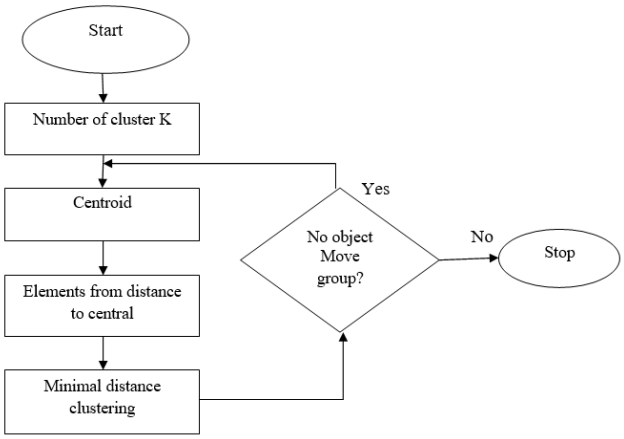

Fig. 7. Flowchart of k-means clustering algorithm

2.2. Hybrid Algorithm

2 or even more constructivist approach are integrated with hybrid algorithms. Hybrid methods are currently prominent since they are able to handle different practical applications that entail cost and risk. They employ distinct algorithms' characteristics. The K-Means++ and Support Vector Machine Algorithms (K++ PSO) for hierarchical clustering were integrated throughout this study. The usual measure in most clustered techniques is the euclidean distance. We however have attempted to analyse alternative techniques, including such City Block & Chebyshev, using alternative euclidean distance.

3. Results and Discussion

Result of the proposed hybrid approach on four reference set including clustering methods, including data from the Intern assessment for teachers, hypothyroidism, seedlings, breast cancer dataset consists.

The data set contains of 150 items and 3 various kinds of categories with 4 characteristics. The figures comprise of three normal seasons of examination of academic performances and 2 consecutive semester of 150 teaching staff. The values was split into three groups (low, medium, high) that were approximately the same size, to produce an ordinary moments. There are 214 occurrences in the endocrine data. Every case includes 5 characteristics comprising a T3-resin reception test, complete thyroid hormone replacement serum, overall serum triodothyronine, baseline TSH and image intensity TSH-value change from the baseline level following 216 milligrams of table shows the comparison testosterone are injected. In each specimen, one among three categories must be classified: Class 1: ordinary (151 occurrences), Class 2: hyperactive (34 occurrences) (31instances).

The collection of information on seeds includes 211 patterns of 3 distinct types of wheat: Kame, Rose and Canada. Every design contains seven geometrical features of grain kernel, including the area, circumference, compaction and kernels lengths, kernel thickness, factor of asymmetries and kernels gap distance.

The data set on cancer includes 683 recordings with 9 characteristics such as width, cell-sized uniform, organogenesis sameness, margin adherence, unicellular size, naked nuclei, dull nucleosides, normal cytoplasm and mitosis. These two groups are normal and abnormal instances (238 records) (445 records).

Table1
Lists specified characteristics of the data set

Data set	Total Samples	No. of classes	No. of attributes	Size of classes
Intern assessment for teachers	150	3	5	49,50,51
Thyroid	216	3	5	151,34,31
Seeds	211	3	7	71,70,70
Tumor of the breast	683	2	9	238,445

Table 2
Evaluation of the objective function score of the 7 clustering methods in the set of data

Distance	K-means	K++	PSO	K_PSO	K++_PSO
Euclidean	1505.563	1505.563	1505.122	1499.193	1494.049
City block	2366.831	2366.627	2338.158	2209.712	2184.583
Chebyshev	1253.935	1230.360	1228.709	1216.684	1211.851

Table 3
Correlation of the best fitness in hypothyroidism set of data for the various cluster methods

Distance	K-means	K++	PSO	K_PSO	K++_PSO
Euclidean	2001.638	2001.638	2250.460	1962.504	1930.335
City block	2985.350	2985.350	3463.442	2929.858	2925.507
Chebyshev	1678.178	1678.178	1752.755	1632.318	1622.337

Table 4
Evaluation of the efficiency of the 7 techniques inside the set of data seedlings

Distance	K-means	K++	PSO	K_PSO	K++_PSO
Euclidean	313.218	313.218	338.983	312.162	312.160
City block	545.622	544.591	672.035	543.684	543.590
Chebyshev	261.506	261.502	286.536	258.017	257.988

Table 5
Evaluation of the efficiency of the 7 techniques inside the breast cancer data set

Distance	K-means	K++	PSO	K_PSO	K++_PSO
Euclidean	2988.429	2988.429	3741.142	2967.179	2966.432
City block	7326.376	7326.376	8243.117	6512.535	6454.469
Chebyshev	1933.128	1933.128	2179.060	1886.596	1880.629

The best performance of the programs is as follows: The particles frequency (p) is 10. The intellectual (c1) and emotional (c2) components are 2.0. The weight of the momentum is 0.91 to 1.41. ALS is declined linearly during the search procedure from 0.91 to 0.41. The preceding Eq is used to compute the dawn.

$$\omega = \omega_{\max} - \frac{\omega_{\max} - \omega_{\min}}{I_{\max}} * I \qquad (5)$$

where ω_{max} and $\omega_{min}$are really the weighted coefficient's starting and last value, correspondingly; ω_{max}= 0.91 and ω_{min}= 0.41; Imax seems to be the highest current iteration; Imax seems to be the present current iteration. The max iteration number is 100. The trials are performed in 10 separate runs for all methods. The error is 0.00001 in iteration (p). The objective of this work is to examine how data collectors use various distance metrics to influence hybrid algorithms. The 100 rounds and 10 separate trials evaluate every method. In this research, basis functions evaluate the success of the grouping of clustered data cluster. The outcomes of various clustering methods for chosen data sets in respect of their efficiency values are compared in table 4 to 8.

Fitness function values: It calculates and agrees the measurement result and inside the group and its cluster centre. It is determined with the Equations 6.

$$\sum_{j=1}^{k} \sum_{x_i \in c_j} d(x_i, c_j) \qquad (6)$$

where, d (xi, cj) seems to be the distances from xi of the piece of data to cj. The minimal significant value shows the greater group integrity. Tables 1 to 5 provide the minimal cost function of the suggested method. 1494.049, 2184.583 and 1211.851 on Intern assessment for teachers data set; 1930.335, 2925.507 and 1622.337 on thyroid data set; 312.160, 543.590 and 257.988 on seeds data set; 2966.432, 6454.469 and 1880.629 on breast cancer data set for Euclidean, City block and Chebyshev distance metrics, respectively. Therefore, in fitness function numbers, the hybrids K++ PSO method outperforms than any other proposed technique. The suggested approach employing the distance Chebyshev generally results superior than those of other feature vectors. It was also noticed.

4. Conclusion

Cluster is a full NP issue grouping pieces of data and that are closer to each other and to individuals. The methods K-means & K-medoids is readily confined to a local optima and susceptible to starting values and bruises. K-means++ gives better results than in other methods. PSO is a probabilistic supervised classifier population. This hybrid approach enhances grouping efficiency. In several clustering methods the distance measure is often used. K++ PSO is developed in this thesis by means of a variety of distance measures, namely City Block and Chebyshev. The comparison of individual methods is assessed by means of best fitness. The

technique presented is evaluated against four traditional scheduling son techniques, including the appraisal of teacher assistants, hypothyroidism, seedlings, Tumor of the breast through the use of various distance measures in fake dataset. Findings from experiments demonstrate that the genetic algorithm worth of the K++ PSO method is superior than the other techniques K-means, K-means++, PSO, K-PSO, K-med PSO. The suggested technique also gives the Chebyshev length excellent results compared to other feature vectors. It is reported.

Competing Interests

The authors have declared that no competing interests exist.

References

Aghabozorgi, S. R., Wah, T. Y., Amini, A., & Saybani, M. R. (2011). A new approach to present prototypes in clustering of time series. In *Proceedings of the International Conference on Data Science (ICDATA)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). **http://eprints.um.edu.my/id/eprint/13448**

Aghabozorgi, S., & Teh, Y. W. (2014). Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications*, *41*(4), 1301-1314. **https://doi.org/10.1016/j.eswa.2013.08.028**

Aghabozorgi, S., Ying Wah, T., Herawan, T., Jalab, H. A., Shaygan, M. A., & Jalali, A. (2014). A hybrid algorithm for clustering of time series data based on affinity search technique. *The Scientific World Journal*, *2014*. **https://doi.org/10.1155/2014/562194**

Aghdasi, T., Vahidi, J., Motameni, H., & Inallou, M. M. (2014). K-harmonic means data clustering using combination of particle swarm optimization and tabu search. *International Journal of Mechatronics, Electrical and Computer Technology*, *4*(11), 485-501.

Akojwar, S. G., & Kshirsagar, P. R. (2016). Performance evolution of optimization techniques for mathematical benchmark functions. *International Journal of Computers*, *1*.

Chuang, L. Y., Lin, Y. D., & Yang, C. H. (2012). Data clustering using chaotic particle swarm optimization. *IAENG International Journal of Computer Science*, *39*(2), 208-213.

Danesh, M., Naghibzadeh, M., Totonchi, M. R. A., Danesh, M., Minaei, B., & Shirgahi, H. (2011). Data clustering based on an efficient hybrid of K-harmonic means, PSO and GA. In *Transactions on computational collective intelligence IV* (pp. 125-140). Springer, Berlin, Heidelberg. **https://doi.org/10.1007/978-3-642-21884-2_2**

Darkins, R., Cooke, E. J., Ghahramani, Z., Kirk, P. D., Wild, D. L., & Savage, R. S. (2013). Accelerating Bayesian hierarchical clustering of time series data with a randomised algorithm. *PloS one*, *8*(4), e59795. **https://doi.org/10.1371/journal.pone.0059795**

Ghassempour, S., Girosi, F., & Maeder, A. (2014). Clustering multivariate time series using hidden Markov models. *International journal of environmental research and public health*, *11*(3), 2741-2763. **https://doi.org/10.3390/ijerph110302741**

H.Kremer, P.Kranen, T.Jansen, T.Seidl, A.Bifet, G.Holmes, B. Pfahringer, An effective evaluation measure for clustering on evolving data streams, in: Proceedings of the 17thACMSIGKDD international conference on Knowledge Discovery and Data Mining, 2011,pp.868–876.

Kshirsagar, P. R., Akojwar, S. G., & Dhanoriya, R. A. M. K. U. M. A. R. (2017). Classification of ECG-signals using artificial neural networks. In *Proceedings of International Conference on Intelligent Technologies and Engineering Systems, Lecture Notes in Electrical Engineering* (Vol. 345).

Kshirsagar, P. R., Manoharan, H., Al-Turjman, F., & Kumar, K. (2020). Design and testing of automated smoke monitoring sensors in vehicles. *IEEE Sensors Journal*.

Kshirsagar, P., & Akojwar, S. (2016, December). Optimization of BPNN parameters using PSO for EEG signals. In *International Conference on Communication and Signal Processing 2016 (ICCASP 2016)* (pp. 384-393). Atlantis Press. **https://dx.doi.org/10.2991/iccasp-16.2017.59**

Kshirsagar, P., Balakrishnan, N., & Yadav, A. D. (2020). Modelling of optimised neural network for classification and prediction of benchmark datasets. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *8*(4), 426-435. **https://doi.org/10.1080/21681163.2019.1711457**

Kshirsagar, P., Chavan, S., & Akojwar, S. (2017). *Brain tumor classification and detection using neural Network*. Scholars' Press.

Kumar, M., Patel, N. R., & Woo, J. (2002, July). Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 557-563). **https://doi.org/10.1145/775047.775129**

Lai, C. P., Chung, P. C., & Tseng, V. S. (2010). A novel two-level clustering method for time series data analysis. *Expert Systems with Applications*, *37*(9), 6319-6326. **https://doi.org/10.1016/j.eswa.2010.02.089**

Madicar, N., Sivaraks, H., Rodpongpun, S., & Ratanamahatana, C. A. (2013). Parameter-free subsequences time series clustering with various-width clusters. In *2013 5th International Conference on Knowledge and Smart Technology (KST)* (pp. 150-155). IEEE. **https://doi.org/10.1109/KST.2013.6512805**

Manoharan, H., Teekaraman, Y., Kshirsagar, P. R., Sundaramurthy, S., & Manoharan, A. (2020). Examining the effect of aquaculture using sensor-based technology with machine learning algorithm. *Aquaculture Research*, *51*(11), 4748-4758. **https://doi.org/10.1111/are.14821**

Ni, L., & Jinhang, S. (2017, October). The analysis and research of clustering algorithm based on PCA. In *2017*

13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI) (pp. 361-365). IEEE.
https://doi.org/10.1109/ICEMI.2017.8265817

Niennattrakul, V., Srisai, D., & Ratanamahatana, C. A. (2012). Shape-based template matching for time series data. *Knowledge-Based Systems*, *26*, 1-8.
https://doi.org/10.1016/j.knosys.2011.04.015

Oh, S., Song, S., Grabowski, G., Zhao, H., & Noonan, J. P. (2013). Time series expression analyses using RNA-seq: a statistical approach. *BioMed research international*, *2013*.
https://doi.org/10.1155/2013/203681

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, *44*(3), 678-693.
https://doi.org/10.1016/j.patcog.2010.09.013

Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, *7*(12), 1-5.

Rakthanmanon, T., Keogh, E. J., Lonardi, S., & Evans, S. (2012). MDL-based time series clustering. *Knowledge and information systems*, *33*(2), 371-399.
https://doi.org/10.1007/s10115-012-0508-7

Ran, L., Yong, Y., & Na, Z. C. (2013). The K-means clustering algorithm based on chaos particle swarm. *Journal of Theoretical and Applied Information Technology*, *48*(2).
https://doi.org/10.1109/JSEN.2020.3044604

S. Akojwar and P. Kshirsagar, "A Novel Probabilistic-PSO Based Learning Algorithm for Optimization of Neural Networks for Benchmark Problems", Wseas Transactions on Electronics, Vol. 7, pp. 79-84, 2016.

Seref, O., Fan, Y. J., & Chaovalitwongse, W. A. (2014). Mathematical programming formulations and algorithms for discrete k-median clustering of time-series data. *INFORMS Journal on Computing*, *26*(1), 160-172.

Sethi, C., & Mishra, G. (2013). A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset. *International Journal of Scientific & Engineering Research*, *4*(6), 1559-1566.

Xia, X., Ye, X., & Zhang, J. (2012, November). Optimal metering plan of measurement and verification for energy efficiency lighting projects. In *2012 Southern African Energy Efficiency Convention (SAEEC)* (pp. 1-8). IEEE.
https://doi.org/10.1109/SAEEC.2012.6408588

Xu, K., Jiang, Y., Tang, M., Yuan, C., & Tang, C. (2013). PRESEE: an MDL/MML algorithm to time-series stream segmenting. *The Scientific World Journal*, *2013*.
https://doi.org/10.1155/2013/386180

Zakaria, J., Mueen, A., & Keogh, E. (2012, December). Clustering time series using unsupervised-shapelets. In *2012 IEEE 12th International Conference on Data Mining* (pp. 785-794). IEEE.
https://doi.org/10.1109/ICDM.2012.26

Zakaria, J., Rotschafer, S., Mueen, A., Razak, K., & Keogh, E. (2012, April). Mining massive archives of mice sounds with symbolized representations. In *Proceedings of the 2012 SIAM International Conference on Data Mining* (pp. 588-599). Society for Industrial and Applied Mathematics.
https://doi.org/10.1137/1.9781611972825.51

Zhang, X., Liu, J., Du, Y., & Lv, T. (2011). A novel clustering method on time series data. *Expert Systems with Applications*, *38*(9), 11891-11900.
https://doi.org/10.1016/j.eswa.2011.03.081