International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

Kleinberg's Hyper-Richness Based Fuzzy Partition Clustering for Efficient Bi-Temporal Data

L. Jaya Singh Dhas, B. Mukunthan

Abstract*: Clustering is the process used for partitioning the total dataset into different classes of similar objects. The group contains knowledge about their members and also helps to understand the structure of the dataset very easily. Clustering the bitemporal data is one of the major tasks in data mining since the bitemporal datasets are very large with various attribute counts. Hence the accurate clustering is still challenging tasks. In order to improve the clustering accuracy with less complexity, Kleinberg's Hyper-richness Bitemporal property based fuzzy c means partition Clustering (KHBP-FCMPC) technique is introduced. The KHBP-FCMPC technique partition the bitemporal dataset into number of possible groups with an improved performance rate based on a distance metric. At first, the 'c' numbers of clusters are initialized. The KHBP-FCMPC technique uses the core data point module and authority sector module to minimize the execution time of clustering the data points. Core data point module served as the centroid of the cluster. Each cluster contains one core data point. After that, the distance is computed with the membership function. The authority sector module assigns the data points into the cluster with minimum distance. After that, the centroid is updated and the process iterated until the convergence is met. Finally, the Kleinberg's Hyper-richness Bitemporal property is applied to verify the total dataset equals the partition of all the data points. This property used to group the entire data points into the cluster with higher accuracy. Experimental evaluation is carried out using a temporal dataset with different factors such as clustering accuracy, false positive rate, time complexity and space complexity with a number of data points. The experimental results show that the proposed KHBP-FCMPC technique increases the bitemporal data clustering accuracy with less false positive rate, time complexity as well as space complexity. Based on the results observations, KHBP-FCMPC technique is more efficient than the state-of-the-art methods.

Keywords: clustering, bitemporal data, Kleinberg's Hyper-richness Bitemporal property, fuzzy c means partition Clustering, Core data point module, authority sector module.

I.INTRODUCTION

Clustering is the data mining method which partitions a whole dataset into different groups based on their similarity. The data within the groups are more similar other than the data in different clusters. In general, clustering of temporal data is more complex since the dimensionality of the dataset is considerably larger. Temporal data is the data which is stored related to the time instances. The temporal data comprises two attributes such as transaction time and valid time.

Manuscript published on 30 August 2019.
***Correspondence Author(s)
L. Jaya Singh Dhas, Research Scholar, Department of Computer Science, Jairams Arts and Science College, Karur – 639003,. Tamilnadu, India.
B. Mukunthan, Research Supervisor, Department of Computer Science, Jairams Arts and Science College, Karur – 639003,Tamilnadu, India.

Transaction time records the time period during which the information stored in the temporal database. Valid time is the time for which an event is true in the real world. These two attributes are united to form bitemporal data. In this research, the clustering techniques are used for bitemporal relational data analysis. In general, the various clustering technique is presented such as the density-based technique, partitioning techniques, Hierarchical techniques, Grid-based methods and so on. Among the several techniques, partitioning based clustering technique is used for bitemporal relational data analysis. The advantage of the partition-based clustering algorithm is it uses an iterative method for generating the clusters where the entire data are grouped. The Bi-weighted ensemble approach was introduced in [1] for partitioning the temporal data through Hidden Markov Model-based approaches. The ensemble clustering approach increased the performances of temporal data clustering but failed to attain high accuracy with less complexity. A Parametric Link among Multinomial Mixtures (PLMM) model was introduced in [2] for partitioning the temporal categorical data. Though the method increased the clustering accuracy, the time complexity was not minimized. An algorithm termed as variational Expectation–Maximization (VEM-DyMix) was developed in [3] for grouping the temporal data. The algorithm increased the clustering and estimation accuracy with minimal computation time but the space complexity was not minimized. A generalized k means clustering technique was introduced in [4] for grouping the similar temporal data through weighted and kernel time. Though the technique minimizes the time complexity, the clustering performance was not increased. Clustering a temporal network was presented in [5] through the topological similarity measure. The approach increased the accurate clustering results but the false positive rate was not minimized in an efficient manner. A finite mixture model was introduced in [6] for dynamic clustering of spatiotemporal data with minimal error. The model discovered the level-based clusters in spatiotemporal data but the time and space complexity were not minimized. A Hidden Markov Model-based hybrid meta-clustering ensemble method was introduced in [7] for increasing the temporal data clustering analysis. The technique failed to solve the more time-consuming problems during clustering analysis. The temporal human activity patterns were recognized using a clustering technique based on temporal similarity [8]. The performance of clustering accuracy was not improved for efficient activity recognition. The Creating Discriminative models were developed in [9] for categorization and partitioning the time series data. Though the model increases the clustering accuracy, the time complexity was not minimized.

A novel hybrid clustering algorithm was introduced in [10] based on the similarity of time series data. The algorithm attained accurate clustering results but the false positive rate was not minimized. The major issues are identified from the above-said literature such as minimum clustering accuracy, high time and space complexity, high false positive rate, failure to attain accurate clustering results and so on. In order to overcome the major issues, an efficient technique called, Kleinberg's Hyper-richness Bitemporal property based fuzzy c means partition Clustering (KHBP-FCMPC) is introduced. The major contributions of the proposed KHBP-FCMPC technique compared to the existing works are summarized as follows,

- To acquire the high clustering accuracy with minimal complexity, KHBP-FCMPC technique is introduced. The KHBP-FCMPC technique uses hyper-richness property based fuzzy 'C' means partition clustering. The 'C' number of clusters and centroids are initialized. The Kleinberg's Hyper-richness uses the core data point module as the centroid of the clusters. The authority sector module computes the distance between the centroid and data points. Then the gradient descent function used to find the minimum distance between the centroid and data points. After that, the authority sector module groups similar data points into the cluster. This process minimizes the time and space complexity.
- To increase the clustering accuracy and minimize the false positive rate, the Kleinberg's Hyper-richness property is applied. The property is employed to correctly group all the similar data points into the cluster.

The paper is ordered as follows. Section 2 reviews the related works. In section 3, the description of the KHBP-FCMPC technique is presented with a neat diagram. Experimental evaluation of proposed KHBP-FCMPC technique and state-of-art methods are discussed with the temporal dataset in section 4. Section 5 provides the experimental results and discussion of certain parameters with the table and graphical representation. Finally, section 6 concludes the proposed work.

II.RELATED WORKS

A novel spatial and temporal similarity measurements based clustering technique was introduced in [11]. The accurate data clustering was not performed with minimal time complexity. An enhanced spatiotemporal clustering technique was introduced in [12] for grouping the spatiotemporal data with less runtime. But the technique failed to minimize the false positive rate during the clustering process. A weighted clustering ensemble algorithm was developed in [13] for grouping the temporal data with high quality. Though the ensemble algorithm minimizes the computational complexity, the algorithm was complex on processing the bitemporal dataset correctly without any prior knowledge. A hierarchical aligned cluster analysis (HACA) was performed in [14] for grouping and visualizing the time series data. This clustering technique failed to minimize the computational complexity in terms of both space and time. An adaptive method was introduced in [15] for grouping the spatiotemporal events by finding the nearest neighbor. The adaptive method failed to redefine the distance between the events for grouping the entire data. Though the method minimizes the time complexity, the space complexity was not minimized. A hierarchical trajectory clustering technique was introduced in [16] for spatiotemporal periodic pattern mining. But, the technique was not suitable for multi-level time periods and complex datasets. A temporal-constrained sub-trajectory cluster analysis was performed in [17] using partition based clustering technique. The cluster analysis was not directly applicable to large datasets. A mixed fuzzy clustering (MFC) algorithm was developed in [18] for grouping the data with dynamic time warping (DTW) distance. Though the algorithm increases the clustering accuracy, the performance of the false positive rate remained unsolved. The Temporal Pattern Miner (TPMiner) and Probabilistic Temporal Pattern Miner (P-TPMiner) were developed in [19] to effectively determine the temporal pattern and probabilistic temporal patterns respectively. The methods outperform well with less runtime and memory usage. But the accurate clustering was not improved. A dynamic stochastic block model was introduced in [20] for grouping the temporal data across the various time steps. But the model was not efficient for grouping the temporal data with less complexity. HMM-based hybrid meta-clustering ensemble was designed in [21] to solve the problems of initialization and model selection for temporal data clustering. But the model did not reduce computational cost. Sequential Subspace Clustering via Temporal Smoothness [22] for minimizing the exponential variance of short term historical data. Efficient Data Stream Clustering with Sliding Windows Based on Locality-Sensitive Hashing was introduced in [23] for improving data stream clustering over sliding windows. But a fixed number of clusters should be specified before clustering. The issues are identified from the above-said methods are addressed by introducing a novel technique called KHBP-FCMPC technique. The processes of KHBP-FCMPC technique are explained in the next section.

III. KLEINBERG'S HYPER-RICHNESS BASED FUZZY C MEANS PARTITION CLUSTERING WITH BITEMPORAL DATA

Bitemporal data clustering is the process of partitioning the huge dataset into dissimilar groups over the time instances. The data points within the group are similar to each other. The processing of large dataset is very difficult using conventional data processing tools. In this case, partitions the large dataset into different groups is essential for identifying the complete knowledge about the data elements and also help the users to understand the structure of the dataset. The conventional clustering technique increases the complexity and not much efficient for processing the large dataset. Therefore, an effective clustering technique is required to group the similar type of data with less complexity. Based on this motivation, Kleinberg's Hyper-Richness Bitemporal property based fuzzy c means Partition Clustering (KHBP-FCMPC) technique is introduced. In KHBP-FCMPC technique, the author Jon Kleinberg's defines the richness property while clustering the data points. The richness property defines that any partition of the bitemporal data is obtained by modifying the distance function.

Based on this property, the clustering is performed to attain higher accuracy and minimal complexity. The architecture of the proposed KHBP-FCMPC technique is illustrated in figure 1.

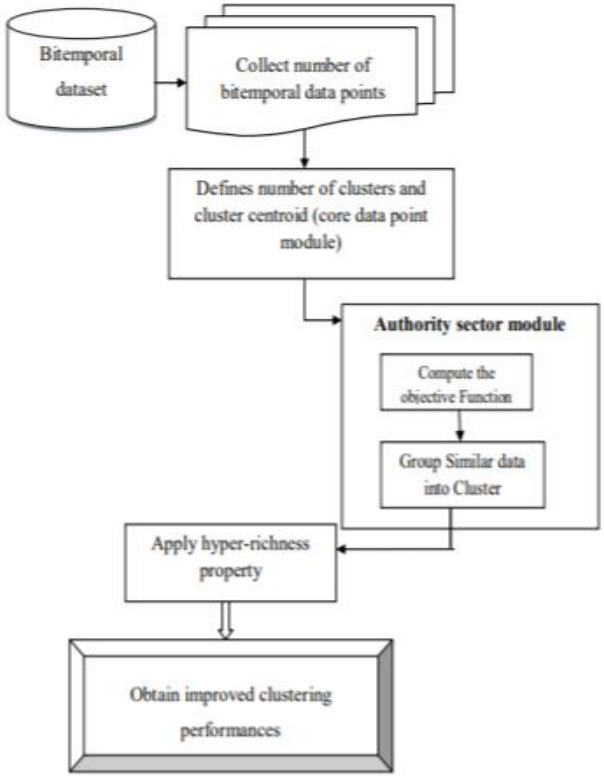

Figure 1: Architecture diagram of the KHBP-FCMPC technique

Figure 1 illustrates an architecture diagram of KHBP-FCMPC technique to group the bitemporal data with high accuracy. Initially, the bitemporal data points are collected from the large dataset. After collecting the data points, the numbers of clusters are initialized randomly for partitioning the data points into different groups. While partitioning the data points, the richness property is applied to find more similar data. The KHBP-FCMPC technique uses the core data point module (i.e. cluster centroid) and authority sector module to minimize the data clustering time. Core data point module served as the center of the cluster. Authority sector module authorizes (i.e. allows) the data points into the cluster based on the distance measure. The clustering process is described as follows. Let us consider the number of bitemporal data collected from the bitemporal dataset $F\ (D)$.

$$t_1, t_2, t_3, \ldots . t_n \in F(D) \qquad (1)$$

From (1),$t_1, t_2, t_3, \ldots . t_n$ represents the number of bitemporal data, $F\ (D)$ represents the bitemporal dataset. Initially, 'c' numbers of clusters $c_1, c_2, c_3, \ldots . c_c$ are initialized. For each cluster, the centroid is computed for grouping data points into the cluster. In KHBP-FCMPC technique, the membership grades are allocated to each data points which specify the degree to which data points belongs to the cluster. The centroid of the cluster defines the mean of entire data points in that cluster. Therefore, the centroid is computed using the following mathematical equations,

$$\alpha_j = \frac{\sum \omega_c(t)^n * t}{\sum \omega_c(t)^n} \qquad (2)$$

From (2),α_j denotes a cluster centroid, any data point 't' has a set of coefficients giving the degree of being in the c^{th} cluster $i.e\ \omega_c\ (t)$, n represents the fuzzifier determines the level of cluster fuzziness. The centroid of the cluster is a core data point module. The objective function of the KHBP-FCMPC technique is expressed as follows,

$$f\ (\mathrm{x}) = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{n} * \left\| t_i - \alpha_j \right\|^2 \qquad (3)$$

From (3), $f\ (x)$ denotes a objective function, μ_{ij} represents the membership value, n represents the fuzzifier, t_i represents the bitemporal data, α_j denotes a cluster centroid, $\left\| t_i - \alpha_j \right\|^2$ represents the Euclidean distance between the data (t_i) and the cluster centroid (α_j). The above equation (3) shows that the clustering algorithm computes the objective function by the multiplication of the membership valuesμ_{ij}. The membership is computed based on the distance function,

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{n} \left(\frac{t_i - \alpha_j}{t_i - \alpha_k} \right)^{\frac{2}{n-1}}} \qquad (4)$$

From (4), μ_{ij} denotes a membership values, t_i represents the bitemporal data, α_j, α_k denotes a centroid of the j^{th}, k^{th} cluster. The KHBP-FCMPC technique uses the gradient descent function to minimize the objective function which is defined as follows,

$$\arg\min f\ (x) \qquad (5)$$

From (5), arguments of the minimum are abbreviated as argmin, $f\ (x)$ represents the objective function. After that, the authority sector module in the proposed technique assigns the data which is closer to the centroid of the cluster. Therefore, the data which is closest to the centroid is said to be a member of that specific cluster. Finally, the cluster centroid is updated to group all the data points into the cluster. After that, the proposed technique uses the Kleinberg's Hyper-Richness Bitemporal property to groups the entire data points into the cluster by changing the distances between the data point and cluster centroid.

Hyper-Richness Bitemporal property:

Let us consider the total bitemporal dataset is $F\ (D)$ and it also considers the pairs (t_i, d) where t_i represents the bitemporal data and d denotes a distance function. The richness property defines the total dataset is equal to the set of all possible partitions which is expressed as follows,

$$F\ (D, t_i, d) = \{p_1, p_2, p_3, \ldots . p_c\} \qquad (6)$$

From (6), $F\ (D, t_i, d)$ represents the total dataset there exist distance functions d, $\{p_1, p_2, p_3, \ldots . p_c\}$ denotes a several possible partitions. $F\ (D, t_i, d)$ is hyper-rich if it is capable of producing all possible partitions i.e. total bitemporal dataset is equal to set of all possible partitions p_c. Hence it is called as hyper-Richness Bitemporal property. By applying this property, the entire similar bipemporal data are correctly grouped into any one of the clusters. As a result, the clustering accuracy is improved and minimized the false positive rate. The flow diagram of the proposed KHBP-FCMPC technique is illustrated in the following figure.

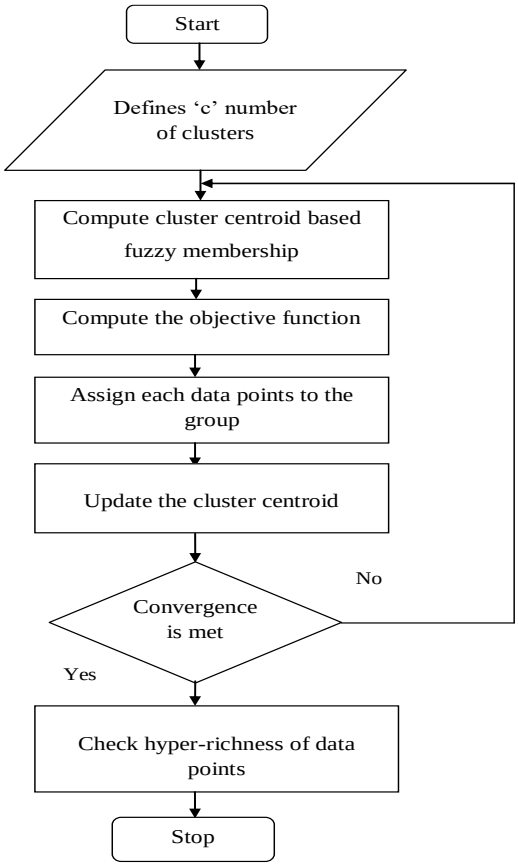

Figure 2: Flow process of proposed KHBP-FCMPC technique

Figure 2 shows the flow process of the proposed KHBP-FCMPC technique to partitions the bitemporal data with higher accuracy based on the distance. The minimum distance among the data point and specific cluster centroid have a high probability to group similar data into the particular cluster. The initial cluster centroid is updated and groups each data points until the convergence is met.

The algorithmic process of KHBP-FCMPC technique is described as follows,

Input : Number of temporal data $t_1, t_2, t_3, \ldots . t_n$
Output: Improves the clustering accuracy
Begin
1. Initialize 'c' number of clusters
2. Compute the cluster centroid
3. Compute the objective function $f(x)$ based on fuzzy membership function
4. Find minimum distance $arg\ min\ f(x)$
5. Group data points into the clusters
6. Update the cluster centroid α_j
7. **If** convergence is not met **then**
8. go to step 2
9. **else**
10. check hyper-richness of the data points
11. **end if**
12. Group all the data points into the cluster
End

Algorithm 1: Kleinberg's Hyper-richness Bitemporal property based fuzzy c means partition Clustering

Algorithm 1 describes the clustering of bitemporal data with high accuracy. Initially, 'c' numbers of clusters are randomly initialized. After that, cluster centroid is computed based on mean of all data points. Then the objective function is computed based on distance function and membership value. The minimum distance between the bitemporal data points and centroid is grouped into the specific cluster. This process is iterative until the convergence is met. Finally, the richness property is verified with the total dataset and the partitions. As a result, the data point within the cluster has high similarity and between the clusters has less similarity. The verification result shows the entire data are correctly grouped into the clusters resulting in increases the accuracy.

IV.EXPERIMENTAL EVALUATION

An experimental evaluation of KHBP-FCMPC technique and Bi-weighted ensemble approach [1] and Parametric Link among Multinomial Mixtures (PLMM) [2] are implemented using Java language. The Activity Recognition from Single Chest-Mounted Accelerometer Data Set is used for the experimental evaluation. This dataset is employed for purpose of activity recognition with Bitemporal data. The dataset is collected from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+ChestMounted+Accelerometer#). This dataset gathers the temporal data points from a wearable accelerometer fixed on the chest. Bitemporal data denotes the valid time and transaction time. The dataset characteristics are univariate, sequential and time-Series. The attributes characteristics are real and the association tasks performed by the dataset are clustering and classification. The uncalibrated accelerometer data are gathered from 15 participants presenting 7 dissimilar activities. This dataset provides the challenges for detection and validation of people through the various motion patterns. The data files are separated for each participant. Each file comprises the five columns such as sequential number, x acceleration, y acceleration, z acceleration, and labels. The labels are represented by the numbers. Totally seven different labels are presented for recognizing the activities. In this dataset, the labels are considered as an output result.

Table 1: Labels information

Labels number	Activities
1	Working at computer
2	Standing Up, Walking and Going updown stairs
3	Standing
4	Walking
5	Going UpDown Stairs
6	Walking and Talking with Someone
7	Talking while Standing

The experiments are carried out with different parameters such as clustering accuracy, false positive rate, time complexity, and space complexity. Totally ten various runs are performed with a number of temporal data. For the experimental consideration, the numbers of temporal data are collected from 1000 to 10000. The experimental results are discussed in the next section.

V.PERFORMANCE RESULTS AND DISCUSSION

The results attained from the experimental evaluation of three clustering methods KHBP-FCMPC technique and Bi-weighted ensemble approach [1] and PLMM [2] are discussed in this section with a table and graphical results. The various metrics such as clustering accuracy, false positive rate, time complexity and space complexity are considered for evaluating performance results of three different clustering methods. For each subsection, sample mathematical calculation is presented.

5.1 Performance results of clustering accuracy

Clustering accuracy is defined as the ratio of a number of similar data points are correctly grouped into the cluster to the total number of data points. The clustering accuracy is mathematically formulated as follows,

$$CA = \frac{Number\ of data\ points\ correctly\ grouped}{Number\ of\ data\ points} * 100 \qquad (7)$$

From (7), CA represents the clustering accuracy. The accuracy is measured in percentage (%). Higher the temporal data point clustering accuracy, more efficient the method is said to be. The sample mathematical computation of clustering accuracy is presented as follows,

Sample mathematical calculation for clustering accuracy:

- **Proposed KHBP-FCMPC technique**: Number of data points correctly grouped is 860 and the total number of data points is 1000. Then the clustering accuracy is calculated as follows,

$$clustering\ accuracy = \frac{860}{1000} * 100 = 86\%$$

- **Existing Bi-weighted ensemble approach**: Number of data points correctly grouped is 780 and the total number of data points is 1000. Then the clustering accuracy is calculated as follows,

$$clustering\ accuracy = \frac{780}{1000} * 100 = 78\%$$

- **Existing PLMM**: Number of data points correctly grouped is 730 and the total number of data points is 1000. Then the clustering accuracy is calculated as follows,

$$clustering\ accuracy = \frac{730}{1000} * 100 = 73\%$$

Table 2: Clustering accuracy versus number of data points

Number of data points	Clustering accuracy (%)		
	KHBP-FCMPC	Bi-weighted ensemble approach	PLMM
1000	86	78	73
2000	84	76	72
3000	85	77	74
4000	89	83	78
5000	91	86	82
6000	89	82	77
7000	91	87	82
8000	92	86	81
9000	90	84	79
10000	92	87	84

Table 2 illustrates the temporal data clustering accuracy using three different clustering techniques namely KHBP-FCMPC, Bi-weighted ensemble approach [1] and PLMM [2]. The clustering accuracy is computed based on the number of temporal data points. For the experimental consideration, ten different temporal data points are considered from 1000 to 10000. The above-reported results show that the temporal data clustering accuracy of KHBP-FCMPC technique is considerably increased when compared to existing clustering techniques. The results are plotted in the graph as shown in figure 3.

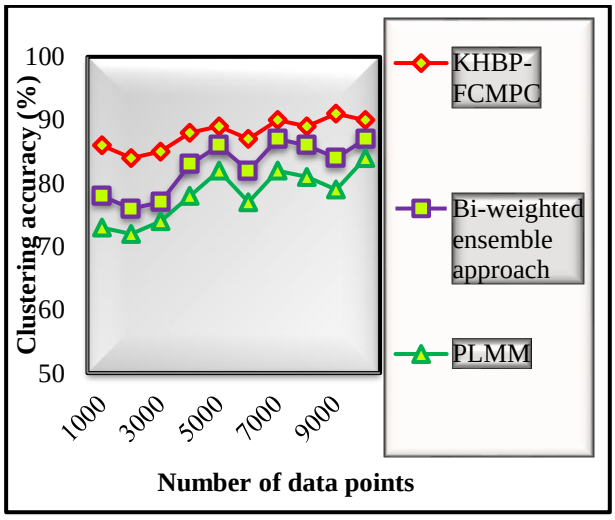

Figure 3: Performance results of clustering accuracy versus number of data points

Figure 3 illustrates the performance results of clustering accuracy based on the number of data points. The numbers of bitemporal data points are collected from the dataset for computing the clustering accuracy. The time series data are taken as input in the 'x' direction and the results of clustering accuracy are attained at the 'y' direction.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The graphical results clearly report that the clustering accuracy of KHBP-FCMPC technique is considerably increased when compared to the existing clustering techniques. This improvement is achieved by a partition based clustering technique. The time series data are collected from the activity recognition dataset. Then the seven clusters are initialized and the centroid of each cluster is computed. Followed by, the distance between the data points and the centroid is calculated to group the similar bitemporal data points into the clusters. The Kleinberg's hyper-richness bitemporal property is applied while clustering the data points to verify all the similar data points are correctly grouped into the clusters. By this way, the seven human activities are correctly recognized with higher accuracy. The ten different results are obtained with various input bitemporal data points. The clustering accuracy results of KHBP-FCMPC technique is compared to the clustering accuracy of conventional techniques. The comparison results clearly report that the clustering accuracy is significantly improved by 7% and 13% when compared to existing Bi-weighted ensemble approach [1] and PLMM [2] respectively.

5.2 Performance results of false positive rate

The false positive rate is defined as the ratio of a number of data points are incorrectly grouped into the cluster to the total number of data points. The mathematical formula for computing the false positive rate is expressed as follows,

$$FPR = \frac{Number\ of data\ points\ incorrectly\ grouped}{Number\ of\ data\ points} * 100 \quad (8)$$

From (8), FPR represents the false positive rate. The false positive rate is measured in terms of percentage (%). The sample mathematical calculation of false positive rate using three different methods are presented as follows,

Sample mathematical calculation for false positive rate:

- **Proposed KHBP-FCMPC technique**: Number of data points incorrectly grouped is 140 and the total number of data points is 1000. Then the false positive rate is computed as follows,

$$False\ positive\ rate = \frac{140}{1000} * 100 = 14\%$$

- **Existing Bi-weighted ensemble approach**: Number of data points incorrectly grouped is 220 and the total number of data points is 1000. Then the false positive rate is calculated as follows,

$$False\ positive\ rate = \frac{220}{1000} * 100 = 22\%$$

- **Existing PLMM**: Number of data points incorrectly grouped is 270 and the total number of data points is 1000. Then the false positive rate is calculated as follows,

$$False\ positive\ rate = \frac{270}{1000} * 100 = 27\%$$

Table 3: False positive rate versus number of data points

Number of data points	False positive rate (%)		
	KHBP-FCMPC	Bi-weighted ensemble approach	PLMM
1000	14	22	27
2000	16	24	28
3000	15	23	26
4000	12	17	22
5000	11	14	18
6000	13	18	23
7000	10	13	18
8000	11	14	19
9000	9	16	21
10000	10	13	16

The performance results of false positive rate versus a number of data points are illustrated in table 3. The false positive rate is calculated to find numbers of data are incorrectly grouped into clusters. The above table shows that the ten different results of false positive rate versus a number of data points. Therefore, the incorrect clustering results of the proposed KHBP-FCMPC technique are significantly minimized when compared to existing clustering techniques. Let us consider the number of data points is 1000, the number of data points are incorrectly grouped is 140 by applying KHBP-FCMPC technique resulting false positive rate is 14%. Similarly, the number of data points are incorrectly grouped are 220 and 270 by applying Bi-weighted ensemble approach [1] and PLMM [2]. The resultant false positive rates are 22% and 27% respectively. Similarly, the nine remaining results are computed using the three different techniques. The performance results are shown in the two-dimensional graph.

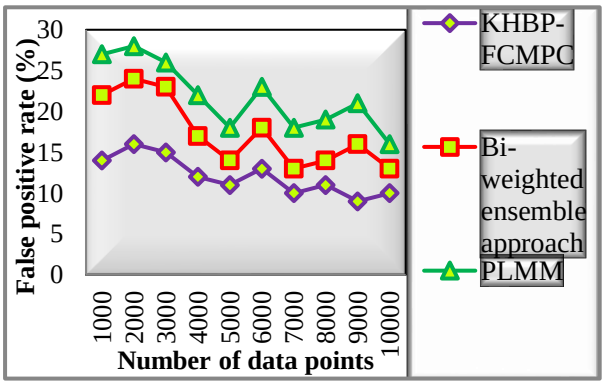

Figure 4: Performance result of false positive rate versus the number of data points

Figure 4 illustrates the experimental results of the false positive rate based on the number of data points using Activity Recognition dataset. By using the Activity Recognition dataset, the different activities performed by the human are identified through the clustering process. In figure 3, three different color lines such as violet, red and green indicate the false positive rate of three different clustering methods namely KHBP-FCMPC technique, Bi-weighted ensemble approach [1] and PLMM [2] respectively.

Retrieval Number: H7095068819/19©BEIESP
DOI: 10.35940/ijitee.H7095.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The graphical results show that the KHBP-FCMPC technique minimizes the false positive rate than the existing methods. This is because the KHBP-FCMPC technique uses the gradient descent function to find the minimum distance between the centroid and data points. This function minimizes the incorrect data point clustering. In addition, the hyper-richness property is applied for verifying the entire bitemporal dataset is correctly partitioned into the seven different classes. In this case, the entire uncalibrated accelerometer data are correctly grouped into the seven types of clusters resulting in minimizes the clustering accuracy. Totally ten various experimental results are obtained with different bitemporal data. The performance results of three methods illustrate that the false positive rate of KHBP-FCMPC technique is minimized by 29% and 44% when compared to existing Bi-weighted ensemble approach [1] and PLMM [2] respectively.

5.3 Performance results of time complexity

Time complexity is defined as the amount of time required to group the similar data point into the cluster. The time complexity is mathematically computed using the following equations,

$$TC = n * t_c\ (group\ one\ \ data\ point) \qquad (9)$$

From (9), TC represents the time complexity, n denotes a number of data points, t_c is the time taken for grouping the one data point. The time complexity is measured in the unit of milliseconds (ms). The calculation of time complexities using three different methods are presented as follows.

Sample Mathematical calculation for time complexity:

- **Proposed KHBP-FCMPC**: Number of data is 1000 and the time taken for grouping the one data is 0.025ms , then the time complexity is mathematically calculated as follows,

$$TC = 1000 * 0.025ms\ = 25ms$$

- **Existing Bi-weighted ensemble approach:** Number of data is 1000 and the time taken for grouping the one data is 0.028ms, then the time complexity is mathematically calculated as follows,

$$TC = 1000 *\ 0.028ms = 28ms$$

- **Existing PLMM**: Number of data is 1000 and the time taken for grouping the one data is 0.034ms , then the time complexity is mathematically calculated as follows,

$$TC = 1000\ *\ 0.034ms = 34ms$$

Let us consider the number of input bitemporal data points is 1000. The proposed KHBP-FCMPC technique takes 25ms for clustering the similar human activity data. The time complexity of the other two clustering techniques Bi-weighted ensemble approach and PLMM are 28ms and 34ms respectively. Similarly, the remaining nine results are computed with a different number of data points. The experimental results are reported in table 4.

Table 4: Time complexity versus number of data points

Number of data points	Time complexity (ms)		
	KHBP-FCMPC	Bi-weighted ensemble approach	PLMM
1000	25	28	34
2000	28	32	38
3000	33	39	45
4000	40	44	52
5000	42	45	50
6000	44	48	54
7000	46	50	56
8000	50	54	60
9000	54	59	64
10000	57	62	68

Table 4 describes the time complexity of the three different clustering techniques KHBP-FCMPC technique, Bi-weighted ensemble approach [1] and PLMM [2]. The above table values clearly show that the time complexity of KHBP-FCMPC technique is minimized compared to the state- of-the-art methods. The various results of time complexity are illustrated in figure 5.

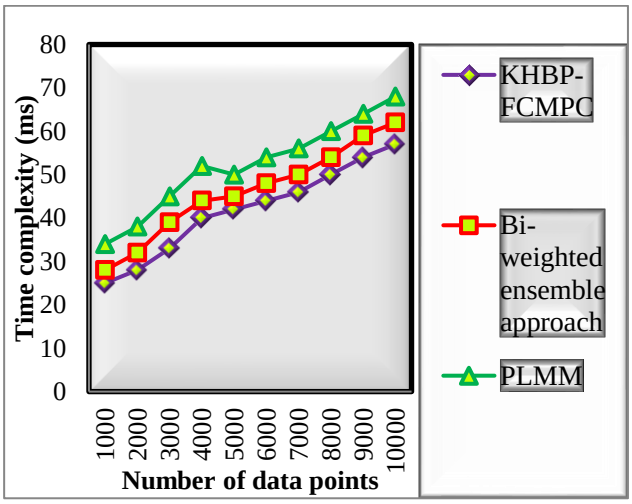

Figure 5: Performance results of time complexity versus number of data points

Figure 5 shows the performance results of time complexity using human activity recognition dataset. The above graphical results clearly illustrate that the proposed KHBP-FCMPC technique minimizes the time complexity while clustering the data points when compared to existing methods. During the clustering process, the KHBP-FCMPC technique uses the two module namely core data point module and authority sector module. The core data point module act as a centroid of the cluster. Then the authority sector module uses the gradient descent function to find the distance between the data points and centroid of the cluster.

The authority sector module assigns the data points into the cluster based on the distance measure. These two modules help to minimize the time complexity. The average of results minimizes the time complexity by 9% and 20% when compared to existing Bi-weighted ensemble approach [1] and PLMM [2] respectively.

5.4 Performance results of space complexity

Space complexity is computed as an amount of storage space consumed for storing similar temporal data points. The formula for computing the space complexity is expressed as follows,

$$SC = n * space\ (storing\ one\ data\ point) \quad (10)$$

From (10), SC represents the space complexity, n denotes a number of data points. Space complexity is measured in the unit of Mega bytes (MB). The method consumes less storage space for storing the similar data. The calculations for space complexity with three different methods are presented as follows.

Sample mathematical calculation for space complexity

- **Proposed KHBP-FCMPC**: Number of data points are 1000 and space for storing the one data point is 0.011MB. Then the space complexity is calculated as follows,

$$SC = 1000 * 0.011MB = 11MB$$

- **Existing Bi-weighted ensemble approach:** Number of data points are 1000 and space for storing the one data point is 0.013. Then the space complexity is calculated as follows,

$$SC = 1000 * 0.013\ MB = 13MB$$

- **Existing PLMM:** Number of data points are 1000 and space for storing the one data point is 0.015MB. Then the space complexity is calculated as follows,

$$SC = 1000 * 0.015MB = 15MB$$

Table 5: Space complexity versus number of data points

Number of data points	Space complexity (MB)		
	KHBP-FCMPC	Bi-weighted ensemble approach	PLMM
1000	11	13	15
2000	13	15	18
3000	14	16	19
4000	15	18	20
5000	16	19	21
6000	17	21	23
7000	18	20	22
8000	21	23	26
9000	22	24	27
10000	23	25	28

Table 5 describes the performance results of space complexity with respect to a number of data points. The bitemporal data points are taken from 1000 to 10000. The above table value clearly illustrates that the space complexity of the KHBP-FCMPC technique is significantly minimized than the existing methods. The results are illustrated in figure 6.

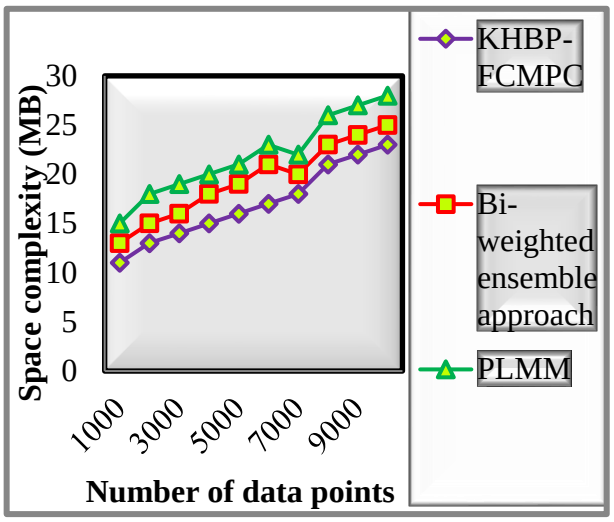

Figure 6: Performance results of space complexity versus number of data points

Figure 6 depicts the experimental results of the space complexity with respect to a number of data points. The above figure confirms that the amount of storage space of the proposed KHBP-FCMPC technique is considerably reduced. While processing the large volume of data points, high storage space is required resulting in increasing the dimensionality. Therefore the proposed algorithm partitions the large dataset into different clusters for minimizing the dimensionality. As a result, the less amount of memory space is exploited for storing bitemporal data points. This is because of the KHBP-FCMPC technique only groups the similar data points into the clusters. The similar data point is identified through the distance measure. Then the gradient descent function only finds the closest data points from the cluster centroid. Thus, the similar temporal (i.e., time series) data are grouped into dissimilar clusters. This process of KHBP-FCMPC technique minimizes the space complexity.

Let us consider, a number of data points is 1000, an amount of space consumed for storing the bitemporal data points are 11MB using KHBP-FCMPC technique. The space complexity of the Bi-weighted ensemble approach [1] and PLMM [2] are 13MB and15MB respectively. Similarly, the nine remaining runs are performed and take the average of ten results. The average results show that the space complexity of KHBP-FCMPC technique is considerably minimized by 13% and 23% when compared to state-of-the-art methods. The above experimental results show that the KHBP-FCMPC technique improves the computation of bitemporal data in terms of high clustering accuracy and less false positive rate, time complexity as well as space complexity.

Retrieval Number: H7095068819/19©BEIESP
DOI: 10.35940/ijitee.H7095.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

VI.CONCLUSION

An efficient technique called Kleinberg's Hyper-richness Bitemporal property based fuzzy c means partition Clustering (KHBP-FCMPC) is developed for improving the bitemporal clustering accuracy with minimal complexity. The Hyper-richness property-based clustering is performed to group the similar data points into the cluster through the distance measure. The gradient descent function finds the data points with minimum distance. Then the authority sector module assigns the similar data points into the cluster. This process minimizes the space and time complexity while clustering the large volume of data points. The richness property also applied for verifying the data points correctly grouped into the clusters. Therefore, the richness property used in KHBP-FCMPC technique increases the clustering accuracy and minimizes the false positive rate. The experimental evaluation is performed using Activity Recognition from Single Chest-Mounted Accelerometer Dataset with various parameters such as clustering accuracy, false positive rate, time complexity and space complexity. The experimental results discussed that the performance of KHBP-FCMPC technique improves the clustering accuracy and minimizes the false positive rate, time complexity as well as space complexity when compared to state-of-art methods.

REFERENCES

1. YunYang and Jianmin Jiang, "Bi-weighted ensemble via HMM-based approaches for temporal data clustering", Pattern Recognition, Elsevier, Volume 76, April 2018, Pages 391 - 403.
2. Md. Abul Hasnat, Julien Velcin, Stephane Bonnevay, Julien Jacques, "Evolutionary clustering for categorical data using parametric links among multinomial mixture models", Econometrics and Statistics, Elsevier, Volume 3, July 2017, Pages 141-159.
3. Hani El Assaad, Allou Saméa, Gérard Govaert and Patrice Aknina, "A variational Expectation–Maximization algorithm for temporal data clustering", Computational Statistics & Data Analysis, Elsevier, Volume 103, November 2016, Pages 206 - 228.
4. Saeid Soheily-Khah Ahlame, Douzal-Chouakria, EricGaussier, "Generalized k-means-based clustering for temporal data under weighted and kernel time warp", Pattern Recognition Letters, Elsevier, Volume 75, May 2016, Pages 63 - 69.
5. Joseph Crawford and Tijana Milenković, "ClueNet: Clustering a temporal network based on topological similarity rather than denseness", PLoS ONE, Volume 13, Issue 5, 2018, Pages 1 - 25.
6. Lucia Paci and Francesco Finazzi, "Dynamic model-based clustering for spatio-temporal data", Statistics and Computing, Springer, March 2018, Volume 28, Issue 2, Pages 359 – 374.
7. Initialization and Model Selection for HMM-Based Hybrid [7] Yun Yang and Jianmin Jiang, "Adaptive Bi-Weighting Toward Automatic Meta-Clustering Ensembles", IEEE Transactions on Cybernetics, Volume 49, Issue 5, 2018, Pages 1 – 12.
8. Yongping Zhang znd Lun Liu, "Understanding temporal pattern of human activities using Temporal Areas of Interest", Applied Geography, Elsevier, Volume 94, May 2018, Pages 95-106.
9. Nazanin Asadi , Abdolreza Mirzaei , Ehsan Haghshenas, "Creating Discriminative Models for Time Series Classification and Clustering by HMM Ensembles", IEEE Transactions on Cybernetics , Volume 46 , Issue 12 , 2016 , Pages 2899 – 2910.
10. Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique", The Scientific World Journal, Hindawi Publishing Corporation, Volume 2014, March 2014, Pages 1 - 12.
11. Xin Yao , Di Zhu ,Yong Gao , Lun Wu , Pengcheng Zhang , Yu Liu, "A Stepwise Spatio-Temporal Flow Clustering Method for Discovering Mobility Trends", IEEE Access, Volume 6, August 2018, Pages 44666 – 44675.
12. K.P. Agrawal, Sanjay Garg, Shashikant Sharma and Pinkal Patel, "Development and validation of OPTICS based spatio-temporal clustering technique", Information Sciences, Elsevier, Volume 369, November 2016, Pages 388 - 401.
13. Yun Yang and Ke Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 2, 2011, Pages 307 – 320.
14. Feng Zhou , Fernando De la Torre , Jessica K. Hodgins, "Hierarchical Aligned Cluster Analysis for Temporal Clustering of Human Motion", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 35 , Issue 3 , March 2013, Pages 582 – 596.
15. Zhilin Li, Qiliang Liu, Jianbo Tang, Min Deng, "An adaptive method for clustering spatio-temporal events", Transactions in GIS, Wiley online library, Volume 22, Issue 1, February 2018, Pages 323 - 347.
16. Dongzhi Zhang, Kyungmi Lee and Ickjai Lee "Hierarchical Trajectory Clustering for Spatio-temporal Periodic Pattern Mining", Expert Systems with Applications, Elsevier, Volume 92, February 2018, Pages 1 - 11.
17. Nikos Pelekis , Panagiotis Tampakis, Marios Vodas, Christos Doulkeridis, Yannis Theodoridis, "On temporal-constrained sub-trajectory cluster analysis", Data Mining and Knowledge Discovery, Springer, Volume 31, Issue 5, September 2017, Pages 1294 – 1330.
18. Cátia M. Salgado, Marta C. Ferreira and Susana M. Vieira , "Mixed Fuzzy Clustering for Misaligned Time Series", IEEE Transactions on Fuzzy Systems , Volume 25, Issue No. 6, December 2017, Pages 1777 – 1794.
19. Yi-Cheng Chen, Wen-Chih Peng and Suh-Yin Lee, "Mining Temporal Patterns in Time Interval-Based Data", IEEE Transactions on Knowledge and Data Engineering , Volume 27 , Issue 12 , 2015, Pages 3318 – 3331.
20. Catherine Matias and Vincent Miele, "Statistical clustering of temporal networks through a dynamic stochastic block model", Journal of Royal Statistical Society, Volume 79, Issue 4, September 2017, Pages 1119 -1141.
21. Yun Yang and Jianmin Jiang, "HMM-based hybrid meta-clustering ensemble for temporal data", Pattern Recognition, Elsevier, Volume 56, January 2014, Pages 299 – 310.
22. Haijun Liu, Jian cheng, and Feng wang, **"**Sequential Subspace Clustering via Temporal Smoothness for Sequential Data Segmentation" Journal of IEEE Transactions on Image Processing, vol. 27, issue 2, Feb 2018, pp. 866 – 878.
23. Jonghem youn, Junho shim, Sang-Goo Lee, "Efficient Data Stream Clustering With Sliding Windows Based on Locality-Sensitive Hashing" Journal of IEEE Access, Volume.6, October 2018, Page 63757 – 63776.

AUTHOR PROFILE

L. Jaya Singh Dhas, received his Bachelor of Science in Computer Science from Madurai Kamaraj University - Madurai, India in the year 1991 and Master of Computer Applications from Bharathidhasan University - Tiruchirapalli in the year 1996 and M.Phil from Alagappa University - Karikudi in the year 1998. He is currently working as a Assistant Professor, Department of Computer Science, Scott Christian College (Autonomous), Nagercoil since 1998 and his research work focuses on Algorithms, Big Data Analytics, Data Mining.
His Orchid ID **is** https://orcid.org/0000-0002-0136-3941

Dr. B. Mukunthan Ph.D pursed Bachelor of Science (Computer Science) from Bharathiar University, India in 2004 and Master of Computer Applications from Bharathiar University in year 2007 and Ph.D from Anna University - Chennai in 2013. He is currently working as Research Advisor in Department of Computer Science, Bharathidasan University, Tiruchirapalli since 2016. He is a member of IEEE & IEEE computer society since 2009, a life member of the MISTE since 2010. He has published more than 10 research papers in reputed international journals. He is also Microsoft Certified Solution Developer. His main research work focuses on Algorithms, Bioinformatics, Big Data Analytics, Data Mining, IOT and Neural Networks. He also invented a Novel and Efficient online Bioinformatics Tool and filed for patent. He has 12 years of teaching experience and 10 years of Research Experience.
His Orchid ID is https://orcid.org/0000-0001-8452-3164