Heuristic Search Based Feature Selection and Discretive Self-Organized Map Clustering for Spatio-Temporal Pattern Discovery

R.Sarala, V.Saravanan

Abstract*: Spatio-temporal pattern discovery is an essential one in data mining for predictive analytics. Since it manages both space and time information depending on their characteristics and the preferred applications performances. The predictive analytics uses the Spatio-temporal features to discover future outcomes. The several works have been done in the Spatio-temporal pattern discovery. But the accurate pattern discovery is the major challenges. In order to improve the accurate pattern discovery, Heuristic Best-First Search based Discretized Self-Organizing Feature Map (HBFS-DSOFM) Model is introduced. The HBFS-DSOFM model comprises two processes namely, Spatio-temporal feature selection and clustering. Initially, the Heuristic Best-First Search Algorithm is used for selecting the relevant Spatio-temporal features from the large dataset for pattern discovery. Best-first search explores a decision tree for selecting the relevant Spatio-temporal features through the maximum information gain value. After that, the Spatio-temporal data are clustered with the selected features by using Discretized Self-Organizing Feature Mapping Algorithm for Spatio-temporal pattern discovery. In Discretized Self-Organizing Feature Mapping, input spatio-temporal data is connected to the prototype neurons through the synaptic weight. For the clustering process, weights of the neurons (i.e. cluster) are initialized with random values. After that, the Manhattan distance is used to compute the distance between the input vector and cluster weight value. The gradient descent is applied to discover closest distance. The cluster whose weight is closest to the input data is grouped into the particular cluster. Then the weight of the cluster is updated with the previous weight value for grouping the entire data. This clustering process gets iterated until it satisfies termination condition. Finally, the outputs of Spatio-temporal data are combined to form a spatio-temporal pattern for efficient predictive analytics. Experimental evaluation is carried out for El Nino Dataset and taxi trajectory dataset using the factors such as time complexity, clustering accuracy, and false positive rate. The results confirm that the proposed HBFS-DSOFM model increases the Spatio-temporal pattern discovery in terms of high clustering accuracy with a less false positive rate as well as minimum time complexity. Based on the clarification, HBFS-DSOFM model is more efficient than the state-of-the-art methods.

***Keywords:* Spatio-temporal pattern discovery, Spatio-temporal feature selection, Heuristic Best-First Search, information gain, Discretized Self-Organizing Feature Mapping, Manhattan distance, gradient descent**

.

Revised Manuscript Received on July 08, 2019.

Mrs.R.Sarala is a Assistant Professor in Department of Computer Science at KG College of Arts and Science, Coimbatore. Email: srisaitechnologymadurai@gmail.com

Dr. V. Saravanan, Associate Professor & HEAD, Hindusthan College of Arts and Science, Department of IT, Hindusthan College of Arts and Science,Coimbatore - 641 028

I. INTRODUCTION

A Spatio-temporal data has become widespread in several applications like public health, public safety, financial fraud detection, transportation, weather forecasting and so on. A Spatio-temporal database comprises the structural variations in space and time. Unlike the traditional dataset are continuous, boundless, and it has a time- variant data distribution. It is a difficult and complex task to discover the interesting patterns from this database. Therefore an efficient data mining techniques such as clustering and classification are used for solving the above issues. Data mining is the process of extracting the significant patterns from the datasets to extract the information and it transforms into a required structure for future use.

A multi valued decision systems approach was developed in [1] for determining the Spatio-temporal patterns from the time series data. In this approach, the rough set theory was applied to choose the important features from the dataset. The accurate clustering was not carried out to find the Spatio-temporal patterns with less error rate. A hierarchical trajectory clustering based periodic pattern mining (PPM) approach was developed in [2] for finding the various Spatio-temporal patterns. Through the hierarchical clustering approach, it detects more periodic patterns, the time complexity was not minimized.

A forward feature selection and random forest algorithm were introduced in [3] for enhancing the performance of the Spatio-temporal prediction. The algorithm failed to prevent the over-fitting in machine learning applications. Gaussian dissimilarity based Similarity Profiled temporal Association Pattern Mining approach was introduced in [4] for identifying the related temporal patterns and minimizing the dissimilarity using fuzzy approach. The approach does not consider the spatial patterns for efficient predictive analytics.

A Spatio-temporal data classification method was presented in [5] with the multidimensional chronological patterns. But the feature selection was not performed to improve the Spatio-temporal data classification with minimum time. A trajectory clustering approach was developed in [6] for discover the spatial and temporal travel patterns. The approach does not obtain accurate

pattern discovery. A Spatio-temporal variable selection-based support vector regression (VS-SVR) model was developed in [7]. The time complexity in the VS-SVR was not minimized.

A Spatio-temporal forecasting method was developed in [8] with spatial and temporal information. The method does not use other spatial data for time series forecasting of different weather variables such as temperature, pressure and so on. A Spatio-temporal trajectory clustering technique was introduced in [9] to identify the heterogeneous trip patterns with minimum computational cost. The clustering process failed to apply the more practical circumstances.

A nonparametric stochastic de-clustering procedure was designed in [10] for finding the triggering patterns of Spatio-temporal event types. The procedure does not obtain exact computational as well as prediction.

The major problems are identified from the above-said literature are high complexity, failure to detect the accurate pattern, lack of improving the clustering accuracy, failure for selecting the more relevant features, high error rate and so on. In order to resolve the issues, an effective HBFS-DSOFM Model is introduced.

The major contributions of the proposed HBFS-DSOFM Model are summarized as follows,

The proposed HBFS-DSOFM Model increases the Spatio-temporal pattern discovery with less complexity. The heuristic approach searches the relevant features among the several features using information gain value. The information gain is used to split the features set into two different subsets as relevant and irrelevant by constructing the decision tree. The feature with maximum information gain value is selected as a relevant. These relevant features are selected for pattern discovery and irrelevant features are removed. This process minimizing the time complexity.

The novel Discretized Self-Organizing Feature Map based clustering approach is applied for Spatio-temporal pattern mining with high accuracy and less false positive rate. In the training phase, discretized representations of input space of the training Spatio-temporal data are collected from the dataset. In Map phase, the input data are mapped into cluster through the synaptic weights. Then the gradient descent function finds the minimum distance between the cluster weight and the input vector for grouping similar data with high accuracy. For each iteration, the weights are updated to group the entire data into the specific cluster. This helps to minimize the false positive rate.

The remainder of the paper is arranged into five various sections. In Section 2, related works along with the discussion based on their capabilities and limitations are discussed. In Section 3, the proposed HBFS-DSOFM Model is described with a neat diagram. In section 4, the experimental evaluation is presented with the relevant datasets. The experimental results are discussed with the several parameters are presented in Section 5. Finally, the conclusion of the paper is presented in Section 6.

Related works

Association rule mining (ARM) was introduced in [11] for identifying the Spatio-temporal patterns. But it failed to effectively provide the location-based information. A Spatio-Temporal - Ordering Points was presented in [12] to discover clustering structure. Though the approach generates the spatial-temporal clusters, the performance was not improved. A cascading Spatio-temporal pattern (CSTP) discovery was presented in [13]. But it failed to apply the other real data analysis including climate/ weather data sets for discovering the patterns for predictive analytics.

Temporal Pattern Miner (TPMiner) and Probabilistic Temporal Pattern Miner (P-TPMiner) were presented in [14] to determine the various interval-based sequential patterns. The approach failed to find the closed patterns and it does not find the spatial information. A new constraint programming model was developed in [15] for detecting the temporal and spatial patterns. But, the model does not accurately detect the patterns since it failed to use any clustering method. Detection of Spatio-temporal patterns was presented in [16] from the location-based social networks. But it failed to detect the temporal dimension of the data-set.

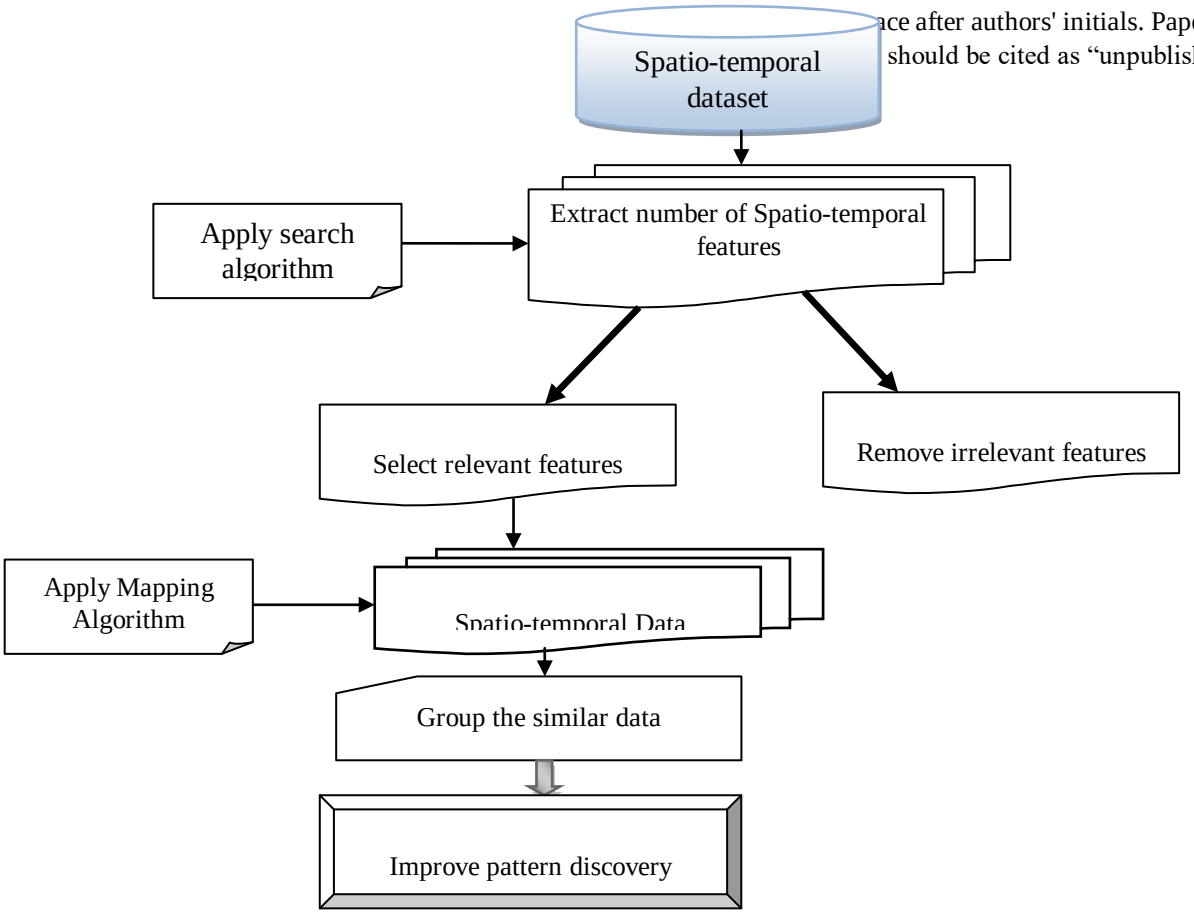
An effective Pattern discovery method called T-pattern analysis was introduced in [17] for identifying the temporal patterns. This analysis was not applied in geographic research for finding the patterns in complex Spatio-temporal data. Socio-spatio-temporal important locations (SSTIL) and SSTIL mining problem were addressed in [18] with spatial and temporal data. But, the accurate spatio-temporal pattern discovery was not attained since it failed to select the relevant features.

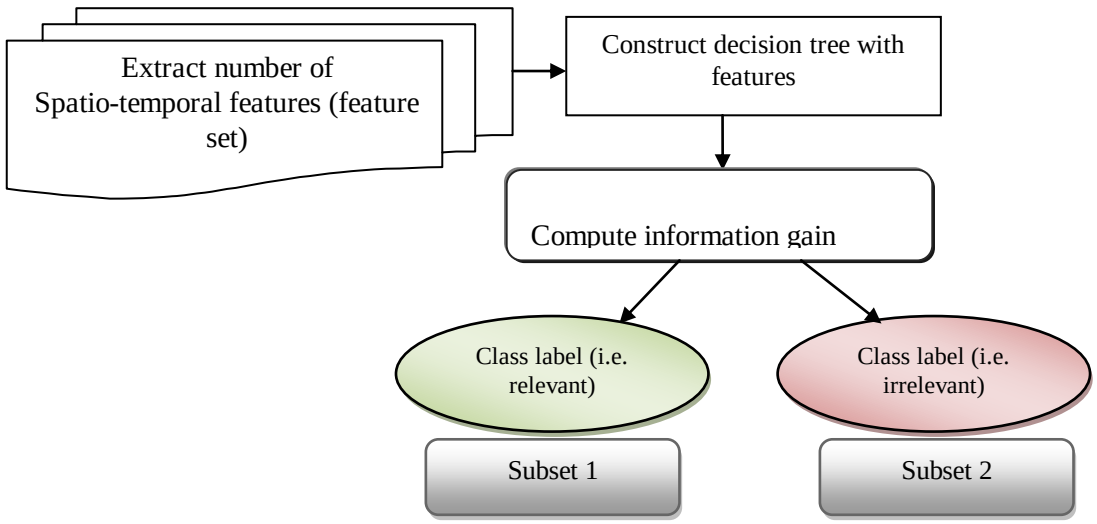
1. Heuristic Best-First Search based Discretized Self-Organizing Feature Map for Spatio-temporal pattern discovery

Figure 1 shows the flow process of the proposed HBFS-DSOFM model to improve the pattern discovery with the relevant features. The Spatio-temporal dataset is taken as input for accurate pattern discovery. From the spatio-temporal dataset, ‘n’ numbers of features are extracted. The proposed HBFS-DSOFM model includes the two processes namely feature selection and clustering. The Heuristic Best-First Search Algorithm is applied for selecting the relevant features and removing the irrelevant features. With the selected features, the spatio-temporal data are clustered using a Discretized Self-Organizing Feature Mapping technique. As a result of clustering, the spatio-temporal patterns are effectively detected for

Footnote).[1] Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

Please note that the references at the end of this document are in the preferred referencing style. Give all authors' names; do not use "*et al.*" unless there are six authors or ace after authors' initials. Papers that have not should be cited as "unpublished" [4]. Papers





predicting
... ." Number footnotes separately in superscripts (Insert |

been submitted for publication should be cited as "submitted for publication" [5]. Papers that have been accepted for publication, but not yet specified for an issue should be cited as "to be published" [6]. Please give affiliations and Figure 2 shows the flow process of Heuristic Best-First Search algorithm to split the total feature set into two subsets. The Heuristic Best-First Search algorithm is a binary decision tree. Each binary decision tree comprises the root node, branch node, and leaf nodes. The root node has the input features. The branch node processes the extracted features based on the information gain value. The leaf node in the tree contains two subsets such as relevant or irrelevant. Let us consider the number of Spatio-temporal features in the dataset is expressed in the following equations,

$$F = \{f_1, f_2, f_3, \dots f_n\} \in D^{st} \quad (1)$$

From (1),F denotes a set of original features $f_1, f_2, f_3, \dots f_n$ in the Spatio-temporal dataset D^{st}. Among the several features, the Best-First Search algorithm begins with the original feature set $'S'$ as the root node information gain of the attribute (i.e fe are removed.

addresses for private communications [7].

Capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation

observed

Then it selects the relevant features which has the largest information gain value. The informaion gain is computed as follows,

$$g(f) = h(F) - \sum_{c_F \in c_1, c_2} \frac{|c_F|}{|F|} * h(c_F) \quad (2)$$

From (2),g (f) represents the information gain of the features, $h(F)$ represents the entropy of the total set, $h(c_F)$ represents the entropy of the subsets,$|c_F|$ denotes two subsets c_1, c_2 (i.e. relevant or irrelevant). $|F|$ denotes a total feature set. The feature with maximum information gain value makes a decision to discover the relevant or irrelevant features. Finally the feature selection is obtained at the le Clusters vant features are selected for Spatio-tempo ery and the irrelevant res

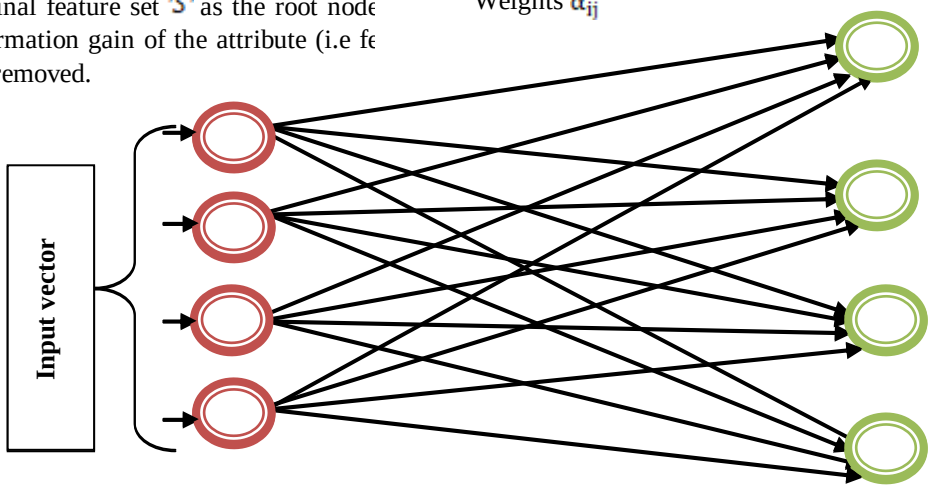


Figure 3 structure of the Discretized Self-Organizing Feature Map clustering

As shown in figure 3, the structure of the Discretized Self-Organizing Feature Map clustering is illustrated. The structure comprises the two modes of operation namely training and mapping. In the training phase, the numbers of Spatio-temporal data are collected from the dataset. In the mapping phase, the Spatio-temporal data from the input space is clustered by finding the node (i.e. cluster) with the closest weight vector to the input space vector. These input vectors comprise the number of Spatio-temporal data which is collected from the dataset.

$$d_i = d_1, d_2 d_3 \dots . d_n \in D^{st} \quad (3)$$

From (3), d_i denotes an input vector and D^{st} denotes a Spatio-temporal dataset. Every input data is linked with the output through the synaptic weight vector which is expressed as follows,

$$\alpha_j = \alpha_1, \alpha_2, \alpha_3 \dots, \alpha_n \quad (4)$$

From (4), α_j denotes synaptic weights. The input data is randomly selected and traverses the each node in the map. The map phase groups the data from the input space through the distance measure. The distance between the input and the weight of the cluster is computed using Manhattan distance formula.

$$D_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} |d_i - \alpha_j| \quad (5)$$

From (5), D_{ij} denotes a Manhattan distance between the input data d_i and the cluster weight vector α_j. After computing the distance measure, finding the winner of the neuron (i.e. cluster) whose weight is closest to the input value using gradient descent function. The gradient descent function is employed for discover the minimum of a function.

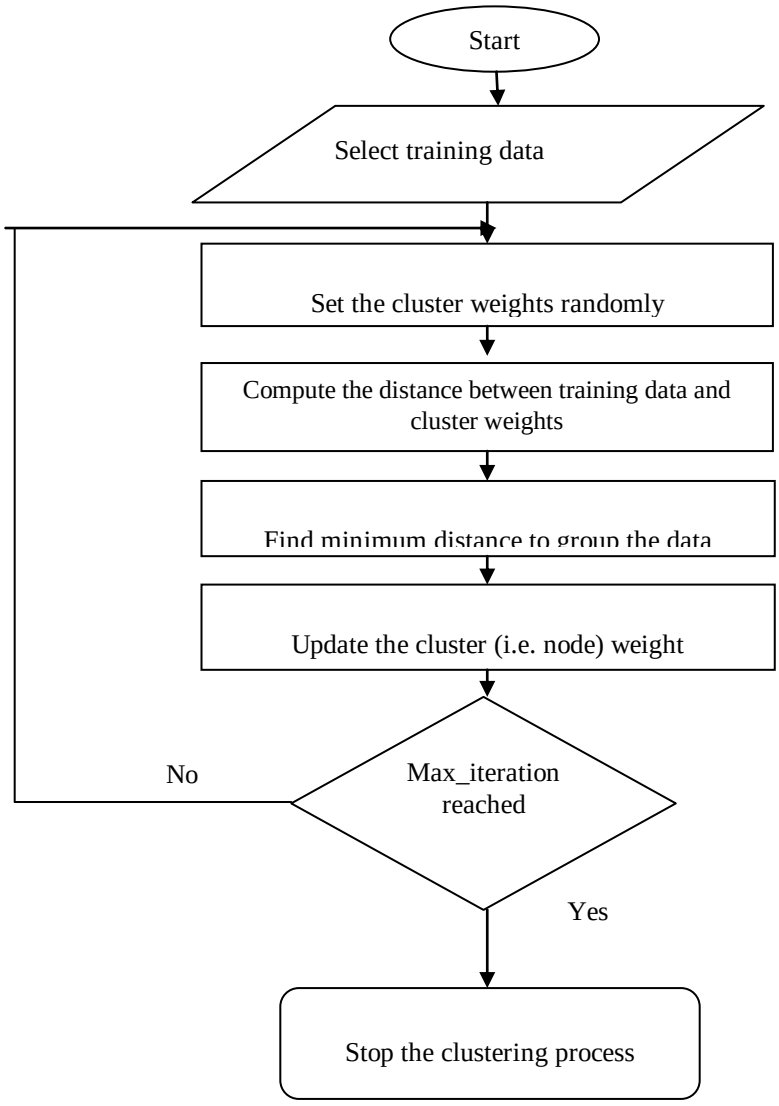
$$F(x) = \arg\min D_{ij} \quad (6)$$

From (6),$F(x)$ denotes a gradient descent function, arg min stands an argument of the minimum at which the distance D_{ij} value are minimized. By this way, the similar

spatio-temporal data are grouped into the clusters. After clustering the similar data, the weights of the cluster is updated by pulling them closer to the input vector.

$$\alpha_j\,(t+1) = \alpha_j(t) + \varphi\,(p,q,t).\gamma\,(t).\,(d_i - \alpha_j(t)) \quad (7)$$

From (7), $\alpha_j\,(t+1)$ denotes an updated weight of the cluster, $\alpha_j(t)$ denotes a current weight of the cluster, $\varphi\,(p,q,t)$ is the restraint caused by distance from the winner usually called the neighborhood function $\gamma\,(t)$ represents the learning rate, d_i denotes a input data. The updating process is used for grouping the entire input data into the any of cluster with the less false positive rate. This clustering process continued until the entire data are grouped into the specific cluster. The flow chart of the DSOFM model is described as follows,



INPUT: SPATIO TEMPORAL DATASET D^{st}, SPATIO-TEMPORAL FEATURES $f_1, f_2, f_3 \ldots f_n$, TRAINING DATA $d_1, d_2, d_3, \ldots. d_n$
OUTPUT: IMPROVE SPATIO-TEMPORAL PATTERN DISCOVERY
BEGIN
\\ **FEATURE SELECTION**

1. EXTRACT NUMBER OF FEATURES $f_1, f_2, f_3 \ldots f_i$
2. **FOR EACH**
3. COMPUTE THE INFORMATION GAIN
4. SELECT RELEVANT FEATURES
5. REMOVE IRRELEVANT FEATURES
6. **END FOR**

\\ **SPATIO-TEMPORAL DATA CLUSTERING**

7. SELECT $d_1, d_2, d_3, \ldots. d_i$ IN INPUT LAYER
8. SET SYNAPTIC WEIGHTS TO CLUSTER

9. **FOR EACH**

10. **FOR EACH NODE**

COMPUTE MANHATTAN DISTANCE

11. FIND MINIMUM DISTANCE $arg\,i$
12. UPDATE NODE WEIGHT
13. **END FOR**
14. **END FOR**
15. **IF (MAX_ITERATION REACHED) THEN**
16. STOP CLUSTERING PROCESS
17. **ELSE**
18. GO TO STEP 9
19. **END IF**

END

Algorithm 1 Heuristic Best-First Search based Discretized Self-Organizing Feature mapping

Algorithm 1 clearly describes the best first search decision tree based feature selection and the DSOM based clustering. For each feature, the information gain is computed. Based on the information gain, the relevant features are selected for pattern discovery with minimum complexity. After selecting the features, the Discretized self-organizing feature map is used for mapping the input vector to the output units (i.e. cluster) through the synaptic weight. The mapping process is carried out by computing the distance between the input and weight vector of the cluster. Then the gradient descent function discovers the minimum distance to group the data into that particular cluster. After that, the weight of the output node is updated by pulling them closer to the input vector. By this way, the entire data are grouped into the cluster to form the patterns. These patterns are used for predicting future outcomes.

The above said two processes are experimented with the conventional methods to show the performance of the proposed HBFS-DSOFM Model than the existing methods.

2. experimental settings

Experimental evaluations of proposed HBFS-DSOFM Model and existing methods namely multivalued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] are implemented using Java language with two various datasets such as El Nino Dataset and Taxi Service Trajectory Dataset. The El Nino Dataset is taken from UCI Machine Learning Repository. The dataset comprises the oceanographic and weather surface data collected from a series of buoys positioned all over the equatorial Pacific. This dataset comprises the 12 attributes and 178080 instances for performing certain tasks. The attribute characteristics are an integer and real. The dataset characteristics are spatio-temporal. The relevant features are selected from the dataset and performing the pattern discovery using the clustering model.

The Taxi Service Trajectory dataset is taken from the UCI Machine Learning Repository. This dataset explains the Trajectory carried out by the 442 taxis running in Porto, in Portugal city. The dataset comprises the 9 attributes and 1710671 instances. The associated tasks of this dataset are clustering and casual discovery. The attribute characteristics are real and the dataset characteristics are multivariate, sequential, time series, domain-theory. The experiments are carried out with these two datasets information with different parameters such as time complexity, clustering accuracy and false positive rate.

3. Results and discussion

The results and discussion of the proposed HBFS-DSOFM Model and existing methods namely multivalued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] with the different parameters such as time complexity, clustering accuracy, and false positive rate are described in this section. Performance results are evaluated with the help of table values and graphical representations. For each subsection, the sample mathematical computation is presented for both El Nino Dataset and Taxi Service Trajectory Dataset.

5.1 Impact of time complexity

Time complexity is defined as the amount of time is required to select the relevant Spatio-temporal features from the dataset. The mathematical formula for calculating the feature selection time is computed as follows,

$$TC = n * t(selecting\ one\ relevant\ feature) \quad (8)$$

From (8), TC denotes a time complexity, n represents a number of features, t denotes a time for selecting the relevant feature. The time complexity is measured in terms of milliseconds (ms).

Sample calculation for time complexity using El Nino Dataset:

- Proposed HBFS-DSOFM Model: No. of the feature is 2 and the time for selecting one feature is 2.2ms, then the time complexity is calculated as follows,

$$TC = 2 * 2.2ms = 4.4ms \approx 4ms$$

- Existing multivalued decision systems: No. of the feature is 2 and the time for selecting one

feature is 3.5ms, then the time complexity is calculated as follows,

$$TC = 2 * 3.5ms = 7ms$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of the feature is 2 and the time for selecting one feature is 2.9ms, then the time complexity is calculated as follows,

$$TC = 2 * 2.9ms = 5.8ms \approx 6ms$$

Sample calculation for time complexity using Taxi Service Trajectory dataset:

- Proposed HBFS-DSOFM Model: No. of the feature is 3 and the time for selecting the one feature is 2.3ms, then the time complexity is calculated as follows,

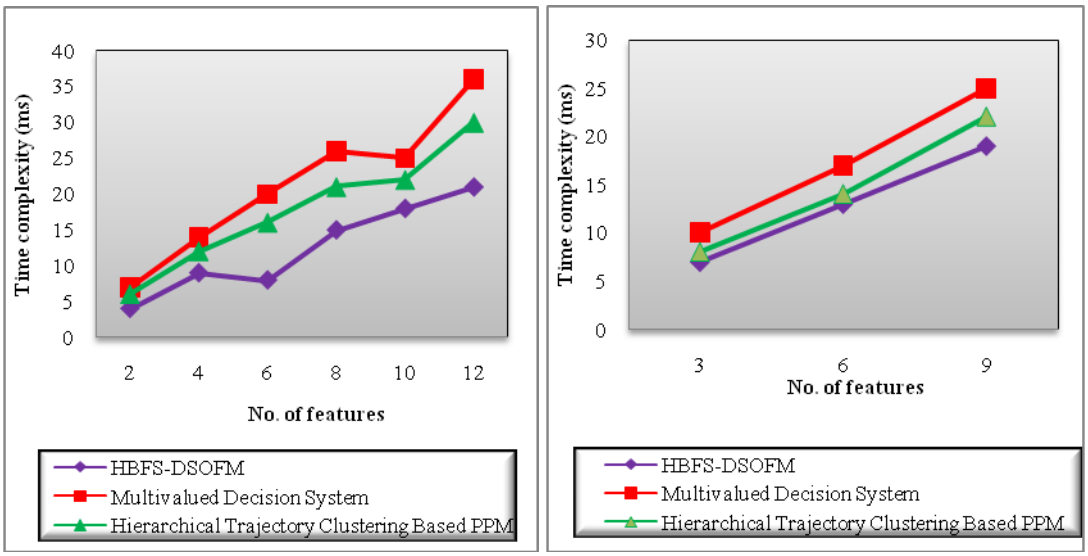
$$TC = 3 * 2.3ms = 6.9ms \approx 7m_s$$

- Existing multivalued decision systems: No. of the feature the is 3 and the time for selecting one feature is 3.4ms, then the time complexity is calculated as follows

$$TC = 3 * 3.4ms = 10.2ms \approx 10ms$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of the feature is 3 and the time for selecting the one feature is 2.5ms, then the time complexity is calculated as follows,

$$TC = 3 * 2.5ms = 7.5ms \approx 8ms$$



(a) (b)

Figure 5 (a) Time complexity versus no. of features using El Nino Dataset (b) Time complexity versus no. of features using Taxi Service Trajectory dataset

Figure 5 (a) (b) shows the experimental results of time complexity versus a number of features using El Nino Dataset and Taxi Service Trajectory dataset. The numbers of spatio-temporal features are taken from this two dataset for calculating the time complexity. The above graphical results confirm that the proposed HBFS-DSOFM Model minimizes the time complexity in the relevant feature selection than the existing multi valued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2]. This significant improvement is achieved by applying a Heuristic Best-First Search approach. This is a decision tree approach where the root node has features. Then the information gain of each feature is computed and separated into the two different subsets such as relevant or irrelevant. Based on the information gain value, the relevant features are selected from the subsets for Spatio-temporal patterns discovery. By this way, a significant feature is selected with minimum time. By applying El Nino Dataset, six different results of the time complexity are attained for three different methods. The results confirm that the HBFS-DSOFM Model minimizes the time complexity by 42% when compared to existing multi valued decision systems [1]. In addition, time complexities of the feature selection are considerably minimized by 31% than the existing Hierarchical Trajectory Clustering Based PPM [2].

Let us consider the Taxi Service Trajectory dataset; three different runs are carryout with three different methods. For each run, the various time complexity results are attained. The average of three results clearly obvious that the HBFS-DSOFM Model comparatively minimized by 26% and 11% than the existing multi valued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] respectively.

3.2 Impact of clustering accuracy

Clustering accuracy is defined as the numbers of Spatio-temporal data are correctly grouped to the total number of data for pattern discovery. The mathematical formula for calculating the clustering accuracy is expressed as follows,

$$CA = \frac{no.of\ data\ correctly\ grouped}{no.of\ data} * 100 \quad (9)$$

From (9), CA denotes a clustering accuracy. It is measured in terms of percentage (%).

Sample calculation for clustering accuracy using El Nino Dataset:

PROPOSED HBFS-DSOFM

- total number of data is 500. Then the clustering accuracy is calculated as follows,

$$CA = \frac{445}{500} * 100 = 89\%$$

- Existing multivalued decision systems: No. of data correctly grouped is 386 and the total number of data is 500. Then the clustering accuracy is calculated as follows,

$$CA = \frac{386}{500} * 100 = 77.2\% \approx 77\%$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of data correctly grouped is 421 and the total number of data is 500. Then the clustering accuracy is calculated as follows

$$CA = \frac{421}{500} * 100 = 84.2\% \approx 84\%$$

Sample calculation for clustering accuracy using Taxi Service Trajectory dataset:

- Proposed HBFS-DSOFM Model: No. of data correctly grouped is 869 and the total no. of data is 1000. Then the clustering accuracy is calculated as follows,

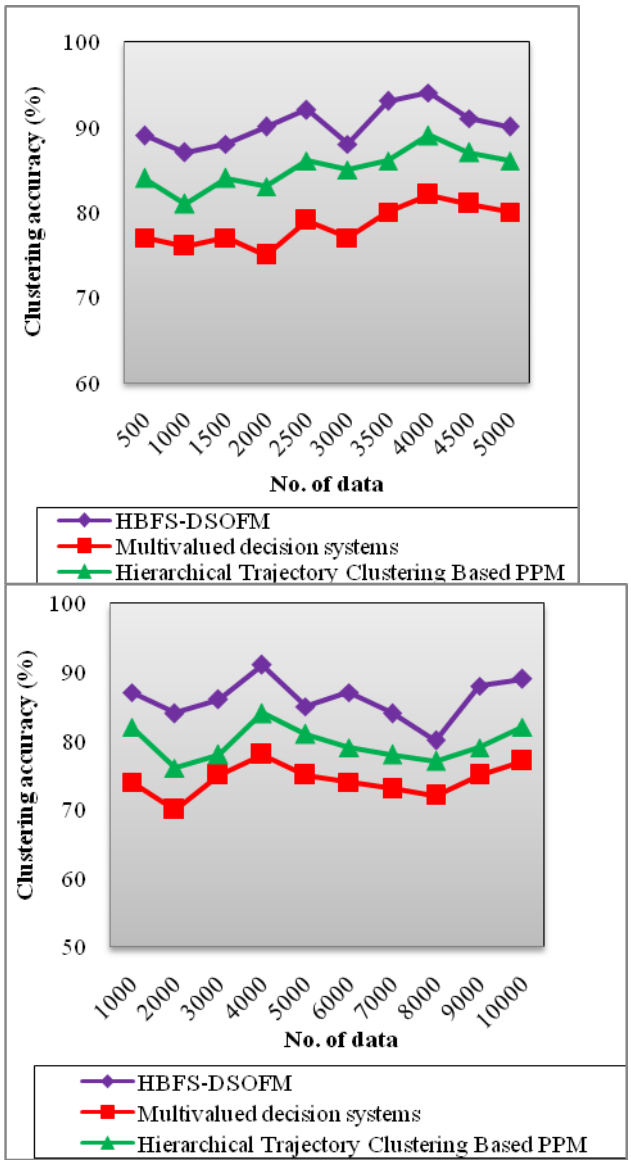
$$CA = \frac{869}{1000} * 100 = 86.9\% \approx 87\%$$

- Existing multivalued decision systems: No. of data correctly grouped is 735 and the total no. of data is 1000. Then the clustering accuracy is calculated as follows,

$$CA = \frac{735}{1000} * 100 = 73.5\% \approx 74\%$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of data correctly grouped is 823 and the no. of data is 1000. Then the clustering accuracy is calculated as follows

$$CA = \frac{823}{1000} * 100 = 82.3\% \approx 82\%$$



(a)

(b)

Figure 6 (a) clustering accuracy versus no. of data using El Nino Dataset (b) clustering accuracy versus no. of data using Taxi Service Trajectory dataset

Figure 6 (a) (b) illustrates the experimental results of clustering accuracy versus a number of data using two dataset El Nino Dataset and Taxi Service Trajectory dataset. The number of Spatio-temporal data is collected from these two datasets. These data are taken as input in 'x' direction whereas the clustering results of three methods are attained at 'y' direction.

By applying El Nino Dataset, the weather data are collected from the equatorial Pacific. For the experimental consideration, the number of data is taken from 500 to 5000. Figure 6 (a) (b) clearly shows that the performance of clustering accuracy is increased using HBFS-DSOFM Model than the conventional clustering methods. The HBFS-DSOFM model effectively groups the similar data into the clusters with the selected relevant features. The Discretized self-organized feature map is used for mapping the input data to the

cluster through the weight value. The mapping is performed through the distance computation. The Manhattan distance is computed between the input data and the weight vector. Then the DSOFM finds the minimum distance and grouped into that specific cluster. By this way, the entire data are grouped into the particular cluster. The grouped data are used for pattern discovery. Ten various evaluations are obtained with different results. The evaluation results clearly show that the clustering accuracy of HBFS-DSOFM model is improved by 15% and 6% when compared to the conventional clustering methods existing multi valued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] respectively.

Similarly, taxi Service Trajectory dataset is applied for calculating the clustering accuracy with the number of data taken from 1000 to 10000. The above graphical representation illustrates the performance results of clustering accuracy of three different methods of HBFS-DSOFM Model, existing multi valued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2]. The DSOFM algorithm is applied to group similar data with the selected features for discovering the periodic patterns to Spatio-temporal trajectories because of the different characteristics including location and temporal data. This helps for improving the clustering accuracy by 16% and 8% than the existing methods.

Impact of false positive rate

The false positive rate is defined as the numbers of (no. of) Spatio-temporal data are incorrectly grouped to the total number of data. The formula for computing the false positive rate is expressed as follows,

$$FPR = \frac{no.of\ data\ incorrectly\ grouped}{no.of\ data} * 100 \quad (10)$$

From (10), FPR represents the false positive rate and it is measured in terms of percentage (%).

Sample calculation for False Positive Rate using El Nino Dataset:

Proposed HBFS-DSOFM Model: No. of data incorrectly grouped is 55 and the total no. of data is 500. Then the false positive rate is calculated as follows,

$$FPR = \frac{55}{500} * 100 = 11\%$$

- Existing multi valued decision systems: No. of data incorrectly grouped is 114 and the total no. of data is 500. Then the false positive rate is calculated as follows,

$$FPR = \frac{114}{500} * 100 = 22.8\% \approx 23\%$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of data incorrectly grouped is 79 and the total no. of data is 500. Then the false positive rate is calculated as follows,

$$FPR = \frac{79}{500} * 100 = 15.8\% \approx 16\%$$

Sample calculation for False Positive Rate using Taxi Service Trajectory dataset:

- Proposed HBFS-DSOFM Model: No. of data incorrectly grouped is 131 and the total no. of data is 1000. Then the false positive rate is calculated as follows,

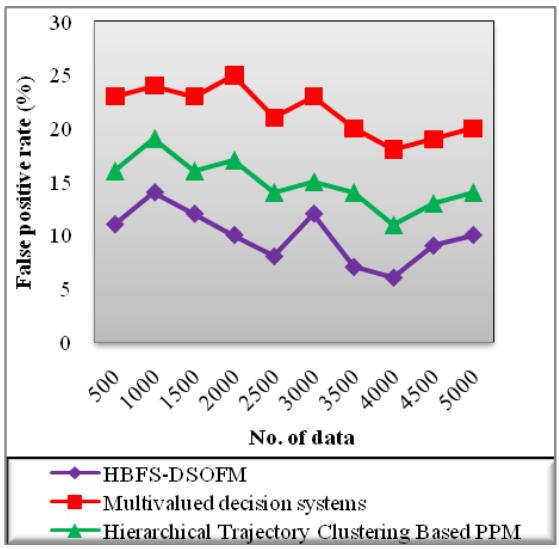
$$FPR = \frac{131}{1000} * 100 = 13.1\%$$

- Existing multi valued decision systems: No. of data incorrectly grouped is 265 and the total no. of data is 1000. Then the false positive rate is calculated as follows,

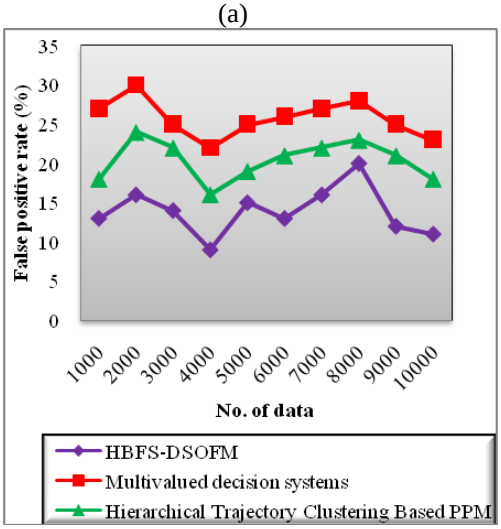
$$FPR = \frac{265}{1000} * 100 = 26.5\% \approx 27\%$$

- Existing Hierarchical Trajectory Clustering Based PPM: No. of data incorrectly grouped is 177 and the total no. of data is 1000. Then the false positive rate is calculated as follows,

$$FPR = \frac{177}{1000} * 100 = 17.7\% \approx 18\%$$



(a)



(b)

Figure 7 (a) false positive rate versus no. of data using El Nino Dataset (b) false positive

rate versus no. of data using Taxi Service Trajectory dataset

Figure 7 (a) (b) portrays the experimental results of the false positive rate with two different datasets based on the number of data. The figure shows that the experimental results of false positive rate of three different methods namely HBFS-DSOFM Model, existing multivalued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] clearly depicted in the above two-dimensional graphical representation. The above figure proves that the false positive rate of HBFS-DSOFM Model is considerably minimized using two datasets compared to conventional methods. This is because, the HBFS-DSOFM Model uses the discretized self-organized map algorithm to find similar cluster members by computing the distance between the input data and cluster weight. Then, the mapping algorithm uses the gradient descent function to find the minimum distance between the input data and the all cluster weight. This function also minimizes the incorrect data clustering. In addition, the weight of the cluster is updated by pulling them closer to the input vector. This process is repeated until all the data objects are correctly grouped into the clusters. This also minimizes the false positive rate.

By applying El Nino Dataset, the false positive result of HBFS-DSOFM Model is significantly minimized by 55% when compared to existing multi valued decision systems [1]. In addition, evaluation results of false positive rate are minimized by 34% using HBFS-DSOFM Model than the existing Hierarchical Trajectory Clustering Based PPM [2].

Similarly, Taxi Service Trajectory dataset is used for computing the false positive rates with ten various data. For each run, the different results are attained. Finally, the proposed results are compared with the existing results. The average of ten various results shows that the HBFS-DSOFM Model comparatively minimizes the false positive rate by 47% and 32% than the existing multi valued decision systems [1] and Hierarchical Trajectory Clustering Based PPM [2] respectively.

The above results and discussion clearly show that the proposed HBFS-DSOFM Model effectively improving the data clustering accuracy with less false positive rate, time complexity. This helps to improve the rate of spatio-temporal pattern with minimal time.

V. CONCLUSION

An efficient model Heuristic Best-First Search based Discretized Self-Organizing Feature Map (HBFS-DSOFM) is developed to analyze the Spatio-temporal data for improving the pattern discovery. The numbers of features are extracted from the Spatio-temporal dataset. Then the Heuristic Best-First Search identifies the relevant features by constructing the decision tree with the information gain value. This process divides the entire feature set into the two subsets namely relevant and irrelevant. This process reduces the time complexity. After that, the Spatio-temporal pattern discovery is carried out using the Discretized Self-Organizing Feature Map. In the mapping phase, the input vector is mapped into the available cluster weights value through the distance measure. Then the minimum distance between the input vector and cluster weight has a higher chance for grouping the data into that specific cluster. As a result of the clustering process, Spatio-temporal patterns are identified for finding future outcomes. This helps to improve the clustering accuracy and minimize the false positive rate. The experimental evaluation is carried out with two different datasets El Nino dataset and taxi trajectory dataset with certain parameters such as time complexity, clustering accuracy and false positive rate. The performance of the clustering accuracy is improved with minimum time complexity as well as false positive rate than the state-of-the-art methods.

REFERENCES

1. Sanchita Mal-Sarkar, Iftikhar U. Sikder, Vijay K. Konangi, "Spatio-temporal Pattern discovery in sensor data: A multivalued decision systems approach", Knowledge-Based Systems, Elsevier, Volume 109, Pages 137–146, 2016
2. Dongzhi Zhang, Kyungmi Lee and Ickjai Lee "Hierarchical Trajectory Clustering for Spatio-temporal Periodic Pattern Mining", Expert Systems with Applications, Elsevier, Volume 92, Pages 1-11, February 2018
3. Hanna Meyer, Christoph Reudenbach, Tomislav Hengl, Marwan Katurji, Thomas Nauss, "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation", Environmental Modelling & Software, Elsevier, Volume 101, 2018, Pages 1-9
4. Shadi A. Aljawarneh, Vangipuram Radhakrishna, Puligadda, Veereswara Kumar, Vinjamuri Janaki, "G-SPAMINE: An approach to discover temporal association patterns and trends in internet of things", Future Generation Computer Systems, Elsevier, Volume 74, 2017, Pages 430-443
5. Yoann Pitarch , Dino Ienco , Elodie Vintrou, Agnès Bégué, Anne Laurent, Pascal Poncelet, Michel Sala, Maguelonne Teisseire, "Spatio-temporal data classification through multi dimensional sequential patterns: Application to crop mapping in complex landscape", Engineering Applications of Artificial Intelligence, Volume 37, 2015, Pages 91–102
6. Jiwon Kima and Hani S.Mahmassani, "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories", Transportation Research Part C: Emerging Technologies, Elsevier, Volume 59, 2015, Pages 375-390
7. Yanyan Xu, Hui Chen, Qing-Jie Kong, Xi Zhai and Yuncai Liu, "Urban traffic flow prediction: a spatio-temporal variable selection based approach", Journal of Advanced Transportation, Volume 50, 2016, Pages 489–506
8. Akin Tascikaraoglu, "Evaluation of spatio-temporal forecasting methods in various smart city applications", Renewable and Sustainable Energy Reviews, Elsevier, Volume 82, Part 1, 2018, Pages 424-435
9. Zihan Hong , Ying Chen , Hani S. Mahmassani, "Recognizing Network Trip Patterns Using a Spatio-Temporal Vehicle Trajectory Clustering Algorithm", IEEE Transactions on Intelligent Transportation Systems , Volume 19 , Issue 8 , 2018, Pages 2548 - 2557
10. Berna Bakır Batu , Tuğba Taşkaya Temizel , H. Şebnem Düzgün, "A Non-Parametric Algorithm for Discovering Triggering Patterns of Spatio-Temporal Event Types", IEEE Transactions on Knowledge and Data Engineering , Volume 29 , Issue 12 , 2017, Pages 2629 – 2642
11. S. Khoshahval , M. Farnaghi, M. Taleai, "Spatio-Temporal Pattern Mining On Trajectory Data Using ARM", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-4/W4, 2017, Pages 395-399
12. K.P.Agrawal, Sanjay Garg, Shashikant Sharma, Pinkal, Patel, "Development and validation of OPTICS based spatio-temporal clustering technique", Information Sciences, Elsevier, Volume 369, 2016, Pages 388-401
13. Pradeep Mohan , Shashi Shekhar , James A. Shine , James P. Rogers,

"Cascading Spatio-Temporal Pattern Discovery", IEEE Transactions on Knowledge and Data Engineering , Volume 24 , Issue 11 , 2012, Pages 1977 – 1992
14. 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
15. Jian Xu, Liang Tang , Chunqiu Zeng , Tao Li, "Pattern discovery via constraint programming", Knowledge-Based Systems, Elsevier, Volume 94, 2016, pages 23–32
16. J. Béjar, S. Álvarez, D. García, I. Gómez, L. Oliva, A. Tejeda & J. Vázquez-Salceda, "Discovery of spatio-temporal patterns from location-based social networks", Journal of Experimental & Theoretical Artificial Intelligence, Volume 28, Issue 1-2, 2016, Pages 313-329
17. Donna J. Peuquet, Anthony C. Robinson, Samuel Stehle, Franklin A. Hardisty and Wei Luo, "A method for discovery and analysis of temporal patterns in complex event data", International Journal of Geographical Information Science, Volume 29, Issue 9, 2015, Pages 1588-1611
18. Mete Celik and Ahmet Sakir Dokuz, "Discovering Socio-Spatio-Temporal Important Locations of Social Media Users", Journal of Computational Science, Elsevier, Volume 22, September 2017, Pages 85-98
19. Adina Iftimi, Marie-Colette van Lieshout, Francisco Montes, "A multi-scale area-interaction model for spatio-temporal point patterns", Spatial Statistics, Elsevier, Volume 26, August 2018, Pages 38-55
20. Ilias Fountalis, Annalisa Bracco, Constantine Dovrolis, "Spatio-temporal network analysis for studying climate patterns", Climate Dynamics, Springer, Volume 42, Issue 3–4, 2014, Pages 879–899

Authors Profile

Mrs.R.Sarala is a Assistant Professor in Department of Computer Science at KG College of Arts and Science, Coimbatore. She did her MCA degree at the Madras University, Chennai. She started her teaching profession at KG College of Arts and Science, Coimbatore in 2012. Her teaching areas are Data Structures, Computer Networks, Data Mining, Visual Programming, Software Engineering and Software Project Management, C Programming and Java Programming. She did her M.Phil. at Bharathiar University, Coimbatore. She is currently pursuing Ph.D. at Bharathiar University under the guidance of Dr.V.Saravanan. She has over 6 publications in international referred journals. She has presented papers at various conferences and published book.

Dr.V. Saravanan is a Professor and Head on Department of Information Technology at HindusthanCollege of Arts and Science, Coimbatore. He did his M.Sc in computer science at the Bharadhidasan University, Trichy. And he did his MCA at the Periyar University, Salem. He started his teaching profession atThanthai Hans RoeverCollege Perambalur, Trichy in 1999. Later in 2004 he joined in Hindusthan College of Arts and Science, Coimbatore. His teaching areas are Networking and mobile computing as it is his main area of interest. He did his M.Phil and Ph.D in at ManonmaniamSundaranar university, Tirunelveli. His Ph.D was on Wireless Networking in video streaming. He has over 36 publications in internationals referred journals. He has presented research papers at conferences, published articles and papers in various journals. He is renowned key note address in both national and international conferences.