FULLY RECURRENT DEEP NEURAL LEARNING BASED UNCERTAIN DATA CLUSTERING

[1]MURUGANANTHAM S, [2]ELANGO N M

[1]Assistant Professor, Department of Computer Technology and Information Technology,
Kongu Arts and Science College (Autonomous), Erode-638107, Tamil Nadu, India,
[2]Associate Professor, School of Information Technology and Engineering, VIT University,
Vellore-632014, Tamil Nadu, India
E-mail: [1]muruganandham.s@gmail.com, [2]nmeoxford@yahoo.com

ABSTRACT

Uncertain data is the data that deviate from the exact and original values. Clustering the uncertain data is an imperative task in data mining since it provides unpredicted results. Many clustering techniques have been developed for processing the uncertain data but the robustness and accuracy were not increased. In order to increase the clustering accuracy, a Fully Recurrent Deep Neural Learning based X-means Data Clustering (FRDNL-XDC) technique is proposed for mining uncertain data with minimal time. In this technique, deep learning approach based X-means clustering is used to partition the whole dataset into different clusters for mining the uncertain data. The 'x' number of clusters and centroids are initialized randomly in the hidden layer. The bivariate correlation coefficient is used to find the correlation between the centroid of the cluster and the data to be clustered. Sigmoid activation function is used at the output layer where it maps the data with the centroid of the cluster to form certain data using Akaike information criterion with higher accuracy and minimum time. In FRDNL-XDC technique, feedback connections formulate to group the similar data with the minimal false positive rate. For validation, the proposed FRDNL-XDC technique is compared with the existing clustering techniques namely UKmeans clustering mechanism and UK- medoids-SMDM. The results obtained on El nino dataset show that the proposed FRDNL-XDC technique increases the clustering accuracy with less false positive rate and time complexity than the state-of-the-art methods.

Keywords: *Uncertain Data Mining, Clustering, Fully Recurrent Deep Neural Learning, X-Means Data Clustering, Bivariate Correlation Coefficient, Sigmoid Activation Function, Akaike Information Criterion.*

1. INTRODUCTION

Clustering is the process of dividing the large dataset into a set of groups where one cluster comprises the related data and the other cluster comprises dissimilar data. The conventional clustering algorithms such as k-means, sub-space clustering, fuzzy c-means clustering mainly focus on certain data. However, due to a variety of reasons like randomness in data production and gathering, vagueness in physical measurement and data staling, uncertain data happens naturally from several real applications such as biomedical measurement, sensor networking, and meteorological forecasting and so on. Uncertain data brings new challenges to state-of-the-art clustering algorithms.

Uncertain data in K-means clustering mechanism namely UKmeans was introduced in [1] for improving the clustering performance. The mechanism does not provide satisfactory results in the cluster formation and failed to find the accurate cluster number for clustering the uncertain objects. UK-medoids-self-adapted mixture distance measure adapted mixture distance measure required more amount of time to attain better clustering results.

A Fuzzy Clustering Large Applications based on Randomized Search (FCLARANS) was developed in [3] for handling the uncertain data with different partitions. The FCLARANS computationally more expensive and involves a number of user-defined parameters. Therefore, the performance of uncertain data clustering over the large data was not enough. Grouping of Large Uncertain Graphs Using Neighborhood Information was presented in [4]. The graph-based approach failed to handle the huge uncertain dataset as the accuracy of which also effects the eventual clustering formation. Further, this approach increases the computational cost since parallel implementation was not performed.

A Large margin clustering technique was presented in [5] using uncertain data and probability distribution relationships. But it failed to use deep learning for improving the clustering accuracy. A modified version of the standard k-medoids (Hk-medoids) algorithm was designed in [6] to increase the clustering results and minimize the running time. But the algorithm does not minimize the error rate as it employs the used fused similarities between objects. A Kullback-Leibler divergence was developed in [7] for partitioning the data into the different clusters with large data sets and high accuracy. The method does not minimize the problems on uncertain data depends on distribution similarity.

A Type-2 fuzzy C-means (GT2 FCM) algorithm was introduced in [8] for minimizing the uncertainty with more training data. The algorithm does not use any criterion function to cluster the entire data. A new density-based uncertain data clustering algorithms were developed in [9] for grouping the uncertain data with high accuracy and efficiency. The algorithm does not minimize the time complexity during the clustering process as it does not utilizes the computation methods for clustering uncertain data streams.

K-median and K-means clustering algorithms were introduced in [10] for minimizing the time and space complexity as well as improving the performance of clustering accuracy. This algorithm increases the error rate as robust clustering results was not obtained.

From the existing algorithms in the literature, some key issues such as less clustering accuracy, more time complexity, high false positive rate have been identified. To achieve high clustering accuracy with minimal time, FRDNL-XDC technique is introduced. The FRDNL-XDC technique uses deep learning approach based ‘x’ means clustering to partition the whole dataset into different clusters for mining the uncertain data. The ‘x’ number of clusters and centroids are initialized randomly in the hidden layer. The bivariate correlation is measured between the data and cluster centroids. Based on the correlation measure, the similar data are grouped into a cluster. To minimize the false positive rate, the deep learning approach performs the recurrent process and the sigmoid activation function is used at the output layer. The activation function is used to check whether all data are grouped into the cluster or not. If any of the data are not a member of the clusters, a deep learning approach uses the Akaike information criterion to find the maximum probability between the data and cluster centroid. This result helps to group all data into that particular cluster.

The paper is organized as follows. In section 2, the reviews related to the uncertain data clustering are presented. In section 3, a new deep learning algorithm is proposed for clustering the uncertain data. In section 4, the experimental setting is presented with the dataset. The results attained from the experiential valuation are presented in section 5. Finally, section 6 provides conclusions of the research work.

2. RELATED WORKS

An improved voting method was developed [11] for handling the uncertain data in the subspace. The method failed to use the data mining techniques for handling the uncertain data and space complexity of this method was more as it does not employ data structure to store the voting map of uncertain data. K medoid clustering algorithm based on probability Distribution similarity was developed in [12] for partitioning the uncertain data. Though the algorithm minimizes the time complexity, but clustering accuracy was not enhanced as it employs Kullback-Leibler divergence as the similarity measurement.

A Maximum Likelihood Estimation based uncertain data clustering was presented in [13] for increasing the performances of clustering task. The clustering approach does not minimize the complexity while clustering the uncertain data because it uses Maximum Likelihood Estimation. A Voronoi Diagrams and R-Tree Indexing technique was introduced in [14] for partitioning the uncertain data with high efficiency. The technique does not find the entire data were grouped into the cluster for minimizing the error rate.

In [15], a new distributed uncertain data clustering algorithm was introduced for attaining better results with minimum execution time. The number of clusters formation and the clustering performance was not improved. A game-theoretic rough set (GTRS) approach was introduced [16] for clustering the data in the presence of uncertainty due to a missing value. The approach failed to consider the various evaluation functions for computing the correlation between an object and a cluster.

A Heuristic model was introduced in [17] for clustering the uncertain objects with high accuracy. Though, the model minimizes the execution time. But, incorrect data clustering was not solved as it utilizes the Euclidean distance as similarity measure. A Regularized Robust Broad Learning System was presented in [18] for improving the modeling performance in the uncertain data environment. But the system failed to handle the large data set with less time complexity for mining the uncertain data. A hierarchical clustering of uncertain data was developed in [19] with the information-theoretic approach. The method failed to provide accurate clustering results while handling the large dataset as it determines the pair-wise distances of the uncertain objects. A density-based clustering approach of big probabilistic graphs was developed in [20] based on neighborhood information. But, the designed strategy does not improve the clustering accuracy where clustering process was performed based on the nodes' degree and neighborhood information.

An enhanced graph based clustering algorithm was designed in [21] to obtain efficient gene or protein clusters over uncertain and noisy data. However, weather dataset was not considered in this algorithm. Real-time Density-based Clustering (RTDBStream) was presented in [22] for clustering data streams. But, uncertain data was not resolved in this method.

The problems identified from the above clustering techniques are addressed by introducing a new technique called FRDNL-XDC. The process of FRDNL-XDC technique is described in the next section.

3. FULLY RECURRENT DEEP NEURAL LEARNING BASED X-MEANS DATA CLUSTERING TECHNIQUE FOR UNCERTAIN DATA MINING

Data mining is the process of detecting and examining the large data sets to find a significant pattern. Various effective algorithms have been developed to provide significant results using the data mining techniques for determining the uncertain data. These data mining methods help to attain the high-quality results. Based on this motivation, an efficient technique called Fully Recurrent Deep Neural Learning based X-means Data Clustering (FRDNL-XDC) is introduced. The FRDNL-XDC technique is proposed by combining the Fully Recurrent Deep Neural Learning in X-means Clustering on the contrary to existing works.

Deep neural learning is a type of machine learning technique that comprises several processing layers. The output of previous layer is exploited as input to the next successive layer. The fully recurrent deep neural learning includes three different layers such as an input layer, hidden layer and output layer. In deep learning, more than one hidden layers are used for clustering the uncertain data in an accurate manner. The proposed FRDNL-XDC technique uses the two hidden layers for handling the uncertain data. The hidden unit performs the clustering process repeatedly from the input unit and provides the efficient results to the output layer. The input layer is completely interconnected to the hidden and output layer through the adjustable weight connections. Hence, the name is called fully recurrent deep neural learning. As shown in figure 1, flow processes of uncertain data clustering are illustrated to partition the large dataset into a number of groups. Initially, several uncertain data is collected from the dataset. The fully recurrent deep neural learning technique is applied to cluster the uncertain data using three different layers. The numbers of data are taken in the input layer. The proposed technique FRDNL-XDC technique has two hidden layers for processing the input data. In the first hidden layer, the number of clusters and the centroid is

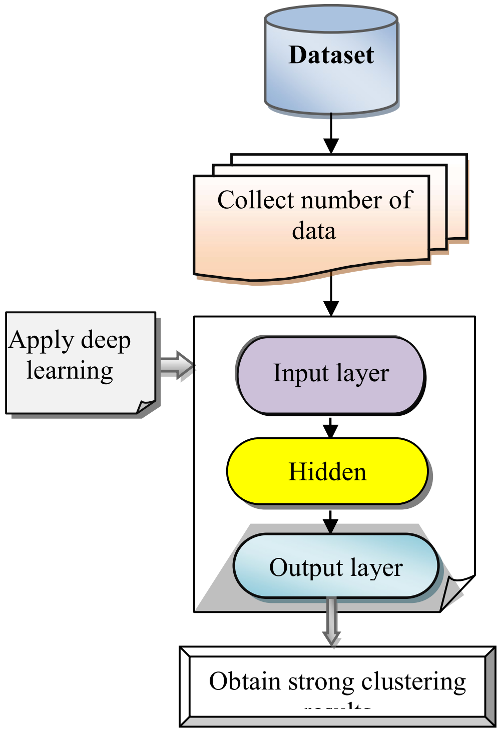

Figure 1: Flow Process of Frdnl-Xdc Technique

initialized. In the second hidden layer, the correlation between the centroid and the data is computed to find the cluster member. The deep learning technique uses sigmoid activation function at the output layer to group the entire data into any of the clusters with higher accuracy and minimum time.

Let us consider the number of data $d_1, d_2, d_3, \ldots . d_n$ extracted from the dataset. Then the three different layers are constructed with the number of data as shown in figure 2.

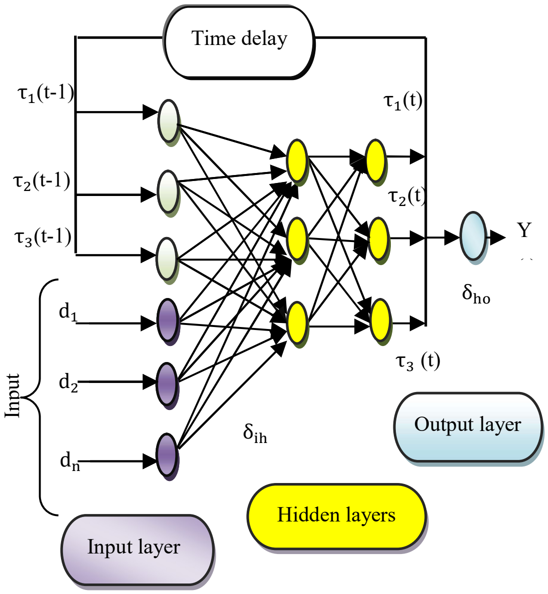

Figure 2: Fully Recurrent Deep Neural Learning

Figure 2 shows the fully recurrent deep neural learning for uncertain data clustering with high accuracy. The network comprises the neuron-like nodes positioned into consecutive layers and each node in a particular layer is linked in a one-way connection to each other node in the next consecutive layer. The input nodes receiving data $d_1, d_2, d_3, \ldots . d_n$. The output node provides the results at the output layer. The hidden nodes transform the input uncertain data from input to output. The fully recurrent deep neural learning comprises the unit-delay feedback connections from the hidden layer output to the input layer $\tau_1(t-1), \tau_2(t-1), \tau_3(t-1)$for deeply learning the input data. The one layer is and the other layers are connected through the weights. The weights between the input and hidden layers are represented asδ_{ih}. Similarity, the weight between the hidden and output layer is represented as δ_{ho}.

In first hidden layer, the 'x' means clustering is applied to define the number of cluster and cluster centroid randomly.

$$cl_j = cl_1, cl_2, cl_3, \ldots . cl_x \qquad j = 1,2,3 \ldots . x \qquad (1)$$

In (1), cl_j denotes a cluster centroid. After defining the cluster centroid, the correlation between the cluster centroid and data is computed in the second hidden layer using Bivariate correlation coefficient. This correlation is the statistical measure to find the linear relationship between two variables (i.e. cluster centroid and data) using below.

$$\beta\left(d_i, cl_j\right) = \frac{\sum d_i * cl_j - \frac{(\sum d_i)(\sum cl_j)}{n}}{\sqrt{\sum d_i^{\,2} - \frac{(\sum d_i)^2}{n}} * \sqrt{\sum cl_j^{\,2} - \frac{(\sum cl_j)^2}{n}}}$$

$$i = 1,2,3, \ldots . n\,, j = 1,2,3 \ldots \ldots, x \qquad (2)$$

In (2), βrepresents the bivariate correlation coefficient. 'n' represents a number of data in the dataset. Here, $\sum d_i * cl_j$ represents the sum of the product of paired score of data and cluster centroid whereas $d_i^{\,2}$ denotes a squared score of data and $cl_j^{\,2}$ denotes a squared score of cluster centroid. The bivariate correlation coefficient provides the two results such as "+1" and "-1". If the result is '+1', then the data and the cluster centroid is more correlated. The correlation coefficient provides '-1'then the data and the cluster centroid is not similar. Therefore, the data with more similar to the cluster centroid is grouped into that particular cluster.

For each time interval 't', the output of the hidden unit is feedback into the network together with the inputs. The recurrent process of the hidden layer is mathematically expressed as follows,

$$\tau(t) = \delta_{ih} i(t) + \delta_h \, \tau(t-1)) \qquad (3)$$

In (3),$\tau(t)$represents the output of the hidden layer at a time 't'. Here, '$\tau(t-1)$' represents the output of hidden layer at the time '$t-1$' and 'δ_h' denotes a weights of the hidden layer, δ_{ih} denotes a weight between input and hidden layer, $i(t)$ represents the input. The hidden layers with recurrent results are fed into the output layers. The resultant clustering results at the output layer is expressed as follows,

$$Y(t) = \rho\,(\tau(t) * \delta_{ho}) \qquad (4)$$

In (4), $Y(t)$ represents the output, ρ denotes activation function, δ_{ho} represents the weight between the hidden and output layer, $\tau(t)$

denotes a hidden layer output.The sigmoid activation function is used at the output layer to find all the data are correctly grouped or not.

$$\rho = \frac{1}{1+\exp(d_i)} \qquad (5)$$

The Sigmoid Function is an S-shape curve. The main reason why FRDNL-XDC technique employs sigmoid function is because it output exists between 0 and 1. This helps for FRDNL-XDC technique to predict the probability as an output of clustering result.

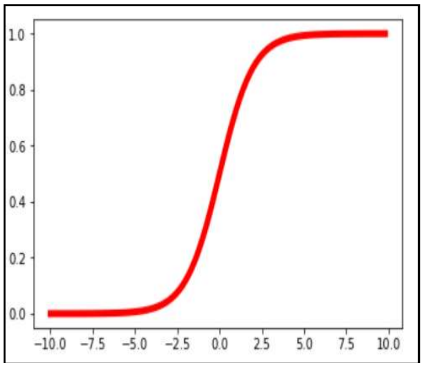

Figure 3: sigmoid activation function output

From the above figure 3, the activation function $'\rho'$ provides the '1' denotes that the entire are correctly grouped into the clusters whereas '0' refers that the data are not grouped into the any of the cluster.

If any of the data are not grouped into the clusters, then FRDNL-XDC technique improves cluster tasks by continually performing the subdivision until all the data are grouped. The Akaike information criterion (AIC) is used for improving the cluster assignment and minimizing the information lost. This helps to group the entire data into any one of the clusters through the maximum likelihood similarity. The Akaike information criterion is calculated as follows,

$$AIC = 2\left(1 - log\, p\left(d_i \middle| cl_j\right)\right) \qquad (6)$$

In (6), AIC represents an Akaike information criterion function, p denotes a probability function, $d_i$denotes an uncertain data and cl_j denotes the cluster centroid. $p\left(d_i \middle| cl_j\right)$ denotes a maximum likelihood of a function. The likelihood of a function illustrates that the maximum probability of the data to becomes a member of the particular cluster. As a result, all the uncertain data are grouped into any one of the cluster based on maximum likelihood. This process gets iterated until the entire data are grouped. This helps to increase the clustering accuracy and minimizes the false positive rate. The algorithmic process of uncertain data clustering is described as follows,

```
// Fully Recurrent Deep Neural Learning based X-means Data Clustering Algorithm
Input: Number of uncertain data, $d_1, d_2, d_3, \ldots. d_n$,
Output: Improve uncertain data clustering accuracy
Begin
    Collect$d_1, d_2, d_3, \ldots. d_n \in D$
        Apply $d_i$ to the input layer
        Fed into the hidden layer
        define the 'x' number of cluster and centroids $cl_j$
        for each $d_i$ and $cl_j$
            Compute the bivariate correlation $\beta\left(d_i, cl_j\right)$
            if $\left(\beta\left(d_i, cl_j\right) = +1\right)$ then
                positive correlation between $d_i$ and $cl_j$
                Group data into cluster
        else
            Negative correlation between $d_i$ and $cl_j$
            data does not said to be cluster member
        end if
        If any data not a member of cluster
            Compute $AIC$ to group the data into cluster
        else
            Stop the clustering process
        end if
        Group all the data into clusters
    end for
end
```

Algorithm 1: Fully Recurrent Deep Neural Learning based X-means Data Clustering

Algorithm 1 describes the uncertain data clustering using FRDNL-XDC. At first, the number of data is collected from the dataset. The number of data is fed into the input layer. In the first hidden layer, the clusters and cluster centroids are defined. For each data and cluster centroid, the correlation is computed. The high correlation results used to identify the cluster member. Otherwise, the data does not move into the particular cluster. The hidden layer outputs are fed into the output layer. In that layer, the sigmoid activation function is applied to detect all the data are grouped into the clusters. If any data are not a member of the particular cluster, the criterion is verified and finds the maximum likelihood probability between the data and cluster centroid. The likelihood function used for grouping the data into the particular cluster. Otherwise, the clustering process is stopped. This process improves the clustering accuracy and minimizes the false positive rate.

4. EXPERIMENTAL SETTINGS

An experimental evaluation of FRDNL-XDC technique and UKmeans clustering mechanism [1] and UK- medoids-SMDM [2] are implemented using Java language. For the experimental consideration, the El Nino dataset is used for mining the uncertain data. The El Nino dataset is taken from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/El+ Nino). The dataset comprises the spatial information taken from the series of buoys positioned over the equatorial Pacific region. The dataset comprises 12 attributes and 178080 instances. The attributes are observation, year, month, day, date, latitude, longitude, zonal winds, meridional winds, humidity, air temperature and sea surface temperature. The attribute characteristics are an integer and real and the characteristics are spatiotemporal. The data variables are taken from the year 1980 for a few locations. The remaining data were taken in different locations are rainfall, solar radiation, current levels, and subsurface temperatures. By using the above-collected data, the uncertain data is determined by clustering the similar data. The experiments are carried out with different parameters such as clustering accuracy, false positive rate and time complexity with a number of data.

5. RESULTS AND DISCUSSION

Results and discussion of FRDNL-XDC technique and existing clustering methods namely UKmeans clustering mechanism [1] and UK-medoids-SMDM [2] are described in this section with tabular values and graphical results. The various performance metrics such as clustering accuracy, false positive rate and time complexity are considered for evaluating the performance results of three different clustering methods. For each parametric result, the proposed results are compared with the existing clustering results. These experimental results are discussed in the following subsections.

5.1 Performance Results Of Clustering Accuracy

Clustering accuracy is measured as the ratio of a number of data that are correctly grouped into the correct cluster to the total number of data. It is measured in terms of percentage (%). If ‘n’ denotes a total number of data in the dataset then the mathematical formula for calculating the clustering accuracy is expressed as,

$$clustering\ accuracy = \frac{Number\ of\ data\ are\ correctly\ clustered}{n} * 100 \qquad (7)$$

Table 1: Clustering Accuracy Versus The Number Of Data

No. of data	Clustering accuracy (%)		
	FRDNL-XDC	UKmeans clustering mechanism	UK-medoids SMDM
1000	87	83	80
2000	84	81	76
3000	86	77	74
4000	84	80	75
5000	82	79	74
6000	85	81	76
7000	87	84	80
8000	90	86	82
9000	89	85	81
10000	88	83	78

Table 1 describes the clustering accuracy versus a number of data with three different clustering techniques FRDNL-XDC technique, UKmeans clustering mechanism [1] and UK- medoids-SMDM [2]. For the experimental consideration, the number of data is varied from 1000 to 10000. The table values clearly show that the clustering accuracy using FRDNL-XDC technique is improved than the conventional clustering techniques. The experimental results are illustrated in figure 4.

Figure 4 depicts the performance results of clustering accuracy versus a number of data with three different clustering methods. The graph shows ten different clustering results for three different techniques.

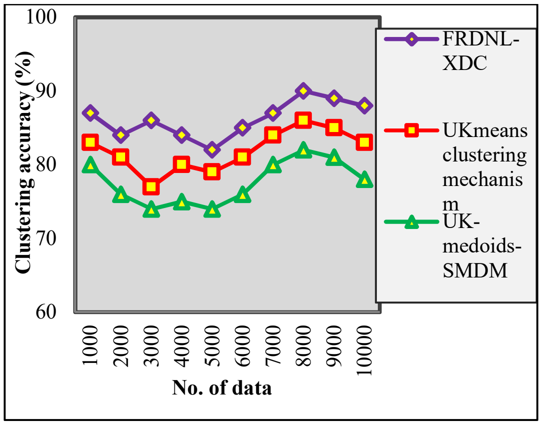

Figure 4: Performance Results Of Clustering Accuracy Versus Number Of Data

In figure 4, the data are taken as input in 'x' direction. The corresponding clustering results are attained at 'y' direction. The above results confirm that the accurate clustering results are attained using FRDNL-XDC technique than the conventional technique. This significant improvement is achieved by applying a deep learning based clustering technique. In FRDNL-XDC technique, the input weather data are taken in the input layer. Then it is fed into the first hidden layer for processing the input data. The number of clusters and the centroids is defined at the first hidden layer. In the second layer, the correlation between the data and cluster centroid is computed to group the similar weather data like air temperature data, humidity, and surface temperature data. This process is fed back into the input layer for grouping the entire data. In the final output layer, the clustering results are attained and it verified with the activation function. The activation function provides that the data are correctly grouped into the cluster or not. As a result, the FRDNL-XDC technique increases the clustering accuracy.

Let us consider the number of data is 1000. Then the clustering accuracy of FRDNL-XDC technique is 87% whereas the clustering accuracy of UKmeans clustering mechanism [1] and UK- medoids-SMDM [2] are 83% and 80% respectively. Similarly, the remaining nine runs are carried out and compare the results of proposed and existing clustering techniques. After the comparison, the average of 10 results is taken to show the performance of the proposed technique. The comparison results clearly show that the clustering accuracy using FRDNL-XDC technique is improved by 5% and 11% when compared to existing UKmeans clustering mechanism [1] and UK- medoids-SMDM [2] respectively.

5.2 Performance Results of False Positive Rate

The false positive rate is computed as the ratio of the number of data is incorrectly grouped into the particular cluster to the total number of data. The false positive rate is computed as follows,

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

From (8) FPR represents the false positive rate and'n' represents the number of data. The false positive rate is measured in terms of percentage (%). Lower the false positive rate, more efficient the method is said to be. Table 2 describes the experimental results of the false positive rate with respect to a number of data.

Table 2: False Positive Rate Versus The Number Of Data

No. of data	False positive rate (%)		
	FRDNL-XDC	UKmeans clustering mechanism	UK-medoids-SMDM
1000	13	17	20
2000	16	19	24
3000	14	23	26
4000	16	20	25
5000	18	21	26
6000	15	19	24
7000	13	16	20
8000	10	14	18
9000	11	15	19
10000	12	17	22

For computing the false positive rate, the number of data is taken from 1000 to 10000. Totally, 10 various runs are carried out to show the experimental results of the proposed FRDNL-XDC technique and existing clustering techniques such as UKmeans clustering mechanism [1] and UK-medoids-SMDM [2]. The experimental results clearly obvious that the false positive rate is considerably minimized using FRDNL-XDC technique than the conventional clustering methods. The graphical representation of the FRDNL-XDC technique is shown in figure 5. As shown in figure 5, the experimental results of the false positive rate with respect to a number of data using weather dataset. By applying weather dataset, the different surface data are collected from the pacific region with different years. For the purpose of grouping, the uncertain data are removed.

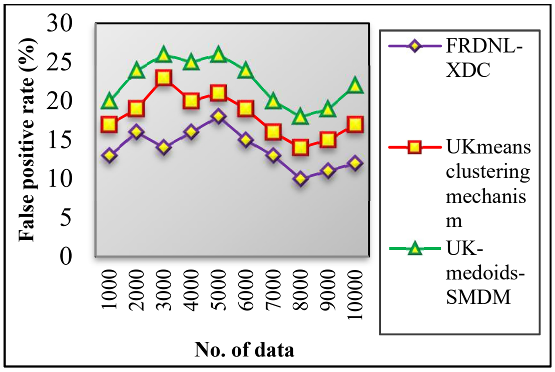

Figure 5: Performance Results Of False Positive Rate Versus The Number Of Data

In figure 5, three different colors of lines are shows that the false positive results of three different methods. The green color line indicates the false positive results of FRDNL-XDC technique. The red color and violet color lines indicate the false positive rate of UKmeans clustering mechanism [1] and UK- medoids-SMDM [2] respectively. The above graphical results show that the FRDNL-XDC technique minimizes the false positive rate than the existing methods.

This is because, the FRDNL-XDC technique perform recurrent processes at different time period for improving the clustering accuracy. The activation function at the output layer used to find the data which is not a member of the cluster. In this case, the Akaike information criterion function is calculated to find the maximum likelihood between the data and cluster centroid. Then the data is grouped into that particular cluster. This helps to minimize the error during the clustering process.

Totally ten various experimental results are attained with different data. The performance results of the proposed FRDNL-XDC technique are compared to the conventional clustering methods. The comparison results show that the FRDNL-XDC technique minimizes the false positive rate by 24% when compared to existing UKmeans clustering mechanism [1].

In addition, the false positive results of FRDNL-XDC technique is minimized by 39% when compared to existing UK- medoids-SMDM [2].

5.3 Performance Results of Time Complexity

Time complexity is defined as the amount of time required for grouping the data into different clusters. Time complexity is measured in terms of milliseconds (ms). The formula for computing the time complexity is expressed as follows,

$$TC = N * time\ (clustering\ the\ data\ objects) \qquad (9)$$

From (9), TC denotes the time complexity and 'N' denotes the number of data. The time complexity of the three different clustering methods is described in table 3. For various input data, ten different results are attained. The above experimental results clearly illustrate that the proposed FRDNL-XDC technique minimizes the time complexity when compared to existing UKmeans clustering mechanism [1] and UK-medoids-SMDM [2]. This significant result show thatthe time complexity of FRDNL-XDC technique is minimized when compared to the conventional clustering techniques.

Table 3: Time Complexity Versus The Number Of Data

No. of data	Time complexity (ms)		
	FRDNL-XDC	UKmeans clustering mechanism	UK-medoids-SMDM
1000	23	32	41
2000	26	34	40
3000	30	36	45
4000	36	44	48
5000	35	41	46
6000	43	49	55
7000	49	53	57
8000	52	56	62
9000	53	59	64
10000	57	62	68

Figure 6 illustrates the experimental results of time complexity using El Nino dataset with a number of data. The numbers of data are taken as an input for calculating the time complexity. The graphical result obviously shows that the proposed FRDNL-XDC technique minimizes the time complexity when compared to existing methods. The FRDNL-XDC technique performs the deep learning process for mining the uncertain data through the clustering process. In a deep learning process, three different layers are considered for processing the large data. This helps to minimize the time complexity. In addition, x means clustering technique is applied for grouping the similar data into the particular cluster through the correlation measure.

The bivariate correlation coefficient provides two results such as positive and negative similarity. The positive similarity results are used for grouping the more similar data into the clusters with minimum time. The averages of ten different results of the time complexity are computed resulting minimizes time complexity by 15% and 25% when compared to existing UKmeans clustering mechanism [1] and UK- medoids-SMDM [2] respectively. The above experimental results show that the proposed FRDNL-XDC technique efficiently increasing the clustering accuracy and minimizing the false positive rate as well as time complexity.

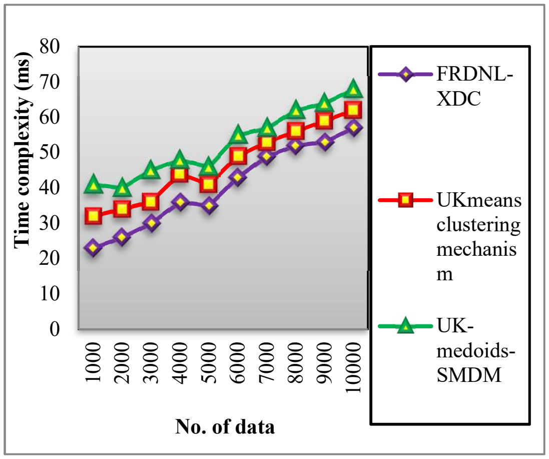

Figure 6: Performance Results of Time Complexity versus the Number of Data

Based on the result evaluation of FRDNL-XDC technique for uncertain data mining, it can be clearly seen that the model differs from prior works in three points:

- The model differs from prior works which applying Fully Recurrent Deep Neural Learning in order to increase the clustering performance of 'x' means clustering as compared to existing works to effectively mine the uncertain data.
- The time complexity of FRDNL-XDC technique is reduced as it employs bivariate correlation on the contrary to state-of-the-art works. This bivariate correlation is estimated between each data and cluster centroids. According to the correlation measure, the similar data are grouped into a cluster with minimal amount of time.
- The FRDNL-XDC technique used sigmoid activation function and Akaike information criterion on the contrary to conventional work to reduce the incorrct data clustering. The activation function is utilized in FRDNL-XDC technique to make sure all data are grouped into the cluster or not. If any of the data are not a member of the clusters, a FRDNL-XDC technique exploits the Akaike information criterion to determine the maximum probability between the data and cluster centroid.

6. CONCLUSION

An efficient technique called Fully Recurrent Deep Neural Learning based X-means Data Clustering (FRDNL-XDC) is developed for mining uncertain data with minimal time complexity. The numbers of data are collected from the dataset. The FRDNL-XDC technique uses deep learning with the numbers of data and the 'X' means clustering technique. The deep learning approach comprises the different layers and groups the similar data into different clusters using bivariate correlation coefficient. The correlation coefficient result improves the clustering accuracy with minimum time complexity. In addition, the sigmoid activation function is used at the output layer for accurately grouping the data into the cluster. Finally, the Akaike information criterion is used to group the data into that cluster. This helps to minimize the false positive rate. The experimental assessment is carried out using El Nino dataset with different parameters such as clustering accuracy, false positive rate and time complexity. The results show that the performance of FRDNL-XDC technique increases the clustering accuracy and minimizes the time complexity as well as the false positive rate when compared to state-of-art methods.

REFERENCES:

[1] Chuan-Ming Liu, Zhendong Niu and Kuan-Teng Liao, "Mechanisms to Improve Clustering Uncertain Data with UKmeans", Data & Knowledge Engineering, Elsevier, Vol. 116, July 2018, pp. 61-79

[2] Han Liu, Xianchao Zhang, Xiaotong Zhang, Yi Cui, "Self-adapted mixture distance measure for clustering uncertain data",Knowledge-Based Systems, Elsevier, Vol. 126, 2017, pp. 33–47

[3] Sampreeti Ghos and Sushmita Mitra, "Clustering large data with uncertainty", Applied Soft Computing, Elsevier, Vol. 13, 2013, pp. 1639–1645

[4] Zahid Halim, Muhammad Waqas, Abdul Rauf Baig, Ahmar Rashid, "Efficient Clustering of Large Uncertain Graphs Using Neighborhood

Information", International Journal of Approximate Reasoning, Elsevier, Vol. 90, November 2017, pp. 274-291

[5] Lei Xu, Qinghua Hu, Edward Hung, Baowen Chen, Xu Tan, Changrui Liao, "Large margin clustering on uncertain data by considering probability distribution similarity", Neurocomputing, Elsevier, Vol. 158, 2015, pp. 81-89

[6] Aalaa Mojahed and Beatriz de la Iglesia, "An adaptive version of k-medoids to deal with the uncertainty in clustering heterogeneous data using an intermediary fusion approach", Knowledge and Information Systems, Vol. 50, No. 1,2017, pp. 27–52

[7] Bin Jiang, Jian Pei , Yufei Tao , Xuemin Lin, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions on Knowledge and Data Engineering , Vol. 25 , No. 4 , 2013, pp. 751 – 763

[8] Ondrej Linda and Milos Manic, "General Type-2 Fuzzy C-Means Algorithm for Uncertain Fuzzy Clustering", IEEE Transactions on Fuzzy Systems, Vol. 20 , No. 5 , 2012 , pp. 883 – 897

[9] Xianchao Zhang, Han Liu, Xiaotong Zhang, "Novel density-based and hierarchical density-based clustering algorithms for uncertain data", Neural Networks, Elsevier, Vol. 93, 2017, pp. 240-255

[10] Inhua Li, Shiji Song, Yuli Zhang, and Zhen Zhou, "Robust K-Median and K-Means Clustering Algorithms for Incomplete Data", Mathematical Problems in Engineering, Hindawi Publishing Corporation, Vol. 2016, October 2016, pp. 1-8

[11] Leandro A.F.Fernandes and Manuel M.Oliveira, "Handling uncertain data in subspace detection", Pattern Recognition, Elsevier, Vol. 47, No. 10, 2014, pp. 3225-324

[12] Vandana Dubey, A. A Nikose, "Technique for Clustering Uncertain Data Based On Probability Distribution Similarity", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, No. 7, July 2015, pp. 2629- 2633

[13] Thierry Denoeux, "Maximum Likelihood Estimation from Uncertain Data in the Belief Function Framework", IEEE Transactions on Knowledge and Data Engineering , Vol. 25 , No. 1 , 2013, pp. 119 – 130

[14] Ben Kao , Sau Dan Lee , Foris K.F. Lee , David W. Cheung , Wai-Shing Ho, "Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index", IEEE Transactions on Knowledge and Data Engineering, Vol. 22 , No. 9 , 2010, pp. 1219 – 1233

[15] Jin Zhou, Long Chen, C. L. Philip Chen, Yingxu Wang and Han-Xiong Li, "Uncertain Data Clustering in Distributed Peer-to-Peer Networks", IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 6, June 2018, pp. 2392 – 2406

[16] Mohammad Khan Afridi, Nouman Azam, Jing Tao, Yao Eisa Alanazi, "A three-way clustering approach for handling missing data using GTRS", International Journal of Approximate Reasoning, Elsevier, Vol. 98, 2018, pp. 11-24

[17] Lei Xu ,Qinghua Hu, Edward Hunga, Chi-Cheong Szeto, "A heuristic approach to effective and efficient clustering on uncertain objects", Knowledge-Based Systems, Elsevier, Vol. 66, August 2014, pp. 112-125

[18] Jun-Wei Ji and C.L.Philip Chen, "Regularized robust Broad Learning System for uncertain data modeling", Neuro computing, Elsevier, Vol. 322, 2018, pp. 58-69

[19] Francesco Gullo, Giovanni Ponti, Andrea Tagarelli and Sergio Greco, "An information-theoretic approach to hierarchical clustering of uncertain data", Information Sciences, Elsevier, Vol. 402, September 2017, pp. 199-215

[20] Zahid Halim and Jamal Hussain Khattak, "Density-based clustering of big probabilistic graphs", Evolving Systems, Springer, March 2018, pp. 1–18

[21] Shahana Bano, K.Rajasekhara Rao, "Graph Based Gene/Protein Prediction and Clustering Over Uncertain Medical Databases", Journal of Theoretical and Applied Information Technology, Vol. 82, No. 3, December 2015, pp. 347-352

[22] K. Shyam Sunder Reddy, C. Shoba Bindu, "Rtdbstream: A Real-Time Density-Based Clustering For Evolving Data Streams", Journal of Theoretical and Applied Information Technology, Vol. 96, No. 12, June 2018, pp. 3973-3983