Sørensen-Dice Similarity Indexing based Weighted Iterative Clustering for Big Data Analytics

KalyanaSaravanan Annathurai[1] and Tamilarasi Angamuthu[2]
[1]Department of Computer Science and Engineering, Kongu Engineering College, India
[2]Department of Computer Applications, Kongu Engineering College, India

Abstract: *Big data is a collection of large volume of data and extract similar data points from large dataset. Clustering is an essential data mining technique for examining large volume of data. Several techniques have been developed for handling big dataset. However, with much time consumption and space complexity, accuracy is said to be compromised. In order to improve clustering accuracy with less complexity, Sørensen-Dice Indexing based Weighted Iterative X-means Clustering (SDI-WIXC) technique is introduced. SDI-WIXC technique is used for grouping the similar data points with higher clustering accuracy and minimal time. First, number of data points is collected from big dataset. Then, along with the weight value, the given dataset is partitioned into 'X' number of clusters. Next, based on the similarity measure, Weighted Iterated X-means Clustering (WIXC) is applied for clustering data points. Sørensen-Dice Indexing Process is used for measuring similarity between cluster weight value and data points. Upon similarity found between weight value of cluster and data point, data points are grouped into a specific cluster. Besides, the WIXC method also improves the cluster assignments through repeated subdivision using Bayesian probability criterion. This in turn helps to group all data points and hence, improving the clustering accuracy. Experimental evaluation is carried out with number of factors such as clustering accuracy, clustering time and space complexity with respect to the number of data points. The experimental results reported that the proposed SDI-WIXC technique obtains high clustering accuracy with minimum time as well as space complexity*.

Keywords: *Bayesian probability criterion, big data analytics, sørensen-dice indexing process, weighted iterated x-means clustering*.

Received August 3, 2019; accepted May 9, 2020
https://doi.org/10.34028/iajit/19/1/2

1. Introduction

Big Data analytics is the method of gathering and analyzing the large sets of data points for extracting the valuable information. Conventionally, large data is handled through several data mining techniques. But the techniques failed to obtain the dimension reduction. Therefore, the clustering is the significant data mining technique used for big data analysis. The clustering is a task of grouping a similar data points to minimize the curse of dimensionality.

Scalable Random Sampling with Iterative Optimization Fuzzy C-Means algorithm (SRSIO-FCM) was introduced in [1] to deal with the big data. The SRSIO-FCM algorithm minimizes the time and space complexity. The SRSIO-FCM algorithm failed to determine the number of clusters for the specific dataset and the quality of cluster was not improved. A High-Order CFS algorithm (HOCFS) was designed in [2] to group the various data by integrating the Clustering by Fast Search (CFS) clustering algorithm as well as the dropout deep learning approach. The algorithm failed to cluster the complex heterogeneous data.

K-means clustering algorithm and MapReduce technique were designed in [3] for processing the large-scale data and obtain efficient, robust and scalable clustering results. The technique does not minimize the space complexity of large-scale data clustering. A Competitive K-Means (CK-Means) clustering technique was presented in [4] for performing clusters analysis on large datasets. The CK-Means clustering technique failed to improve the performance of accurate cluster analysis over large datasets.

An extending standard Random Forests technique was developed in [5] for processing the big data in offline or online contexts. While handling the big data, the space complexity remained unsolved. A Clustering-based Collaborative Filtering (ClubCF) technique was introduced in [6] for big data applications. The clustering technique minimizes the online computation time but it failed to group the more similar services.

Adapting k-means algorithm was developed in [7] for grouping the big data. The algorithm improves the big data analysis with highly scalable and accuracy. It does not work well with clusters of various size and density. In [8], Clustering Visual Objects in Pair wise Similarity Matrix (CVOIPSM) was constructed for improving the grouping accuracy of similar objects from both small and large datasets.

A spatiotemporal indexing approach was introduced in [9] for efficiently handling the big climate data with MapReduce. The approach does not improve its performance for handling more generic scenarios. Spectral Ensemble Clustering (SEC) algorithm was designed in [10] for improving the big data clustering and minimizing the time and space complexity. The algorithm failed to group the more similar data into a cluster thus minimizes the clustering accuracy.

The issues identified from the above literature such as minimum clustering accuracy, failed to cluster complex data, uncover hidden patterns, more space and time complexity, failed to improve the clustering performance with big data and so on. In order to resolve the issues, an efficient technique called Sørensen Dice Indexing based Weighted Iterative X-means Clustering (SDI-WIXC) technique is introduced.

The contribution of the SDI-WIXC technique is described as follows.

1. SDI-WIXC technique is to improve the clustering accuracy and to reduce the time as well as space complexity. It is achieved by applying the Weighted Iterative X-means clustering technique for examining thecollected meteorological data. WIXC technique measures the similarity for data points of oceanographic and surface meteorological readings. Based on the similarity measure of meteorological readings, the similar data points are grouped into a cluster. This helps to minimize the space complexity.
2. The weighted iterative 'X' means clustering function repeatedly and verifies whether all the data points are grouped. If any data point does not group into a cluster, Bayesian probability criterion is used to find a higher probability of data points to become a cluster member. This helps to improve the clustering accuracy.

The rest of the paper is organized as follows. Section 2 briefly discusses the related works. Section 3 provides the description of SDI-WIXC technique with the algorithm in detail. In section 4, experimental evaluation of proposed and existing state-of-art methods are described. Section 5 provides the results and discussion of certain parameters with different datasets. Finally, the conclusion of the research work is presented in section 6.

2. Related Work

Fast Constrained Spectral Clustering (CSC) and Cluster Ensemble with Random Projection were developed in [11] to greatly improve the clustering accuracy. The CSC algorithm does not provide the different weights for basic partitions to obtain high clustering performances. K-means algorithm was introduced in [12] for enhancing the accuracy. The designed algorithm was combined with Initial Centroid Selection Optimization technique and genetic algorithm. Chi-square similarity was evaluated for grouping the data set. The self-organizing map was applied for increasing the clustering process. But, the clustering performance was not improved.

A Weighted kernel Possibilistic C-Means (wkPCM) clustering technique was introduced in [13] for grouping the big data into different clusters. The clustering technique failed to improve the clustering performance of the different types of big data. Principal component analysis based on variance covariance structure was designed in [14]. However, the space complexity was not minimized by principal component analysis.

Density-based Clustering Algorithm with an Improved Version (DENCLUE-IM) was introduced in [15] for grouping the data points. The DENCLUE-IM does not use an efficient clustering algorithm that does not satisfy all the principle of big data. Certain permutation and combination-based approaches for big data analysis was provided in [16]. The existing methods did not obtain distributions for linear combinations. But, the error rate was not minimized by permutation and combination-based approaches.

Interrelationship between the number of labels of the fuzzy variables and the scarcity of the data was analyzed [17] to perform data sampling in MapReduce. By setting high number of Map processes, the number of labels to represent the problem was considerably. Clustering of big spatiotemporal data was performed [18] in multiple Euclidean spaces using centroid-based clustering. The clustering technique failed to optimize the number of clusters based on similarity measure.

A new K‑Means modified inter and Intra Clustering (KM-I2C) algorithm was designed in [19] to group the data points by using distance metric. The KM-I2C algorithm does not improve the clustering performance using large datasets. Yet another model-based clustering using discriminate analysis with regularization approach was designed in [20] with the objective of addressing optimal variable selection in high dimensional dataset. However, the time consumption for clustering was not minimized.

Random Sampling And Consensus (RANSAC) technique were introduced in [21] for big data clustering. The technique improves the accuracy of reduced complexity but the performance of technique was not improved. Fast kernel matrix computation was performed in [25] using k means clustering algorithm for big data clustering. The algorithm does not group the more similar data with minimum time complexity.

Fuzzy Consensus Clustering (FCC) technique was introduced in [23] for big data clustering. The FCC technique failed to estimate the number of clusters and improve the cluster validity.

In [24], a new Fuzzy C-Means (FCM) clustering algorithm with random projection was introduced to

minimize the dimensionality. The FCM clustering algorithm does not improve the clustering accuracy. An efficient distributed density-based clustering of big data using Hadoop called Cludoop algorithm was introduced in [26]. The clustering algorithm does not improve the cluster quality since it has more time to group the data points. The issues identified from the above-said methods are overcome by introducing an SDI-WIXC technique. The description of the SDI-WIXC technique is presented in the following section.

3. The Reflective Process

Clustering is considered to be one of the significant data mining techniques for big data analysis, where huge amount of similar data is grouped.

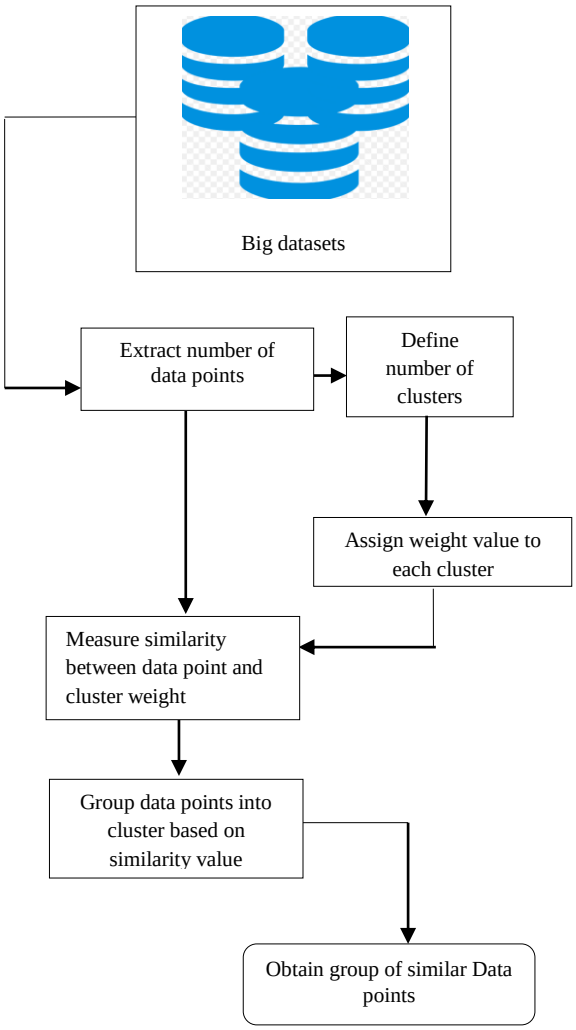


Figure 1. Architecture diagram of the SDI-WIXC technique.

Big data involves huge data volume and also complex in nature. In other words, big data mining is the ability to mine significant information from huge dataset. Several clustering techniques have been introduced to extract information from large dataset. But with the enormous data, processing is said to be complicated due to limitations in storage. To address these issues in this work, a new data mining technique called SDI-WIXC is investigated for two different big datasets, Historical hourly weather data 2012-2017 and Weather dataset. Figure 2 represents the clustering of air temperature and subsurface temperature data in Historical Hourly Weather Data 2012-2017.

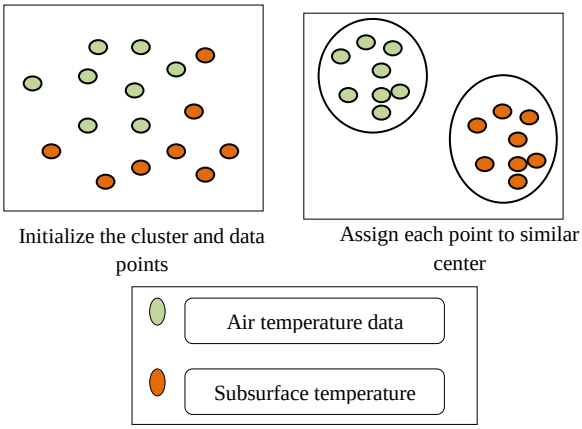


Figure 2. clustering of air temperature and subsurface temperature data in historical hourly weather data 2012-2017.

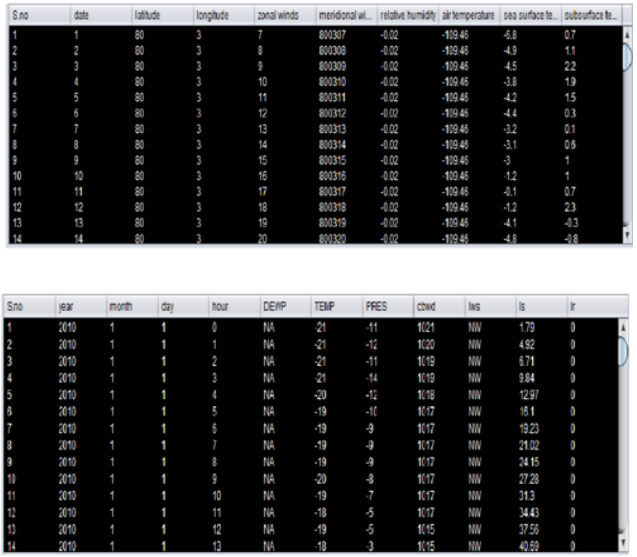
The Sørensen-Dice Indexing is a method for measuring similarity of two variables. It was developed by the botanists thorvald sørensen and lee raymond dice. Hence the name is called as Sørensen-Dice Indexing. According to the results of similarity, clustering process is carried out using Weighted Iterative X-means technique. These processes are described in following sections. Figure 1 shows the architecture diagram of SDI-WIXC technique to group similar data points into different clusters for minimizing dimensionality for two different big datasets, Historical hourly weather data 2012-2017 and Weather dataset. The proposed SDI-WIXC technique efficiently groups data points based on similarity value.

Initially, with the help of two input big dataset, Historical hourly weather data 2012-2017 and Weather dataset, number of data points is extracted. Then for '*X*' number of clusters, weights are assigned. Next, the similarity between the data points and cluster weight is measured using Sørensen-Dice Indexing. Based on the similarity value, the data points are assigned to a specific group. At first, initialize the data points and cluster center. Consider X number of clusters where X=2. Assign each point to the similar center or cluster. The weight value for each cluster is assigned.

Let us consider the number of data points is extracted from the big dataset is described as follows,

$$D_i = D_1, D_2, D_3 \ldots, D_n \in R^d \quad (1)$$

From (1), D_i denotes a set of data points $d_1, d_2, d_3, \ldots, d_n$ and R^d is the big dataset. The extracted data points for Historical Hourly Weather Data 2012-2017 and Weather dataset is given below.



S.no	date	latitude	longitude	zonal winds	meridional wi...	relative humidity	air temperature	sea surface te...	subsurface te...
1	1	80	3	7	800307	-0.02	-109.46	-6.8	0.7
2	2	80	3	8	800308	-0.02	-109.46	-4.9	1.1
3	3	80	3	9	800309	-0.02	-109.46	-4.5	2.2
4	4	80	3	10	800310	-0.02	-109.46	-3.8	1.9
5	5	80	3	11	800311	-0.02	-109.46	-4.2	1.5
6	6	80	3	12	800312	-0.02	-109.46	-4.4	0.3
7	7	80	3	13	800313	-0.02	-109.46	-3.2	0.1
8	8	80	3	14	800314	-0.02	-109.46	-3.1	0.6
9	9	80	3	15	800315	-0.02	-109.46	-3	1
10	10	80	3	16	800316	-0.02	-109.46	-1.2	1
11	11	80	3	17	800317	-0.02	-109.46	-0.1	0.7
12	12	80	3	18	800318	-0.02	-109.46	-1.2	2.3
13	13	80	3	19	800319	-0.02	-109.46	-4.1	-0.3
14	14	80	3	20	800320	-0.02	-109.46	-4.8	-0.8

S.no	year	month	day	hour	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
1	2010	1	1	0	NA	-21	-11	1021	NW	1.79	0
2	2010	1	1	1	NA	-21	-12	1020	NW	4.92	0
3	2010	1	1	2	NA	-21	-11	1019	NW	6.71	0
4	2010	1	1	3	NA	-21	-14	1019	NW	9.84	0
5	2010	1	1	4	NA	-20	-12	1018	NW	12.97	0
6	2010	1	1	5	NA	-19	-10	1017	NW	16.1	0
7	2010	1	1	6	NA	-19	-9	1017	NW	19.23	0
8	2010	1	1	7	NA	-19	-9	1017	NW	21.02	0
9	2010	1	1	8	NA	-19	-9	1017	NW	24.15	0
10	2010	1	1	9	NA	-20	-8	1017	NW	27.28	0
11	2010	1	1	10	NA	-19	-7	1017	NW	31.3	0
12	2010	1	1	11	NA	-18	-5	1017	NW	34.43	0
13	2010	1	1	12	NA	-19	-5	1015	NW	37.55	0
14	2010	1	1	13	NA	-18	-3	1015	NW	40.69	0

Figure 3. Feature extraction for weather dataset and historical hourly weather data 2012-2017.

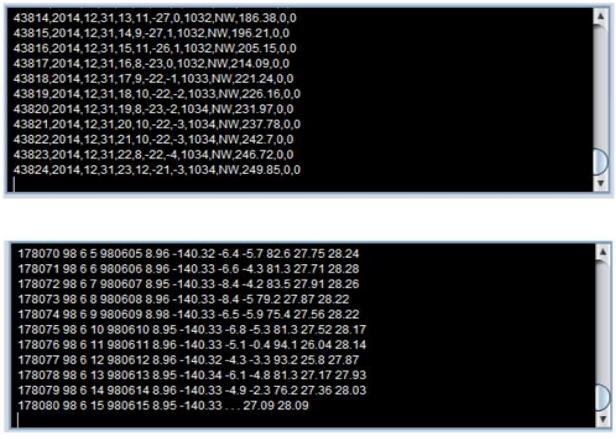
From the above Figure 3, with the given big dataset Weather dataset and Historical Hourly Weather Data 2012-2017. By applying the SDI-WIXC technique, '*X*' number of clusters is initialized. By applying the SDI-WIXC technique, the '*X*' number of clusters is initialized.

$$C_i = c_1, c_2, c_3, ..., c_x \tag{2}$$

From (2) C_i denotes a cluster set and C_1, C_2, $C_3 \ldots C_x$ represents '*X*' number of clusters. Then the weight value is assigned randomly to each cluster.

$$w_t = \{c_1, c_2, c_3, ..., c_x\} \tag{3}$$

From (3), w_t denotes a weight value to each cluster. After allocating the weight value to each cluster, data points D_i are grouped. Clustering is a separation of data points into different groups of similar objects. Each group, called a cluster which includes data points that are similar to one another and different to other groups. The similarity measurement for two datasets is given below in Figure 4.



```
43814,2014,12,31,13,11,-27,0,1032,NW,186.38,0,0
43815,2014,12,31,14,9,-27,1,1032,NW,196.21,0,0
43816,2014,12,31,15,11,-26,1,1032,NW,205.15,0,0
43817,2014,12,31,16,8,-23,0,1032,NW,214.09,0,0
43818,2014,12,31,17,9,-22,-1,1033,NW,221.24,0,0
43819,2014,12,31,18,10,-22,-2,1033,NW,226.16,0,0
43820,2014,12,31,19,8,-23,-2,1034,NW,231.97,0,0
43821,2014,12,31,20,10,-22,-3,1034,NW,237.78,0,0
43822,2014,12,31,21,10,-22,-3,1034,NW,242.7,0,0
43823,2014,12,31,22,8,-22,-4,1034,NW,246.72,0,0
43824,2014,12,31,23,12,-21,-3,1034,NW,249.85,0,0
```



```
178070 98 6 5 980605 8.96 -140.32 -6.4 -5.7 82.6 27.75 28.24
178071 98 6 6 980606 8.96 -140.33 -6.6 -4.3 81.3 27.71 28.28
178072 98 6 7 980607 8.95 -140.33 -8.4 -4.2 83.5 27.91 28.26
178073 98 6 8 980608 8.96 -140.33 -8.4 -5 79.2 27.87 28.22
178074 98 6 9 980609 8.98 -140.33 -6.5 -5.9 75.4 27.56 28.22
178075 98 6 10 980610 8.95 -140.33 -6.8 -5.3 81.3 27.52 28.17
178076 98 6 11 980611 8.96 -140.33 -5.1 -0.4 94.1 26.04 28.14
178077 98 6 12 980612 8.96 -140.32 -4.3 -3.3 93.2 25.8 27.87
178078 98 6 13 980613 8.95 -140.34 -6.1 -4.8 81.3 27.17 27.93
178079 98 6 14 980614 8.96 -140.33 -4.9 -2.3 76.2 27.36 28.03
178080 98 6 15 980615 8.95 -140.33 . . . 27.09 28.09
```

Figure 4. Calculate similarities in Weather dataset and historical hourly weather data 2012-2017.

The Sørensen-Dice Indexing technique is applied for measuring the similarity [26]. The similarity between data points and a weight value of cluster is measured as follows,

$$\rho = \log_{10} 10^2 * \beta \tag{4}$$

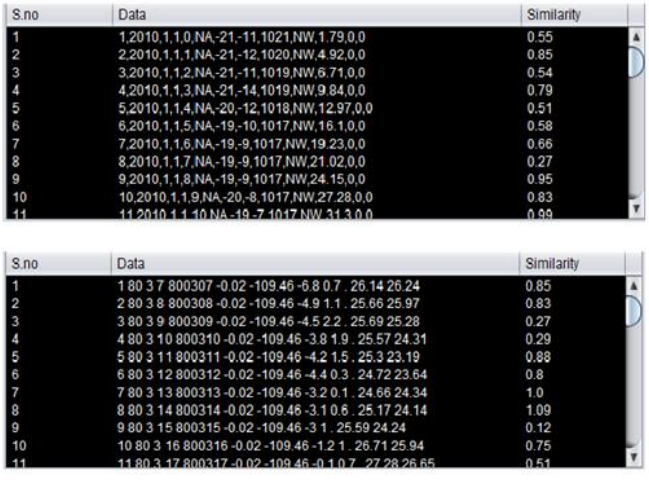
$$\beta = \frac{D_i \cap C_{w_t}}{|D_i| * |C_{w_t}|} \tag{5}$$

The above Equation (5), β represents the ratio function of the mutual dependence between data points D_i and a weight value of cluster C_{w_t} to the cardinalities of the two sets. The intersection symbol '$\cap$' denotes a mutual dependence between the data points and a weight value of cluster with cardinalities of two sets represented by $|D_i|$ and $|C_{w_t}|$. Cardinality is a measure of the number of similar data points of the given set. Finally, the above Equation (5) is substituted to the Equation (6) to get the final coefficient results,

$$\rho = \log_{10} 10^2 * \beta * \left(\frac{D_i \cap C_{w_t}}{|D_i| * |C_{w_t}|} \right) \tag{6}$$

From (6), ρ denotes a Sørensen-Dice similarity coefficient, D_i represents data points in the big dataset, C_{w_t} denotes a weight value of the cluster.

Figure 5 shows the similarity measurement between data point and cluster weight for historical hourly weather data 2012-2017 and weather dataset.



S.no	Data	Similarity
1	1,2010,1,1,0,NA,-21,-11,1021,NW,1.79,0,0	0.55
2	2,2010,1,1,1,NA,-21,-12,1020,NW,4.92,0,0	0.85
3	3,2010,1,1,2,NA,-21,-11,1019,NW,6.71,0,0	0.54
4	4,2010,1,1,3,NA,-21,-14,1019,NW,9.84,0,0	0.79
5	5,2010,1,1,4,NA,-20,-12,1018,NW,12.97,0,0	0.51
6	6,2010,1,1,5,NA,-19,-10,1017,NW,16.1,0,0	0.58
7	7,2010,1,1,6,NA,-19,-9,1017,NW,19.23,0,0	0.66
8	8,2010,1,1,7,NA,-19,-9,1017,NW,21.02,0,0	0.27
9	9,2010,1,1,8,NA,-19,-9,1017,NW,24.15,0,0	0.95
10	10,2010,1,1,9,NA,-20,-8,1017,NW,27.28,0,0	0.83
11	11,2010,1,1,10,NA,-19,-7,1017,NW,31.3,0,0	0.99

S.no	Data	Similarity
1	1 80 3 7 800307 -0.02 -109.46 -6.8 0.7 . 26.14 26.24	0.85
2	2 80 3 8 800308 -0.02 -109.46 -4.9 1.1 . 25.66 25.97	0.83
3	3 80 3 9 800309 -0.02 -109.46 -4.5 2.2 . 25.69 25.28	0.27
4	4 80 3 10 800310 -0.02 -109.46 -3.8 1.9 . 25.57 24.31	0.29
5	5 80 3 11 800311 -0.02 -109.46 -4.2 1.5 . 25.3 23.19	0.88
6	6 80 3 12 800312 -0.02 -109.46 -4.4 0.3 . 24.72 23.64	0.8
7	7 80 3 13 800313 -0.02 -109.46 -3.2 0.1 . 24.66 24.34	1.0
8	8 80 3 14 800314 -0.02 -109.46 -3.1 0.6 . 25.17 24.14	1.09
9	9 80 3 15 800315 -0.02 -109.46 -3 1 . 25.59 24.24	0.12
10	10 80 3 16 800316 -0.02 -109.46 -1.2 1 . 26.71 25.94	0.75
11	11 80 3 17 800317 -0.02 -109.46 -0.1 0.7 . 27.28 26.65	0.51

Figure 5. Similarity between data point and cluster in weather dataset and Historical Hourly Weather Data 2012-2017.

The similarity 'ρ' denotes a Sørensen-Dice similarity coefficient that provides similarity value between 0 and 1. When the weight value of cluster and data point is similar, then the output value is '1'. When the two weight value is dissimilar, the output value is '0'. After measuring the similarity, more similar data points are grouped into the cluster. If a data points does not belong to any of the cluster, then weighted iterative x-means clustering is applied in the proposed work to

improve cluster assignments by repeatedly attempting subdivision, using Bayesian probability criterion. This helps to assign entire data points into specific cluster based on the likelihood. The Bayesian probability criterion [26] is defined as follows,

$$\delta = -2 \log P(D_i \mid C_{w_t}) + \log(n) \qquad (7)$$

From (7), 'δ'denotes Bayesian probability criterion function, 'P' denotes a probability function, 'D_i denotes the number of data points and 'C_{w_t} denotes weight value of cluster and 'n' represents the number of data points. Here '$P(D_i \mid C_{w_t})$ denotes a maximized value of the likelihood function of the data points and cluster. The likelihood shows that the higher probability of data points becomes a member of particular cluster. As a result, all data points are grouped into any one of cluster based on their weight value. This process is repeated until the entire data points in big dataset are grouped. This in turn helps to improve the clustering accuracy as well as reduce the time consumption. By applying the clustering technique, the big dataset is divided into various groups based on maximum likelihood probability. Figures 6 and 7 illustrates the grouping of data points into cluster using weather dataset and historical hourly weather data 2012-2017 respectively. As illustrated in the above two Figures 6 and 7, as the grouping of data points into cluster are performed using Bayesian probability criterion function, the data points within a cluster is said to be more similar. Flow chart of Sørensen-Dice Indexing based WIXC is described as follows,

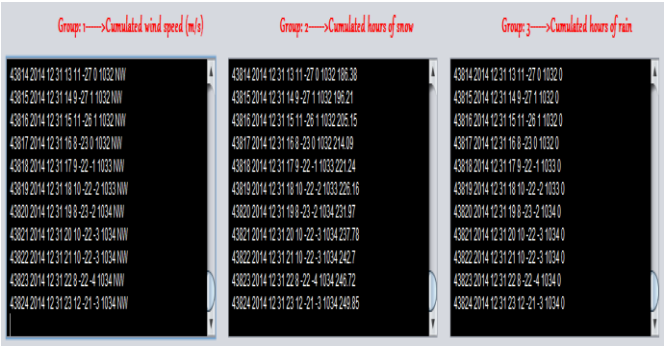


Figure 6. Grouping of data points into cluster (weather dataset).

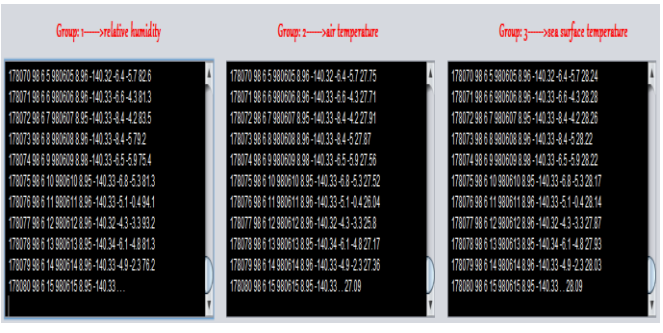


Figure 7. Grouping of data points into cluster (historical hourly weather data 2012-2017).

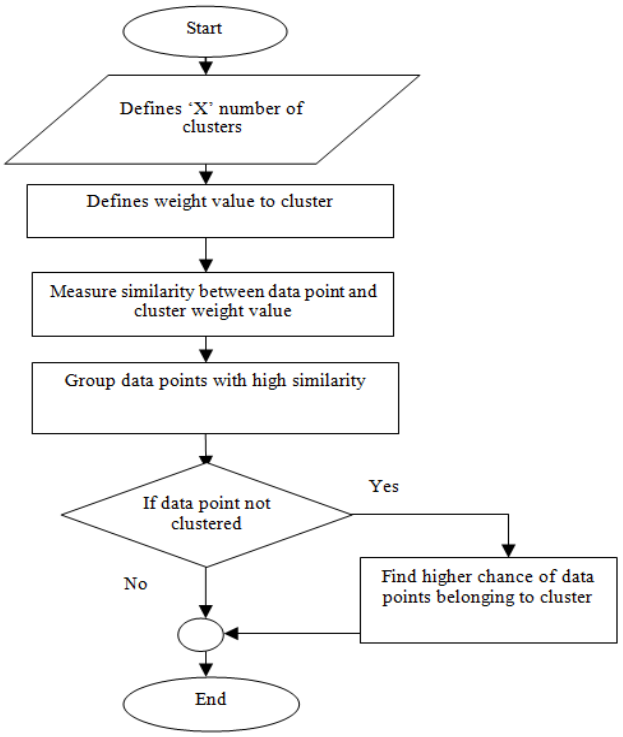


Figure 8. Flowchart of SDI-WIXC technique.

Figure 8 shows the flowchart of the SDI-WIXC technique to improve the clustering processes. The clusters and weight values are initialized randomly. After that, the similarity between data points and cluster is computed to group the more related data points. For example, the first cluster (air temperature) includes the data points of 32.1, 32.2, 32.3, 32.4, 32.6, 32.7, 32.8, and 32.9. The weight value of first cluster is 32. Therefore32.1, 32.2, 32.3, and 32.4 data points are grouped as cluster. The remaining points in the first cluster is grouped to second cluster by calculating Bayesian probability criterion to find the maximum likelihood of point for grouping as a cluster member. Maximum likelihood is determined on 32.6, 32.7, 32.8, 32.9 data points (32.6≈33) and grouped to second cluster.

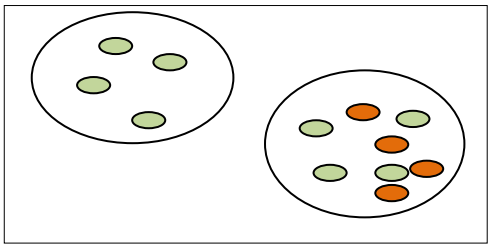


Figure 9. Example for clustering the more related points.

This helps to partition all data points into 'X' clusters. If any data points move into the group, then the Bayesian probability is measured. As a result, all the data points are grouped into a particular cluster which is illustrated in Figure 9. The algorithmic process of SDI-WIXC technique is described as follows, Algorithm (1) describes the clustering of data points to minimize dimensionality. The proposed SDI-

WIXC technique effectively partitions data points into different clusters for the given big dataset. Initially, 'X' number of clusters and their weight values are randomly initialized for the given dimensional space. For each data point, the similarity between data point and cluster weight value is calculated for grouping more related data points. Then the weighted iterative 'X' means clustering repeatedly checks the data points until higher probability to become a cluster member using Bayesian probability criterion is arrived at. Similarly, entire data points in big dataset are partitioned resulting in improvement of the clustering accuracy and therefore minimize the space complexity. The performance analysis of the proposed SDI-WIXC technique is explained in the next sections.

Algorithm 1: Sørensen-Dice Indexing based Weighted Iterative X-means Clustering

Input: Number of data points $D_1, D_2, D_3 ..., D_n$
Output: Improve clustering accuracy
Begin
1. Initialize the 'X' number of cluster C_x
2. Assign weight W_t *to cluster* C_x
3.For each data points in dataset R^d
4. Measure similarity between the cluster weight W_t *and data point* D_i
4.If similarity coefficient $\rho = 1$ *then*
5. W_t *and* D_i *are more similar*
6. else
7. W_t *and* D_i *are dissimilar*
8. end if
9. Similar data points D_i *grouped into cluster* C_x
10. If D_i *not moved into a group then*
11. Repeatedly attempting subdivision to group the data points using Bayesian probability criterion' δ *'*
12. else
13. Stop the clustering process
14. end if
15. end for
end

4. Experimental Evaluation

Experimental evaluation of the SDI-WIXC technique is implemented using java language with 3.4 GHz Intel Core i3processor, 4GB RAM, and windows 7 platform. It uses two big datasets namely Historical Hourly Weather Data 2012-2017 and Weather dataset. The historical hourly weather data 2012-2017 is taken from the, Kaggle https://www.kaggle.com/selfishgene/historical-hourly-weather data/data?select=weather_description.csv [22].

The objective of the dataset is to elevate this small talk to medium talk. The weather is tremendous for demonstrating kinds of concepts because it contains periodic temporal structure with two different periods (daily and yearly).The dataset includes 5 years of high temporal resolution data of different weather attributes like temperature, humidity, air pressure, etc., The data is accessible for 6 Israeli cities, 30 US and Canadian Cities. Every attribute has own file. The rows are time axis and the columns are different cities. For every city, the country, latitude and longitude information are stored in a separate file. The dataset includes 226 column and 45254 instances in 7 files.

The other dataset is a Weather dataset is taken from the Kaggle https://www.kaggle.com/muthuj7/weather-dataset. The dataset contains 12 attributes and it has 96454 instances. The attributes are formatted date, summary, precip type, temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, land cover, pressure and daily summary. The characteristics of the attributes are decimals and strings.

The experimental is carried out with four different methods proposed SDI-WIXC technique and existing methods namely FCC [1], KM-I2C [2], SRSIO-FCM [3].

5. Results and Discussion

The experimental results of the SDI-WIXC technique and existing methods FCC [1], KM‑I2C [2] and SRSIO-FCM [3] are described in this section. The result analysis is performed with different parameters such as clustering accuracy, clustering time and space complexity based on the number of data points in the given big dataset. These parameters are evaluated and show the improvement of proposed SDI-WIXC technique.

Performance analysis of clustering accuracy: Clustering accuracy is measured as the number of similar data points in big dataset is correctly grouped into a particular cluster. This in turn narrows big data to the most relevant information and therefore helpful in analyzing business decisions. The clustering accuracy is calculated using following mathematical equation,

$$CA = \frac{X}{n} * 100 \tag{8}$$

From (8), *CA* represents a clustering accuracy and '*n*' denotes a total number of data points whereas '*X*' denotes a number of similar data points correctly grouped. The clustering accuracy is measured in terms of percentage (%).

Sample calculation using historical hourly weather data 2012-2017

- SDI-WIXC technique: Number of similar data points correctly grouped is 4000, and the total number of data points is 4500. Then the clustering accuracy is calculated as follows,

$$CA = \frac{4000}{4500} * 100 = 89\%$$

- FCC: Number of similar data points correctly grouped is 3600, and the total number of data point 4500. Then the clustering accuracy is,

$$CA = \frac{3600}{4500} * 100 = 80\%$$

- KM-I2C: Number of similar data points correctly grouped is 3400, and the total number of data point 4500. Then the clustering accuracy is,

$$CA = \frac{3400}{4500} * 100 = 76\%$$

- SRSIO-FCM: Number of similar data points correctly grouped is 3100, and the total number of data point 4500. Then the clustering accuracy is,

$$CA = \frac{3100}{4500} * 100 = 69\%$$

Sample calculation using Weather dataset

- SDI-WIXC technique: Number of similar data points correctly grouped is 42, and the total number of data points is 9500. Then the clustering accuracy is calculated as follows,

$$CA = \frac{8500}{9500} * 100 = 89\%$$

- FCC: Number of similar data points correctly grouped is 8500, and the total number of data point 9500. Then the clustering accuracy is,

$$CA = \frac{8000}{9500} * 100 = 84\%$$

- KM-I2C: Number of similar data points correctly grouped is 8000, and the total number of data point 9500. Then the clustering accuracy is,

$$CA = \frac{7600}{9500} * 100 = 80\%$$

- SRSIO-FCM: Number of similar data points correctly grouped is 31, and the total number of data point 9500. Then the clustering accuracy is,

$$CA = \frac{7200}{9500} * 100 = 76\%$$

The performances of clustering accuracy with two different datasets are described and the experimental results are shown in Table 2.

Table 1. Tabulation for clustering accuracy using historical hourly weather data 2012-2017.

Number of data points	Clustering accuracy (%)			
	SDI-WIXC	FCC	KM-I2C	SRSIO-FCM
4500	89	80	76	69
9000	91	82	78	71
13500	92	85	80	73
18000	94	87	83	75
22500	92	84	81	72
27000	91	82	79	70
31500	89	80	77	68
36000	90	83	79	70
40500	92	86	81	73
45000	93	88	83	76

Tables 1 and 2 describes the performance results of clustering accuracy with two different big weather datasets. Table 1 shows the clustering accuracy using historical hourly weather data 2012-2017 with a number of data points taken from 4500 to 45000. Table 2 shows the clustering accuracy using weather dataset. The number of data points taken from the weather dataset is varied from 9500 to 95000 that include the hourly information for different time periods. The clustering accuracy of four different methods namely SDI-WIXC technique and existing methods FCC [1], KM-I2C [2] and SRSIO-FCM [3] are clearly described in Tables 1 and 2. The table values clearly show that the clustering accuracy of SDI-WIXC technique is improved using two datasets. By using historical hourly weather data 2012-2017, the proposed SDI-WIXC technique groups weather data points for further processing. The SDI-WIXC technique uses the weighted iterative 'x' means clustering technique to group weather data points based on the similarity measure. By applying clustering technique, 'X' number of clusters and their weight values are initialized randomly. Then clustering technique measures the similarity between distributed data points or meteorological data and weight value of cluster for the two datasets. The similarity is measured using Sørensen dice indexing technique. Then the Sørensen dice similarity coefficient provides positive and negative similarity. The positive similarity between data points and their values are more appropriate to particular cluster weight. Then the data points to become a member of the particular cluster. If any data point does not group into a cluster, SDI-WIXC technique uses Bayesian probability criterion. It helps to maximize the probability of data points to group the particular cluster. This helps to improve the clustering accuracy using Historical Hourly Weather Data 2012-2017. The clustering accuracy using SDI-WIXC technique is increased by 9%, 15% and 27% when compared to existing FCC [1], KM-I2C [2] and SRSIO-FCM [3] respectively.

Table 2. Tabulation for clustering accuracy using weather dataset.

Number of data points	Clustering accuracy (%)			
	SDI-WIXC	FCC	KM-I2C	SRSIO-FCM
9500	89	84	80	76
19000	92	86	82	78
28500	94	87	85	81`
38000	93	85	83	79
47500	91	82	80	76
57000	89	80	78	74
66500	91	82	79	77
76000	93	84	81	79
85500	92	83	80	78
95000	94	86	82	79

Let us consider the weather dataset. By using this dataset, meteorological data points are collected and grouped into a different cluster as temperature data, pressure data, snow data, and rainy data and so on. The positive similarity of the above said data points and predefined cluster weights are used to identify the cluster members. Based on positive similarity, numbers of temperature data points, pressure data points, snow data points, rainy data points, wind speed

data points, wind direction data points are grouped into that particular cluster. As a result, the comparison results show that the proposed SDI-WIXC technique improves the clustering accuracy by 9%, 13% and 18% than the existing methods FCC [1], KM-I2C [2] and SRSIO-FCM [3]. From the above-said discussion, the proposed SDI-WIXC technique improves clustering accuracy using two datasets.

Performance analysis of clustering time: Clustering time is measured as the amount of time taken to group the data points into different clusters based on the similarity. The clustering time is measured using following equation,

$$CT = n * t(gropuingthedatapo\,\text{int}\, s) \quad (9)$$

From (9), CT denotes clustering time and 'n' denotes total number of data points. It is measured in terms of milliseconds (ms).

Sample calculation using Historical Hourly Weather Data 2012-2017

- SDI-WIXC technique: Number of data points is 4500, and time for grouping one data point is 0.002msThen the clustering time is calculated as follows,

$$CT = 4500 * 0.002 = 9ms$$

- FCC: Total number of data points is 4500 and time for grouping one data point is 0.0023ms. Then the clustering time is,

$$CT = 4500 * 0.0023 = 10.35ms$$

- KM-I2C: Total number of data points is 4500 and time for grouping one data point is 0.0025ms. Then the clustering time is,

$$CT = 4500 * 0.0025 = 11.25ms$$

- SRSIO-FCM: Total number of data points is 4500 and time for grouping one data point is 0.0032ms. Then the clustering time is,

$$CT = 4500 * 0.0032 = 14.4ms$$

Sample calculation using Weather dataset

- SDI-WIXC technique: Number of data points is 9500, and time for grouping one data point is 0.003ms. Then the clustering time is calculated as follows,

$$CT = 9500 * 0.003 = 28.5ms$$

- FCC: Number of data points is 9500, and time for grouping one data point is 0.0034ms. Then the clustering time is calculated as follows,

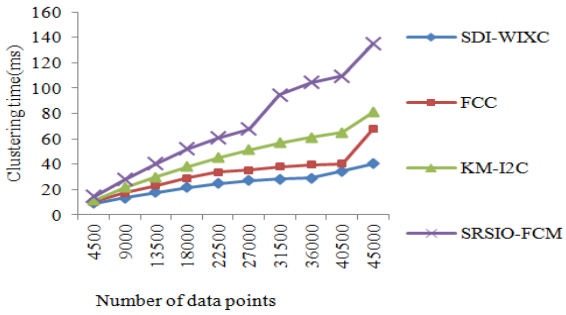
$$CT = 9500 * 0.0034 = 32.3ms$$

- KM-I2C: Number of data points is 9500, and time for grouping one data point is 0.0036ms. Then the clustering time is calculated as follows,

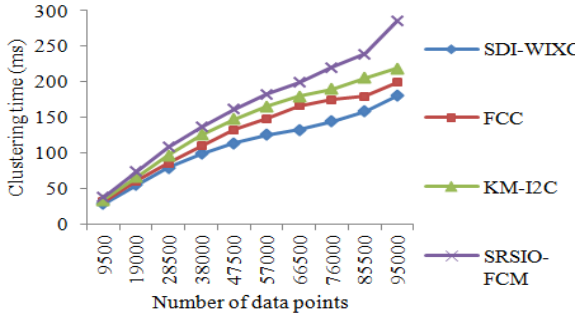
$$CT = 9500 * 0.0036 = 34.2ms$$

- SRSIO-FCM: Number of data points is 9500, and time for grouping one data point is 0.004ms. Then the clustering time is calculated as follows,

$$CT = 9500 * 0.004 = 38ms$$



a) Historical hourly weather data 2012-2017.



b) Weather dataset.

Figure 10. Clustering time vs. number of data points obtained.

Figure 10-a) and 10-b) shows the performance results of clustering time using Historical Hourly Weather Data 2012-2017 and weather dataset. The number of data points taken as an input for measuring the clustering time. The above figures clearly illustrate that the proposed SDI-WIXC technique minimizes the clustering time than the existing methods. The historical hourly weather data 2012-2017 is used for experimental evaluation. The different weather data points are collected and grouped into the particular cluster. The conventional technique groups the data points based on the distance measure. It takes more time to group the data points. On the contrary, the SDI-WIXC technique takes minimum time to group the data points through the similarity measure. With this, resultant data weather conditions are said to be predicted in an efficient manner. When the weight value of cluster and data point is similar, the Sørensen-Dice similarity coefficient provides the positive similarity i.e. Otherwise it provides the negative similarity. As a result, this similarity measure improves the clustering accuracy and minimizes the clustering time. With this positive and negative similarity, relationship between the weather conditions in different places and also fluctuations between zonal and meridional winds is said to be measured effectively. The comparisons of proposed and existing methods show that the SDI-WIXC technique

minimizes the clustering time by 24%, 43% and 60% than the existing FCC [1], KM-I2C [2] and SRSIO-FCM [3] respectively.

By applying the Weather dataset, the performance of clustering time is obtained. The meteorological data points taken from this datasets are temperature, pressure, wind speed data, snow data, and rainy data. "Big Data is a number of both structured and unstructured data points that is so large that it's difficult to process and it takes more time. While accessing big weather data, the time-efficiency of clustering is necessary to improve the clustering quality. Hence the groupings of the weather data points are carried out for further processing according to dew point, temperature and pressure separately. An SDI-WIXC technique processes the large weather dataset by partitioning into different groups where each group has different data points. Let us consider the temperature data is grouped into the particular cluster. It shows that the temperature data points and the cluster weight are more similar. Similarly, other data points such as pressure, wind, rainy and snow data are grouped into a particular cluster. This process takes minimum clustering time. The clustering time of SDI-WIXC is significantly reduced by 13%, 21% and 30% when compared to existing FCC [1], KM-I2C [2] and SRSIO-FCM [3] respectively.

5.1. Performance Analysis of Space Complexity

Space complexity is defined as the amount of memory space required to store the data points. It is measured as follows,

$$SC = n * memory(Storingonedatapoints) \quad (10)$$

From (10), *n* denotes a number of data points, *SC* denotes a space complexity. The space complexity is measured in terms of Mega Bytes (MB).

Sample calculation using historical hourly weather data 2012-2017

- SDI-WIXC technique: Number of data points is 4500, and space for storing one data point is 0.0021MB, then the space complexity is calculated as follows,

$$SC = 4500 * 0.0021 = 9.45MB$$

- FCC: Total number of data points is 4500 and space for storing one data point is 0.0026MB, then the space complexity is,

$$SC = 4500 * 0.0026 = 11.7MB$$

- KM-I2C: Total number of data points is 4500 and space for storing one data point is 0.003MB, then the space complexity is,

$$SC = 4500 * 0.003 = 13.5MB$$

- SRSIO-FCM: Total number of data points is 4500 and space for storing one data point is 0.0038MB, then the space complexity is,

$$SC = 4500 * 0.0038 = 17.1MB$$

Sample calculation using Weather dataset

- SDI-WIXC: Total number of data points is 9500, and space for storing one data point is 0.0026. Then the space complexity is calculated as follows,

$$SC = 9500 * 0.0026 = 24.7MB$$

- FCC: Total number of data points is 9500, and space for storing one data point is 0.003. Then the space complexity is calculated as follows,

$$SC = 9500 * 0.003 = 28.5MB$$

- KM-I2C: Total number of data points is 9500, and space for storing one data point is 0.0034. Then the space complexity is calculated as follows,

$$SC = 9500 * 0.0034 = 32.3MB$$

- SRSIO-FCM: Total number of data points is 9500, and space for storing one data point is 0.0039. Then the space complexity is calculated as follows,

$$SC = 9500 * 0.0039 = 37.05MB$$

Tables 3 and 4 shows the space complexity using two different datasets, namely historical hourly weather data 2012-2017 and weather dataset. The experimental results of four different methods namely SDI-WIXC technique and existing methods FCC [1], KM-I2C [2] and SRSIO-FCM [3] are clearly described in table 3 and table 4. The above table clearly shows that the space complexity using SDI-WIXC technique is minimized than the other existing methods. The SDI-WIXC technique efficiently groups the more similar data points into one cluster. The historical hourly weather data 2012-2017 is used for big data analytics. In order to handle a large number of weather data, the clustering technique is applied for dimensionality reduction. The total data is divided into dissimilar groups based on the Sørensen-Dice similarity coefficient results. As a result, SDI-WIXC technique minimizes the space complexity by grouping the similar weather data points. After performing comparison, the space complexity of SDI-WIXC technique is reduced by 19%, 36% and 49% when compared to existing FCC [1], KM-I2C [2] and SRSIO-FCM [3] respectively.

Table 3. Tabulation for Space complexity using historical hourly weather data 2012-2017.

Number of data points	Space complexity (MB)			
	SDI-WIXC	FCC	KM‑I2C	SRSIO-FCM
4500	9.45	11.7	13.5	17.1
9000	18	21.6	26.1	32.4
13500	25.65	29.7	36.45	47.25
18000	32.4	37.8	46.8	57.6
22500	36	42.75	56.25	67.5
27000	40.5	48.6	62.1	78.3
31500	40.95	50.4	69.3	88.2
36000	43.2	54	72	97.2
40500	44.55	56.7	72.9	101.25
45000	45	67.5	81	90

Table 4. Tabulation for Space complexity using weather dataset.

Number of data points	Space complexity (MB)			
	SDI-WIXC	FCC	KM‑I2C	SRSIO-FCM
9500	24.7	28.5	32.3	37.05
19000	47.5	60.8	66.5	74.1
28500	68.4	85.5	96.9	108.3
38000	87.4	110.2	125.4	136.8
47500	104.5	133	147.25	161.5
57000	114	148.2	165.3	182.4
66500	119.7	166.25	179.55	199.5
76000	129.2	174.8	190	220.4
85500	149.625	179.55	205.2	239.4
95000	180.5	199.5	218.5	285

By applying weather dataset for big data analytics, the similarity based weighted iterative clustering technique is used to group the similar data points. This process takes minimum storage space for storing the number of data points. Totally 10 different runs are performed. For each iteration, the number of input data points is varied. After performing the ten runs, the comparison between the proposed and three existing methods are carried out. The comparison results show that the space complexity of SDI-WIXC technique is considerably minimized by 20%, 28% and 37% than the existing FCC [1], KM‑I2C [2] and SRSIO-FCM [3] respectively. From the above results discussion, the similar data points are correctly grouped into the cluster with less time and space complexity.

6. Conclusions

An efficient technique called SDI-WIXC technique is applied for grouping the similar data points with higher clustering accuracy and minimal time. The number of data points is collected from the big dataset. By applying weighted iterative X-means clustering, the number of clusters and their weight values are initialized randomly. Then the similarity of each data points and the cluster weight is measured using Sørensen-Dice Indexing. Based on the similarity measure, the more similar data point has a maximum probability to become a member of that particular cluster. This helps group the more similar data point in the cluster with minimum time. The separation of similar data points minimizes the space complexity. Therefore, proposed SDI-WIXC technique is efficiently overcomes the problems of clustering performance, space complexity and time complexity. Experimental evaluation of proposed SDI-WIXC technique and existing methods are carried out with two different big weather datasets namely Historical Hourly Weather Data 2012-2017and weather dataset. The performance results of proposed SDI-WIXC technique improves the clustering accuracy of weather data points and minimizes the clustering time as well as space complexity. Though the accuracy is improved, the computational cost is not reduced by proposed SDI-WIXC technique. In addition, the error rate is reduced further by introducing the ensemble clustering techniques.

References

[1] Bharill N., Tiwari A., and Malviya A., "Fuzzy-Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark," *IEEE Transactions on Big Data*, vol. 2, no. 4, pp .339-352, 2016.

[2] Bu F., Chen Z., Li p., Tang T., and Zhang Y., "A High-Order CFS Algorithm for Clustering Big Data," *Mobile Information Systems*, vol. 2016, pp.1-8, 2016.

[3] Cui X., Zhu P., Yang X., Li K., and Ji C., "Optimized Big Data K-Means Clustering using Mapreduce," *The Journal of Supercomputing*, vol. 7, no. 3, pp. 1249-1259, 2014.

[4] Esteves R., Hacker T., and Rong C., "A New Approach for Accurate Distributed Cluster Analysis for Big Data: Competitive K-Means," *International Journal of Big Data Intelligence*, vol. 1, no. 1-2, pp. 50-64, 2014.

[5] Genuer R., Poggi J., Tuleau-Malot C., and Villa-Vialaneix N., "Random Forests for Big Data," *Big Data Research*, vol. 9, pp. 28-46, 2017.

[6] Hu R., Dou W., and Liu J., "ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 302-313, 2014.

[7] Jain M. and Verma C., "Adapting K-Means for Clustering in Big Data," *International Journal of Computer Applications*, vol. 101, no. 1, pp. 19-24, 2014.

[8] Kuru K. and Khan W., "Novel Hybrid Object-Based Non-Parametric Clustering Approach for Grouping Similar Objects in Specific Visual Domains," *Applied Soft Computing*, vol. 62, pp. 667-701, 2018.

[9] Li Z., Hu F., Schnase J., Duffy D., Lee T., Bowen M., and Yang S., "A Spatiotemporal Indexing Approach for Efficient Processing of Big Array-

Based Climate Data With Mapreduce," *International Journal of Geographical Information Science*, vol. 31, no. 1, pp. 1-19, 2016.

[10] Liu H., Wu J., Liu T., Tao D., Fu Y., "Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 1129-1143, 2017.

[11] Liu W., Ye M., Wei J., and Hu X., "Fast Constrained Spectral Clustering and Cluster Ensemble with Random Projection," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1-14, 2017.

[12] Maghawry A., Omar Y., and Badr A., "Self-Organizing Map vs Initial Centroid Selection Optimization to Enhance K-Means with Genetic Algorithm to Cluster Transcribed Broadcast News Documents," *The International Arab Journal of Information Technology*, vol. 17, no. 3, pp. 316-324, 2020.

[13] McParland D. and Gormley I., "Model Based Clustering for Mixed Data: Clustmdm," *Advances in Data Analysis and Classification*, vol. 10, no. 2, pp. 155-169, 2016.

[14] Rehioui H., Idrissi A., Abourezq M., and Zegrari F., "DENCLUE-IM: A New Approach for Big Data Clustering," *Procedia Computer Science*, vol. 83, pp. 560-567, 2016.

[15] Rikhtehgaran R. and Kazemi I., "The Determination of Uncertainty Levels In Robust Clustering of Subjects with Longitudinal Observations Using the Dirichlet Process Mixture," *Advances in Data Analysis and Classification*, vol. 10, no. 4, pp. 541-562, 2016.

[16] Scrucca L. and Raftery A., "Improved Initialisation of Model-Based Clustering Using Gaussian Hierarchical Partitions," *Advances in Data Analysis and Classification*, vol. 9, no. 4, pp. 447-460, 2015.

[17] Shao W., Salim F., Song A., and Bouguettaya A., "Clustering Big Spatiotemporal-Interval Data," *IEEE Transactions on Big Data*, vol. 2, no. 3, pp. 90-203, 2016.

[18] Sreedhar C., Kasiviswanath N., Reddy P., "Clustering Large Datasets Using K-Means Modified Inter and Intra Clustering (KM-I2C) in Hadoop," *Journal of Big Data*, vol. 4, no. 27, pp. 1-19, 2017.

[19] Tortora C., Summa M., Marino M., Palumbo F., "Factor Probabilistic Distance Clustering (FPDC): A New Clustering Method," *Advances in Data Analysis and Classification*, vol. 10, no. 4, pp. 441-464, 2016.

[20] Traganitis P., Slavakis K., and Giannakis G., "Sketch and Validate for Big Data Clustering," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 678 -690, 2015.

[21] Tsapanos N., Tefas A., Nikolaidis N., Iosifidis A, and Pitas I., "Fast Kernel Matrix Computation for Big Data Clustering," *Procedia Computer Science*, vol. 51, pp. 2445-2452, 2015.

[22] Wikipedia, Bayesian Information Criterion, https://en.wikipedia.org/wiki/Bayesian_information_criterion#:~:text=In%20statistics%2C%20the%20Bayesian%20information,the%20lowest%20BIC%20is%20preferred, Last Visited, 2021.

[23] Wu J., Wu Z., Cao J., Liu H., Chen G., and Zhang Y., "Fuzzy Consensus Clustering With Applications on Big Data," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1430-1445. 2017.

[24] Ye M., Liu W., Wei J., and Hu X., "Fuzzy *c*-Means and Cluster Ensemble with Random Projection for Big Data Clustering," *Mathematical Problems in Engineering*, vol. 2016, pp. 1-13, 2016.

[25] Yu Y., Zhao J., Wang X., Wang Q., and Zhang Y., "An Efficient Distributed Density-Based Clustering for Big Data Using Hadoop," *International Journal of Distributed Sensor Networks*, vol. 2015, pp. 1-13, 2015.

[26] Zhang Q. and Chen Z., "A Weighted Kernel Possibilistic C-Means Algorithm Based on Cloud Computing for Clustering Big Data," *International Journal of Communication Systems*, vol. 27, no. 9, pp. 1378-1391, 2014.

KalyanaSaravanan Annathurai is currently working as an Assistant Professor in the School of computer technology and applications at Kongu Engineering College PerunduraiTamilnadu India. He obtained Msc (Applied Science-Computer Technology) from Barathiyar University and M.E. (Software Engineering) at Anna University (University Campus). Currently pursuing his research in Anna University. He has more than 13years of teaching experience. Also he is responsible for Industry institute participation cell in the department consultancy activities. Established Internet of Things Lab and High Performance Computing Lab (Parallel Programming) in the department for the regular practice and research Also centre for excellence in IoT in collaboration with C-DAC Bangalore at KEC .He has organized several workshops on, Big Data Analytics, Web Technology, Cloud security, R-Toolsin his Institution. Presented several papers at National and International conferences in india and abroad. He has guided a number of research-oriented as well as application oriented projects.

Tamilarasi Angamuthu currently working as a Professor and Head in Department of Computer Applications in Kongu Engineering College. She has more than 25 years of experience in teaching and research in mathematics and computer science. She guided more than twenty scholars and publications in various reputed journals. Also she was a doctoral committee member/Thesis evaluator for the various universities in India and abroad.