VOL. 10, NO. 5, MARCH 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

CLUSTERING WITH SHARED NEAREST NEIGHBOR-UNSCENTED TRANSFORM BASED ESTIMATION

M. Ravichandran[1] and A. Shanmugam[2]
[1]Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamil Nadu, India
[2]Department of Electronics and Communication Engineering, S.N.S College of Technology, Coimbatore, Tamil Nadu, India
E-Mail: mraveen2007@gmail.com

ABSTRACT

Subspace clustering developed from the group of cluster objects in all subspaces of a dataset. When clustering high dimensional objects, the accuracy and efficiency of traditional clustering algorithms are very poor, because data objects may belong to diverse clusters in different subspaces comprised of different combinations of dimensions. To overcome the above issue, we are going to implement a new technique termed Opportunistic Subspace and Estimated Clustering (OSEC) model on high Dimensional Data to improve the accuracy in the search retrieval.Still to improve the quality of clustering hubness is a mechanism related to vector-space data deliberated by the propensity of certain data points also referred to as the hubs with a miniature distance to numerous added data points in high dimensional spaces which is associated to the phenomenon of distance concentration. The performance of hubness on high dimensional data has an incapable impact on many machine learning tasks namely classification, nearest neighbor, outlier detection and clustering. Hubness is a newly unexplored problem of machine learning in high dimensional data spaces, which fails in automatically determining the number of clusters in the data. Subspace clustering discovers the efficient cluster validation but problem of hubness is not discussed effectively. To overcome clustering based hubness problem with sub spacing, high dimensionality of data employs the nearest neighbor machine learning methods. Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method is developed to overcome the hubness problem with determination of cluster data. The core objective of SNNC is to find the number of cluster points such that the points within a cluster are more similar to each other than to other points in a different cluster. SNNC-UT estimates the relative density, i.e., probability density, in a nearest region and obtains a more robust definition of density. SNNC-UT handle overlapping situations based on the unscented transform and calculate the statistical distance of a random variable which undergoes a nonlinear transformation. The experimental performance of SNNC-UT and k-nearest neighbor hubness in clustering is evaluated in terms of clustering quality, distance measurement ratio, clustering time, and energy consumption.

Keywords: unscented transform, shared nearest neighbor clustering, high dimensional data, cluster validation, hubness, distance measure.

1. INTRODUCTION

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. Cluster is used to group items that seem to fall naturally together. Various types of clustering: hierarchical (nested) versus partitioned (un-nested), exclusive versus overlapping versus opportunistic, and complete versus partial. Clustering is an unverified learning process that partitions data such that similar data items grouped together in sets referred to as clusters. This activity is important for condensing and identifying patterns in data.

Clustering is supposed as an unverified process in which the process of organizing objects into groups whose members are similar in some way. The authority of clustering results has to be evaluated by discovery the optimal number of clusters that best fits the given data set. Clustering objects in high dimensional spaces may detain to clustering the objects in subspaces which may be of diverse dimensions. In a number of recent periodicals the hubness event has been described and surveyed as a general problem of machine learning in high dimensional data spaces. Hubs are data points which occur frequently in nearest neighbor lists of numerous additional data points. The effect is mainly difficult in algorithms for likeness search, as the same similar objects are found over and over again. Data mining is the process of extracting the data from huge high dimensional databases, used as technology to produce the required information.

The tendency of high-dimensional data enclose points hubs as shown in [1] that frequently occur in k-nearest neighbor lists of other points. Hubness successfully subjugated in clustering within a high-dimensional data cluster. Hub objects have minute distance to a remarkably large number of data points, and anti-hubs are far from all other data points. It is connected to the attention of distances as shown in [20] which impairs the contrast of distances in high dimensional spaces. An association to the hubness phenomenon at the instance of their revision was not extremely well known.

Hub songs express less perceptual connection to the songs which are very closed according to an audio likeness function, than non-hub songs in [15]. Recognizing groups of genes that manifest comparable expression patterns in such enormous quantity of information is critical in the examination of gene expression time series. The hubness of oceanographic data as shown in [17] used to visualize and detect both prototypical sensors/locations, as well as uncertain and potentially mistaken ones. In [9] work, present an integrated examination of microarray

information using relationship mining and clustering that conclude intrinsic grouping based on co-occurrence patterns in such data.

Data mining methods are used to expect future data trends, approximate its scope, and used as a reliable information in the decision making process. Some of the functions related to data mining include association, relationship, prediction, clustering, classification, analysis, trends, outliers and divergence investigation, and similarity and dissimilarity analysis. The origins of experience showing it is an intrinsic property of data distributions in high-dimensional vector space highly related to dimensionality reduction. Nearest neighbors in high dimensional data as demonstrated in [6] discover its influence on wide range of machine-learning tasks openly or ultimately based on measuring distances.

Direct application of sparse coding guide to different information by integrating distribution distance estimates for the embedded data. The algorithm prevails over the shortcomings of the sparse coding algorithm in [11] on synthetic data and achieves enhanced analytical performance on real world chemical toxicity transfer learning task. Sparse coding fails to provide a high-quality starting point for addressing the complex task of knowledge transfer from numerous heterogeneous data sources.

Several feature based and class based measures are used to analyze the statistical characteristics of the training partitions. To assess the effectiveness of dissimilar types of training partitions a huge number of disjoint training partitions with distinctive distributions was generated. Then, these training partitions were empirically estimated in [10] and their collision was evolved using the performance of the system by utilizing the feature-based and class-based measures. Mounting mechanisms are more sophisticated methods and events for training data subsets with overlaps would fail to unite the filter based data partitioning approach using a wrapper based method.

Multi-Cluster Feature Selection (MCFS), for unsupervised feature selection in [16] select those features such that the multi-cluster structure of the data is best conserved. The corresponding optimization problem professionally resolve as it only involves a sparse eigen-problem and a L1-regularized least squares in an efficient manner.

One of regularly used data mining method to discover patterns or groupings of data is clustering. Clustering is the separation of data into objects that have similarities. Showing the data into smaller clusters to make the data becomes much simpler, however, due to the loss of imperative piece of data, the cluster has to be investigated and evaluated. Using multiple viewpoints, more revealing estimation of similarity fails in defining alternative forms for the relative similarity that combine the relative similarities according to the different viewpoints in [12].

Cluster analysis systematizes analogous points in the data into groups called clusters. Popular clustering algorithms use the inactive environment of data to discover an optimal solution. When changes are made to the dataset, these algorithms have to be run on the complete dataset to update feasible changes in clustering, involving important unnecessary computations. Cluster specially undertake problem using the restricted redundant computations and distribute a clustering identical to the original clustering. The speed-up gain will have tremendous collision in scenarios where changes to dataset are rather frequent.

An important improvement in text clustering tasks in [2] is based on a detailed similarity measurement and on a generic seeds construction strategy and is broadly functional to other clustering difficulty domains. The individual impact of the similarity metric is further effectual in text clustering. The uncertainty information is important for not only the identification of the assignment of data points as illustrated in [19], but also that of the suitable projections transversely in which the data is clustered.

Graph Clustering is the assignment of grouping vertices of a graph into clusters so that the vertices in the identical clusters are similar. It's different from the clustering of sets of graphs based on structural similarity. Graph Clustering are separated into two groups based on the measures recognize the clusters and computing the reserve predefined among vertices. The algorithms consider connectivity and shared neighbors discretely. Certain algorithms also consider the attributes of vertices. But in most conditions, the original source is only a graph and vertices have no additional attributes value.

An implementation of approximate kNN-based spatial clustering algorithm using the K-d tree as shown [5] is achieved using spatial clustering, and compares its presentation to the brute-force approach. Brute-force search is a very universal problem-solving technique that consists of methodically enumerating all probable candidates for the explanation and checking whether each candidate satisfy the problem's statement. The bottom-up method searches for density-based mutual subspace clusters in [14] to systematically go after a low-dimensional subspace to high-dimensional ones.

Non-convex matrix decomposition problem in closed form of [18] involves a novel polynomial threshold operator on the remarkable values of the data matrix, which requires minimum reduction. The divergence problem experience varies in diverse subspace cardinalities as shown in [7] without considering the problem. The previous works operated with a density threshold to recognize the dense regions in all subspaces, which resulted in the failure of clustering accuracy level in varied subspace cardinalities.

Linear Discriminant Analysis (LDA) within a co-training scheme in [13] developed labels cultured robotically in one view to discover discriminative subspaces in another. The effectiveness was demonstrated empirically under scenarios where the conditional autonomy assumption is in addition fully fulfilled or only partially satisfied. Optimization model that well describe

the optimization process using a new clustering algorithm FG-k-means optimizes the optimization model. The algorithm in [8] is an addition to k-means that adds two extra steps to routinely compute the two types of subspace weights.

K-means type algorithm which is linear with respect to the number of the data points in [4] is also useful in describing the cluster formation in terms of attributes contribution to dissimilar clusters. These clusters with the attribute weights have better understanding of cluster formation. To understand the relationship between user's defined value and number of desired clusters, it fails in achieving enhanced results by optimizing these values. Research in subspace clustering method has a bunch of prospective to be developed in [3]. Densities of each object neighbors with Min Points occur in accordance with differed density level of every entity neighbors. More in-depth study fails to narrate preprocessing, dimension reduction, and outlier detection of subspace clustering method.

In proposed work, SNNC-UT using bike sharing dataset addresses the class discovery problem. Typically in clustering, the objective is to find the number of clusters points such that the points within a cluster are more similar to each other than to other points in a different cluster. Using the bike sharing data analysis, the hourly and daily counts of rental bikes are present that exhibiting similar patterns of expression are clustered together. Several statistical measures have been developed to calculate the similarity between bike share system with the corresponding weather and seasonal information.

Using SNNC-UT algorithm, correlation coefficient calculate the similarity of hourly and daily count of rental bikes across different time series with the obtained statistic captures similarity in shape of the expression profile. A correct and competent distance computation is based on the unscented transform estimation which is regularly used to obtain an enhanced result with the distance being evaluated using the bike sharing data-set. The experimental results indicate that SNNC-UT estimation method outperforms previously suggested k-nearest neighbor hubness using clustering method. The contributions of Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method include the following:

1. To overcome the hubness problem with determination of cluster data.
2. To find the number of cluster points such that the points within a cluster are more similar to each other than to other points in a different cluster.
3. To handle overlapping situations based on the unscented transform and calculate the statistical distance of a random variable which undergoes a nonlinear transformation.

The rest of the paper is structured as follows. Section 2 describes about the diverse form of existing work with their limitations. Section 3 describes about the Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method to automatically determine the cluster points. Section 4 presents the effective results on the simulation parameter to compute the statistical distance. Section 5 evaluated the performance with the table values and graph form. The final section summarizes a valuable solution with SNNC-UT estimation method for providing number of cluster points automatically with nearest distance measure.

2. PROPOSED OPPORTUNISTIC SUBSPACE AND ESTIMATED CLUSTERING ON HIGH DIMENSIONAL DATA MODEL

The proposed work is efficiently designed for estimating the clusters in high dimensional by adapting the Opportunistic Subspace and Estimated Clustering (OSEC) model.

The architecture diagram of the proposed Opportunistic Subspace and Estimated Clustering (OSEC) model is shown in Figure-1. The proposed opportunistic subspace clustering is processed under different input, intermediate and output processes. The input process takes the high dimensional data. We select the Difference Subspace Clustering algorithm as the basic clustering algorithm for initialization, which take over the advantages of opportunistic type clustering algorithms such as easiness of calculation, effortlessness, and can covenant with noise and overlap clusters. Our proposed system consists of two modules namely:

- ✓ Clustering module
- ✓ Attack Detection module

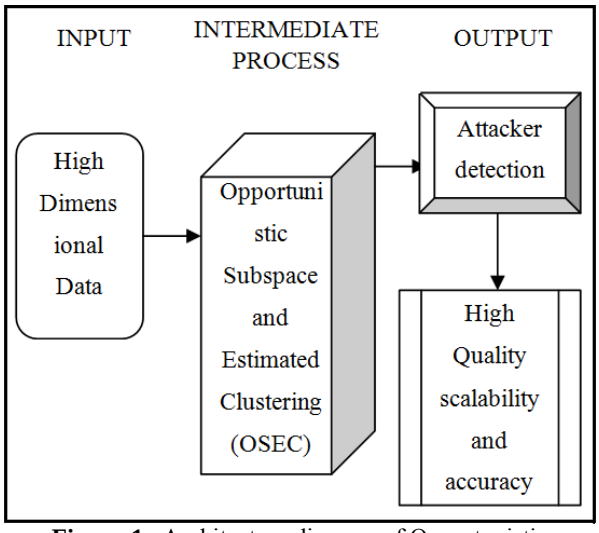

Figure-1. Architecture diagram of Opportunistic Subspace and Estimated Clustering (OSEC) model.

The clustering module is supplementary separated into two sub modules, which are

- Initialization using Difference Subspace clustering
- Clustering using opportunistic logic

Difference subspace clustering

Subspace clustering try to find clusters in different subspaces within a real dataset. This means that a data summit might fit in to multiple clusters, each

accessible in a different subspace. Subspace algorithm determines the each cluster center and then determines all their centroid points. Frequently in high dimensional data, several dimensions may be inappropriate and can cover accessible clusters in noisy data. Subspace clustering algorithms usually restrict the investigation for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces.

Steps to perform the difference subspace clustering

Step 1:	Highest point Data object selected as initial cluster center
Step 2:	Neighborhood data objects are removed from the initial cluster center
Step 3:	Goto Step 1 and Step 2 until the data points are within the radii of cluster

Algorithm 1: Initialization using Difference Subspace clustering:

Input: Dataset Y= {y1, y2, …, yn} ∈ Real $_d$
(Initialize the centre E (0) and set E $_{max}$ =k)

Step-1: For each $y_i \in Y$, compute the mass index

$$\mathbf{F_i} = \sum_{j=1}^{n} \exp o[-\| yi-yj \|^2 /(.5q_a)^2$$

Step-2: Let F_{c1}=max {Fi,i=1,2,…,n},then select y_{c1} as the initial cluster center;
Step-3: Repeat the step if y_{ck} be the k_{th} cluster center, and the mass index be F_{ck}.
Step-4: For each yi ∈ Y, update the mass index,

$$\mathbf{F_i = F_i - F\,ck} \sum_{j=1}^{n} \exp o[-\| yi-yck \|^2 /(5q_b)^2$$

Step-5: For each yi ∈ Y, until F_{ck} +1 / Fc1 < δ, where qa, qb and δ need pre assignment.

$$q_a = q_b = 1/2^{\min_k \{ {}^{\max}_i \{ \| y1 - yk \| \} \}}$$

In general, each phase tests the number of clusters 'c' between C_{min} and C_{max}. The cluster centers should be initialized at the establishment of running the difference algorithm. In difference clustering algorithm, the production order of the cluster centers is determined by the mass index. The superior of mass index is that the earlier of cluster center generated. Thus, at each step the top 'c' cluster centers can be selected as the new initialization cluster centers, and there is no need to re-initialize the cluster centers. After initialization of the centroid, the next step is to concern the opportunistic algorithm to get hold of the association degree for every data point with deference to each cluster.

3. SHARED NEAREST NEIGHBOR CLUSTERING BASED ON UNSCENTED TRANSFORM ESTIMATION

Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method is developed to find the number of cluster points within a cluster. SNNC-UT estimates the relative density of the nearest region and obtains a more robust definition of density. Unscented Transform estimation calculates the statistical distance of a random variable which experiences a nonlinear transformation. The flow diagram of SNNC-UT estimation method is shown in Figure-2.

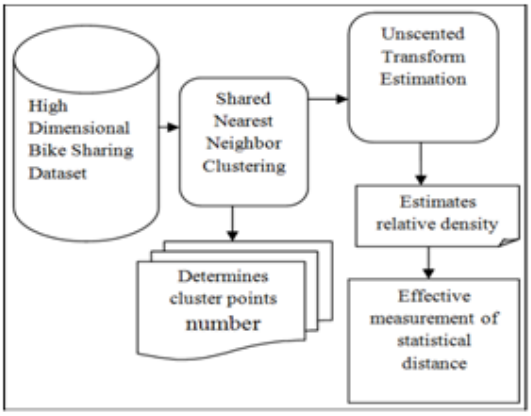

Figure-2. Flow diagram of SNNC-UT method.

Figure-2 describes the Shared Nearest Neighbor Clustering using the bike sharing high dimensional data. The strength of SNNC is based on the number of shared neighbors and obtains a more robust definition of density. Clustering based on nearest neighbor is relatively insensitive to the high dimensionality with capability to handle ratios of different density. In order to evaluate the system accuracy, different cluster points are measured using supervised clustering.

Unscented Transform (UT) estimation approximates a Gaussian density with a set of deterministically selected cluster points. The cluster points completely capture the mean and covariance of the distance distribution measure. With the nonlinearly transformed in UT estimation, the new cluster points entirely confine the mean and covariance of the new density forms. The unscented transform estimation method with SNNC calculates the information of an arbitrary variable which experience a nonlinear transformation.

3.1 The shared nearest neighbor clustering phenomenon

Shared Nearest Neighbor Clustering is an aspect that involves correlation coefficient which is one of the core information used to count the number of cluster points in bike sharing data analysis. A similarity matrix is constructed for the entire bike sharing data set based on the correlation coefficients across a time series. Let the

density of a bike share system ‘ be defined as the sum of the similarity of its neighbors and represented as equation (1).

$$density\ (b_i) = \sum_{j=1}^{n} count \quad (1)$$

Where, n is the number of neighbors, similarity threshold of cluster points ‘i’ and ‘j’. Using the shared nearest neighbor the neighbor processing of two most dense points **(b_i)** are measured to a cluster and assigned with the bike share system to the appropriate clusters. Additionally, the neighbors of two most dense points **(b_i and b_j)** are assigned to the same cluster if both the bike share neighbors are greater than a given shared nearest neighbor threshold. If **(b_i)** and **(b_j)** in the bike share system represents the highest density identified from equation (1), then the number of shared nearest neighbors is obtained from the following equation,

$$Shared\ nearest\ neighbor(b_i, b_j) = Size\ (NN\ (b_i) \cap N.$$

(2)

From the above Equation (2), and represent the nearest neighbor lists of **(b_i)** and **(b_j)** which are greater than or equal to given similarity threshold cluster points respectively. The use of shared nearest neighbor measure is justified by the fact that the presence of shared neighbors between two dense bike share means that the density around the dense bike share is similar. Hence, it should be included in the same cluster along with their neighbors. The neighbors of bike share system are identified by building a nearest neighbor mask for the dense bike share using vertical approach. The Figure-2 illustrates shared nearest neighbor graph where density is number of points per unit volume.

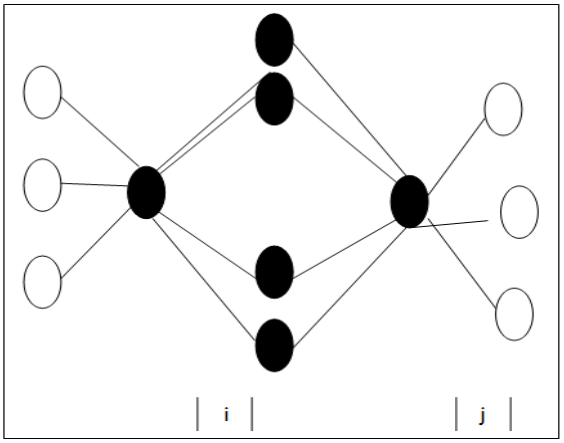

Figure-3. Shared nearest neighbor graph.

The Figure-3 represents the nearest neighbor mask of a bike share **(b_i)** with a bit pattern of 1’s and 0’s, with 1 representing with a bike share with its neighbor to **(b_j)** and 0 otherwise. Consider that the neighbors of bike share points are identified, then the shared nearest neighbors between them is nothing but the root count of “AND” operation. The nearest neighbor masks of the two points **(b_i)** and **(b_j)** are computed vertically using the following equation (3).

$$Average\ count\ of\ Shared\ nearest\ neighbor(b_i, b_j) = RootCountpoint: (NNp(b_i)\ ^\wedge\ NNp(b_j)$$

(3)

Where, $(NNp(b_i)\ ^\wedge N$ are the Nearest Neighbor masks of two points (and (respectively. Algorithm 1 shows the algorithmic steps of Shared Nearest Neighbor Clustering.

Begin
Step 1: **Compute** the connection matrix with data points
Step 2: **Find** the Shared Nearest Neighbor at density of each cluster points.
Step 3: **Form** Cluster based on density value.
Step 4: **Let** ‘M’ denote the similarity matrix that holds only the shared similar nearest neighbors.
Step 5: **Construct** the shared nearest neighbor graph from the connection matrix using eqn (2)
Step 6: **Find** nearest neighbor mask of two points using eqn (3).
End

Algorithm-2: Shared Nearest Neighbor Clustering.

In the case of SNNC algorithms, it identifies the nearest neighbors of the dense data points and check to see whether the neighbors of are present in the neighborhood of . Scanning of neighborhoods on the bike share system is required to obtain the nearest neighbor points.

3.2 Processing of unscented transform

Once the shared nearest neighbor has been identified, the processing of unscented transform starts. The unscented transform overlapping circumstances is handled for manipulative the information which undergoes a nonlinear transformation. In cases, the SNNC processing noise is in Gaussian form, it performs the better particle filter with Unscented Transform. The UT uses the initial order term of Taylor expansion for non-linear function. The UT uses the true nonlinear function and estimates the distribution of the function output to get hold of estimation and distance measure between the nearest neighbors.

The unscented transform estimation include the n-dimensional normal random points u ~ f (u) = S (α, Σ).

SNNC estimates the expectation of g (u) in such a way that

$$\int f(u)_g \quad (4)$$

Where, $f($ are cluster points for the position f and g. Using the unscented transform approach the following set of cluster points is chosen. These cluster points totally detain the true mean and variance of the normal distribution f (u). The uniform distribution over the cluster points has mean μ and covariance matrix Σ. Given the cluster points, let us define the following estimation as,

$$\int f(u)g(u)du \sim \frac{1}{2d}\sum_{i=}^{2d} \quad (5)$$

Although UT estimation uses only a small number of points and verified that if g (u) is a quadratic function then the estimation is precise. The basic UT estimation is generalized with mean of the Gaussian distribution μ and also included in the set of cluster points. Moreover, scaling parameters are included that provide an extra degree of freedom to control the scaling of the cluster points towards μ. The covariance matrices of the UT components of 'f' have the following form,

$$\sum_i = covartancs\,(\sigma_{i1}^2, \quad (6)$$

Where i=1…n. The structure of the covariance matrix of the components of 'f' and 'g' are utilized to decrease the complexity of computing on the cluster points. Algorithm 3 given below shows the detailed evaluation of shared nearest neighbor clustering based on unscented transform.

Procedure: Shared nearest Neighbor Clustering based on Unscented Transform

Input: All data points (i.e.,) bike sharing dataset

Output: Clustering of similar points with distance measure

Begin
Let and denotes two dense bike sharing data analysis
While (Unprocessed data>0) do
Most Dense Data←Find Dense Data Value (
Processed Data add←Most Dense Data
Get Nearest Neighbor (Most Dense Data, SNN threshold)
Processed Data add←Nearest Neighbor (Most Dense Data)
End If
End If
If non Neighbors = (Most Dense Data) Then
No is data add← Most Dense Data
Else If
$RootCountpoint\ (NNp(\ b_i)\hat{}N$ > Shared Nearest neighbor Threshold
then
Cluster Nearest Neighbor ()←
$Size\ (NN\ (b_i) \cup N$
Processed Data add←Nearest Neighbor (Most Dense Data)
End If
End If
End While
// Assign Boundary Result to clusters
For i=0 to boundary data Size ()
Neighbors (Boundary Size)
If NN (boundary size data) cluster
Cluster[i] ←Boundary data
Else
Cluster[i] ←Boundary data based on similarity
End If
End For
End

Algorithm-3. Shared nearest neighbor clustering based on unscented transform.

The algorithm for the shared nearest neighbor clustering algorithm is initiated by identifying the two most dense bike sharing **(b_i)** and **(b_j)** from the unprocessed data analysis. Initially, checking is performed on the neighbors. The neighbors of the dense bike sharing are obtained using the neighbor point which is a basic bit pattern that has a 1-bit if a neighbor is present and 0-bit if not. In SNNC-UT estimation method, similarity matrix is built and correlation coefficient of bike sharing is identified. The two bike sharing system with highest density is evaluated using SNNC algorithm unless the unprocessed data vectors are processed.

If there are no shared neighbors, then label both bike sharing data are marked as noise, else measures are taken to check if the neighbors of **(b_i)** and **(b_j)** and share neighbors are greater than the given shared nearest neighbor threshold. If they share neighbors, the neighbors of both bike sharing are merged and assigned them to a cluster. Cluster points are automatically determined using the merge operation. If data points of **(b_i)** and **(b_j)** and do not share any neighbors, then each dense data points is processed by checking. If it is a bike sharing system, then its neighbors are processed and clustered together, else it is labeled as a boundary result. If **(b_i)** has neighbors and if

one of its neighbors has higher density than **(b_i)** then **(b_j)** is defined as the boundary result. The processed and the unprocessed data points in SNNC estimation method are updated and the process is repeated until all the data points are processed.

4. EXPERIMENTAL EVALUATION SET UP OF SNNC-UT ESTIMATION METHOD

Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method is evaluated using JAVA with WEKA tool to estimate the performance. For evaluation purpose, comparison is performed on the proposed SNNC-UT estimation method with the existing k-nearest neighbor hubness in clustering. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization and is also suited for mounting new machine learning schemes.

Bike Sharing Dataset extracted from UCI repository contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bike share system with the equivalent weather and seasonal information. Bike sharing systems are new generation data where complete process from membership, rental and return back has happen to be automatic. Through these systems, user is able to effortlessly rent a bike from an exacting position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is self-possessed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Cluster Quality achieved by shared nearest neighbor algorithms that diverge considerably in their notion of what constitutes a cluster and how to efficiently find them. Cluster quality, a density threshold depend on the particular data set and intended use of the result is measured in terms of percentage (%). Clustering time is the average amount of time consumed to perform the clustering operation based on the shared nearest neighbor, measured in terms of seconds. Energy utilization is definite as the amount of energy consumed to automatically determine the cluster points in the data, measured in terms of Kilo joules (KJ).

$$\text{Energy consumption} = Ts^2$$

Where, 'T'=Total no. of data for processing and's' represents the speed of clustering effect of information. Distance measure ratio states the distance between points, where it groups the points into some number of clusters. So that members of a cluster are in some sense as close to each other as possible, measured in terms of millimeters (mm). Accuracy is defined as the degree of closeness of measurements to that quantity's actual true value, measured in terms of percentage (%). Computation cost is defined as the amount it takes to perform the bike sharing computation based on clustering.

5. PERFORMANCE RESULT OF SHARED NEAREST NEIGHBOR CLUSTERING

Shared Nearest Neighbor Clustering based on Unscented Transform (SNNC-UT) estimation method is compared against the existing k-nearest neighbor hubness in clustering using the JAVA programming. The table given below shows the experimental values and graph illustrates the pictorial form of SNNC-UT method against k-nearest neighbor hubness. Table-1 shows the value of the SNNC-UT method and existing method cluster quality value.

Table-1. Tabulation of Cluster quality.

Iterations	Cluster quality (%)	
	Existing k-nearest neighbor hubness	SNNC-UT method
5	78	90
10	79	91
15	80	93
20	81	93
25	82	95
30	84	95
35	85	96
40	87	97

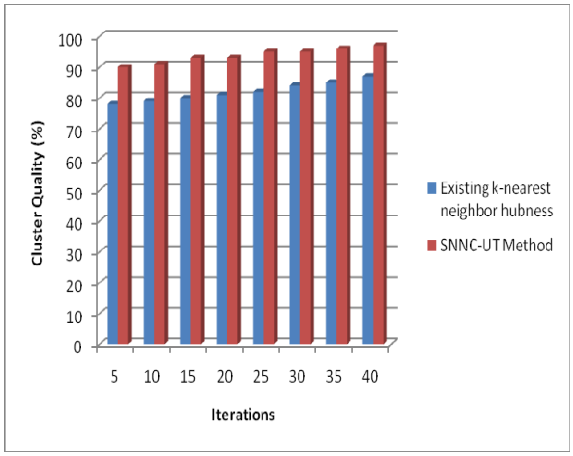

Figure-4. Cluster quality measure**.**

Figure-4 describes the cluster quality measure based on the iterations taken place in the SNNC-UT method and k-nearest neighbor hubness in clustering.
As the iterations increased, the quality of the cluster is also improved to 10 – 15% in SNNC-UT method when compared with the k-nearest neighbor hubness in clustering [1]. The cluster quality is improved using SNNC-UT method because of when processing of two most dense points (assigns the bike share system to the appropriate clusters. Specifically, assigning the neighbors of two most dense points to the same cluster improves the shared nearest neighbor threshold value.

Table-2. Tabulation for clustering time.

Neighbor Size (KB)	Clustering Time (sec)	
	Existing k-nearest neighbor hubness	SNNC-UT method
23	212	201
45	351	335
55	485	471
32	528	515
97	616	599
112	786	755

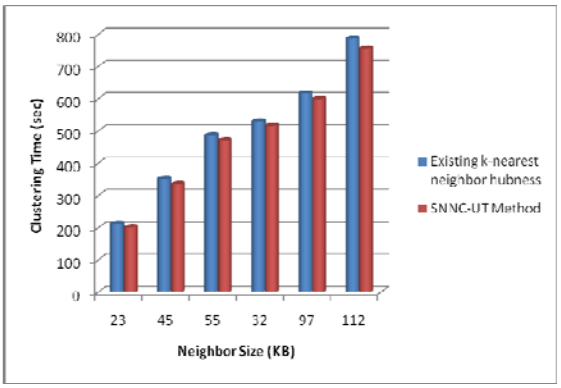

Figure-5. Clustering time measure.

The time taken to cluster the similar data points using the shared nearest neighbor is shown in Figure-5. As the neighbor size varies, the clustering time taken is also diverged. The SNNC-UT method consumes 2 – 6 % lesser time for clustering when compared with the k-nearest neighbor hubness, because the presence of shared neighbors between two dense bike points reduces time consumed. Neighbor size is measured in terms of Kilo Bytes (KB).

Table-3. Tabulation of energy consumption.

No. of attributes	Energy Consumption (KJ)	
	Existing k-nearest neighbor hubness	SNNC-UT method
2	1562	1323
4	1845	1585
6	1250	1124
8	1383	1154
10	1620	1420
12	1730	1444
14	1858	1593

Table-4.3 describes the energy consumption based on the SNNC-UT Method and existing k-nearest neighbor hubness. As the attribute range varies, energy consumption is measured in terms Kilo Joules (KJ).

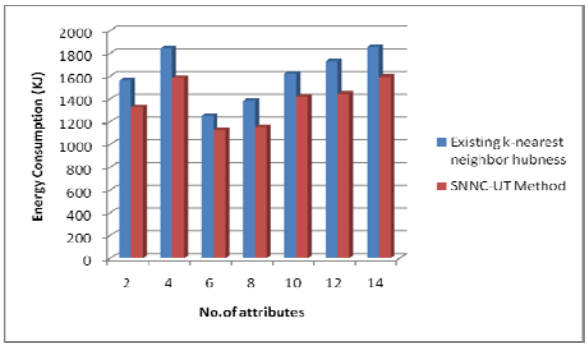

Figure-6. Energy consumption measure.

Figure-6 describes the energy consumption based on the attribute count. As the attribute count increases, energy consumption is decreased to 10 – 20% in SNNC-UT Method when compared with the existing k-nearest neighbor hubness. The consumption of energy is reduced in the SNNC-UT method by using the shared approach. Shared nearest neighbor develops the connection matrix 'M' to reduce the energy usage drastically. Connection matrix also constructs the shared nearest neighbor graph for effective processing.

Table-4. Tabulation of distance measurement ratio.

Cluster group	Distance measurement ratio (millimeter)	
	Existing k-nearest neighbor hubness	SNNC-UT Method
3	260	220
6	240	210
9	360	320
12	380	335
15	445	390
18	555	495
21	670	590

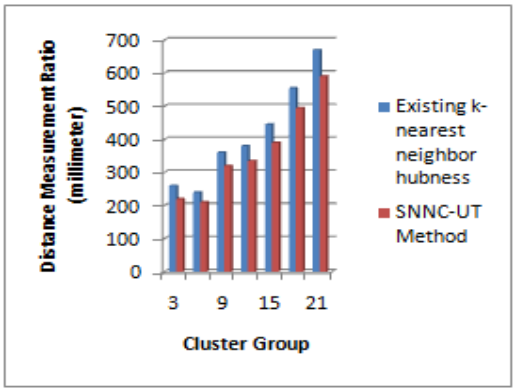

Figure-7. Measure of distance ratio.

Figure-7 describes the distance ratio measurement based on the cluster group. As the cluster group varies from 3 to 21, the distance measurement is also computed based on the count.

SNNC-UT method consumes 12 – 20 % reduced distance when compared with the k-nearest neighbor hubness in clustering due to unscented transform estimation. The distance measure is reduced in SNNC-UT method because it performs the better particle filter based on Unscented Transform that in turn reduces the distance.

Table-5. Tabulation of system accuracy.

% of Mutual clusters	System accuracy (%)	
	Existing k-nearest neighbor hubness	SNNC-UT method
40	86	91
50	85	88
60	89	94
70	88	84
80	91	95
90	86	90
100	90	94

Table-4.5 describes the system accuracy based on the mutual clustering. As the % of mutual cluster ranges from 40, 50 … 100, the system accuracy is also achieved, measured and illustrated in terms of percentage (%).

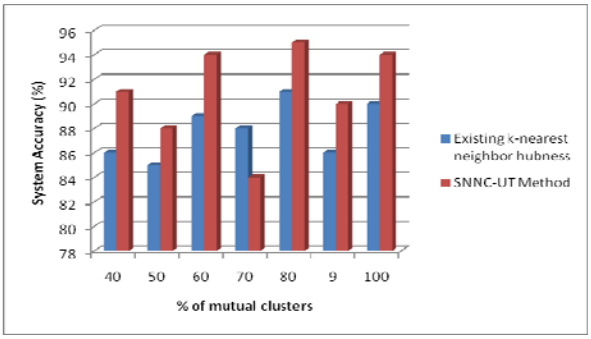

Figure-8. Measure of system accuracy.

Figure-8 describes the system accuracy on the SNNC-UT Method and existing k-nearest neighbor hubness in clustering. System accuracy of SNNC-UT method is 2- 5 % improved because, SNNC-UT method uses the true nonlinear function and estimates the distribution of the function output. Non-linear function compute the random points f (u) = S (α, Σ) for improved accuracy rate.

Table-6. Tabulation of computational cost.

Cluster point size	Communication cost (sec)	
	Existing k-nearest neighbor hubness	SNNC-UT method
10	119	112
20	126	120
30	148	141
40	156	148
50	166	157
60	172	164

Table-6 describes the computational cost based on the cluster point size. Cluster point size ranges from 10, 20….60. The computational cost is measured in terms of seconds (sec). As the time increases, cost consumed is also improved gradually.

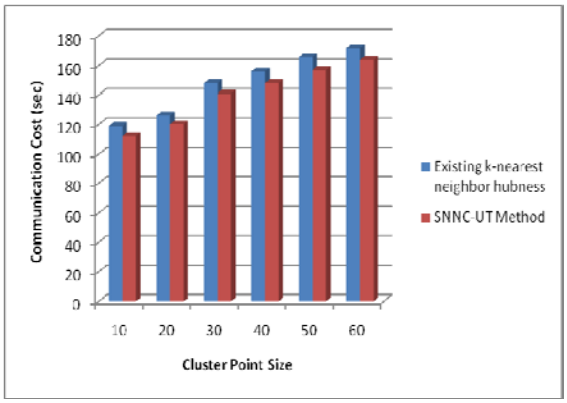

Figure-9. Computational cost measure.

Figure-9 describes the computational cost based on the cluster point size. The computational cost of the SNNC-UT Method is decreased approximately from 5 to 10 % when compared with the existing k-nearest neighbor hubness. The covariance matrix of the components of 'f' and 'g' are utilized to decrease the complexity in computational cost on the cluster points. The uniform distribution over the cluster points consumes lesser time, which gradually leads to the minimization of the computational cost.

Finally, it is being observed that the Shared Nearest Neighbor Clustering based on Unscented Transform estimation method calculates the similarity of hourly and daily count of rental bikes across different time series. Shared nearest neighbor automatically determine the clusters points within a cluster and with the different cluster. The statistic distance measure captures similarity based on the unscented transform estimation which is consistently used to obtain an enhanced result.

6. CONCLUSIONS

SNNC-UT method presented a new clustering algorithm based on density and shared nearest neighbor measure using the bike sharing data. The cluster point counts are identified instantaneously using shared nearest neighbor present in the data space. The kind of cluster analysis on bike sharing data is extremely useful in determining automatically based on the density. Unscented Transform estimation achieves the best results on distance measure utilizing the performance improvement with Gaussians model. A correct and competent distance computation is based on the unscented transform estimation which is regularly used to obtain an enhanced result. The similarity between bike share system with the corresponding weather and seasonal information are developed. The experimental result of SNNC-UT estimation method using the bike sharing dataset contains

information of hourly and daily count of rental bikes. SNNC-UT method attains the improved cluster quality, minimal time consumption for clustering, 4.142 % increased system accuracy, decreases the consumption of energy, reduced distance ratio and minimal computational cost.

REFERENCES

[1] Nenad Tomasev., Milo S. Radovanovic., Dunja Mladenic and Mirjana Ivanovic. Revised 2013. The Role of Hubness in Clustering High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering.1-12.

[2] Renchu Guan., Xiaohu Shi., Maurizio Marchese., Chen Yang. and Yanchun Liang. 2011. Text Clustering with Seeds Affinity Propagation. IEEE Transactions on Knowledge and Data Engineering,. 23(4):627-637.

[3] Rahmat Widia Sembiring., Jasni Mohamad Zain. 2011. Cluster Evaluation of Density Based Subspace clustering. Journal of Computing. 2(11): 1-6.

[4] Amir Ahmad., Lipika Dey. 2011. A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. Pattern Recognition Letters. Elsevier Journal. 1062-1069.

[5] Dr. Mohammed Otair. 2013. Approximate K-Nearest neighbor based spatial clustering using K-D tree. International Journal of Database Management Systems (IJDMS), 5(1):97-108.

[6] Milos Radovanovic., Alexandros Nanopoulos. and Mirjana Ivanovic. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. Journal of Machine Learning Research. pp. 2487-2531.

[7] Yi-Hong Chu., Jen-Wei Huang., Kun-Ta Chuang., De-Nian Yang. and Ming-Syan Chen. 2010. Density Conscious Subspace Clustering for High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering. 22(1): 16-30.

[8] Xiaojun Chen., YunmingYe., XiaofeiXu. and Joshua Zhexue Huang. 2012. A feature group weighting method for subspace clustering of high-dimensional data.Pattern Recognition, Elsevier journal.434 – 446.

[9] Rosy Das., D. K. Bhattacharyya. and K. Kalita. 2008. A Frequent Itemset– Nearest Neighbor Based Approach for Clustering Gene Expression Data. Citeseer.73-78.

[10] Rozita A. Dara., Masoud Makrehchi. and Mohamed S. Kamel. 2010. Filter- Based Data Partitioning for Training Multiple Classifier Systems. IEEE Transactions on Knowledge and Data Engineering. 22(4):508-522.

[11] Brian Quanz., Jun (Luke) Huan. and Meenakshi Mishra. 2012. Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue. IEEE Transactions on Knowledge and Data Engineering. 24(10):769 – 779.

[12] Duc Thang Nguyen., Lihui Chen. and Chee Keong Chan. 2012. Clustering with Multi viewpoint-Based Similarity Measure. IEEE Transactions on Knowledge and Data Engineering 24(6): 988-1001.

[13] Xuran Zhao., Nicholas Evans., Jean-Luc Dugelay. 2013. A subspace co-training framework for multi-view clustering. Pattern Recognition Letters, Elsevier Journal.73-82.

[14] Ming Hua. and JianPei. 2012. Clustering in applications with multiple at a source - A mutual Subspace clustering approach, Neurocomputing, Elsevier Journal.133-144.

[15] Arthur Flexer., Dominik Schnitzer Jan Schluter. ISMIR 2012. A mirex meta-analysis of hubness in audio music similarity.175-180.2012.

[16] Deng Cai., Chiyuan Zhang., Xiaofei He. 2010. Unsupervised Feature Selection for Multi-Cluster Data.ACM journal.333-342.

[17] Nenad Tomasev. and Dunja Mladenic. 2011. Exploring the hubness - related properties of oceanographic sensor data. Artificial Intelligence Laboratory.

[18] Rene Vidal., Paolo Favaro. 2013. Low rank subspace clustering (LRSC). Pattern Recognition Letters, Elsevier Journal, 47-61.

[19] Charu C. Aggarwal. 2009. On High Dimensional Projected Clustering of Uncertain Data Streams. IEEE Transaction on Data Engineering.1152-1154.

[20] Arthur Flexer., Dominik Schnitzer. 2013. Can Shared Nearest Neighbors Reduce Hubness in High - Dimensional Spaces. Journal of Machine Learning Research.460 - 467.