Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System

Sadia Nawrin
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Md Rahatur Rahman
Simplexhub Ltd.
Dhaka,
Bangladesh

Shamim Akhter
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

***Abstract*—Traffic Management System (TMS) is used to improve traffic flow by integrating information from different data repositories and online sensors, detecting incidents and taking actions on traffic routing. In general, two decision making systems-weights updating and forecasting are integrated inside the TMS. The models need numerous data sets for making appropriate decisions. To determine the dynamic road weights in TMS, four (4) different environmental attributes are considered, which are directly or indirectly related to increase the traffic jam– rain fall, temperature, wind, and humidity. In addition, peak hour is taken as an additional attribute. Usually, the data sets are classified by instinct method. However, optimum classification on data sets is vital to improve the decision accuracy of the TMS. Collected data sets have no class label and thus, cluster based unsupervised classifications (partitioning, hierarchical, grid-based, density-based) can be used to find optimum number of classifications in each attribute, and expected to improve the performance of the TMS. Two most popular and frequently used classifiers are hierarchical clustering and partition clustering. K-means is simple, easy to implement, and easy to interpret the clustering results. It is also faster, because the order of time complexity is linear with the number of data. Thus, in this paper we are going to demonstrate the performance of partition k-means and hierarchical k-means with their implementations by Davies Boulder Index (DBI), Dunn Index (DI), Silhouette Coefficient (SC) methods to outline the optimal number classifications (features) inside each attribute of TMS data sets. Subsequently, the optimal classes are validated by using WSS (within sum of square) errors and correlation methods. The validation results conclude that k-means with DI performs better in all attributes of TMS data sets and provides more accurate optimum classification numbers. Thereafter, the dynamic road weights for TMS are generated and classified using the combined k-means and DI method.**

Keywords—Traffic Management System (TMS); Data Clustering; K-means; Hierarchical Clustering; Cluster Validation

I. Introduction

A new low cost, flexible, maintainable, and secure internet-based traffic management system with real time bi-directional communication was proposed and implemented (in [1][2][3][4]) to assist and reduce the traffic situation. To determine the dynamic road weights in TMS, four (4) different environmental attributes - rain fall, temperature, wind, and humidity are considered. Rainfall is one of the most influential weather attributes to determine the road congestion in metro city, as the road segments are submerged due to the heavy rains, and makes slower traffic movements. The heat released from the engines, air-conditioners of the traffic stacked vehicles, may raise the overall temperature of the area. Thus, the current temperature helps to classify traffic congestion status of a particular road segment. Gusts of wind have direct influence on road safety and that pushes to slower vehicle movement. In addition, temperature, wind and humidity have direct influence to predict the future rainfall in a particular area. Peak hour is one of the most influential attributes to cause traffic congestion in metro cities. Thus, these four (4) environmental attributes and peak hour have direct or indirect relationship on traffic congestion as well as vehicle movement and influence to choose them as decision making parameters.

The value of these attributes (features) are intelligently crawled by search engine, with metadata indexing (title, description, keyword etc.), directly from the multiple data feeds (like web site, RSS feeds, web service etc.) from the web page in [5]. Crawled data are simplified (structured) and stored in a historic table. However, the number of attributes can be changed according to the system requirements. We collect more than two (2) years or 750 days (1/12/2006 to 20/12/2008) data of five features from the web page in [5].

Initially, decision tree (DT) [1] [2] [3] was used to classify road weights and weighted moving average analytic was implemented to estimate or predict feature values in DT [28][29] based system and achieved 16.45% accuracy. However, the model data sets were classified by instinct method. Cluster based classifications (K-means, Locality-Sensitive Hashing (LSH) etc.) can be used to find optimum number of classifications in each feature and can improve the performance of the TMS. With this hypothesis, we implement two unsupervised clustering techniques partition k-means and hierarchical k-means. There are several methods (internal/external) to measure similarity between two clustering steps and used to compare how well different data clustering algorithms perform on a set of data. Only internal methods - Davies Boulder Index (DBI), Dunn Index (DI), and Silhouette Coefficient (SC) - are used to choose the optimum number of classification, as they do not have any external information. Subsequently, the optimal classes are cross-validated by using statistical analytics - correlation and Within Sum of Square (WSS) errors.

Results highlight that Dunn Index (DI) performs better for both partition k-means and hierarchical k-means algorithms by providing minimum Sum of Square Error (SSE) for all environmental attributes. However, the optimum numbers of classifications are generated by both algorithms, for each environmental attribute, differs in their numbers. Both algorithms are compared by computing the correlation values on their optimal number of clusters for each attribute. The correlation values of partition k-means algorithm are higher than the correlations of hierarchical k-means algorithm for all attributes. The validation results conclude that the combination of the k-means with Dunn Index performs better and provides more accurate optimum classification number(s) on environmental data set. Thereafter, the dynamic road weights for TMS are generated and classified with these combined algorithms.

II. Related Works

Integrating intelligence technologies in transportation system including intelligent and effective route planning to reduce travel time, reliable estimation of traffic congestion, accident and/or hazard detection etc., can help to reduce both fuel consumption and the associated emission of greenhouse gases. However, this kind of Intelligent Transportation System (ITS) requires collecting and modeling tremendous amount of continuous data from all road segments, in different time domains, for everyday in a year, and is a complex task. In addition, analytical decision making on optimum route planning requires high data processing and centralized computation. Data mining techniques, especially clustering, are involved to shape the unstructured data to a structural formulation and make easier decision making system for ITS problems.

Traffic flow data is used in [31] to detect the traffic status and predict the traffic patterns from historical database. Two different data mining techniques-cluster analysis and classification analysis are used in the historical data prediction model. Classified road features are used to estimate traffic flows in [32]. Functional Data Analysis (FDA) is used in [33] to analyze the daily traffic flow. A comparative study on different data mining techniques to classify traffic congestion is done in [34]. It examines J48 Decision Tree, Artificial Neural Network, Support Vector Machine (SVM), PART and K-Nearest Neighborhood to classify future traffic status and concludes J48 Decision Tree algorithm has the best performance.

In our previous works, traffic management data attributes were worked with DT (decision tree) [1] [2] [3] (Fig.1) and Neural Network (NN) [4]. NN performs better than DT. However, these works did not perform any recognized data mining or classification technique to the environmental data sets. Rather, they classified data according to the intuitive guesses. Thus, the proposed TMS is suffering from optimal data classification strategies.

There are many available methods/techniques used to classify data sets. In [12], optimal cluster numbers are determined based on the intra-cluster and inter-cluster distance measurements. They use Davies-Bouldin index and Dunn's index methods for classifying both synthetic and natural images.

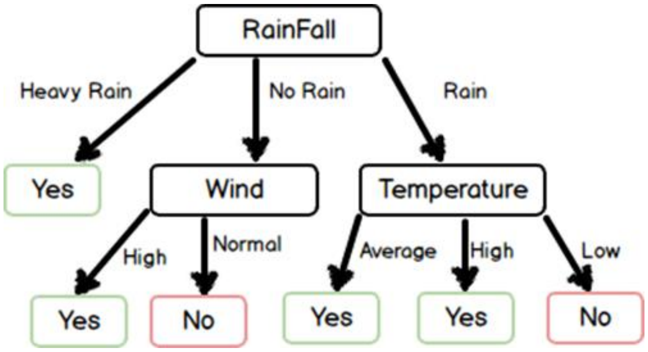


Fig. 1. Decision Tree Using ID3 Algorithm

Paper [13] evaluates the performance of three clustering algorithms (hard k-means, single linkage, and a simulated annealing) and determining the number of clusters using four methods-Davies-Bouldin index, Dunn’s index, Calinski-Harabasz index and index I. Paper [14] compares three clustering algorithms- agglomerative hierarchical clustering k-means algorithm, bisecting k-means algorithm and standard k-means. Results indicate that the bisecting k-means technique performance better than other two.

In [15], authors discuss and compare the various clustering methods to find the best and fix the optimal number of clusters over three (3) structured datasets. They use three (3) different clustering algorithms- hierarchical, k-means, PAM and three (3) internal optimal clustering methods- connectivity, silhouette and dunn.

It is common and popular to apply hierarchical or partition clustering on classification problems [16]. K-means is simple, easier to implement and provide linear order complexity. Thus, partition k-means and hierarchical k-means algorithms are used to classify the TMS data sets and their optimum classification numbers are determined by three (3) different cluster validity indexes- Davies Boulder Index (DBI), Dunn Index (DI), Silhouette Coefficient (SC).

III. Classification Techniques

There are many industrial problems identified as classification problems. For examples, stock market prediction, weather forecasting, bankruptcy prediction, medical diagnosis, speech recognition, character recognitions to name a few [6-10]. Classifications are typically classified into three broad categories- supervised, unsupervised and reinforce learning [11]. Supervised learning is used when the data class label are known. Unsupervised learning (cluster analysis) is applicable on unknown class label datasets. Reinforcement learning is the problem of getting an agent to act in the world to maximize its rewards. In this paper, TMS data sets have no class label thus falls in unsupervised learning category. This section describes the algorithms and methods- those are used for clustering in this paper. Notations and their descriptions are listed in Table I.

A. Hierarchical Clustering

Hierarchical clustering constructs a hierarchy of clusters (dendrogram). Dendrogram is a process that captures whether the order in which clusters are merged (bottom-up view) or clusters are split (top-down view). There are two variant of hierarchical clustering methods (in fig. 2.): i) Agglomerative Hierarchical clustering algorithm (HAC) or AGNES (bottom-up approaches), ii) Divisive Hierarchical clustering algorithm (HDC) or DIANA (top-down approaches). In this paper, we implement the divisive hierarchical cluster to classify the feature data, as it has less computational cost compare to AGNES. We stop our iteration when optimal clustering number is reached.

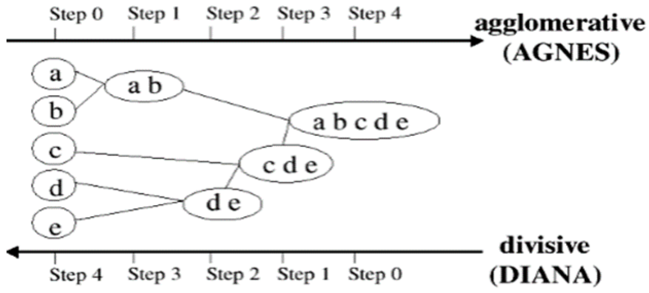


Fig. 2. Hierarchical clustering structure

1) Divisive Hierarchical Clustering Algorithm: Division Hierarchical Clustering Algorithm (HDC) or DIANA (DivisiveANAly) [17] is a variant of hierarchical clustering. It starts evaluation from the top with all data in one cluster (fig. 3) and then split using flat clustering algorithm such as k-means clustering.

Algorithm:

- **a.** Initially all items belong to one cluster C_i=0.
- **b.** Split C_i into sub-clusters, C_{i+1} and C_{i+2}.
- **c.** Apply K-mean on C_{i+1} and C_{i+2}.
- **d.** Increment the value of i.
- **e.** Repeat steps b, c and d until the desired cluster structure is obtained.

Node 0 containing the whole data set

C_1=2 input nodes 1-2.

C_2=3 input nodes-> 2- 4 (1 spilt into 2 sub group-3 and 4).

C_3=4 input nodes ->3-6 (1 spilt into 2 sub group-5 & 6).

Do until C_{kmax}not reached where C_{kmax} is maximum number of clusters.

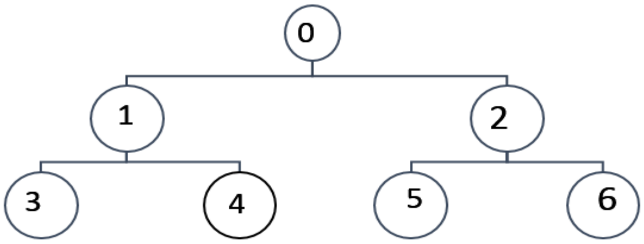


Fig. 3. Splitting node in DIANA

B. Partitional Clustering

Partitional clustering determines a flat clustering into k clusters with minimal costs. It partitions data set into k clusters and assigns the object to their nearest centers. Here (in fig. 4), k is the number of centroids.

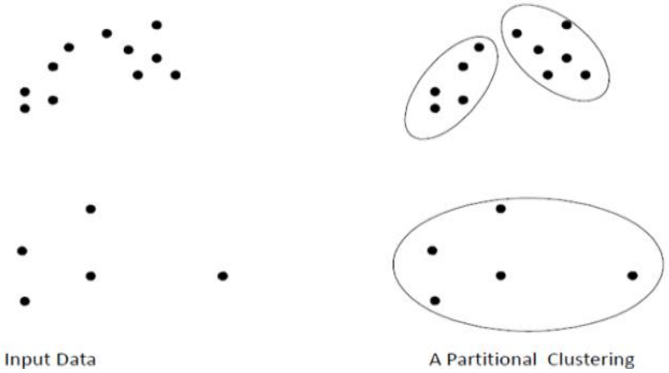


Fig. 4. Partitional clustering

1) K-means Clustering Algorithm: K-means clustering [18][19][27] aims to partition data into k clusters. K-means is the most popular non-hierarchical iterative clustering algorithm (Fig.5). The basic idea of k-means is to start with an initial partition and assign data objects to cluster so that the squared error decreases.

Algorithm:

- **a.** Randomly initialize k center from the set of data point $\{X^d=x_1^d, x_2^d, x_3^d \ldots x_n^d\}$.
- **b.** Assign each point to their nearest center using Manhattan distance measure.

$$\min_{0\leq i<n_c}(d(x^d, v_i^d)) \quad (1)$$

- **c.** Compute the centroid for each cluster by averaging the data objects belonging to the cluster, assign it as a new cluster center.

$$v_{inew}^d = \frac{1}{n_i}\sum_{k=1}^{n_i} d(x^d, v_{iold}^d) \quad (2)$$

- **d.** Re-assign all the data points to its new center.
- **e.** Repeat b, c and d steps until all the cluster centers do not change anymore otherwise stop.

TABLE. I. LIST OF SYMBOLS AND THEIR DESCRIPTION

SL No	Symbol/Notation	Description
1.	n_c	Number of total cluster
2.	C_i	ithcluster
3.	d(x,y)	Manhattan distance between two data element
4.	n_i	Number of element in the ith cluster
5.	v_i	Value of the center of the ithcluster
6.	d(v_i,v_j)	Distance between two center
7.	S_i	Variance of ithcluster
8.	C_{kmax}	Maximum number of cluster
9.	d	No of dimension

IV. OPTIMAL CLUSTERING METHODS

Clustering validity indexes [20][21][22][23] are usually defined by combining compactness and separability of the clusters. Compactness measures closeness of cluster elements. A common measure of compactness is variance. Separability indicates how distinct two clusters are. Basically, there are two types of validity techniques used for clustering evaluation-external criteria and internal criteria [30]. External criteria are used for categorized data clustering. No internal information is needed for internal criteria. It evaluates the quality of clusters, using only the data and without referencing to the external information. There are so many methods to measure the quality of the clustering-Davies-Bouldin index, Dunn index, CH index, Elbow method, X-means clustering, Information Criterion Approach, Information Theoretic Approach, Silhouette method, and cross-validation. The used TMS data do not have any external information and thus influences to use internal measure or criteria for clustering validation.

1) Davies-Bouldin Index: Davies Bouldin (DB) index [20][21] measures the average similarity between each cluster and its most similar one. Lower value of DB Index indicates that clusters are tight compact and well separated which reflects better clustering. The goal of this index is to achieve minimum within-cluster variance and maximum between cluster separations. It measures similarity of cluster (R_{ij}) by variance of a cluster (S_i) and separation of cluster (d_{ij}) by distance between two clusters (v_i and v_j). The formulae of DB index are-

$$S_i = \frac{1}{n_i - 1} \sum_{x \epsilon c_i} d(x, v_i)^2 \tag{3}$$

$$d_{ij} = d(v_i, v_j) \tag{4}$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \tag{5}$$

$$R_i = \max_{0 \le j < n_c, i \ne j}(R_{ij}) \ , i = 1 \cdots \cdots n_c R_i \ge 0 \tag{6}$$

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \tag{7}$$

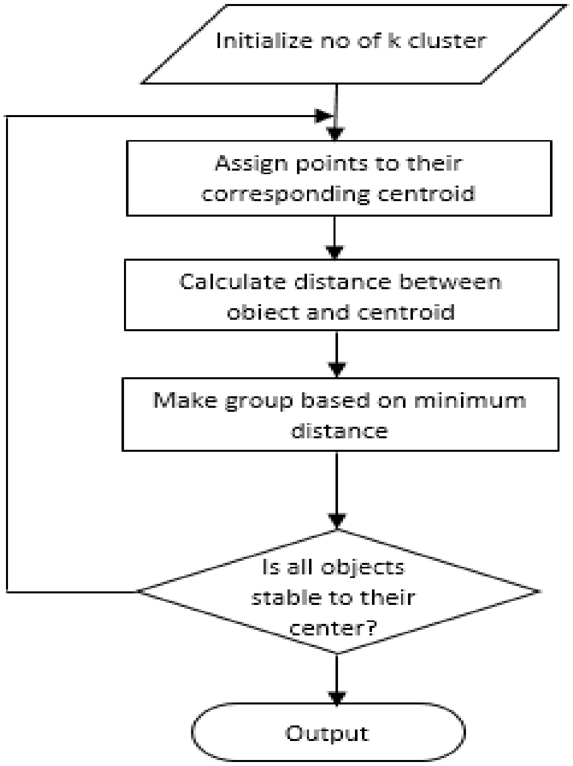

Initialize no of k cluster
Assign points to their corresponding centroid
Calculate distance between obiect and centroid
Make group based on minimum distance
Is all objects stable to their center?
Output

Fig. 5. Flow chart of K-mean Algorithm

2) Dunn Index: The value of Dunn index (DI) [21] is expected to large if clusters of the data set are well separated. If the dataset has compact and well-separated clusters, the distance between the clusters is expected to be larged and the diameter of the clusters is expected to be smaller. The clusters are compact and well separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance. Large value of Dunn index indicates the compact and well-separated clusters. The formulae of Dunn index are-

$$D = \frac{\min_{0 \le i < n_c, 0 \le j < n_c, i \ne j}(d(C_i, C_j))}{\max_{0 \le k < n_c}(diam(C_i))} \tag{8}$$

Where,

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j}(d(x, y)) \tag{9}$$

$$diam(C_i) = \max_{x, y \in C_i}(d(x, y)) \tag{10}$$

3) Silhouette Coefficient: Silhouette Coefficient (SC) [22] [23][24] shows- how well the objects can fit within the cluster. It measures the quality of the cluster by ranging between -1 and 1. A value near to one (1) indicates that the point x is affected to the right cluster. There are two terms- cohesion and separation. Cohesion is intra clustering distance, and separation is distance between cluster centroids. A(x) is the average dissimilarity between x and all other points of its cluster. B(x) is the minimum dissimilarity x and its nearest cluster. A cluster which has a value near -1, indicates that the point should be affected to another cluster. The formulae of SC are-

$$a(x) = \frac{1}{n_i - 1} \sum_{y \in C_i y \neq x} d(x, y) \qquad (11)$$

$$b(x) = \min_{j, j \neq i} [\frac{1}{n_j} \sum_{y \in C_j} d(x, y)] \qquad (12)$$

$$SC = \frac{1}{n_c} \cdot \sum_i \{\frac{1}{n_i} \sum_{x \in C_i} \{\frac{b(x) - a(x)}{\max[b(x), a(x)]}\}\} \qquad (13)$$

V. Cluster Validation Methods

A. *Correlation:* An effective clustering algorithm needs a suitable measurement of similarity or dissimilarity. Correlation (in Fig. 6) computes the similarity matrix and incident matrix (also called occurrence matrix) to measure the correlation between the data and its cluster [25]. Higher value of correlation indicates that the points belong to the same cluster (very close to each other), and reflects good clustering. The formula of correlation is-

$$r = \frac{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1, j=1}^{n} (c_{ij} - c)^2}} \qquad (14)$$

Here,

r =correlation of the data and its cluster,

Distance matrix, D= {d_{11},d_{22},d_{33}, …,d_{nn}},

Incident matrix C= {c_{11},c_{22},c_{33}, …,c_{nn}},

$\bar{d}$ =mean of the distance matrix,

and $\bar{c}$ =mean of the incident matrix.

B. *Distance Matrix :* It is also called similarity matrix, an nxn two dimensional matrix -where n is the number of elements in a data set. d(x, y) distance or dissimilarity between objects x and y. Fig. 7 represents distance matrix.

$$d(x,y)=|x-y| \qquad (15)$$

C. *Incidence Matrix:* An incidence matrix is a matrix that shows the relationship between two classes of objects. It is an nxn matrix where n is the total number of data set. If the object x and the object y belongs to the same cluster then Ixy=1 and if the object x and the object y belongs to the different cluster then Ixy=0.

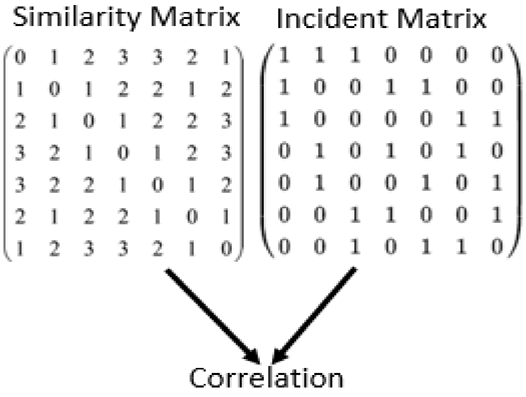
D. *Manhattan Matrix :* Manhattan distance is the absolute distance between two points. Let, the objects x = (x_1, ...,x_d) and y = (y_1, ..., y_d) then the Manhattan distance between the two objects is,

$$d(x, y) = \sum_{i=1}^{d} |x_i - y_i| \qquad (16)$$
$$where, d = dimension$$

In this work, we use Manhattan distance as a distance measurement technique.

E. *Within Sum of Square Error (WSS) :* WSS is also called Sum of Squared Error (SSE) [26]. Sum of Square Error (SSE) or within sum of square cluster error (WSS) is widely used for criteria measuring. The value of SSE is high, indicates high error, which means poor quality cluster. Good clustering aims for minimum value of SSE. The formula of within Sum of Square Error is-

$$SSE\ or\ WSS = \sum_{k=1}^{n_c} \sum_{x_i \in C_i} d(x_i, V_i) \qquad (17)$$



$$\begin{pmatrix} 0 & 1 & 2 & 3 & 3 & 2 & 1 \\ 1 & 0 & 1 & 2 & 2 & 1 & 2 \\ 2 & 1 & 0 & 1 & 2 & 2 & 3 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 \\ 3 & 2 & 2 & 1 & 0 & 1 & 2 \\ 2 & 1 & 2 & 2 & 1 & 0 & 1 \\ 1 & 2 & 3 & 3 & 2 & 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Fig. 6. Correlation

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Fig. 7. Distance Matrix

VI. Experimental Results

Based on the above algorithms and methods, data are formulated to determine the optimal classes in each feature, road weight and verify better algorithm. Experiments in Table 2 and 3 are generated from the 750 days (1/12/2006 to 20/12/2008) collected data from [5] and presented in Fig. 8.

1) Results of Divisive Hierarchical Method: the sum of square error (SSE) of all features using divisive hierarchical cluster with Davies-Bouldin index, Dunn index and Silhouette index are presented in Table II. Shaded block (in Table II) indicates the minimum value of SSE. This table represents the optimal cluster size of each feature using three methods and also presents that Dunn index minimizes the SSE values in all cases. Thus, we conclude that Dunn index performs better for HDC to find optimal cluster. Thus, the optimal classes of each feature using HDC are – Rainfall (k=2), Temperature (k=2), Wind (k=3), Humidity (k=5) and Peak hour (k=4).

TABLE. II. OPTIMAL NUMBER OF CLUSTER AND VALUE OF SSE OF RAINFALL, TEMPERATURE, WIND, HUMIDITY AND PEAK HOUR USING HDC WITH DB, DUNN AND SC INDICES

Feature / Method	Rainfall		Temperature		Wind		Humidity		Peak hour	
	Optimal k	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***
DB	2	25051.32	2	3690.40	2	9301.24	3	31152.06	3	99507.77
Dunn	2	25051.32	2	3690.40	3	4850.62	5	11231.18	4	49786.87
SC	2	25051.32	2	3690.40	3	4850.62	3	31152.06	2	183452.98
Optimal k	2		2		3		5		4	

TABLE. III. OPTIMAL NUMBER OF CLUSTER AND ITS VALUE OF SSE OF RAINFALL, TEMPERATURE, WIND, HUMIDITY AND PEAK HOUR USING K-MEAN WITH DB, DUNN AND SC INDICES

Feature / Method	Rainfall		Temperature		Wind		Humidity		Peak hour	
	Optimal k	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***
DB	3	9574.97	2	3690.40	2	9301.24	3	23146.38	2	183452.98
Dunn	3	9574.97	3	2106.56	4	4657.67	6	6339.761	5	49541.539
SC	2	25051.32	2	3690.40	2	9301.24	3	23146.38	2	183452.98
Optimal k	3		3		4		6		5	

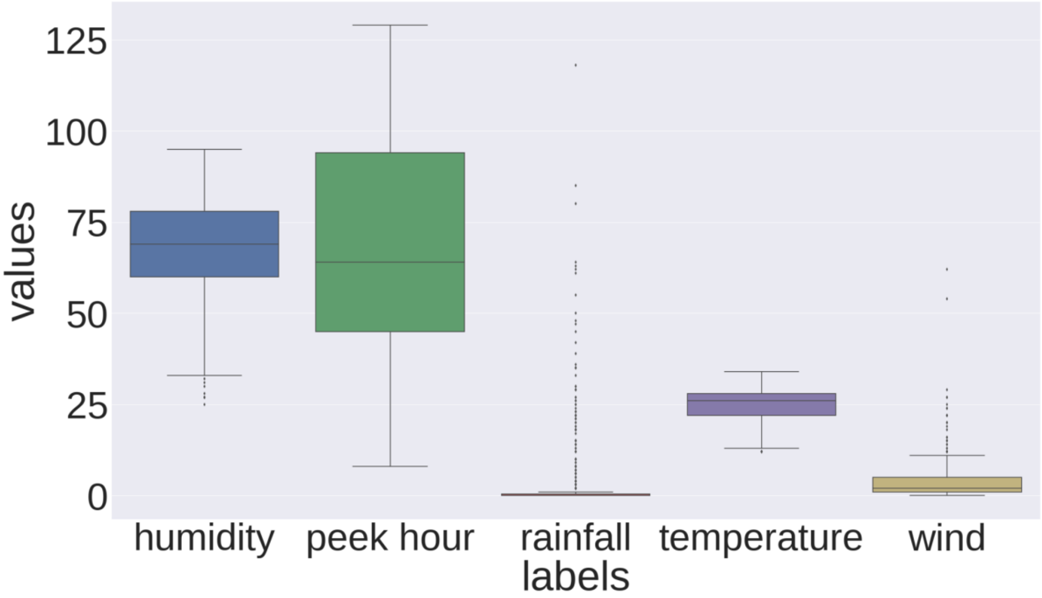


Fig. 8. Collected data from ACCU Weather[5]

TABLE. IV. COMPARISON THE CORRELATION BETWEEN TWO ALGORITHMS

Algorithm Feature	**Hierarchical**		**K mean**	
	Optimal k cluster Using Dunn index	***Correlation***	***Optimal k cluster Using Dunn index***	***Correlation***
Rainfall	2	-0.801	3	-0.789
Temperature	2	-0.736	3	-0.721
Wind	3	-0.580	4	-0.405
Humidity	5	-0.555	6	-0.521
Peak hour	4	-0.639	5	-0.578

TABLE. V. CLUSTER SIZE AND DUNN INDEX VALUE OF THE ROAD WEIGHT

No of cluster k	**Dunn index value**
2	0.08
3	0.09
4	0.10
5	0.11
6	0.11
7	0.13
8	0.12
9	0.12

2) Result of K-means Clustering Method: The sum of square error (SSE) [26] of all features using k-means clustering algorithm with Davies Bouldin index, Dunn index and Silhouette index are presented in Table III. Shaded block (in Table III) indicates the minimum value of SSE. Table III reflects that Dunn index provides minimum value of the SSE in all features. Thus, we conclude that Dunn index performs better for k-means algorithm to find optimal cluster numbers. The optimal classes of each feature using k-means are – Rainfall (k=3), Temperature (k=3), Wind (k=4), Humidity (k=6) and Peak hour (k=5).

3) Comparison of HDC and K-means: Hierarchical clustering and K-means clustering are compared by computing the correlation on their optimal cluster numbers in each feature. It is clear from Table IV that the correlations of K-means are higher than the correlations of HDC, for all features. Thus, we conclude k-means performs better than HDC.

TABLE. VI. SAMPLE ROAD WEIGHT CLUSTERING RESULT

Data	**Rainfall**	**Temperature**	**Wind**	**Humidity**	**Peak hour**	**Road weight**
1	0	1	0	3	1	0
2	0	1	0	4	3	5
3	0	0	0	3	2	2
4	0	0	0	3	3	5
5	0	2	1	4	0	4
6	0	2	2	4	1	6
7	0	2	0	3	1	0
8	0	2	1	4	3	3
9	0	2	0	3	0	4
10	0	2	1	5	0	4

4) Optimal Cluster of Road Weight : From the previous experiments it is clear that k-means with Dunn index performs better for all features. Thus, for the classification of the road weight k-means with Dunn index can be chosen. Table V shows the no of cluster of road weight and Dunn index value of that corresponding cluster. This table represents that maximum value of Dunn index achieves in k=7. Thus, the optimal cluster size of road weight is seven (7) and there should be seven (7) different type of classes for road weight updates. Table VI presents some sample experimental results of road weight updates.

VII. CONCLUSION AND FUTURE WORKS

In this section, we summarize our work. The features data are collected from the external feeds (like web site, RSS feed, web service etc.) for classifying data. We cluster the data using two approaches (partition k-means and hierarchical k-means) and find the optimal number of clusters for each feature using Davies-Bouldin index, Dunn index and Silhouette coefficient. Thereafter, conclusion has been drawn which algorithm is better for which feature data and then find the optimal number of clusters of road weights with the input of the measured five (5) feature clusters.

In future, we can also measure validity of the classes by other probabilistic and statistical methods. Dunn index method needs lots of computational cost. Improvement on the computation cost and error of the cluster building procedure can be reduced using other statistical models. At present, we are not considering other characteristics of environmental and road status such as: accidents, road works, etc. Roads and Highway authorities in Bangladesh does not provide/publish any road construction, maintenance status and thus, these attributes will be considered in our future research direction.

Online multi data feeds capability supports the proposed model to be connected with different Social Medias (facebook, twitters etc.), and collects necessary information (mishap, disaster situations), and uses analytical tools to make proper decisions. However, special consideration is required on internet securities as all of the information is available on the internet. Recently, deep learning (DL) techniques are also used to solve unsupervised clustering problem. Interpolation of deep leaning is much complex than k-means. In addition, deep learning works with multi-layer data representation and sometimes degrades the performance due to the limited amount of data. It addresses over fitting problem also. Thus, a comparative study with simple k-means and DL is required and will be applied in near future.

Still, the proposed TMS is in construction phase and cover small road networks. City level broader area will be considered in near future. A GSM and GPS based micro controller with different embedded sensors is in developing phase. This device will help to collect real time environmental data at an instant time.

ACKNOWLEDGEMENTS

East West University (EWU), Bangladesh is gratefully acknowledged for providing article processing charge (APC) to cover the costs of publication.

REFERENCES

[1] Rahman, M. R. and Akhter, S. 2015. Real time bi-directional traffic management support system with GPS and websocket, Proc. of the 15th IEEE International Conference on Computer and Information Technology (CIT-2015). Liverpool. UK. 26-28 Oct. 2015.

[2] Rahman, M. R. and Akhter, S. 2015. Bidirectional traffic management with multiple data feeds for dynamic route computation and prediction system. *International Journal of Intelligent Computing Research (IJICR)*. Special Issue. Volume 7. Issue 2. ISSN: 2042 4655. Mar/2015. http://infonomics-society.ie/ijicr/

[3] Rahman, M. R. and Akhter, S. 2015. Bi-directional traffic management support system with decision tree based dynamic routing, *Proc. of 10th International Conference for Internet Technology and Secured Transactions, ICITST 2015*. London. United Kingdom. December 14-16. 2015.

[4] Akhter,S. Rahman, M.R., and Islam M. A. 2016. Neural Network (NN) based route computation for bi-directional traffic management system. *International Journal of Applied Evolutionary Computation- special issue on Emerging Research Trend in Computing and Communication Technologies.* Volume 7. Issue 4.

[5] AccuWeather. (2016, March 7). Retrieved from http://www.accuweather.com/en/bd/dhaka/28143/january-weather/28143?monyr=1%2F1%2F2016&view=table

[6] Moghadassi, F. Parvizian, and S. Hosseini. 2009. A new approach based on artificial neural networks for prediction of high pressure vapor-liquid equilibrium. *Australian Journal of Basic and Applied Sciences*. Vol. 3. No. 3. pp. 1851-1862. 2009.

[7] Asadi, R., Mustapha, N., Sulaiman, N. and Shiri, N. 2009. New supervised multi layer feed forward neural network model to accelerate classification with high accuracy. *European Journal of Scientific Research*. Vol. 33. No. 1. 2009. pp.163-178.

[8] Pelliccioni, R. Cotroneo, and F. Pung 2010. Optimization of neural net training using patterns selected by cluster analysis: a case-study of ozone prediction level", *8th Conference on Artificial Intelligence and its Applications to the Environmental Sciences*, 2010.

[9] Kayri, M. and Çokluk, Ö. 2010. Using multinomial logistic regression analysis In artificial neural network: an application, *Ozean Journal of Applied Sciences*. Vol. 3. No. 2. 2010.

[10] Khan, U., Bandopadhyaya, T. K. and Sharma, S. 2009. Classification of Stocks Using Self Organizing Map. *International Journal of Soft Computing Applications*. Issue 4. pp.19-24.

[11] Hagan, M.T., Demuth, H.B., and Beale, M. 1996. *Neural Network Design.*PWS publishing company, Boston, Massachusetts.

[12] Ray, S., and Turi, R., H. 1999. Determination of number of clusters in K-means clustering and application in color image segmentation, *Published in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*. Pg. 137-143. Culcutta, India.

[13] Maulik, U., and Bandyopadhyay, S., “Performance Evaluation of Some Clustering Algorithms and Validity Indices”, Published in Journal IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 24 Issue 12, Pg.1650-1654, December 2002.

[14] Steinbach, M., Karypis, G., and Kumar, V., “A Comparison of Document Clustering Techniques”, Published in the 6th ACM SIGKDD, World Text Mining Conference, (2000).

[15] Begum, S., F., Kaliyamurthie, K., P., and Rajesh, A., “Comparative Study of Clustering Methods over Ill-Structured Datasets using Validity Indices”, published in Indian Journal of Science and Technology, Vol 9(12), March 2016.

[16] Reddy, C.K. and Vinzamuri, B., “A Survey of Partitional and Hierarchical Clustering Algorithms”, Data Clustering: Algorithms and Applications. CRC Press 2014, ISBN 978-1-46-655821-2 (pg.88-91).

[17] Hierarchical-clustering-algorithm available at: https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm,15-5-2016.

[18] Ray, S., and Turi, R., H., “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation”, Published in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (Pg. 137-143).

[19] K-mean Algorithm Available at: http://www.academia.edu/3438357/K_Means_Algorithm_Example,12-4-2016.

[20] Kovács, F., Legány, C. and Babos, A., “Cluster Validity Measurement Techniques” ,AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (2006), ISBN:111-2222-33-9 (pg. 388-393).

[21] Davies, D.L. and Bouldin, D.W., “A Cluster Separation Measure”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

[22] Dunn, J.C., “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” J. Cybernetics, vol. 3, 1973, (pg. 32-57).

[23] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J.,”Understanding of Internal Clustering Validation Measures”, Proceeding ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining ISBN: 978-0-7695-4256-0 (Pg. 911-916).

[24] Zoubi, M., and Rawi, M., “An efficient approach for computing silhouette coefficients,” Journal of Computer Science 4,(2008) (pg.252–255).

[25] Correlation Definition available at:

www.cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_basics.pdf, 23-4-2016.

[26] Sum of square error at:https://hlab.stanford.edu/brian/error_sum_of_squares.html,15-5-2016.

[27] Pelleg, D., Moore, A.W., “X-means: Extending K-means with Efficient Estimation of the Number of Clusters”, Proceeding ICML '00

Proceedings of the Seventeenth International Conference on Machine Learning, ISBN: 1-55860-707-2 (Pg.727-734).

[28] ID3 Decision Tree Algorithm - Part 1 at: http://www.codeproject.com/Articles/259241/ID-Decision-Tree-Algorithm-Part

[29] https://datajobs.com/data-science-repo/Decision-Trees-[Rokach-and-Maimon].pdf

[30] Liu,Y., Li,Z., Xiong,H., Gao,X.,Wu,J., “Understanding of Internal Clustering Validation Measures”, Proc. of IEEE International Conference on Data Mining,2010 http://datamining.rutgers.edu/publication/internalmeasures.pdf

[31] Wang, Y., Chen, Y., Qin, M. and Zhu, Y., “Dynamic traffic prediction based on traffic flow mining,” in Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA 0'6), vol. 2, pp. 6078–6081, Dalian, China, June 2006.

[32] Caceres, N., Romero, L. M. and Benitez, F. G., “Estimating traffic flow profiles according to a relative attractiveness factor”, Procedia—Social and Behavioral Sciences, vol. 54, pp. 1115–1124, 2012.

[33] Guardiola, I. G., Leon, T. and Mallor, F., “A functional approach to monitor and recognize patterns of daily traffic profiles”, Transportation Research, Part B: Methodological, vol. 65, pp. 119–136, 2014.

[34] Mirakhorli, A., "A Comparative Study: Utilizing Data Mining Techniques to Classify Traffic Congestion Status", UNLV Theses, Dissertations, Professional Papers, and Capstones. Paper 2197.