International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 3
ISSN: 2088-5334

Enhanced Manhattan-based Clustering using Fuzzy C-Means Algorithm for High Dimensional Datasets

Joven A. Tolentino[#, *], Bobby D.Gerardo[#]

[#]*Technological Institute of the Philippines, Quezon City, Philippines*
E-mail: jatolentino@tau.edu.ph; bobby.gerardo@gmail.com

[*]*Tarlac Agricultural University, Tarlac, Philippines*

***Abstract*—The problem of mining a high dimensional data includes a high computational cost, a high dimensional dataset composed of thousands of attribute and or instances. The efficiency of an algorithm, specifically, its speed is oftentimes sacrificed when this kind of dataset is supplied to the algorithm. Fuzzy C-Means algorithm is one which suffers from this problem. This clustering algorithm requires high computational resources as it processes whether low or high dimensional data. Netflix data rating, small round blue cell tumors (SRBCTs) and Colon Cancer (52,308, and 2,000 of attributes and 1500, 83 and 62 of instances respectively) dataset were identified as a high dimensional dataset. As such, the Manhattan distance measure employing the trigonometric function was used to enhance the fuzzy c-means algorithm. Results show an increase on the efficiency of processing large amount of data using the Netflix ,Colon cancer and SRCBT an (39,296, 38,952 and 85,774 milliseconds to complete the different clusters, respectively) average of 54,674 milliseconds while Manhattan distance measure took an average of (36,858, 36,501 and 82,86 milliseconds, respectively) 52,703 milliseconds for the entire dataset to cluster. On the other hand, the enhanced Manhattan distance measure took (33,216, 32,368 and 81,125 milliseconds, respectively) 48,903 seconds on clustering the datasets. Given the said result, the enhanced Manhattan distance measure is 11% more efficient compared to Euclidean distance measure and 7% more efficient than the Manhattan distance measure respectively.**

***Keywords*— fuzzy C-Means; high dimensional dataset; Manhattan distance; clustering.**

I. INTRODUCTION

The high dimensional dataset is common nowadays due to the colossal amount of information being gathered electronically by varying information systems. Movies, medical health record, and agricultural dataset can be observed to be as high dimensional dataset. Duplication of records, multiple attributes and thousands number of records were categorized as high dimensional datasets, and most of the data mining algorithms suffer low accuracy and high computational cost in processing when a high dimensional dataset was supplied [1]. This high dimensional dataset can also be observed to know what this dataset shows and implies.

A common technique to observe this dataset is using clustering. Clustering splits a large amount of data and performs grouping considering the similarities of the individual data supplied [2]. However, several clustering algorithms suffer from high computational cost and one of which is the Fuzzy C-Means algorithm.

Fuzzy C-Means also suffers from its accuracy and speed when a dataset contains high dimension or not [3], [4]. The study aims to enhance the Fuzzy C-Means algorithm by changing the distance measure to solve the weakness of the said algorithm. Manhattan distance measure was used since it is also ideal when applied to high dimensional dataset [5]. The trigonometric approach was utilized to the said distance measure since the accuracy of the Manhattan distance measure suffers when centroid and points are connected diagonally [6], [7].

Data mining procedures will also be used to prepare the actual dataset for mining. The computational cost will be observed by testing the algorithm with different distance measures (Euclidean, Manhattan and Enhanced Manhattan) and three different high dimensional datasets (Netflix Movie Rating, Colon Cancer and SRCBT) which will lead on what specific distance measure is faster when applied to the said algorithm.

II. MATERIALS AND METHOD

To investigate the performance of the modified algorithm, Knowledge Discovery Model were used proposed by [8] consisting the step of data selection, data pre-processing, transformation, and data mining. Figure 1 shows the actual process of KDD.

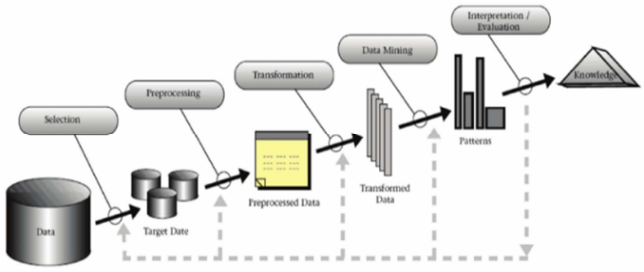

Fig.1 the Knowledge Discover Model

With the KDD model, the dataset should be ideal to be processed from the part of the selection to the step of data mining.

A. Data Selection

Selection of the actual dataset for clustering was done by searching for the appropriate dataset that has high dimensionality. High dimensional datasets are the ones who have multiple fields and thousands of records [1]. The high dimensionality of data is also when dataset features are greater than the number of instances [9]. The Netflix movie rating, small round blue cell tumors (SRBCTs) and Colon Cancer dataset are also categorized as high dimensional considering these definitions. Table 1 shows the number of features of the said datasets.

TABLE I
THE FEATURES AND INSTANCES OF THE DATASETS

Dataset	Features	Instances
Netflix Movie	5	1,500
Colon Cancer	2000	62
SRCBT	2308	83

B. Pre-Processing

The pre-processing technique was also done to prepare the dataset that will be used. This technique reduces the dimensionality of the dataset [10], [11]. The dataset was merged into a file and field were also observed to identify the process needed to be done to reduce its dimensionality. The term discretization technique describes another advantage of this step. In this part, the equal frequency binning was used. This step converts the text into a numeric value. Each instance in the dataset that has the same value are considered as one and converted to a similar numeric value [12]. In Table II, the values of the feature, genre, were discretized to fit the algorithm.

TABLE II
A PORTION OF THE MOVIE DATASET WITH ITS GENRE

No	Movie Title	Genre	Discretised Value
1	Cat Run 2 (2014)	Action	1
2	He Who Dares (2014)	Action	1
3	How to Train Your Dragon 2 (2014)	Action|Adventure|Animation	2
4	Hercules (2014)	Action|Adventure	3
5	Falcon Rising (2014)	Action|Adventure	3
6	Land Ho! (2014)	Adventure|Comedy|Documentary|Mystery	4
8	Seventh Son (2014)	Adventure|Children|Sci-Fi	6

In this process, the field, genre, was discretized to be applicable with clustering. The same process was done for the two remaining datasets (Colon Cancer and SRCBT). The feature class was converted into a numeric value.

C. Transformation

Making the dataset suitable for knowledge discovery requires the dataset to be transformed. The dataset for Netflix movie rating is composed of several tables (Movie, Rating, and Tags) that are connected via Primary Key (PK) and a Foreign Key (FK). A foreign key is several techniques can do a specific property of dataset, which is described by the implementation of the primary key to another data table [13] and merging this dataset. One technique for combining this data table for preparation for data mining is union. The union is the process of identifying the intersection of two or more data table with their PK and FK[14]. Hence, the researcher created a tool for merging the data table into a single dataset concerning the primary key and foreign key.

For the two remaining datasets, features were already normalized aside from the pre-processing technique. Observation of the actual content of the dataset was also needed to be observed thoroughly to see how these datasets were constructed such that the enhanced algorithm can process it. Based on the pre-processing and transformation techniques, the following portions of the datasets of Netflix Movie Rating, Colon Cancer and SRCBT had been derived.

TABLE III
NETFLIX MOVIE DATASET

Rating	UserID	Time Stamp	Genre	MovieID
2.5	53930	1393064439	30	22306
2.5	87813	1387131563	30	22306
3	137200	1398867354	30	22306
4.5	13494	1421295240	114	23623
4	15720	1426647292	114	23623

TABLE IV
COLON CANCER DATASET

FTR1	FTR1	FTR1	FTR to 2000	Class
88.23	39.67	67.83	28.7	2
82.24	85.03	152.2	16.77	1
76.97	224.62	31.23	15.16	2
74.53	67.71	48.34	16.09	1
54.56	223.36	73.1	31.81	2
33.2	91.85	5.88	21.88	1
98.54	54.62	30.54	24.45	2

TABLE V
SRCBT DATASET

FTR1	FTR1	FTR1	FTR to 2308	Class
0.143	0.888	0.068	0.108	2
0.085	0.324	0.635	0.271	1
0.193	0.39	0.378	0.107	3
0.159	0.248	1.164	0.224	4

D. Data Mining

Clustering algorithm will be enforced in this study by using the Fuzzy C-means algorithm. This tool can be used to address its problem on clustering high dimensional datasets. Figure 2 shows the actual process of how Fuzzy C-Means Clustering works.

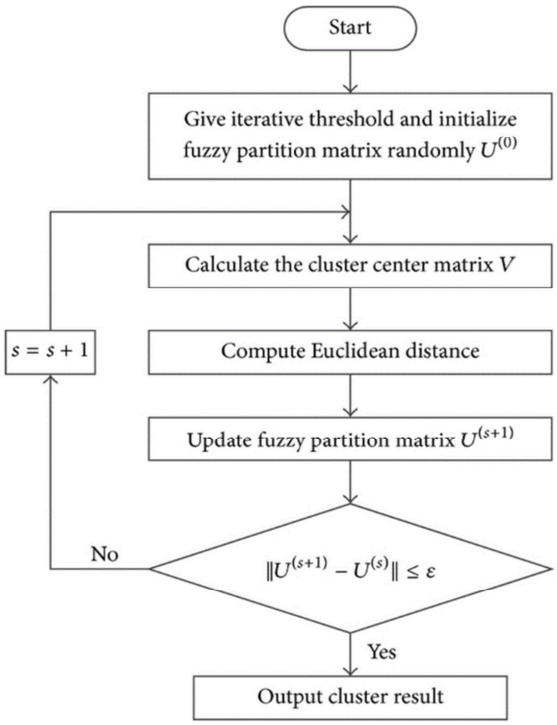

Fig. 2 The actual process of clustering using Fuzzy C-means algorithm

The first step is that Fuzzy C-Means selects the number of cluster and membership functions ranging from zero to one. The calculation of the actual centroid with the corresponding parameter follows. The computation of the actual centroid plays a vital role in creating the clusters[15]. This will identify how many iterations will be done. The third step is to date the actual cluster with the specific distance measure and lastly, validate the result. The iterations take place until convergence is achieved [16]. With this given process of Fuzzy C-Means algorithm, changing the distance measure can improve the performance of the said algorithm.

E. Manhattan Distance Measure

Providing a result with lesser computational cost can be achieved using different strategies. Observing the distance measure used by the algorithm and its performance can be a basis in identifying what distance measure is applicable for the high dimensional dataset. The Manhattan distance measure is commonly used when the point that is generated was vertically or horizontally connected. Selecting an appropriate distance measure plays a vital role in providing a good set of clusters [17]. The study also shows that Manhattan distance measure is more accurate in the calculating distance when the dataset is high dimensional compared to other distance measures [18]. Table VI shows the side by side comparison of several distance measure.

TABLE VI
COMPARISON OF SEVERAL DISTANCE MEASURE.

Distance Measure	Benefits	Drawbacks
Euclidean	Easy to Implement and Test	Results are greatly influenced by variables that have the largest value. Does not work well for Image data, Document Classification
Manhattan	Easily generalized to a higher dimension	Does not work well for image data and document classification
Cosine	Handles both continuous and categorical variables	Does not work well for nominal data
Jaccard Index	Handles both continuous and categorical variables	Does not work well for nominal data

The use of the Manhattan Distance Measure will allow the algorithm to speed up its processing time, although Manhattan distance measure has a problem needed to be addressed.

F. Euclidean Distance Measure

On the other hand by default Euclidean distance measure were used in Fuzzy C-Means, it produces a more accurate result but higher computational cost [19], this is the main reason why the algorithm needed to be improved with the proposed modification conceptualized.

G. Enhancement of the Distance measure

A weakness of the Manhattan distance measure is in terms of clustering points that are connected diagonally. Fig. 3 shows the actual points connecting to the centroid diagonally.

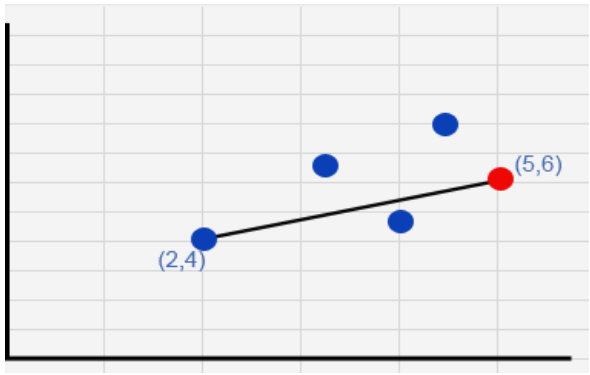

Fig. 3 Centroid and data point are connected diagonally

Employing trigonometric function specifically COSINE Equation 2 was tested in order to check the speed of the said

distance measure when applied to the Fuzzy C-Means algorithm.

$$cosine(\emptyset) = \frac{\text{adjacent}}{hupothenuse} \quad (2)$$

Where adjacent (next to) is to the angle θ and Hypotenuse is the long line, equation 3 shows the actual solution to address the problem of Manhattan Distance.

$$d = \sum \frac{|(x3-x2)+ (y_2-y_3)|}{cosine(\emptyset)} \quad (3)$$

Where y_3 is the point of intersection from the created imaginary line and y_2 is the y-coordinate of the centroids. The difference of y_3 and y_2 will be divided to $cosine(\emptyset)$. $Theta$ $(\emptyset)$ is used since the actual angle is not yet solved. To calculate the actual distance the following, steps were considered.

Step 1. *Create an imaginary line to form a right triangle*

Step 2. *Identify the point of intersection*

Step 3. *Compute the Distance of the Imaginary line using Manhattan. Given that (x2=5,x3=5) and (y2=6,y3=4)*

(5-5)+(6-4)= 2

Step 4. *Compute for the distance*

2/Cosine(53.60)=3.61

The given steps in calculating the actual distance of the centroid to the dataset points may lead to higher accuracy for the Fuzzy C-Means Algorithm when supplied.

H. Fuzzy C-Means

To further test the algorithm, the steps for the distance measure were invoked by the enhanced Manhattan distance measure. By default, Fuzzy C-Means uses Euclidean distance. Fig. 4 shows the actual process of clustering the dataset using the enhanced Manhattan distance.

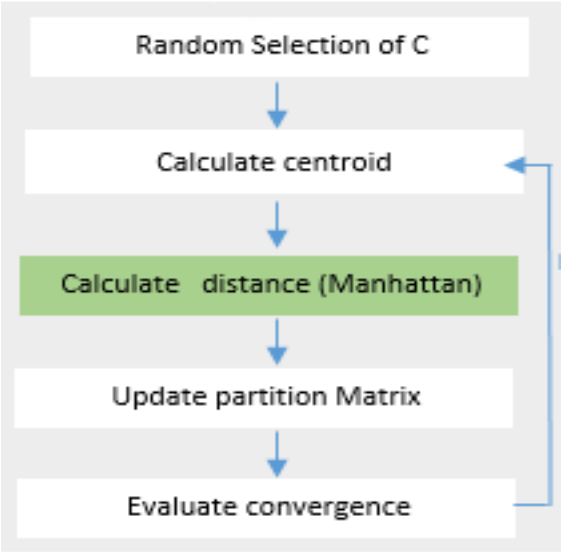

Fig. 4 The actual process of clustering using Fuzzy c-means algorithm.

The following pseudo-code was used to implement the Modified Manhattan distance measure over Fuzzy C-Means Algorithm.

Start
Required Array of Points and Centroid
Declare Distance

For counter=0; to LengthofPoints step 2

If Centroid is equal to Points
Get the absolute difference of points and the centroid
Else
Get the absolute difference of centroid and Imaginary line
Divide the absolute difference to $cosine(\emptyset)$

EndIf

Update distance by adding the difference
Iterate to each column and pair it with (x,y) format and do the calculation for the distance.
End

I. Evaluation

To validate the performance of the said modified algorithm, the duration to complete the process of clustering using Fuzzy C-Means with different distance measures were compared, and the starting points of the program were tracked. The differences were calculated to identify how many milliseconds were needed to complete the actual clustering process. The following pseudo code was used to evaluate the performance of Fuzzy C-Means on applying the three distance measures and three high dimensional datasets.

Start
Get Start time in milliseconds
Declare Threshold=1, iteration=0
While Threshold is not equal to 0
Update value of iteration +1
Assign new center
End While
Elapsed time = end time - start time
End

The process of Fuzzy C-Means clustering stops when the convergence is reached. This means that when the threshold becomes zero, the actual clustering process is finished on clustering [16] and as prescribed by the algorithm threshold use was zero. Tracking the execution time of the program can now be observed along with the behavior of the algorithm when different distance measures and different datasets with high dimensions was applied.

III. Results and Discussion

With the procedure of pre-processing and transformation, the dataset Netflix Movie composed of 4 attributes with 1,500 instances, Colon Cancer having 2000 features and 65 instances and SRCBT 2308 features and 83 instances are

now ready for clustering and comparison of the actual speed of the modified algorithm to the standard Fuzzy C-Means Algorithm. The algorithm was tested by computing the actual time elapsed when the clustering processes were simulated. Table VI showed the actual result of the algorithm when Euclidean and Enhanced Manhattan distance measures were used.

TABLE VII
RESULT OF THE ALGORITHM IN (MS), WHEN EUCLIDEAN AND ENHANCED MANHATTAN IS USED

Dataset	Clusters	Euclidean	Enhanced Manhattan
Netflix Movie	10	39296	33216
Cancer	4	38952	32368
SRCBT	3	85774	81125

Observing the actual result, the Enhanced Manhattan distance measure outperformed the Euclidean distance Measure. To further investigate, the Manhattan distance measure was also used to compare the actual results as shown in Table VIII.

TABLE VIII
RESULT OF THE ALGORITHM IN (MS), WHEN MANHATTAN AND ENHANCED MANHATTAN IS USED

Dataset	Clusters	Manhattan	Enhanced Manhattan
Netflix Movie	10	36858	33216
Cancer	4	36501	32368
SRCBT	3	82860	81125

With the dataset supplied to the Manhattan distance and enhanced Manhattan distance, the result shows that the actual modification decreases the processing time for clustering the three datasets. Comparison of the actual result for clustering using Fuzzy C-Means with the three distance measure is indicated in Figure 5 and 6. The behavior of the algorithm varies on the dataset supplied, especially when it comes to high dimensions.

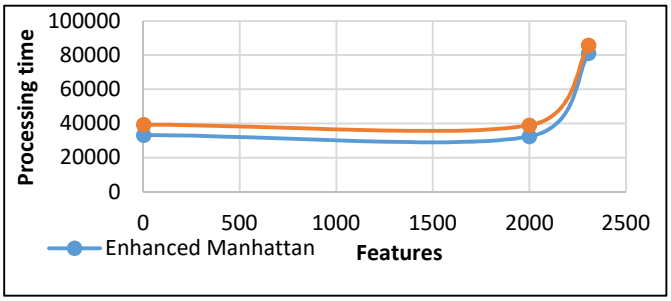

Fig. 5 The comparison of the processing time of Euclidean and Enhanced Manhattan against the Dataset Features.

The trend of the three distance measure plotted along with the number of features and to its processing time showed an improvement when the Enhanced Manhattan Distance measure was supplied. This indicates that the modification can now be applied to the algorithm to increase its speed on clustering high dimensional datasets.

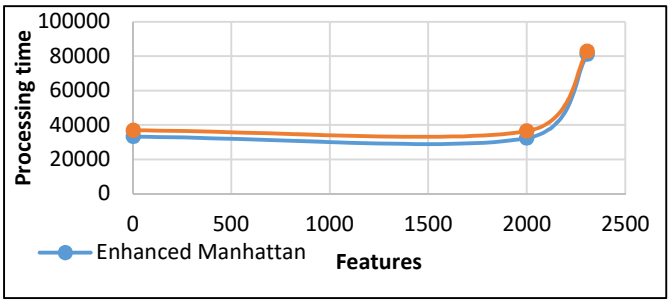

Fig. 6 The comparison of the processing time of Manhattan and Enhanced Manhattan against the Dataset Features.

The trend of the three distance measure plotted along with the number of features and to its processing time showed an improvement when the Enhanced Manhattan Distance measure was supplied. This indicates that the modification can now be applied to the algorithm to increase its speed on clustering high dimensional datasets.

IV. CONCLUSION

Fuzzy C-means algorithm is an algorithm that suffers from high computational cost when a high dimensional dataset is applied. One way to address the said problem is by invoking the distance measure used. In this study, an Enhanced Manhattan-based clustering was used employing trigonometric function to address the issue of Manhattan distance measure.

Results show that an increase in the efficiency in terms of speed of the said algorithm can be observed when using the enhanced Manhattan distance measure. Euclidean distance measure shows that clustering the three datasets such as Netflix Movie Rating, Colon Cancer, and SRBT has a (39,296, 38,952 and 85,774 milliseconds to complete the different clusters, respectively) average of 54,674 milliseconds while Manhattan distance measure took an average of (36,858, 36,501 and 82,86 milliseconds, respectively) 52,703 milliseconds for the entire dataset to cluster. On the other hand, the enhanced Manhattan distance measure took (33,216, 32,368 and 81,125 milliseconds, respectively) 48,903 seconds on clustering the datasets.

Given the said result, the enhanced Manhattan distance measure is 11% more efficient compared to Euclidean distance measure and 7% more efficient than the Manhattan distance measure respectively. While the efficiency increases for the said algorithm, it needs further observation on the behavior of the algorithm in clustering a standard type of dataset. Accuracy also needs to be studied in applying this modified algorithm. Other factors can also be considered to increase the efficiency of the said algorithm.

ACKNOWLEDGMENT

This study would not be possible without the support of the Commission on Higher Education Kto12 Transition Program Unit - Quezon City, Philippines, and the Tarlac Agricultural University Tarlac, Philippines. Gratitude is also extended to the Technological Institute of the Philippines – Quezon, City.

REFERENCES

[1] N. Raksha and R. Alankar, "Detection of fuzzy duplicates in high dimensional datasets," *2016 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI, 2016*, pp. 1423–1428, 2016.

[2] Y. G. Jung, M. S. Kang, and J. Heo, "Clustering performance comparison using K- means and expectation maximization algorithms," *Biotechnol. Biotechnol. Equip. ISSN1310-2818*, vol. 2818, no. October 2015.

[3] Z. Cebeci and F. Yildiz, "Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures," *J. Agric. Informatics*, vol. 6, no. 3, pp. 13–23, 2015.

[4] R. Winkler, F. Klawonn, and R. Kruse, "Problems of Fuzzy c-Means Clustering and Similar Algorithms with High Dimensional Data Sets," *Challenges Interface Data Anal. Comput. Sci. Optim.*, pp. 1–8, 2012.

[5] S. Pandit and S. Gupta, "A Comparative Study On Distance Measuring," *Int. J. Res. Comput. Sci.*, vol. 2, no. 1, pp. 29–31, 2011.

[6] T. K. Mohana, V. Lalitha, L. Kusuma, N. Rahul, and M. Mohan, "Various Distance Metric Methods for Query Based Image Retrieval," vol. 7, no. 3, pp. 5818–5821, 2017.

[7] M. Khan and T. Shah, "A copyright protection using watermarking scheme based on nonlinear permutation and its quality metrics," *Neural Comput. Appl.*, vol. 26, no. 4, pp. 845–855, 2014.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in," vol. 17, no. 3, pp. 37–54, 1996.

[9] J. Zhang and M. Pan, "A high-dimension two-sample test for the mean using a cluster," *Comput. Stat. Data Anal.*, vol. 97, pp. 87–97, 2016.

[10] L. Zhou, "Preprocessing Method before Data Compression of Cloud Platform," pp. 1223–1227, 2017.

[11] M. A. Chaudhari, P. M. Phadatare, P. S. Kudale, R. B. Mohite, and R. P. Petare, "Preprocessing of High Dimensional Dataset for Developing Expert IR System," pp. 417–421, 2015.

[12] Z. Marzuki and F. Ahmad, "Data Mining Discretization Methods and Performances Data Mining Discretization Methods and Performances," no. December, pp. 3–6, 2014.

[13] N. A. Mian and N. A. Zafar, "Key Analysis of Normalization Process using Formal Techniques in DBRE," 2010.

[14] C. Ordonez, "Data Set Preprocessing and Transformation in a Database System," vol. 15, no. 4, pp. 1–19, 2011.

[15] Z. Wang, N. Zhao, W. Wang, R. Tang, and S. Li, "A Fault Diagnosis Approach for Gas Turbine Exhaust Gas Temperature Based on Fuzzy C-Means Clustering and Support Vector Machine," *Math. Probl. Eng.*, vol. 2015, pp. 1–11, 2015.

[16] N. Grover, "A study of various Fuzzy Clustering Algorithms," *Int. J. Eng. Res.*, vol. 5013, no. 3, pp. 177–181, 2014.

[17] L. H. Son, "Generalized picture distance measure and applications to picture fuzzy clustering," *Appl. Soft Comput. J.*, pp. 1–12, 2016.

[18] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," pp. 420–434, 2001.

[19] A. Fahad *et al.*, "A Survey of Clustering Algorithms for Big Data : Taxonomy & Empirical Analysis," 2014.