An Empirical Perusal of Distance Measures for Clustering with Big Data Mining

Kamlesh Kumar Pandey, Diwakar Shukla

***Abstract**: The distance measure is the core idea of data mining techniques such as classification, clustering, and statistical analysis and so on. All clustering taxonomies such as partition, hierarchical, density, grid, model, fuzzy and graphs used to distance measures for the data point's categorization under difference cluster, cluster construction and validation. Big data mining is the advanced concept of data mining respect to the big data dimensions. When traditional clustering algorithm is used under the big data mining the distance measure is needed for scalable under big data mining and support to a huge size dataset, heterogeneous data and sources, and velocity characteristics of the big data. From a theoretically, practically and the existing research perspective, the paper focuses on volume, variety, and velocity big data criterion for identifying a distance measure for the big data mining and recognize how to distance measure works under clustering taxonomy. This study also analyzed all distance measures accuracy with the help of a confusion matrix through clustering.*

***Index Terms**: Big Data, Big Data Mining, Big Data Characteristics, Clustering, Clustering Taxonomy, Distance Measure, Distance Measure Families.*

I. INTRODUCTION

The rapid growth of digital technologies based emerging applications such as internet of things, cloud computing, social network, sensor accessibility, medical application, and other technologies are changing nature of data to big data in the form of the vast amount of data with the heterogeneous format. In general, data are organized into a structured, unstructured and semi-structured format. In essential organization used for structured data a few years back, but nowadays some organization used to structured, unstructured and semi-structured data for interactive user support. Traditional data management technologies are generally handled to structured and unstructured data in limited volume with the homogenous data format. In the big data perspective, data management technologies must be capable of managing structured, unstructured and semi-structured data with high volume, processed any data in less time with high data quality for decision making and support to advance computation technologies [1][2]. Cluster analysis is one of the techniques for data mining and data mining is one of the elementary techniques for the big data analysis which works under the big data environment. Every clustering algorithms have own working process for cluster creation based on similarity and dissimilarity distance function [3]. Distance measures are not only essential to solve the clustering problem, but it is also solved to pattern recognition, classification, retrieval related problems [4], help to the derivation of new distance measure[5], text classification and clustering[6], document content comparison[7], time-series data management [8], uncertain data classification [9] and clustering [l0], bio-cryptic authentication in cloud databases[11], spatial concentration [12], location fingerprinting [13], author profiling[14], combining density [15], heavy aggregation operators [16], analyzing inconsistent information [17], network intrusion anomaly detection [18] for high volume, variety and velocity. The objective of this paper is identifying the best cluster distance measure for cluster creation in the big data mining and this objective is obtained by the six sections. The first section presents how to traditional data changes to big data and what the role is of distance measures. The second section presents the usual background of big data, big data dimension, big data mining, big data storage technologies and clustering concept of identifying the nature and storage format of big data for cluster analysis of the big data mining. The third section describes various distance measure for cluster creation on the bases of existing research. The fourth section gives the distance measure with their necessary properties and big data dimensions. The aim of this section is identifying which distance measure is used as a dissimilarity and similarity function for accurate cluster creation in the big data environment. The fifth section of this paper defines a major taxonomy of cluster with respect to distance measure and three big data dimensions. The sixth section is defined to the distance (dissimilarity) functional efficiency using the partitioning based family K-means algorithm, which is a more popular algorithm for the capability of handling large data set with scalability.

II. BACKGROUND

A. Big data

Nature of Big data is generally defined by Volume (huge scale data set), Variety (heterogeneous data and source), and Velocity (speed of data generation and processing.). These 3V's is known as three dimensions of big data and given by Laney (2001) for identifying an accurate definition of big data [1]. Gartner et.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Kamlesh Kumar Pandey*, Research Scholar, Department of Computer Science and Applications, Dr. HariSingh Gour Vishwavidyalaya, Sagar, India.

Prof. Diwakar Shukla, Dean & HoD, Department of Computer Science and Applications, Dr. HariSingh Gour Vishwavidyalaya, Sagar, India.

al.(2012), summarized these dimensions of big data as "Big data have high volume, high velocity, and high variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." and Tech America Foundation (2012) describes big data as "Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" [19]. The fourth dimension Veracity (data in doubt) is contributed by IBM respect to unreliability, untrustworthiness and uncertainty sources of data and it is related to the Variety. The Veracity measures the completeness and accuracy of data for decision support and confidently dealt with unreliability, untrustworthiness and uncertainty data. The fifth dimension Variability (variation in the data flow rates) is given by SAS and it related to Velocity and Variety. The Variability defines data continually get it from different sources and some sources are defined as a different meaning of the same data. The sixth dimension Value (data in highlight) is introduced by Oracle for attribute identification for the data mining or analysis. The last dimension is Visualization, which is visualized the data mine and analysis' results according to user expectation [1] [19] [20]. Volume, Variety, and Velocity are known as a basic dimension of big data creation and Veracity, Variability, Value, and Visualization are known as a supportable dimension of big data creation. Figure 1 summarized all dimensions of big data.

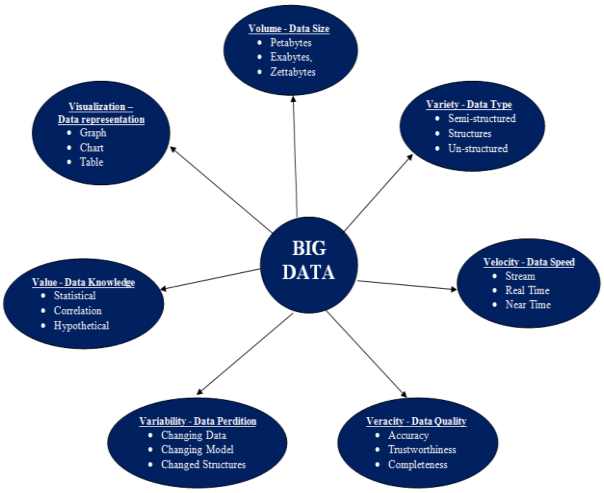

Figure 1 7V's of Big Data

B. Big data mining

Data mining techniques are one of the techniques for the big data analysis, which is discovered to meaningful knowledge, interesting patterns, and hidden relations in the high volume dataset with high accuracy. Traditional data mining techniques are divided into three groups as classification, clustering and association mining. These techniques face multiple challenges such as lack of efficiency, accuracy, scalability, speed when applied big data in the real-time environment. In the big data mining perspective, data mining techniques are capable of mine on data in high-volume, high-variety, and high-velocity data by using statistical methods and machine learning [21] [22].

C. Big data storages for mining

Big data mining objectives to minimize the processing, hardware and storage cost and big data management techniques properly managed for reliable, accessible, and secure data mining. In the Big data storages perspective, data are stored in a file format and database format. Some file base storages are Google File System (Google-GFS), Hadoop Distributed File System (Google-HDFS), Cosmos File System (Microsoft-CFS), Haystack File System (Facebook-HFS), Taobao File System (Taobao-TFS), Fast Distributed File System (Taobao-FDFS). The most popular big data databases are NoSQL, which stores a data in the four formats such as Key-Value pair (Amazon-Dynamo, LinkedIn-Voldemort, Redis, Tokyo-Cabinet, Tokyo-Tyrant, Memcached, Riak, Scalaris), Column-oriented formate (Google-BigTable, Facebook-Cassandra), Document format(MongoDB, SimpleDB, and CouchDB) and Graph-based format(Neo4j, GraphDB, InfoGrdi)[23][24][25]. For the mining of the big data is done by the programming model because the programming model fulfills the big data processing requirement such as fast data loading, response, query processing, utilized storage space, and dynamic workload adapt with the parallel and distributed environment. Traditional parallel and distributed models such as Message Passing Interface (MPI), and Open Multiprocessing (OpenMP) are not suitable for big data mining because it has low scalability, fault tolerance required, and elasticity. Some popular parallel and distributed based models are MapReduce, Graph processing (Google-Pregel, Dryad, Pig Latin, and GraphLab) and Stream processing (Yahoo- S4, ApacheStorm) [25][26].

D. Clustering

Clustering is the unsupervised method of big data mining in a machine learning perspective based on a statistical approach. The aim of the clustering has grouped the data together based on their characteristics, similarities or features. Clustering is providing high homogeneity within a cluster and reduces the data volume in every group [23]. Clustering is useful for finding hidden relations and pattern between two data objects within the same cluster or different clusters. In big data perceptive, for clustering process, many programming models such as Hadoop, Spark and so on is the divide the data volume in various parts for fast execution with heterogeneous data [27]. In based on machine execution, the Big data clustering algorithms are classified into two groups. First one is single machine based clustering techniques which are capable of executing only one machine data set and the second is multiple machines based clustering techniques which capable of executing multiple machine data set. Both clustering classification has used a partition, hierarchical, density, grid, model, fuzzy and graph based clustering taxonomy based on their working process, behaviors and cluster nature for machine execution [28]. Every clustering techniques are required for cluster creation under big data mining must be dealing with large data volume, high dimensional data, provide to quality clusters from binary, numerical and categorical data attributes, the shape of clusters draw in arbitrary shape, handle to noisy itself due to missing, inaccurate or erroneous of data and every cluster should be easy to understand and unambiguous [29][30].

III. DISTANCE MEASURES TAXONOMY

The definition of distance measure is defined four requirements, where first three axioms define as basic property under the distance measure and fourth axioms define distance metrics for distance measures as under the sub-category [31][32][33][34]. This paper used is *dis*() function for defining distance measures requirements, *dis*() function takes as input two distinct data points A and B, and returns their distance value. Distance measure properties can be specified as follows.

Zero distance or reflexivity: if and only if A is equal to B then distance measure must be equal to zero.

$$dis(A,B) = 0 \qquad \text{(A1)}$$

Non-negativity: if and only if A and B greater than or equal to zero and different for each one then distance measure must be positive.

$$dis(A,B) \geq 0 \qquad \text{(A2)}$$

Symmetry: The distance of A and B must be returned the same distance value as the distance of B and A.

$$dis(A,B) = dis(B,A) \qquad \text{(A3)}$$

Triangle inequality or metric inequality: this property considers the third data point C, the sum of any two side *dis*(*A,B*) or *dis*(*B,C*) distance value must be greater than or equal to the remaining side *dis*(*B,C*) or *dis*(*A,B*) distance value .

$$dis(A,C) \leq dis(A,B) + dis(B,C) \qquad \text{(A4)}$$

In present time various distance measure is available for clustering and these distance measure groups under Minkowski, L(1), L(2), Inner product, Shannon's entropy, Combination, Intersection and Fidelity family[4][14][35]. In this section, the paper describes various distance measures under these families [4-18] [32-35].

A. Minkowski family :- This distance measure gave the distance between data point A and B on the bases of the p-value.

$$dis(A,B) = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right)^{\frac{1}{p}} \qquad \text{(Eq.1)}$$

Manhattan distance: This distance measure is a special case of Minkowski family when the p-value fixes as 1. The Manhattan distance observes the absolute distance between two data items and dimensions. This distance measure is known as a city block distance, absolute value distance, rectilinear distance, taxicab distance or L1 distance family and gives cluster shape as hyper-rectangular [36]. Manhattan distance formulation shows as Eq.2 based upon the Eq.1.

$$dis_{manhattan}(A,B) = \left(\sum_{i=1}^{n} |A_i - B_i|^p\right) \qquad \text{(Eq.2)}$$

Euclidean distance: Euclidean distance is another special case of Minkowski family when the p-value fixes as 2. This distance measure is most commonly used for numerical data and gives to the shortest distance between two data points in the Cartesian coordinate system. Euclidean distance performs well when dataset needs to the deployed compact or isolated cluster and it satisfies all distance measure properties [37]. Euclidean distance formulation shows as Eq.3 based upon the Eq.1 and it also is known as L2 distance family.

$$dis_{euclidean}(A,B) = \sqrt[2]{\sum_{i=1}^{n} |A_i - B_i|^2} \qquad \text{(Eq.3)}$$

Minkowski distance: This distance measure is used for normalization on the p-value between 2 to ∞ for the dataset is organized into an isolated or compacted form and needs to good efficient distance. Minkowski distance formulation shows as Eq.4 based upon Eq.1.

$$dis_{minkowski}(A,B) = \sqrt[p]{\sum_{i=1}^{n} |A_i - B_i|^p} \qquad \text{(Eq.4)}$$

Chebyshev distance: This distance measure is used for when two data points are greatest of their absolute magnitude along with data dimension. Chebyshev distance takes less time for distance calculation between two data points by p-value defined as ∞. This distance measure is also known as chessboard distance, maximum metric, Chebyshev distance minimax approximation or *L*∞ . Chebyshev distance formulation shows as Eq.3 based upon Eq.1.

$$dis_{chebyshev}(A,B) = \sqrt[\infty]{\sum_{i=1}^{n} |A_i - B_i|^\infty} = max_{i=1}^{n} |A_i - B_i|^p \qquad \text{(Eq.5)}$$

B. L(1) family:- The Manhattan distance measure faces two difficulties in the respects of distance value. First one is normalization of a distance value and second is related to figure out of small and large distance. The various solutions are available for removing this difficulty and these solutions are known as variants of the L(1) family. Some proposed solutions are shown as Eq. 6 to Eq. 12 on the bases of Eq. 2. Sorensen suggests, assume that all data points are non-negative and normalize the Manhattan distance by the sum of all data components. After normalization Sorensen distance gives the distance value between 0 and 1. Sorensen distance is also known as Czekanowski distance or Bray Curtis distance and their formulation shows as Eq. 6.

Next Variant is Soergel distance which normalizes the Manhattan distance by choosing the max coefficient data point of the data set and Kulczynski distance normalizes the Manhattan distance by choosing the min coefficient data point of the data set their formulation shows as Eq. 7 and Eq. 8 respectively. Next L(1) family member is Motyka distance measure, which it takes the max data point of the data set and the normalizes through the sum of all data components. Motyka distance formulation shows as Eq. 9.

$$dis_{sorensen}(A,B) = \frac{\left(\sum_{i=1}^{n} |A_i - B_i|\right)}{\left(\sum_{i=1}^{n} |A_i + B_i|\right)} \qquad \text{(Eq.6)}$$

$$dis_{soergel}(A,B) = \frac{\left(\sum_{i=1}^{n} |A_i - B_i|\right)}{\left(\sum_{i=1}^{n} \max(A_i B_i)\right)} \qquad \text{(Eq.7)}$$

$$dis_{kulczynski}(A,B) = \frac{\left(\sum_{i=1}^{n} |A_i - B_i|\right)}{\left(\sum_{i=1}^{n} \min(A_i B_i)\right)} \qquad \text{(Eq.8)}$$

$$dis_{motyka}(A,B) = \frac{\left(\sum_{i=1}^{n} \max(A_i B_i)\right)}{\left(\sum_{i=1}^{n} |A_i + B_i|\right)} \qquad \text{(Eq.9)}$$

Next Manhattan distance variance is Canberra, which is given the absolute difference of the individual data point and normalized them based on its summation. Canberra distance formulation shows as Eq. 10. Next Manhattan distance variance family is Lorentzian, which is based on the natural logarithm and their formulation shows as Eq. 11. L(1) family member Wave-Hedges distance, normalizes the difference of each data pair with its max value. The formulation of this function is shown as Eq. 12.

$$dis_{\text{canberra}}(A,B) = \sum_{i=1}^{n} \frac{|A_i - B_i|}{|A_i + B_i|} \qquad \text{(Eq.10)}$$

$$dis_{\text{lorentzian}}(A,B) = \sum_{i=1}^{n} \ln(1 + |A_i - B_i| \qquad \text{(Eq.11)}$$

$$dis_{\text{wavehedge}}(A,B) = \sum_{i=1}^{n} \frac{|A_i - B_i|}{\max(A_i B_i)} \qquad \text{(Eq.12)}$$

C. L(2) or χ² families:- The L(2) family member based on the Euclidian distance and gives the distance value after normalization. This family member is also known as the variance of the L(2). Some proposed L(2) family member are Matusita, Clark, Divergence, Squared Euclidean, squared χ², Pearson χ² Neyman χ² shown as Eq. 13 to Eq. 19 respectively on the basis of Eq. 3. Matusita distance (Eq. 13) is given a distance for the probability-based data point and eliminates data sparsely problems [38].

$$dis_{\text{matusita}}(A,B) = \sqrt{\sum_{i=1}^{n} (\sqrt{A_i} - \sqrt{B_i})^2} \qquad \text{(Eq. 13)}$$

$$dis_{\text{clark}}(A,B) = \sum_{i=1}^{n} \left(\frac{|A_i - B_i|}{(A_i + B_i)}\right)^2 \qquad \text{(Eq. 14)}$$

$$dis_{\text{divergence}}(A,B) = 2\sum_{i=1}^{n} \frac{(A_i - B_i)^2}{(A_i + B_i)^2} \qquad \text{(Eq. 15)}$$

$$dis_{\text{squared\_euclidean}}(A,B) = \left(\sum_{i=1}^{n} (A_i - B_i)^2\right) \qquad \text{(Eq.16)}$$

$$dis_{\text{squared\_chi}}(A,B) = \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{(A_i + B_i)} \qquad \text{(Eq. 17)}$$

$$dis_{\text{pearson\_chi}}(A,B) = \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{B_i} \qquad \text{(Eq. 18)}$$

$$dis_{\text{neyman\_chi}}(A,B) = \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{A_i} \qquad \text{(Eq. 19)}$$

D. Inner product family:- Another well-known distance measure is an inner product. This measure is also known as a scalar product or dot product when it used to the real data point. If the inner product used to the binary data point, then it's called overlap inner product. It is given to a distance of data point on the basis of multiplication and their formulation shows as Eq. 20. Sometimes it returns a normalized distance value because the distance value not interpreted as large or small. Some proposed solution is available for this type of problem such as Cosine, Jaccard, Dice, and Harmonic mean shown as Eq. 21 to Eq. 24 respectively on the basis of Eq. 20. These all distance measure is known as the variance of the Inner product.

$$dis_{\text{inner\_product}}(A,B) = \sum_{i=1}^{n} A_i B_i \qquad \text{(Eq. 20)}$$

$$dis_{\text{cosine}}(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad \text{(Eq. 21)}$$

$$dis_{\text{Jaccard}}(A,B) = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i B_i} \qquad \text{(Eq. 22)}$$

$$dis_{\text{dice}}(A,B) = 1 - \frac{2\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2} \qquad \text{(Eq. 23)}$$

$$dis_{\text{Harmonic\_mean}}(A,B) = 2\sum_{i=1}^{n} \frac{A_i B_i}{A_i + B_i} \qquad \text{(Eq. 24)}$$

D. Shannon's entropy family:- This family is based on probabilistic uncertainty or entropy. Here all data point A_i must be non-negative and their sum is always equal to 1 and logarithm base is fixed as 2 for distance measure. Formulation of Shannon's distance measure is shown as Eq. 25 and it is also known as Kullback–Leibler divergence (KLD), information deviation and relative entropy. Sometime Kullback–Leibler divergence is given to large distance measures between two data points with non-symmetric norms. The various solutions are proposed for this type of problems such as Jeffreys, K divergence, Topsoe, Jensen-Shannon, and Jensen difference shown as Eq. 26 to Eq. 30 respectively on the basis of Eq. 25. These all distance measure is known as the variance of Shannon's entropy.When the KL divergence measure is used as symmetric with the addition method is known as Jeffreys or J divergence and Topsoe distance or information statistics. When the KL divergence measure is used as symmetric with the average method is known as K divergence. Sometime Topsoe distance obtained as Jensen-Shannon distance measure in the form of symmetric when it is divided by 2. The radius of Shannon's entropy is measured by concavity property so it called Jensen difference distance measure.

$$dis_{\text{kullback\_leibler}}(A,B) = \sum_{i=1}^{n} A_i \ln\left(\frac{A_i}{B_i}\right) \qquad \text{(Eq. 25)}$$

$$dis_{\text{jeffreys}}(A,B) = \sum_{i=1}^{n} A_i - B_i \ln\left(\frac{A_i}{B_i}\right) \qquad \text{(Eq. 26)}$$

$$dis_{\text{k\_divergence}}(A,B) = \sum_{i=1}^{n} A_i \ln\left(\frac{2A_i}{A_i - B_i}\right) \qquad \text{(Eq. 27)}$$

$$dis_{\text{topsoe}}(A,B) = \sum_{i=1}^{n} \left(A_i \ln\left(\frac{2A_i}{A_i - B_i}\right) + B_i \ln\left(\frac{2A_i}{A_i - B_i}\right)\right) \qquad \text{(Eq. 28)}$$

$$dis_{\text{jensen\_shannon}}(A,B) = \frac{1}{2}\left(\sum_{i=1}^{n} A_i \ln\left(\frac{2A_i}{A_i - B_i}\right) + \sum_{i=1}^{n} B_i \ln\left(\frac{2A_i}{A_i - B_i}\right)\right)$$

(Eq. 29)

$$dis_{\text{jensen\_difference}}(A,B) = \sum_{i=1}^{n} \left(\frac{A_i \ln A_i + B_i \ln B_i}{2} - \frac{A_i - B_i}{2} \ln \frac{A_i - B_i}{2}\right)$$

(Eq. 30)

E. Combination family: - This family member defines a distance measure on the bases of a combination of two or more distance measures. This family has two popular members that are Taneja, and Kumar Johnson distance measure. The Taneja distance measure gives the distance on the bases of arithmetic and the geometric mean and Kumar Johnson gives distance on the bases of arithmetic and the geometric mean with symmetric χ2. The formulation of Taneja, and Kumar Johnson distance measure is shown as Eq. 31 and Eq. 32.

$$dis_{\text{taneja}}(A,B) = \sum_{i=1}^{n} \frac{A_i + B_i}{2} \ln\left(\frac{A_i + B_i}{2\sqrt{A_i B_i}}\right) \qquad \text{(Eq. 31)}$$

$$dis_{\text{kumar\_johnson}}(A,B) = \sum_{i=1}^{n} \frac{(A_i^2 - B_i^2)^2}{2(A_i B_i)^{\frac{3}{2}}} \qquad \text{(Eq. 32)}$$

F. Intersection family: - This family member defines a distance measure on the basis of intersection between data points.

The formulation of intersection distance measure is shown as (Eq. 33). Various variants are available for this distance measure such as Wave hedges, Ruzicka, Tanimoto, and their formulation is shown as Eq. 34 to Eq. 36 respectively based on normalization.

$$dis_{\text{intersection}}(A,B) = \sum_{i=1}^{n} \min(A_i B_i) \qquad \text{(Eq. 33)}$$

$$dis_{\text{wavehedges}}(A,B) = \sum_{i=1}^{n} \frac{|A_i - B_i|}{\max(A_i B_i)} \qquad \text{(Eq. 34)}$$

$$dis_{\text{ruzicka}}(A,B) = \frac{\sum_{i=1}^{n} \text{mix}(A_i B_i)}{\sum_{i=1}^{n} \max(A_i B_i)} \qquad \text{(Eq. 35)}$$

$$dis_{\text{tanimoto}}(A,B) = \frac{\sum_{i=1}^{n} \max(A_i B_i) - \sum_{i=1}^{n} \text{mix}(A_i B_i)}{\sum_{i=1}^{n} \max(A_i B_i)} \qquad \text{(Eq. 36)}$$

G. Fidelity family or Squared-chord family:- - This family member defines a distance measure on the basis of geometric means. Fidelity distance measure formulation shows as 37. Some variance such as Bhattacharyya, Hellinger, Matusita, Squared-chord and their formulation is shown as Eq. 38 and Eq. 41 respectively. Bhattacharyya, Hellinger, and Matusita distance measure are related to the probability that reason it gave the distance between 0 and 1. If Matusita distance use without square root it is called Squared-chord distance measure.

$$dis_{\text{fidelity}}(A,B) = \sum_{i=1}^{n} \sqrt{A_i B_i} \qquad \text{(Eq. 37)}$$

$$dis_{\text{bhattacharyya}}(A,B) = -ln \sum_{i=1}^{n} \sqrt{A_i B_i} \qquad \text{(Eq. 38)}$$

$$dis_{\text{hellinger}}(A,B) = \sqrt[2]{1 - \sum_{i=1}^{n} \sqrt{A_i B_i}} \qquad \text{(Eq. 39)}$$

$$dis_{\text{matusita}}(A,B) = \sqrt{2 - 2\sum_{i=1}^{n} \sqrt{A_i B_i}} \qquad \text{(Eq. 40)}$$

$$dis_{\text{squared\_chord}}(A,B) = \sum_{i=1}^{n} \left(\sqrt{A_i} - \sqrt{B_i}\right)^2 \qquad \text{(Eq. 41)}$$

IV. COMPARATIVE ANALYSIS OF DISTANCE MEASURES FOR BIG DATA MINING

In the second section of this paper are studied 41 distance measures (Eq.1-Eq.41) under the eight groups. After that this section, paper recognized which distance measure fulfills basic properties (A1-A5) of distance measures with big data dimensions and summarized them in Table 1. Similarity and dissimilarity are two basic properties of cluster construction. The distance function is used for creating a cluster on the basis of the relationship between data points and dimensions with efficiency and similarity function used for creating a cluster on the basis of the data features and natures [29][30]. The distance measure is also known as dissimilarity. If any distance measure satisfying the all distance measure properties such as reflexivity, non-negativity, symmetry, and triangular inequality are known as dissimilarity or distance function otherwise it is known as Similarity function. [32][33]. Some distance measure is not cover to only triangular inequality, but satisfying all other properties, these types of distance measure are known as Semi-metric (distance or similarity function) [39]. For existing research perspective, Table 1 shows which distance measure satisfying the distance axiom or not and this table gave the guidance for distance measure selection under the big data mining on the basis of dissimilarity and similarity function. In the big data mining perspective, every distance function is dealt with high volume, high variety, and high-velocity data environment.

If two data point distance is gathered to a small value that defines a data point has a high relation and they are close to each other otherwise they have high relation and known as similarity [33]. In general, the similarity function is used for categorical data because it's given to small distance value and the dissimilarity function is used for continuous data because it's given to the large distance value. Sometimes data are organized as mixed-mode (a combination of categorical and continuous) then dichotomize is used for separation of categorical and continuous data. According to dichotomize, first used a similarity function for categorical data and rescaling all the data points. After that, replace same data on the basis of their ranks and used to the dissimilarity function for continuous data. If any similarity and dissimilarity function is used for the correlation coefficient value so it is unable to measure the huge size observation or data point. [32][33][34].

Fahad et al., 2014 and Pandove et al., 2015 describes Volume, Velocity, and Variety criteria for designing the clustering algorithm under big data mining. Volume related criteria defines the cluster is must be dealt huge size, high dimensional and noisy of the dataset, Variety related criteria defines the cluster is must be recognized as dataset categorization and clusters Shape, and Velocity related criteria defines the complexity, scalability, and performance of the clustering algorithm during the execution of real dataset. This paper takes for distance measures perspective Volume as deal with the high scale dataset, Variety deal with a data type of clustering, and Velocity deal with a time complexity for identification of big data enable distance measures [29][30].

Distance measure	A1	A2	A3	A4	Type of function	Volume	Variety	Velocity
Minkowski family								
Eq. 2	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 3	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 4	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 5	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
L(1) family								

Eq. 6	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 7	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 8	Yes	Yes	Yes	No	Semi-metric	Medium	Categorical	0(n)
Eq. 9	No	Yes	Yes	Yes	similarity	Medium	Categorical	0(n)
Eq. 10	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 11	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 12	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
L(2) or χ 2 family								
Eq. 13	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 14	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(n)
Eq. 15	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Eq. 16	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Eq. 17	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 18	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 19	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Inner product family								
Eq. 20	Yes	No	Yes	Yes	Similarity	Medium	Categorical	0(3n)
Eq. 21	Yes	No	Yes	Yes	Similarity	Medium	Categorical	0(3n)
Eq. 22	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 23	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 24	Yes	No	Yes	Yes	Similarity	Large	Categorical	0(3n)
Shannon's entropy family								
Eq. 25	Yes	No	No	No	Similarity	Medium	Categorical	0(n)
Eq. 26	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Eq. 27	Yes	No	No	No	Similarity	Medium	Categorical	0(n)
Eq. 28	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Eq. 29	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Eq. 30	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(n)
Combination family								
Eq. 31	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Eq. 32	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Intersection family								
Eq. 33	Yes	No	No	No	Similarity	Medium	Categorical	0(2n)
Eq. 34	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(2n)
Eq. 35	Yes	No	No	No	Similarity	Medium	Categorical	0(2n)
Eq. 36	Yes	Yes	Yes	No	Semi-metric	Large	Categorical	0(2n)
Fidelity family or Squared-chord family								
Eq. 37	Yes	No	No	No	Similarity	Medium	Categorical	0(3n)
Eq. 38	Yes	Yes	Yes	Yes	Distance	Medium	Continuous	0(3n)
Eq. 39	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(3n)
Eq. 40	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(3n)
Eq. 41	Yes	Yes	Yes	Yes	Distance	Large	Continuous	0(3n)

Table 1- Summary of the distance measure properties with big data dimensions [4-18][32-38]

According to table 1 Eq. 2, Eq. 3, Eq. 4, Eq. 5, Eq. 7, Eq. 10, Eq. 11, Eq. 12, Eq. 13, Eq. 14, Eq. 34, Eq. 38, Eq. 39, Eq. 40, Eq. 41 is more suitable and universal for cluster creation under the big data environment and other equation used for comparing two clusters on the basis of similarity. [34]

V. Distance Measures Enable Clustering Taxonomy

The Clustering algorithms are usually classified into the Partition, Hierarchical, Density, Grid, Model, Fuzzy and Graph based clustering taxonomy based on their working process, behaviors and cluster nature. They're basic of cluster creation, data type and distance measurements are given in section A to G inside of this section. [25-30][32-34][39-41]

A. Partitioning based method

This clustering method firstly, user preset the number of clusters and keep the dataset according to their choice, after that relocating the data point from one cluster to another cluster based on the center or mean. This clustering method used distance function as finding the center. Here distance measures work only numerical data but some time creates a cluster according to the categorical data. If data sets have categorical data or a combination of categorical and numerical data then needs to be dichotomized for distance calculation. Some typical algorithms of this kind of clustering algorithm are K-Mean, K-Medoids, K-parameter PAM(Partitioning around medoids), CLARA (Clustering Large Applications), and CLARANS (Clustering Large Applications based upon Randomized Search).

B. Hierarchical based clustering

This clustering method constructs the clusters by recursively splitting the dataset using Agglomerative or Divisive method. Agglomerative method (bottom-up), initially assigns each data into own cluster and start merge operation until the desired cluster is obtained. Divisive method (top-down), initially assign all data into the one cluster and start splitting the cluster into sub cluster until the desired cluster is obtained. The result of both methods is recognized by the two-dimensional diagram as a dendrogram. When performing the merging and division operation for obtaining the desired cluster, it is used some similarity function and normalized distance value. Merging and division of the clusters are done by using the minimum distance (nearest neighbor), maximum distance (farthest neighbor), and average distance method between two cluster members. Some typical algorithms of this kind of clustering algorithm are BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), CURE (Clustering Using Representatives), ROCK (RObust Clustering uses links), Chameleon, ECHIDNA, WARDS, and SNN.

C. Density-based clustering

This clustering method scans the whole spatial databases at one time and automatically detects the cluster on the basis of probability distribution based distance measure. Here distance measurements are used for calculating the core, border and noise point for creating the dense cluster. Some typical algorithms of this kind of clustering algorithm are DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points To Identify the Clustering Structure), Mean-Shift, DENCLUE (DENsity based CLUstEring), and GDBSCAN.

D. Model-based Clustering

This clustering method tries to optimize the dataset according to the mathematical model. Here probability distributions based distance measure used for defining the different type of mathematical model and measuring the parameter of a selected mathematical model for cluster creation. Some typical algorithms of this kind of clustering algorithm are COBWEB, SLINK, SOM (Soft-Organizing feature Map), ART, and EM (Expectation Maximization).

E. Grid-based clustering

This clustering method splits the data space into a finite number of cells in the form of the grid structure. Here distance measure constructing the grid structure, calculating the cell density, and identifying the cluster centers. Some typical algorithms of this kind of clustering algorithm are STING (statistical information Grid approach), CLIQUE, Wave Cluster, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, and STIRR (Sieving through Iterated Relational Reinforcement).

F. Fuzzy-based clustering

This clustering method used a membership function for constructing the cluster. These membership function same as similarity measure if creating a cluster on the basis of the distance between 0 and 1 otherwise it's used to dissimilarity measure. Some typical algorithms of this kind of clustering algorithm are FCM (Fuzzy C-Means), FCS and MM.

G. Graph-based clustering

This clustering method first constructs a graph and then applies a clustering algorithm to partition the graph through a similarity function. Spectral clustering is a variance of Graph-based clustering, which is used for Web-ranking analysis, image segmentation and dimension reduction. Some typical algorithms of this kind of clustering algorithm are CLICK and MST.The second section of this paper described the continuous and categorical data type with their distance measure. If the clustering algorithm is used for the continuous data that is used to dissimilarity function and if the clustering algorithm is used for the categorical data that is used to similarity function. Table 2 shows clustering taxonomies with their distance measure and data type for designing the clustering algorithm under the big data mining [29-30][32-35].

IV. AN EMPIRICAL ANALYSIS OF DISTANCE FUNCTION

In the big data mining perspective, the good clustering algorithm is constructing the clusters in the arbitrary shape, handling to a real dataset with heterogeneous data sources, and easy to scalable according to the datasets, and able to detect and remove any type of outliers in the cluster [34]. Partitioning based method is very efficient for cluster creation in the large and high-dimensional dataset because it has own objective function and goal to minimize them. This method constructs the cluster on convex shapes because it's based on the center point. In addition, the partitioning method is faster than the hierarchical clustering. The density based method is more efficient for arbitrary shapes with spatial dataset. Density-based and grid-based approaches are very useful for creating the cluster in multidimensional dataset space [32].

Clustering algorithm	Distance measure	Volume	Data type (Variety)	Scalability (Velocity)
Partitioning based method	Dissimilarity	Large	Both type or mix	Yes
Hierarchical based method	Similarity	Medium	Categorical	Yes
Density based method	Similarity	Large	Categorical	Yes
Model based method	Dissimilarity	Medium	Both type or mix	No
Grid based method	Dissimilarity	Large	Both type	Yes
Fuzzy based method	Similarity	Large	Continuous	No
Graph based method	Similarity	Medium	Categorical	No

Table 2 Clustering Taxonomy with their Distance measure with big data dimensions

The big data clustering is managing the huge data set with the heterogeneous data type, deal with high dimensional, parallel computation, multidimensional dataset space, scalability and so on. Some typical algorithms of big data clustering are K-means, BIRCH, CLARA, CURE, CLARANS, DBSCAN, and DENCLUE [29]. In this section, the paper describes the distance measure efficiency and tries to find out which distance function is more suitable in big data mining. To achieve distance measure efficiency objective, this paper takes the partitioning based K-Means algorithm and applies all distance measure (Table 1) for cluster constructions. K-Means clustering has the capability for handling large dataset execution with scalability and their variance K-parameter handles mix data type dataset. K-Means algorithm is the top second algorithm for the data mining techniques [42].

Accuracy is the basic factor of validating any research. In this section, the paper finds out the accuracy of all discussed distance measures with the help of Anuran Calls (MFCCs) Data Set and K-Mean clustering algorithm for single machine based big data mining. Anuran Calls (MFCCs) Data Set consist of 7195 real data points with 22 dimensions [43]. This experiment is used in R and RHadoop and systems are configured with an Intel I3 processor, 4 GB DDR3 RAM, 320 GB hard disk and Operating system used are windows 7. Table 3 shows the accuracy of Eq. 2 to Eq. 41 distance measures for selecting distance measures in the big data environment and their graphical accuracy view show as Figure 2. To achieve clustering accuracy, this paper used clustering confusion matrix after cluster creation to help of Eq. 2 to Eq. 41 distance measures.

Distance Measure	**Cluster 1 Element**	**Cluster 2 Element**	**Cluster 3 Element**	**Unbalance Element**	**Accuracy (%)**
Minkowski family					
Eq. 2	4733	539	1675	248	96.55316
Eq. 3	4710	521	1616	348	95.16331
Eq. 4	4690	513	1602	390	94.57957
Eq. 5	4692	520	1602	381	94.70466
L(1) family					
Eq. 6	4684	623	1223	665	90.75747
Eq. 7	4684	624	1284	603	91.61918
Eq. 8	4702	683	1203	607	91.56359
Eq. 9	4694	678	1236	587	91.84156
Eq. 10	4693	678	1247	577	91.98054
Eq. 11	4694	689	1287	525	92.70327
Eq. 12	4693	678	1278	546	92.4114
L(2) or χ 2 family					
Eq. 13	4704	426	1556	509	92.92564
Eq. 14	4704	422	1556	513	92.87005
Eq. 15	4702	419	1554	520	92.77276
Eq. 16	4704	423	1556	512	92.88395
Eq. 17	4698	423	1556	518	92.80056
Eq. 18	4698	423	1554	520	92.77276
Eq. 19	4698	421	1554	522	92.74496
Inner product family					
Eq. 20	4654	423	1503	615	91.4524
Eq. 21	4654	422	1503	616	91.4385
Eq. 22	4669	422	1511	593	91.75817
Eq. 23	4669	421	1504	601	91.64698
Eq. 24	4654	422	1503	616	91.4385
Shannon's entropy family					
Eq. 25	4462	484	1536	713	90.09034
Eq. 26	4462	483	1531	719	90.00695
Eq. 27	4458	483	1539	715	90.06254
Eq. 28	4459	482	1536	718	90.02085
Eq. 29	4459	483	1534	719	90.00695

Eq. 30	4462	483	1531	719	90.00695
Combination family					
Eq. 31	4669	421	1504	601	91.64698
Eq. 32	4669	422	1503	601	91.64698
Intersection family					
Eq. 33	4654	422	1519	600	91.66088
Eq. 34	4652	439	1518	586	91.85546
Eq. 35	4658	422	1518	597	91.70257
Eq. 36	4654	421	1518	602	91.63308
Fidelity or Squared-chord family					
Eq. 37	4623	524	1528	520	92.77276
Eq. 38	4628	526	1524	517	92.81445
Eq. 39	4637	518	1527	513	92.87005
Eq. 40	4637	517	1527	514	92.85615
Eq. 41	4619	512	1524	540	92.49479

Table 3 Accuracy of Distance measure for using K-means algorithm (K=3)

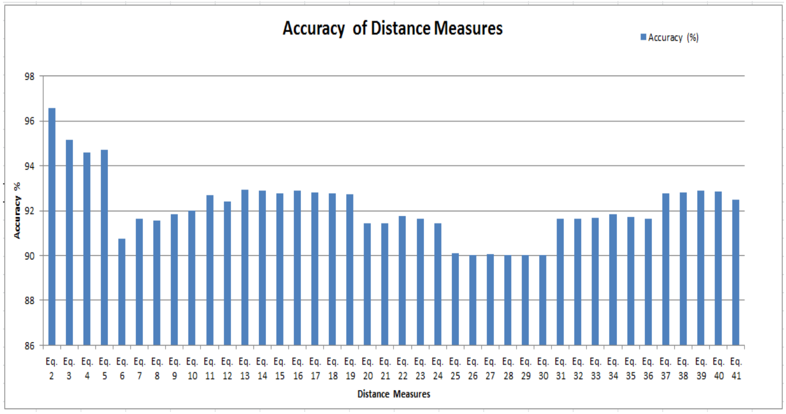

Figure 2 Graphical View of Accuracy of Distance measure

VI. CONCLUSION

This paper reviewed the basic idea of big data, big data mining, big data storages, clustering, and distance taxonomy with respect to the big data dimensions, and defined which distance measures are suitable under the big data mining. The first section describes the natures of big data and distance measures. The second section of this paper gives the background of big data, big data mining, big data storages for mining and clustering. The third section describes distance measures under the Minkowski, L(1), L(2), Inner product, Shannon's entropy, Combination, Intersection, and Fidelity family. The fourth section is identified Eq. 2, Eq. 3, Eq. 4, Eq. 5, Eq. 7, Eq. 10, Eq. 11, Eq. 12, Eq. 13, Eq. 14, Eq. 34, Eq. 38, Eq. 39, Eq. 40, Eq. 41 distance measures are more suitable and scalable for big data mining because it fulfills to big data dimensions and distance property. The fifth section gives details about the cluster creation basis of the distance measures and summarized clustering taxonomy based upon 3 V's of big data such as volume, variety, and velocity criteria with distance measures. The sixth section is important for a practical point of view. This section validates the distance measures accuracy with the help of partition based K-Means algorithm and Anuran Calls (MFCCs) Data Set. This paper finds out distance Eq. 2, Eq. 3, Eq. 4, and Eq. 5 respectively given high accuracy for cluster creation in the all studied distance measures.

REFERENCES

1. Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303. doi:10.1016/j.bushor.2017.01.004
2. Siddiqa, A., Hashem, I. A., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. Journal of Network and Computer Applications, 71, 151-166. doi:10.1016/j.jnca.2016.04.008
3. Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall. ISBN 9780130222787
4. Cha, S. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, 4(1), 300-307. doi:10.1109/icpr.2000.906010
5. Abudalfa, S. I., & Mikki, M. (2013). K-means algorithm with a novel distance measure. Turkish Journal Of Electrical Engineering & Computer Sciences, 21, 1665-1684. doi:10.3906/elk-1010-869
6. Lin, Y., Jiang, J., & Lee, S. (2014). A Similarity Measure for Text Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, 26(7), 1575-1590. doi:10.1109/tkde.2013.19
7. Chim, H., & Deng, X. (2008). Efficient Phrase-Based Document Similarity for Clustering. IEEE Transactions on Knowledge and Data Engineering, 20(9), 1217-1229. doi:10.1109/tkde.2008.50

8. Wang, X., Yu, F., & Pedrycz, W. (2016). An area-based shape distance measure of time series. Applied Soft Computing, 48, 650-659. doi:10.1016/j.asoc.2016.06.033

9. Tavakkol, B., Jeong, M. K., & Albin, S. L. (2017). Object-to-group probabilistic distance measure for uncertain data classification. Neurocomputing, 230, 143-151. doi:10.1016/j.neucom.2016.12.007

10.Liu, H., Zhang, X., Zhang, X., & Cui, Y. (2017). Self-adapted mixture distance measure for clustering uncertain data. Knowledge-Based Systems, 126, 33-47. doi:10.1016/j.knosys.2017.04.002

11.Ramya, R., & Sasikala, T. (2018). A comparative analysis of similarity distance measure functions for biocryptic authentication in cloud databases. Cluster Computing. doi:10.1007/s10586-017-1568-y

12.Marcon, E., & Puech, F. (2017). A typology of distance-based measures of spatial concentration. Regional Science and Urban Economics, 62, 56-67. doi:10.1016/j.regsciurbeco.2016.10.004

13.Moghtadaiee, V., & Dempster, A. G. (2015). Determining the best vector distance measure for use in location fingerprinting. Pervasive and Mobile Computing, 23, 59-79. doi:10.1016/j.pmcj.2014.11.002

14.Kocher, M., & Savoy, J. (2017). Distance measures in author profiling. Information Processing & Management, 53(5), 1103-1119. doi:10.1016/j.ipm.2017.04.004

15.Ikonomakis, E. K., Spyrou, G. M., & Vrahatis, M. N. (2019). Content driven clustering algorithm combining density and distance functions. Pattern Recognition, 87, 190-202. doi:10.1016/j.patcog.2018.10.007

16.Merigó, J. M., Casanovas, M., & Zeng, S. (2014). Distance measures with heavy aggregation operators. Applied Mathematical Modelling, 38(13), 3142-3153. doi:10.1016/j.apm.2013.11.036

17.Grant, J., & Hunter, A. (2017). Analysing inconsistent information using distance-based measures. International Journal of Approximate Reasoning, 89, 3-26. doi:10.1016/j.ijar.2016.04.004

18.Weller-Fahy, D. J., Borghetti, B. J., & Sodemann, A. A. (2015). A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection. IEEE Communications Surveys & Tutorials, 17(1), 70-91. doi:10.1109/comst.2014.2336610

19.Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007

20.Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286. doi:10.1016/j.jbusres.2016.08.001

21.Yaqoob, I., Hashem, I. A., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. International Journal of Information Management, 36(6), 1231-1247. doi:10.1016/j.ijinfomgt.2016.07.009

22.Oussous, A., Benjelloun, F., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431-448. doi:10.1016/j.jksuci.2017.06.001

23.Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19(2), 171-209. doi:10.1007/s11036-013-0489-0

24.Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable Big Data: A survey. Computer Science Review, 17, 70-81. doi:10.1016/j.cosrev.2015.05.002

25.Weichen, W. (2016). Survey of Big Data Storage Technology. Internet of Things and Cloud Computing, 4(3), 28-33. doi:10.11648/j.iotcc.20160403.13

26.Chong, D., & Shi, H. (2015). Big data analytics: A literature review. Journal of Management Analytics, 2(3), 175-201. doi:10.1080/23270012.2015.1082449

27.Khan, N., Yaqoob, I., Hashem, I. A., Inayat, Z., Ali, W. K., Alam, M., . . . Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. The Scientific World Journal, 2014, 1-18. doi:10.1155/2014/712826

28.Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014). Big Data Clustering: A Review. In Murgante B. et al. (eds), International Conference on Computational Science and Its Applications (Vol. 8583, Lecture Notes in Computer Science, pp. 707-720). Springer. doi:10.1007/978-3-319-09156-3_49

29.Pandove, D., & Goel, S. (2015). A comprehensive study on clustering approaches for big data mining. In Proceedings of IEEE 2nd International Conference on Electronics and Communication Systems (pp. 1333-1338). IEEE Xplore Digital Library. doi:10.1109/ecs.2015.7124801

30. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., . . . Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 2(3), 267-279. doi:10.1109/tetc.2014.2330519

31.Nadler, M., & Smith, E. P. (1993). Pattern recognition engineering. New York: John Wiley & Sons, ISBN-13: 978-0471622932

32. Gan, G., Ma, C., & Wu, J. (2007). Data clustering: Theory, algorithms, and applications. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics.

33.Everitt, B. S. (2011). Cluster Analysis (5th ed., Wiley series in probability and statistics). Southern Gate, Chichester, West SussexUnited Kingdom: John Wiley & Sons.ISBN: 978-0-470-74991-3

34.Aggarwal, C. C., & Reddy, C. (2014). Data Clustering Algorithms and Applications. CRC Press Taylor & Francis Group.ISBN 978-1-4665-5822-9

35.Manning, C. D. , Raghavan, P. , & Schütze, H. (2008). Introduction to information retrieval . Cambridge: Cambridge University Press.

36.Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE, 10(12), doi:10.1371/journal.pone.0144059

37.Kumar, V., Chhabra, J. K., & Kumar, D. (2013). Impact of Distance Measures on the Performance of Clustering Algorithms. Intelligent Computing, Networking, and Informatics Advances in Intelligent Systems and Computing, 183-190. doi:10.1007/978-81-322-1665-0_17

38. Selvi, C., & Sivasankar, E. (2018). A novel similarity measure towards effective recommendation using Matusita coefficient for Collaborative Filtering in a sparse dataset. Sādhanā, 43(12). doi:10.1007/s12046-018-0970-3

39.Kaur P., Kaur K. (2017). Comparative Study of Techniques and Issues in Data Clustering. In Saini H., Sayal R., Rawat S. (eds), Innovations in Computer Science and Engineering(Vol. 8, Lecture Notes in Networks and Systems,pp. 1-7). Springer, Singapore. doi:10.1007/978-981-10-3818-1_1

40.Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In Kogan J., Nicholas C., Teboulle M. (eds), Grouping Multidimensional Data (pp. 25-71). Berlin, Heidelberg: Springer. doi:10.1007/3-540-28349-8_2

41.Chen, W., Oliverio, J., Kim, J. H., & Shen, J. (2018). The Modeling and Simulation of Data Clustering Algorithms in Data Mining with Big Data. Journal of Industrial Integration and Management, 12(4), 1-16. doi:10.1142/s2424862218500173

42.Dave, M., & Gianey, H. (2016). Different clustering algorithms for Big Data analytics: A review. In Proceedings of IEEE International Conference System Modeling & Advancement in Research Trends (pp. 328-333). IEEE Xplore Digital Library. doi:10.1109/sysmart.2016.7894544

43.Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

AUTHORS PROFILE

Kamlesh Kumar Pandey Kamlesh Kumar Pandey is pursuing a Ph.D. from Dr. HariSingh Gour Vishwavidyalaya (A Central University), Sagar, India, under the supervision of Prof. Diwakar Shukla. Currently, He is doing research on the design of Big Data Mining algorithms with respect to three dimensions of Big Data. He is the author and co-author of several research papers in International journals and conference such as IEEE, Springer, and others. He has 6 years of teaching and research experience. He awarded Training of Young Scientist in 34th M.P. Young Scientist Congress.

Prof. Diwakar Shukla is presently working as HOD in the Department of Computer science and applications, and Dean in the School of Mathematical and Physical Sciences, Dr. HariSingh Gour Vishwavidyalaya, Sagar, India, and has over 25 years' experience. He obtained M.Sc.(stat.), Ph.D.(stat.) degrees from Banaras Hindu University, Varanasi and served the Devi Ahilya University, Indore, M.P. as a permanent Lecturer from 1989 for nine years and obtained the degree of M.Tech.(Computer Science) from there. He joined Dr. HariSingh Gour Vishwavidyalaya, Sagar as a Reader in statistics in the year 1998. During Ph.D. from BHU, he was junior and senior research fellow of CSIR, New Delhi through Fellowship Examination (NET) of 1983. Till now, he has published more than 75 research papers in national and international journals and participated in more than 35 seminars/conferences at the national level. He also worked as a Professor in the Lucknow University, Lucknow, U.P., for one (from June 2007 to 2008) year and visited abroad to Sydney (Australia) and Shanghai (China) for conference participation and paper presentation. He has supervised fourteen Ph.D. theses in Statistics and Computer Science and seven students are presently enrolled for their doctoral degree under his supervision. He is the author of two books. He is a member of 11 learned bodies of Statistics and Computer Science at the national level. The area of research he works for are Sampling Theory, Graph Theory, Stochastic Modeling, Data mining, Big Data, Operation Research, Computer Network, and Operating Systems.