International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

A Hybrid Data Clustering Technique in Big Data using Machine Learning

K. Sharma[1*], P. Rehan[2]

[1,2]Dept. of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

[*]*Corresponding Author: kapil4ravit@gmail.com, Tel.: +91-8283809270*

DOI: https://doi.org/10.26438/ijcse/v8i1.4047 | Available online at: www.ijcseonline.org

Accepted: 13/Jan/2020, Published: 31/Jan/2020

Abstract— Big Data refers to a huge collection of data like the Banking data, social media data, repository data etc. These types of fields are responsible for day to day relevant data retrieval and processing. Clustering is one of major tasks which are done for data in order to minimize the time delay and efficient information retrieval. In this work we worked on similarity index in the form of cosine and soft cosine to count the total connection with respect to documents in the form of data. Then we use Cosine and Soft Cosine measures as hybrid Similarity algorithm to intakes the threshold policy of K means and co relation linkage property of Linkage clustering and forms new clusters. The cross-validation of the proposed work model has been done using Support Vector Machine followed by K-Mediod to improve the accuracy of clustering. This research work also focuses on different techniques of Clustering as well as classification. This research work mainly focuses on optimizing the clustering performance of the Big Data so that wealthy information can be retrieved with least cost.

Keywords— Data mining, Big data, Clustering, Classification, Support Vector Machine

I. INTRODUCTION

I.I Big Data

Big data actually implies an assortment of very large datasets which cannot be processed easily by implementing traditional computing methods. Big data is not merely a data, it has rather transformed into a comprehensive topic, which encompasses number of tools, procedures and frameworks. In general, big data is a datasets that could not be observed, attained, managed, and administered with hold-style IT and software/hardware components within a bearable period. Big Data technologies pronounce a novel origination of equipment's and constructions, designed to assist numerous organizations to cautiously abstract value from very huge bulks of a widespread diversity of the data by facilitating high-velocity acquisition, innovation, and/or exploration [1]. This realm of Big Data have need of a modification in the computing manner, so that the clients can control both the data saving necessities and the hefty server processing needed to economically evaluate massive volumes of data. Much of this data explosion is the consequence of an intense rise in the equipment's that are sited at the border of the network comprising implanted sensors, smart phones, and tablet computers. This data produces novel chances to "abstract more value" in healthcare, human genomics, funding, oil and gas, exploration, investigation, and several other regions.

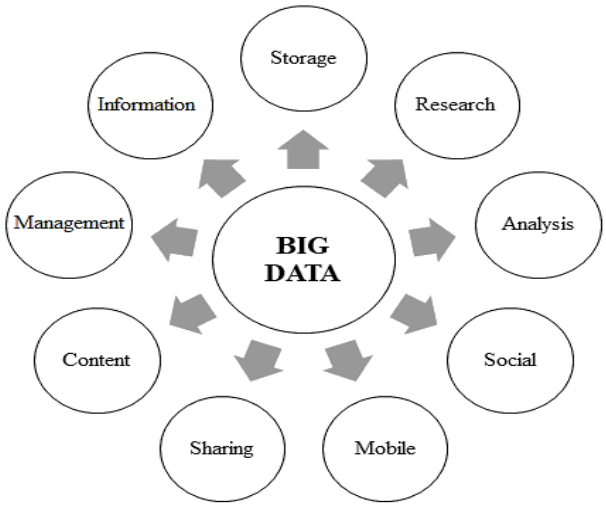

Figure 1: Representing Big Data

Presently, data has developed into significant production element which may perhaps be similar to physical resources and human wealth. Since, the social media, multi-media and IoT are emerging; enterprises will accumulate further facts, leadings to an exponential rise of data volume [2]. Big data will 1have an enormously growing potential in generating values for big business and users.

The three significant constituents for big data that are also acknowledged usually as 3 V's of big data are defined as follows [3]:

- ***Volume*** - currently the data extent is considerably larger in contrast to past data sizes, i.e. exceeding - terabytes and petabytes. The striking range and gradual surge in the data size tags it vigorously hard to save and review by employing the conventional approaches. For instance, Facebook consumes approximately 500-terabytes of data on a daily basis.
- ***Velocity*** - the deployment of the big data is must as it streams the data to obtain the optimum use of its value for time restricted processes.
- ***Variety*** - origination of the big data is primarily based on the diversity of sources. The Conventional systems of databases were proposed to mark lower extents of classified data, smaller amount updates or a steady and feasible data arrangement. However, the spatial data, 3-D data, audio-video, and the cluttered manuscript, comprising account files and social media are also considered as big data.

I.II Data Mining

Data mining is generally pronounced as the method of inspecting the data from different aspects and extracting it into some valued evidence, which can be further used to increase the revenues, budget cuts, etc. It has been considered as most extensively utilized field and merges several researches in areas such as Machine learning, Artificial Intelligence (AI), databases, etc. Previously, the load of mining the valued facts from the categorized data has been accomplished by the analysts. However, due to the ever-increasing volume of data in the current businesses and sciences, there is an emerging requirement of the computer-based provisions for this assignment. Since, the data sets have established in extent and intricacy, the analyst works over more difficult and problematical tools. The whole procedure of employing the computer-based approach, encompassing separate means for knowledge discovery from the listed data, is frequently termed as data mining [3].

1.1.1 Data Mining Techniques

The utmost significant goal of any data mining process is to discover valuable knowledge which is simply implied in large data sets. There are five essential classes of errands that are involved with data mining:

1) **Association Rule** - allows detection of interdependencies in the midst of distinct variables in huge databases. It permits revealing of the concealed patterns contained by the data acquired from the huge databases, that can be used for identifying the variables inside the data and the co-occurrences of diverse variables which appear with the highest frequencies
2) **Clustering Analysis**- usually explained as a rule of identifying data sets that are comparable in order to be familiar with the dissimilarities along with the similarities inside the database. The clusters have explicit features in general which can be employed further to improve targeting algorithms.
3) **Classification Analysis** - is an efficient rule for accomplishing noteworthy and appropriate information concerning the data and metadata. The classification analysis supports in differentiating that to which of a set of groups, diverse varieties of data fit in. It is strictly linked to cluster analysis as the classification can be applied to cluster data.
4) **Regression Analysis** - attempts to describe the reliance between the variables. It undertakes a one-way unusual result from single variable to the response of another variable. Liberated variables can be persuaded by each other however it does not deduce that this reliance is mutually as is the case with relationship analysis. A regression analysis can disclose that one variable is reliant on another but not vice-versa.
5) **Anomaly or Outlier Detection** -leads to the quest for data objects in a dataset that do not bear a resemblance to a proposed pattern or supposed behavior. Anomalies are correspondingly called as outliers, exceptions, surprises or pollutants and they usually extract critical and actionable information. An outlier is an entity which deviates considerably from the usual average within a dataset or an association of data. It is statistically alienated from the rest of the data and consequently, the outlier implies that something is out of the usual and demands further study.

I.III Clustering in Data Mining

Clustering is basically the process of grouping the entities in different classes called as clusters comprising similar objects. A cluster is group of entities that are similar and belongs to the same class. In the process of clustering, firstly the set of data is partitioned in groups on the basis of data similarity and afterward labels are assigned to those groups. Clustering is generally used in several applications such as data analysis, pattern recognition, image processing, etc. [4]. Clustering helps the marketers to determine different groups in their customer base, on the basis of which they can distinguish their customers group based on the buying history. Clustering is very useful in detecting the credit card fraud. It also helps in classifying the documents on the web for the purpose of information finding.

In data mining, clustering functions as a means to achieve awareness into the scattering of data to witness the features of each cluster. Clustering plays an incomparable share in quite a few applications of the data mining i.e. information retrieval and scientific data exploration, spatial database applications, text mining. The articles of investigation in the clustering procedure could be people, wages, views, software units and many others. The features of such articles required to be presented carefully as these features are the primary variables of the problem and their selection considerably affects the products of clustering algorithm [5].

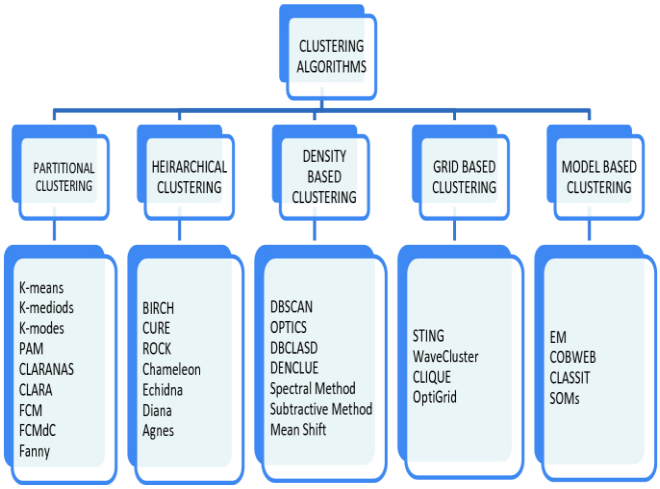

Figure 2: Different types of Clustering Algorithms

I.IV Classification in Data Mining

Classification is a method that organizes the unstructured data into structured classes or groups, which in-turn helps the customers in knowledge discovery and furthermore provides intellectual decision making. There are two stages in classification, i.e. training and testing. Training is a learning process stage in which huge training data sets are provided and investigation takes place then instructions and designs are generated. Afterwards, execution of the second phase initiates. Assessment or testing of data sets and archives the precision of a classification patterns. The classification techniques are employed in big data to classify the data sets according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyse and store. Classification procedure can be categorized into five basic categories on the basis of diverse mathematical models i.e. rule-based, distance-based, decision tree-based, statistical-based, and neural network-based. The classification analysis helps to identify which type of data belongs to which category.

The following are the types of supervised learning classification algorithms. These algorithms are initially studied well and then carefully chosen to fulfil the purpose of this research work. A brief introduction and their working are presented as follows:

I.V Support Vector Machine

In the supervised machine learning environment the training data consist of a set of training examples, where each example is a pair consisting of an input and an anticipated output value. A supervised learning algorithm examines the training data and then predicts the correct output categorization for given data-set input. Support Vector Machine (SVM) was introduced by Vapnik in 1979 and it is primarily defined as a supervised learning approach that utilizes a subcategory of training point which is also acknowledged as support vectors in order to categorize diverse entities [6]. SVM basically finds the best possible linear decision surface concerning the two classes. The biased arrangement of the support vectors is generally known as decision surface. In other words, the nature of the margin among the two classes is decided by the support vectors.

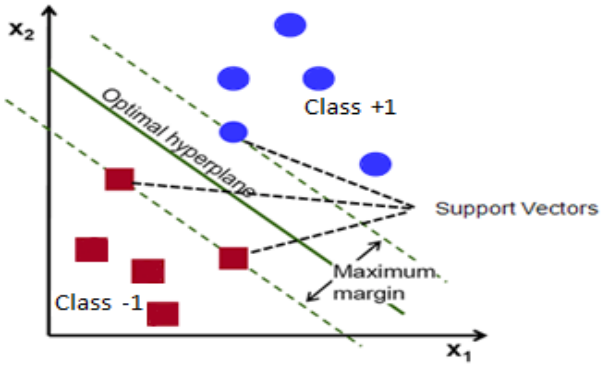

Figure 3: Support Vector Machine's architecture

SVM mainly determines a hyperplane or a surface that surge the margin between the positive and negative annotations, for a predefined class. From the figure shown above, it is clear that the red boxes and blue circle are the support vectors that are generally the observations which maintain hyperplane on both sides [7]. Primarily, the hyperplane is well-defined as a line in 2D, however in 3D, it is defined as plane. In the bigger dimensions i.e. above3D, it is labeled as hyperplane. Margin is basically the gap between hyperplane and nearest data point. However, if that gap is doubled, then, it would be equal to the margin. Support Vector Machines are flexible for diverse decision function. SVMs are quite useful when very large dimensional gaps are concerned. Moreover, SVM's also seemed to be quite useful in case where the number of dimensions turn out to be bigger than the current number of samples.

II. RELATED WORK

R. Tamilselvi et al [2013], provided an overview of the data mining, types of data mining and different application involved with the data mining. Data mining is also called as knowledge discovery which discovers the patterns and delicate relationships in data and infers rule that allows future predictions. Afterwards, several tasks such as Association rules, Decision tree, clustering, and certain related algorithms are included along with the merits and demerits of those related algorithms. Finally, the different application domains of the decision trees and clustering algorithms are presented. This survey has not covered all classification techniques for understanding the estimate of the accuracy of the classification rules. [8]

Adil Fahad et al [2014], presented the notions and algorithms associated with the clustering, altogether with a brief survey on the up-to-date algorithms of clustering as well as their comparative analysis based on both theoretical and empirical perception. A categorizing framework is developed on the basis of the properties that are highlighted in the previous researches. Afterwards, experiments are conducted comparing the most representative algorithm from all categories utilizing large number of real data sets. The effectiveness of the algorithm is measured through several performance metrics. In conclusion the best performing clustering algorithms for the big data are highlighted in this work. The investigations of ensembles for single and individual clustering by means for the accuracy and stability has not been covered. The author has not defined the setting of parameters for every clustering algorithm. [9]

Lisbeth Rodríguez-Mazahua et al [2016], offered a comprehensive analysis of Big Data mechanisms for identification of the primary complications, tools, presentation area and evolving classes of Big Data. The study within the field of big data is emerging immensely these days. In order to meet the objective of this study, the authors have studied 457 papers to investigate and categorize the concepts that already exist in the field of big data. This evaluated research work recommends associated material to researchers concerning significant functioning in study and Big Data application in miscellaneous practical regions. The research has not covered the aspects like provision of more specialised framework review and PLs (Programming language) for the analysis of big data for extending the information for the limitations, advantages for every framework and the language for giving the guidelines with the concept proof for exposing the functionality of every PL or some framework. [10]

T. Sanjana et al [2016], analysed a number of algorithms that are used for the purpose of clustering in the big data processing. As per the research work, it has been discovered that ORCLUS, BIRCH and CLIQUE are the algorithms of clustering that can be utilized to detect outliers in big data. It has also been suggested that the algorithms like CURE and ROCK can be applied on the categorical data to generate clusters having arbitrary shape. Also, the non-convex shaped clusters can be obtained by utilizing the algorithms like COBWEB and CLASSIT, on the model based statistical data. At last, this work has also discovered that the algorithms like 1STING, OPTIGRID, PROCLUS and ORCLUS can be applied on the spatial data to acquire arbitrary shaped cluster. This survey has analyzed different algorithms for the data clustering for the processing in Big Data. The survey has focused on the identification of the outliers in large data sets by using the different algorithms but some parameters like variety, velocity and value are not according to the requirement and there are a lot of possibilities to improve that factor by using a classifier along with the clustering algorithm. [11]

V. W. Ajin et al [2016], provided a complete study about the various clustering algorithms bearing the big data principles. The foremost objective of the clustering methods is to sort the data into different groups so that the related data entities can be assembled inside the similar group based on similarity, potential and activities. Furthermore, several varieties of the clustering methods are deliberated within this study, along with reasonable investigation of the utmost frequently employed and efficient algorithms i.e. FCM, BIRCH, K-Means, CLIQUE, etc. are disclosed based on the Big Data views. The study has not covered the single clustering algorithms being accurate and stable as compared to the individual and single clustering for the ensembles. The incorporation of distribution system for performance improvement and existing algorithms efficiency for big data has not been taken place. For every clustering algorithm, the appropriate parameter setting has not been discussed. [12]

Ahmed Oussous et al [2017], deliberated a survey which is entirely based on the recent technologies that are emerged for big data. The big data characteristics have been thoroughly studied together with the trials that are raised up in the big data computing systems. The constituents and tools that are utilized in every layer of big data platform have been 8accentuated and furthermore equated the tools and allocations which is primarily based on their proficiencies, profits and confines. Further big data a system are categorized on the basis of their service area and features that are delivered to consumers. In conclusion, this survey offers a wide-ranging knowledge about the structural design, practices and methodologies that are presently tailed in the computing system of big data. The author has lacked in covering the creation of next generation infrastructure, different areas like platform tools, domain specific tools and data organization. [13]

III. METHODOLOGY

There are two kind of clustering which is observed in big data framework namely reference based clustering and non-reference based clustering. The reference based clustering always contains a reference data cluster for each document. As for example, the user will always know that the he is uploading a file or data against Football. The core work in this type of clustering is referred as data management where algorithm like Map-Reduce is helpful. The area of interest of this research work is a non-reference based clustering in which there is no previous reference available at the initial stage. In such a scenario, the data is clustered utilizing the relation between the elements or data files which are to be uploaded. The proposed structure utilizes two similarity indexes to form a new similarity measure.

Following similarity measures are utilized

a) Cosine Similarity
b) Soft Cosine Similarity

Cosine Similarity

It is the dot product divided to its magnitude values.

$$Cos = \frac{A.B}{|A||B|}$$

Soft Cosine

It is the dot product of the squares of the individual elements divided by the sum of each element value present in the data file.

$$Soft\ Cosine = \frac{Sum(A).Sum(B)}{|Sum(A)||Sum(B)|}$$

The following results are obtained for cosine and soft cosine similarities.

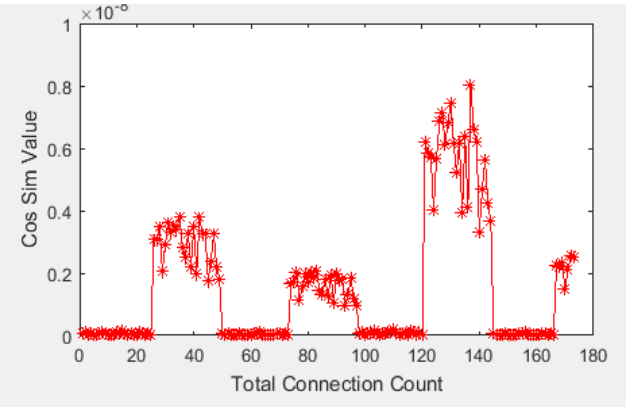

Figure 4: Cosine Similarity

Total Similarity Values = where n is total number of documents.

Soft-cosine will have the same count as that of cosine similarity.

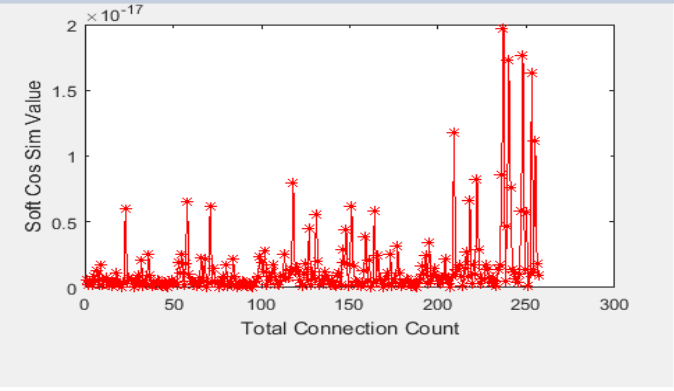

Figure 5: Soft Cosine Similarity

The hybrid similarity is attained utilizing the combination of cosine similarity and soft cosine similarity.

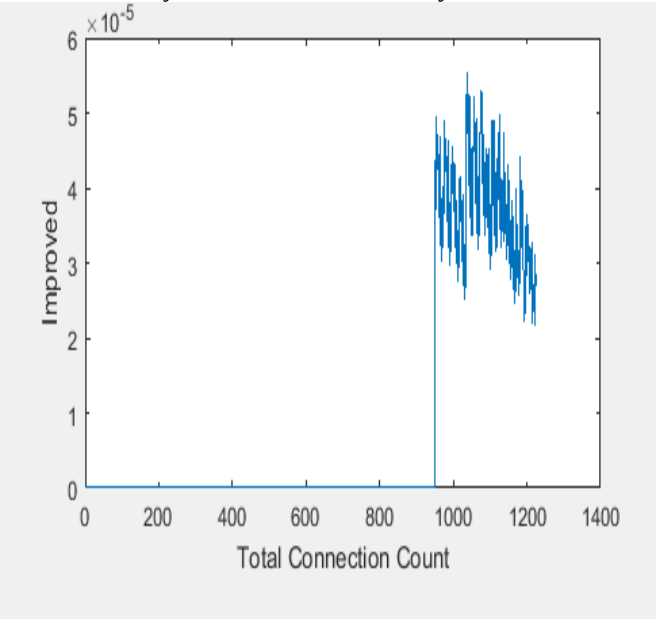

Figure 6: Hybrid Similarity

$$HYBRID\ SIM = COSINE + SOFT\ COSINE$$

Based on this hybrid similarity, a new algorithm is proposed which intakes the features of K means and Linkage Similarity. The brief is as follows

Let us consider that that there are 6 data files for clustering. The hybrid relation will be calculated as follows

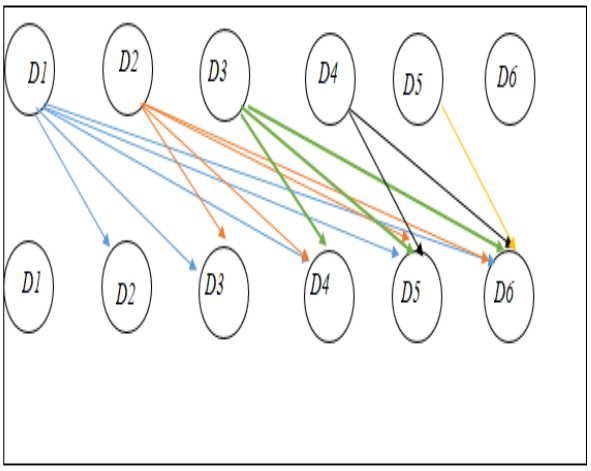

Figure 7: Hybrid Relations between data files

IV. RESULTS AND DISCUSSION

Based on the similarity measures, following pseudo code is applied for clustering

cluster1=[];


```
cluster2=[];
cluster1(1)=1;
[r,c] = size(improved);
improved = improved * 10000;
lastdoc = connection(r,2);
doccount = 2;
cluster1count = 1;
cluster2count = 0;
while doccount < lastdoc
currentdoc = doccount;
refdoc = connection(doccount - 1,2);
fs = refdoc - 1;
refsim = improved(fs);
s = find(connection(:,1) ==
currentdoc);
found = 0;
foundcount = 0;
for j = 1:numel(s)
if improved(s(j)) < refsim
found = 1;
foundcount = foundcount + 1;
end
end
if foundcount < 15
cluster1count = cluster1count + 1;
cluster1(cluster1count) = currentdoc;
else
cluster2count = cluster2count + 1;
cluster2(cluster2count) = currentdoc;
end
doccount = doccount + 1;
end
```

The proposed clustering algorithm intakes the threshold policy of K means and co relation linkage property of Linkage clustering and forms two clusters 1 and 2.

CROSS-VALIDATION

The cross-validation of the proposed work model has been done using Support Vector Machine followed by K-medoid.

Support Vector Machine (SVM)

SVM has been characterized as a learning algorithm optimized for solving complex problems. SVM is efficient in differentiating the two clusters very well. It is used for classification purposes. The general description of SVM is provided in the following section.

SVM is a type of supervised machine learning. An algorithm in which, a set of data is provide as a training data that is belonging to one of a number of categories, the SVM training algorithm constructs a model that predicts the category of the new example. SVM has a stronger ability to summarize problems, which is the goal of statistical learning. Statistical learning theory provides an outline for studying the problem of acquiring knowledge, making predictions, and making decisions based on a set of data. In the statistical learning theory, the problem formula for supervised learning is as follows.

Let us provide a set of training data { $(a_1, b_1) \ldots\ldots\ldots$ in samples as per the unknown probability function and loss function p(a, b) and V(b, f(a)) that is used to determined error for a defined function a, f(a) is predicted instead of the actual value b. The problem is to find a function f which minimizes the error expectation of novel data by finding a function f that minimizes the expected error, which is defined as

$$\int V(b,f(a)p(a,b)dadb$$

Early machine learning algorithms were designed to learn the depiction of uncomplicated functions. Therefore, the main aim of learning is to output a hypothesis that performs the accurate classification of the training data, as well as the early learning algorithm is developed to determined this precise fit to the data . The ability to correctly classify data that is not in the training set is called its generalization.

The basic concept of SVM can be explained using the following four points.

- The Separating hyper plane.
- The maximum margin hyper plane.
- Soft margin.
- The Kernel function

People are able to easily distinguish between different data types as given for every sample, but it is very difficult to distinguish and represents for a computer system. Figure 3.4 has two distinct data types and the researcher's aim is to classify these two data type. In this case, it is very easy to visually classify by naked eye because it can be in the visual field. However, a single line that separates these data types can be used to refer to these two different classes.

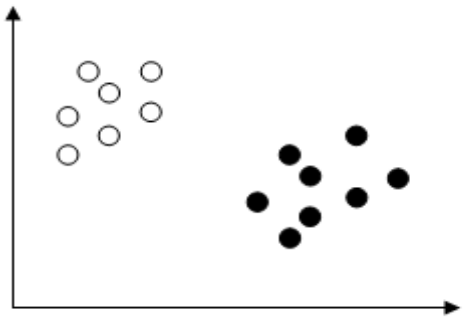

Figure 8: (a) Two types of dataset

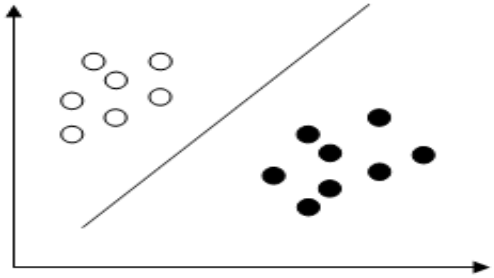

Figure 8: (b) Classification

A single line is inserted between two datasets for the classification purpose of 2D data and shown in figure 3.4 (b). The classification of data type in 1-D can be represented in figure 8 (a). The 3D data separation of a given data is separated by hyper plane as shown in figure 8 (b).

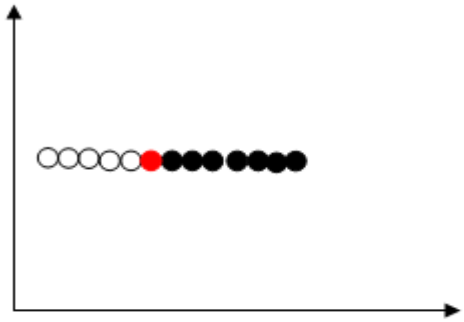

Figure: 9 (a) Linear Classification

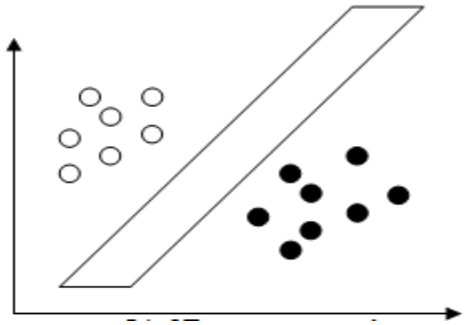

Figure: 9 (b) 3D Repersentation

SVM Architecture used in the research work is:

$$1.\ SVM_{population} = Group_{Elemen}$$

$$2.\ For\ each\ vector\ in\ SVM_{population}.\ Element_{Value}$$

$$3.\ SVM_{Kernel} = F_{SVM}\left(SVM_{threshold}, Present_{SVM}\right)$$

$Kernel_{value}$	$1\ \ If\ Present_{SVM}\ satisfies\ F_{SVM}$ $0\ \ Otherwise$

The categorization of each data element in Cluster 1 and Cluster 2 has been considered, a kernel value is designed which classifies the hybrid similarity values and average similarity depends upon the kernel value of the entire population. If the kernel value returns 1 then there will be no adjustments within the cluster elements else cluster elements has been shifted to form a cluster. The proposed method determines the average data elements in the different clusters. The average similarity value obtained from the SVM is used as the input weight of the cross-validation. The generated document value worked as a target label for the input layer. According to a supervised learning algorithm, the entire training search space work as a test set.

V. CONCLUSION AND FUTURE SCOPE

The proposed algorithms are dependent on the new hypothetical outcome with major enhancement for developing it practical. Cosine similarity in hybridization with soft cosine similarity along with K-mean and SVM approach is used to determine the similarity between the text documents.

It is concluded that a new work model has been proposed using different algorithms. The hybridization of cosine similarity and soft cosine is implemented for better results. The proposed algorithms for soft cosine and cosine similarity are hybridized using the clustering formation algorithm. The values attained from these techniques are used for hybridization. The proposed approach is evaluated using the SVM which identifies the segregated values used for the formation of clusters. The groups formed due to classification and regression process. In addition, the proposed algorithm further evaluated using the SVM algorithm to validate the results. There is still need of improvement in terms of accuracy and processing time of clustering with machine learning methods. We can use more machine learning concepts to improve and automated clustering concepts.

REFERENCES

[1] Dipti Shikha Singh and Garima Singh, "*Big Data: A Review*", International Research Journal of Engineering and Technology (IRJET), Vol. **04**, No. **04**, pp. 822-824, **2017**

[2] Richa Gupta, Sunny Gupta, and Anuradha Singhal, "*Big data: overview*" International Journal of Computer Trends and Technology (IJCTT), Vol. **9**, No. **5**, pp. 266-268**, 2014**

[3] S. Gnanapriya, R. Suganya, G. Sumithra Devi, and M. Suresh Kumar, "*Data Mining Concepts and Techniques*", Data Mining and Knowledge Engineering, Vol. **2**, no. **9**, pp: 256-263, **2010**
[4] T. Sajana, CM Sheela Rani, and K. V. Narayana, "*A survey on clustering techniques for big data mining*", Indian Journal of Science and Technology, Vol. **9**, no. **3**, **2016**.
[5] V. W. Ajin, and Lekshmy D. Kumar, "*Big data and clustering algorithms*", In IEEE International Conference on Research Advances in Integrated Navigation Systems (RAINS), pp. **1-5, 2016.**
[6] Raj Kumar, and Rajesh Verma, "*Classification algorithms for data mining - A survey*", In the International Journal of the Innovations in Engineering and Technology (IJIET), vol. **1**, no. **2**, pp: 7-14. **2012.**
[7] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih, "*Big Data Technologies: A Survey*", Journal of King Saud University-Computer and Information Sciences, **2017**.
[8] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, and Abdelaziz Bouras, "*A survey of clustering algorithms for big data: Taxonomy and empirical analysis*", IEEE transactions on emerging topics in computing, Vol. **2**, no. **3**, pp: 267-279**, 2014**
[9] G. Kesavaraj, and S. Sukumaran. "*A study on classification techniques in data mining*." In IEEE Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp**. 1**-**7**. **2013.**
[10] R. Tamilselvi and S. Kalaiselvi, "*An Overview of Data Mining Techniques and Applications*", International Journal of Science and Research (IJSR), Vol. **2**, No. **2**, pp. 506-509**, 2013.**
[11] Praful Koturwar, Sheetal Girase, and Debajyoti Mukhopadhyay, "*A survey of classification techniques in the area of big data*", arXiv preprint arXiv: 1503.07477, **2015**.
[12] V. W. Ajin, and Lekshmy D. Kumar, "*Big data and clustering algorithms*", In IEEE International Conference on Research Advances in Integrated Navigation Systems (RAINS), pp. **1-5**. **2016.**
[13] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih, "*Big Data Technologies: A Survey*", Journal of King Saud University-Computer and Information Sciences, **2017.**

Authors Profile

Kapil Sharma received a Bachelor degree in Computer Science and Engineering in 2011 and a Master degree in Computer Science & Engineering in 2013. He is currently pursuing Ph.D in Computer Science & Engineering at RIMT University. He is with Guru Nanak Dev Engineering College, Ludhiana (Punjab) India as Assistant Professor, in CSE department from September, 2015 to till date. He has published 15 papers in reputed international journals. His research area includes data mining and big data.

Palak Rehan received a Bachelor degree in Computer Science and Engineering in 2014 and a Master degree in Computer Science & Engineering in 2017. She is with Guru Nanak Dev Engineering College, Ludhiana (Punjab) India as Assistant Professor, in CSE department since 2017. Her research area includes Natural language processing and data mining.

