An Extended Hybrid Clustering Method Utilzing Svm As Cross Validator

Kapil Sharma, Satish Saini

***Abstract:** The massive data accumulation from the internet creates attention for the researchers. The data collected in the form of structured and unstructured data. The structured data consists of messages, transactions, conversations, etc. while unstructured represents video and audio clips. This essentially manages the raw data problem in which unreferenced clustering is used. A hybrid approach is proposed using Cosine Similarity and soft cosine. A novel clustering technique is designed which is cross-validated using the Support Vector Machine (SVM). The validated approach is further verified by using K- means clustering. The clustering results have been further evaluated using parameters precision, recall, and F-measure. The evaluated results show the improvement in precision and recall accuracy due to hybridization of cosine similarity and soft cosine techniques.*

***Keywords**: About four key words or phrases in alphabetical order, separated by commas.*

I. INTRODUCTION

Due to the excess use of social network, the problem of accumulation a large amount of data surges. It becomes severe in the coming years. There is a large count of advertisements that hampers the internet. In addition, we are depending upon information technology in various aspects of our daily life. The data usually accumulates during uploading, downloading, and storing the information in the internet platform [1]. The massive data gathered from various user communication devices such as smartphones, tablets, and other media devices. The conventional and modern data emerging technologies differ in various concern. The digital data becomes more complex as its quantity is larger and managing the computation technology is difficult. The data gathered in the form of structured and un-structured data. The information in the unorganized form comes from Twitter, Facebook, and various other social sites. On the other hand, the organized data comes from websites interacting with user [2]. The data is characterized in the form of variability and complexity. In a way to manage complex information, different tools and techniques were used to analyze the big data. Therefore, the big data concept has been also represented in the form of weighted graphs to solve data issues. Fig.1 represents the architecture of big data.

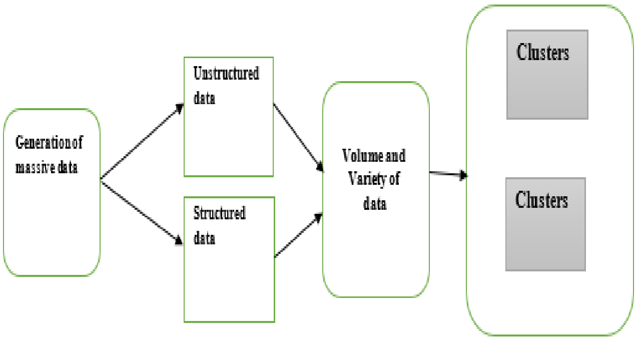

Fig.. 1 Big data Architecture

A big data network is stated as a network in which a number of people indicated as p and different sources have been depicted as s. An edge-weighted graph is defined as C (M, G, Y). Here, the cluster of nodes represented by M,$|M| = p$, and directed relationships given as G, $M \subseteq G \times G, |M| =$ s, and Y represents the edges weight related with edges of G. There is a huge sources of data on the internet. Big data converges new opportunities for the business ventures to enhance activities using big data [3]. This leads to form an exercise in which separation of data is essential. Therefore, a concept of big data clustering comes in concern in which big data is divided into a set of objects which forms a groups. The data of similar nature is grouped and other group form on the basis of other parameters. The whole groups is known as clusters. Clustering is used for the analysis of data in a statistical form. The main tasks of clustering is realizing the data clustering problem using multi-objective function [4]. The unprecedented volumes of data pose a serious problem of big data. There is a need of scalable and effective tools for analyzing the data in an exploratory manner. There are various applications of clustering in various fields such as marketing, medicine, bioinformatics, anomaly detection etc. [5]. Clustering is a notion in which collection of objects were grouped based on the similarity index. A constructive and collaborative approach representing the clusters is that abstract points are collected together and distances among points depicted as similarities [6]. The similar nature depends upon the closeness among the points. Thus, points in the same cluster represent similarity while points in the different cluster has been dissimilar in nature. The different techniques of clustering such as spectral, centroid-based, parallel K-means, agglomerative, and informative-theoretic are synthesized for different applications [7-9]. But, applications of the modern world has been sorted using top order architecture. The modern clustering approaches based on metaheuristic techniques [10]. This simply means quick access of data, whenever the user demands the information.

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

Kapil Sharma*, Assistant Professor, Guru Nanak Dev Engineering Collage, Ludhiana, Panjab, India. Email: kapil4ravit@gmail.com

Dr. Satish Saini, Professor, Electronics and Communication Engineering School of Engineering, RIMT University, Mandi Gobindgarh, Panjab, India.

Clustering measure similarity between different groups [11]. An example of k-means clustering is shown in Fig. 2.

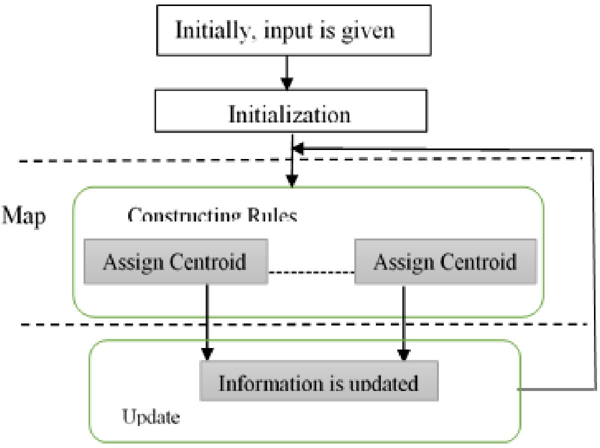

Fig. 2 Example of K means clustering

Fig. 2 demonstrates the example of k-means clustering. The functioning of clustering shows that each node in the network answers the assigned task of the patterns. Once the input patterns were assigned to the respective clusters, the update function will automatically compute the new means of clustering. The update and assign automatically perform its task continuously until halt condition met.

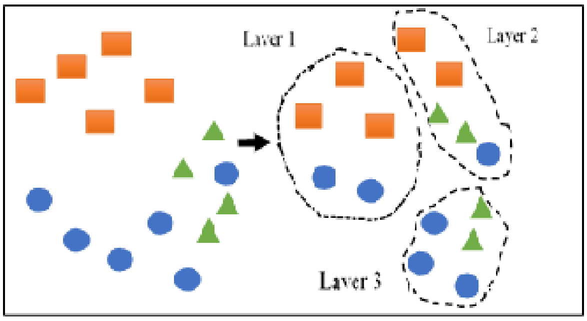

Fig. 3 Clusters with their data layers

Fig. 3 depicts the formation of clusters with their data layer m having data server M (m1, m2, m3…….mn) where m is the number of clusters with the layer. The definition of M can be demonstrated as follows:

Definition 1: A Clustering is a set of clusters with data layer m having data server M where m is a subset of $\int_0^n \frac{m_{data}}{Number\ of\ Sources}$ where m data is the information collected by the clusters through internet.

The formation of the clusters depends upon the user query to the data server, which determines the availability of the information and the appropriate cluster is formed to respond to the query of the user. If the search is appropriate, then transition time is very large and if it is inappropriate then a term known as True Positive Rate (TPR) remains low for the entire time.

Lemma 1: If the flow of data is normal then the formation of clusters to arrange the data is easy as shown in Fig. 3. Thus, the number of clusters created to arrange the data is infinite. There are two types of situations:

Situation 1: A reference elements or reference cluster Relement persists for each data element Delement such that $D_{\text{element}} \in R_{element}$. In other case, it is $D_{\text{element}} \notin R_{element}$. In this scenario, categorization of Delement becomes easy and simple.

Situation 2: When there is not any reference cluster for the Delement or there is not any preliminary cluster present prior in the system.

The issues created due to situation 2 raises serious concerns for data categorization. Hence, data can be categorized based on another element present in the system.

Definition 2: Let M be a file list having M= {M1, M2, M3……..Mn} where M1, M2, M3……..Mn are the different records. It is noted that intersect (Mfirst, Mnext) must not empty, if there is a relation between Mfirst and Mnext.

There is a big issue of clustering for managing data. So, researchers focussed on hybrid techniques to manage a large volume of data [12]. The data categorization must be cross-validated. This paper uses Support Vector Machine for cross-validation in the clustering.

This paper focuses on Situation 2 with relation factor as depicted in Definition 2. The rest of the paper is organized as follows. Section 2 represents the proposed model. The results have been evaluated in section 3. This paper is finally concluded in section 3.

II. PROPOSED METHODOLOGY

The proposed methodology is divided into three sections such as:

1. Relation Finder
2. Cluster Organization
3. Cross-Validation

A. Relation Finder

In this section, Situation 2 and Definition 2 has been considered, when there is no reference cluster available in the architecture. A relation exists between elements of list M, and all the elements can put together.

Fig. 4 represents data records with its relation value. It is seen in the Fig. that (M1, M2, M5, M8) are very close to each other and hence place in a cluster and (M3, M4, M6, M7) has been placed in another cluster or group. It is seen that M9 is equidistant from cluster 1 and cluster 2. Therefore performed using the SVM. But, here is a problem as data record does not observe with measuring distance normally. Hence, co-relation is required for evaluation.

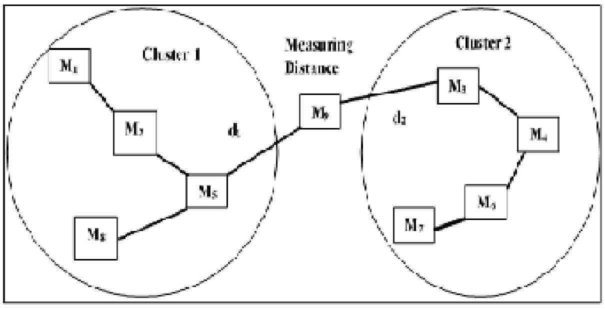

Fig. 4 represents the measurement of a distance between data of similar nature

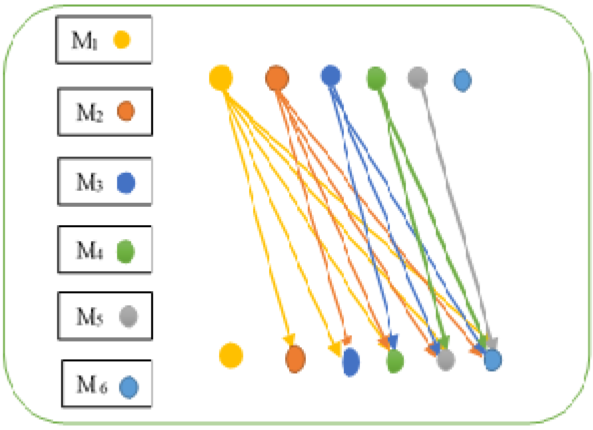

Fig. 5 Co-relation between data elements

For n number of data records, we use the formula as given to compute the relation values.

$$n * \frac{n-1}{2} \qquad (1)$$

More specifically, Relation, R can be evaluated if $M_{current} \,!= M_{next}$ and R $(M_{current}, M_{next})$ or R $(M_{next}, M_{current})$ has not been available in the present R list. Fig. 5 shows the relationships between data elements in which relation still computed. There are 6 documents, hence total of 15 relationships will be determined as given below:- {M1-M2, M1-M3, M1- M4, M1-M5, M1-M6, M2-M3, M2-M4, M2-M5, M2-M6, M3-M4, M3-M5, M3- M6, M4-M5, M4-M6, M5-M6}

The hybrid approach of Cosine similarity and soft cosine has been used for the evaluation of relation R. There is a probability that data elements may be similar in nature due to similar data patterns. Soft cosine technique measures the similarity between two vectors. The hybrid technique provides consistent results as per the researchers [13]. The physical distance between two data elements can be easily measured but, this technique does not good to measure the behavior of the data elements. Fig. 6 represents the measure of distance d1 between M5 and M9.

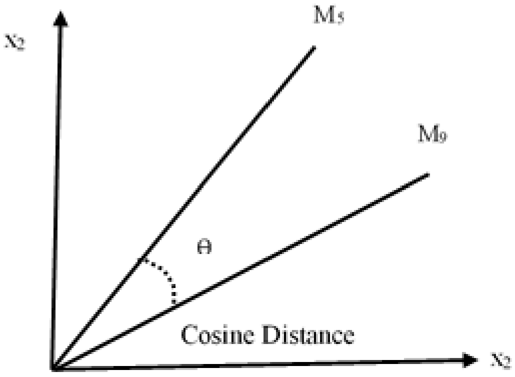

Fig. 6 Measure of cosine distance

The Cosine relation between two vectors can be computed using Algorithm 1 for the proposed work model.

B. ALGORITHM 1: COSINE SIMILARITY RELATION EVALUATION

```
1. $Cos_{relation} = functionCosRelation(data_{files})$
2. $Cosine_{catalogue_{relation}} = [\,];$ Empty array determines similarity index
3. $Sim_{count} = 0;$ Identified relations are computed
4. $For\ y = 0\ to\ datarecords.count$ For 1 to total number of data records.
    5. $Present_{catalogue} = datarecords\ (y);$ Document I
    6. $For\ z = I + 1\ totaldatarecords.count$
        7. $P = \left|Cosine_{catalogue}(Present_{catalogue}) - \cos(datarecords(z))\right|;$
        8. $Cosine_{catalogue_{relation}}\ [Relation_{count}, 0] = present\_catalogue$ ; There are three columns in the similarity measure.
        9. $Cos_{catalogue_{relation}}\ [Relation_{count}, 1] = datarecords(z);$
        10. $Cos_{catalogue_{relation}}\ [Relation_{count}, 2] = P;$ The similarity value
        11. $Sim_{count} = Sim_{count} + 1;$ Count is incremented
    12. $End\ for$;
13. $End\ for;$
12. $End\ if;$
```

The data records taken as input and similarity of cosine relation has been evaluated. The evaluated relation has been saved and uses further to make a hybrid relation with the soft cosine. In other words, the soft similarity is equivalent to standard similarity when there is no equivalence. Thus, soft cosine measure has been used. The soft relation between data records has been evaluated which is depicted in Algorithm 2. In addition, the cosine relation count is similar or equal to the soft cosine relation.

C. ALGORITHM 2: Soft Cosine Relation

```
1. $Soft\ Cosine_{relation} = FunctionSoftCosineRelation(data_{records})$
2. $//SoftCosine_{catalogue_{relation}} = [\,];//Similarity\ Index\ calculation\ using\ empty\ array$
3. $//\ Sim_{count} = 0;//identified\ relations\ Count$
4. $For\ s = 0\ to\ datarecords.count\ //\ Total\ no.of\ data\ records$

    5. $Present_{catalogue} = data_{record}(s)$;
    6. $For\ z = I + 1\ totaldatarecords.count//\ Next\ series$
        7. $P = \left|SoftCosine_{catalogue}(Present_{catalogue}) - \cos(datarecords(z))\right|;$
        8. $SoftCosine_{catalogue_{relation}}\ [Relation_{count}, 0] = present\_catalogue$ ; There are three columns in the similarity measure.
        9. $Cos_{catalogue_{relation}}\ [Relation_{count}, 1] = datarecords(z);$
        10. $Cos_{catalogue_{relation}}\ [Relation_{count}, 2] = P;$ The similarity value
        11. $Sim_{count} = Sim_{count} + 1;$ Count is incremented
```

```
12. $End\ for;$
13. $End\ for;$
14. $End\ function;$
```

The hybridization of Cosine Similarity and Soft Cosine will provide the desired hybrid value [6]. The relationship value will be further use for clustering. Fig. 7 depicts the relationship hybridization value.

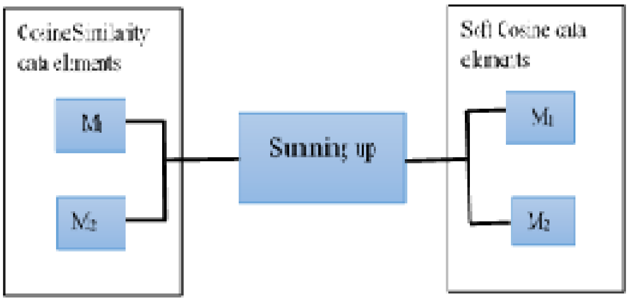

Fig. 7 Hybridization of Cosine similarity and Soft cosine measure

D. Formation of Clustering

The formation of clustering relies on the closest value during evaluation. The proposed algorithm depicts the position of the reference attribute of each data element. The developed algorithm the combined value, whether it is greater or less than the referred value. The cluster is formed by placing the first data file or record in Catalogue 1. The working of the proposed algorithm for clustering has been given in Algorithm 3.

ALGORITHM 3 : CLUSTERING

```
$Group\ A, Group\ B = Function\ Clusters$
$(Co-Relation\ identifies)$   (1)
1. $Group\ A =$
$[\ ]; Initially\ empty\ array\ has\ been\ generated$
2. $Group\ B = [\ ];\ //Group\ B\ also\ empty$
3. $GroupelementsA = 0;//Number\ of\ elements\ in\ cluster\ A$
4. $GroupelementsB = 0;//Number\ of\ elements\ in\ clusterB$
5. $First\ Relation\ Identifies = Co-Relation\ identifies\ (1,1)//$
6. $For\ each\ connection\ identifies\ in\ Co-Relation\ //\ Detect\ same\ data\ file$
7. $For\ each\ Similar_{Connection} = Find\begin{pmatrix}Co-Relation\ identifies\ (:,1)\\ ==1st\ identification\ mark\end{pmatrix};$
8. $d.//complete\ list$
9. $b.//determine\ other\ connection\ of\ identified\ above$
10. $Similar_{Connection_{Value}} = Sum(All_{Similar_{Connection}})$
$//\ Determines\ total\ connections$

11. $.//Compute\ the\ average\ connection\ value$
12. $If\ Avg_{Value}\ of\ Co-Relation > Similar\ Connection$
$//\ find\ the\ greater\ connection$
13. $.//Identify\ value\ stored\ in\ the\ Group\ A$
14. $Group\ A[GroupelementsA] = Present\ DataRecords;$
15. $GroupelementsA = GroupelementsA + 1;$
16. $Else$
17. $GroupB[GroupelementsB] = Present\ DataRecords;$
18. $GroupelementsA = GroupelementsB + 1;$
19. $End\ for\ each$
15. $End\ Algorithm$
```

E. 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.

SVM Architecture

```
1. $SVM_{population} = Group_{Element_{Count}}$
2. $For\ each\ vector\ in\ SVM_{population}.Element_{Value}$
3. $SVM_{Kernel} = F_{SVM}\ (SVM_{threshold}, Present_{SVM})$
```

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

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 woked as a target label for the input layer. According to a supervised learning algorithm, the entire training search space work as a test set.

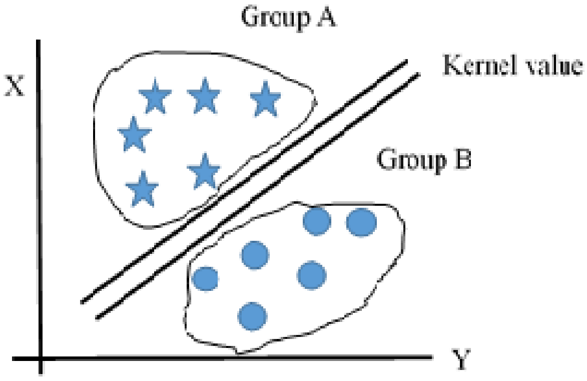

Fig. 8 Hybrid Similarity formation

It can be seen in Fig. 8 that values of Group A and Group B have been classified based on its similarity. The kernel value is generated based on the similarity between data elements. The kernel value also works for the classification of the non-linear search space.

The average value generated from Group A and Group B has been summed up for hybridization.

ALGORITHM 4 CROSS-VALIDATION

```
1. $Cross_{Validation} = Fun\ Cross_{validation}(Input_{set}, Target_{Label})$
2. $Total_{Simulation\ Rounds} = 800;$
3. While $Traine\ d_{Record}.Validation < Minimum_{Set\ kernel}$
4. Kernel Value ++
5. End While;
6. While $Trained_{structure}.F - Measure! = Validation_{Regression} F - Measure$
7. End While
8. $Test_{set} = Trained_{set};$
9. $Simulation_{Labels} = Simulate(Trained_{set}, Test_{set})$

10. For each Ker in $Simulation_{Label}$
11. If Ker! = $Target_{Label}$;
12. Transfer Group;
13. End if;
```

The ordinal measure of the algorithm 4 given as:

Table 1 Ordinal Measure of SVM

Designated Structure	Kernel Value
Cross-Validation Parameter	F measure
Total Simulations	800
Criteria for Satisfaction	Gradient

III. RESULTS AND DISCUSSION

The results of the proposed work model have been identified using the parameters precision, recall, and F-measure for SVM.

a) $$Precision = \frac{True_{adjustments}}{True_{adjustments} + False_{adjustments}}$$

b) $$Recall = \frac{True_{adjustments}}{True_{adjustments} + True_{left\ adjustments}}$$

c) $$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The proposed algorithm has been identified using the predefined labeled data set which has been passed as a test set. It has been used to calculate precision. The data set has been used to evaluate the results.

A. Data Set

US Census Demographic data set is the US Census Bureau data report which covers the data not from New York, but includes data set from entire countries.

All the demographic data has been included which estimates the US Census level from 5 year American Survey. The data collected from DP03 and DP05 of 5 years estimates.

The data includes:-

1. acs2015_census_tract_data.csv: Data for each census tract in the US, including DC and Puerto Rico.

2. acs2015_county_data.csv: Data for each county or county equivalent in the US, including DC and Puerto Rico.

The link to download the data was https://www.kaggle.com/muonneutrino/us-census-demographic-data.

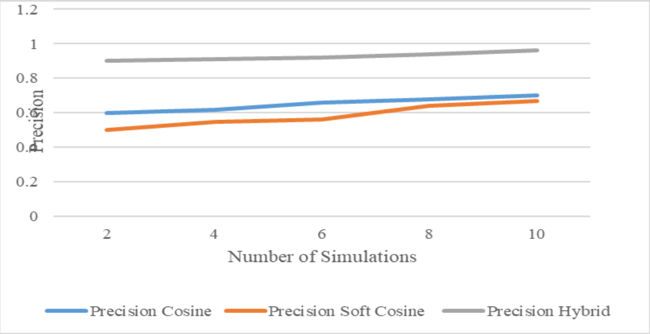

Fig. 9 Determination of Precision

The precision characteristic determines the precision for different simulations. There are 6 aspects passed for similarity vector and related indexes are evaluated. Fig.9 depicts the precision determination for different simulations. It is seen in the graph that the maximum value for precision in case of cosine and soft cosine almost same which is 0.7. The minimum precision attained for cosine similarity and soft cosine is 0.6 and 0.5 respectively. The maximum precision attained from hybridization of both techniques is approximately 0.9. It is clear that the value of precision almost remains constant for the number of simulations. The proposed model shows that there is an improvement in a precision value from 0.7 to 0.96. Thus it provides $0.96 - 0.7 * 100 = 26\ \%$ accuracy in case of cosine similarity and $0.96 - 0.68 * 100 = 28\%$ accuracy in case of soft cosine.

Similarly, Fig. 10 depicts the measurement of recall. The parameter recall that maximum recall for soft cosine and cosine similarity is 0.78 and 0.72 respectively. The minimum recall for soft cosine and cosine similarity measure is 0.64 and 0.7 respectively. On the other hand, the maximum recall value attained due to hybridization of both techniques is approximately 0.96. Hence the parameter recall has been improved with 18 % accuracy in case of soft cosine and 24% revamped using the cosine similarity.

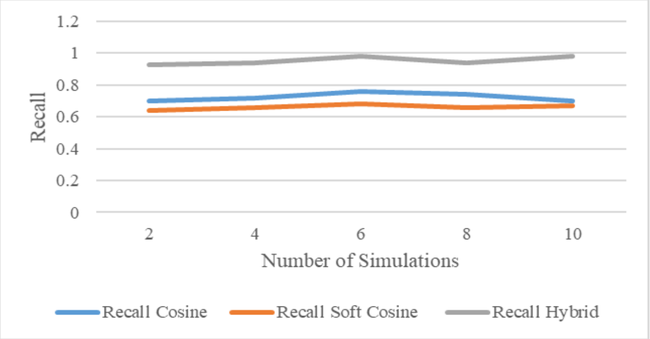

Fig. 10 Recall Measurement

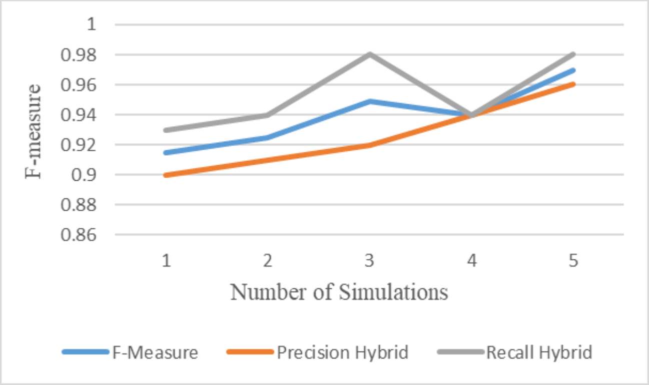

Fig. 11 F-Measurement

In the same manner, Fig. 11 depicts the F-measurement. The graph shows that maximum F-measure obtained at 0.95 for 3 simulations. The graph follows the increasing trend. The minimum F-measure has been obtained at 0.93.

IV. CONCLUSION

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. In addition, different parameters such as precision, recall, and F- measure are used for further evaluating the results. The outcomes for precision provides 26% and 28% accuracy from cosine similarity and soft cosine respectively. On the other hand, recall parameter attained 18% and 24% accuracy in soft cosine and cosine similarity respectively.

REFERENCES

1. Huda, M., Maseleno, A., Atmotiyoso, P., Siregar, M., Ahmad, R., Jasmi, K. and Muhamad, N., Big data emerging technology: insights into an innovative environment for online learning resources. International Journal of Emerging Technologies in Learning (iJET), 13(1), pp.23-36, 2018.
2. Ghani, N.A., Hamid, S., Hashem, I.A.T. and Ahmed, E., Social media big data analytics: A survey. Computers in Human Behavior, 2018.
3. Zhao, X., Liang, J. and Dang, C., A stratified sampling based clustering algorithm for large-scale data. Knowledge-Based Systems, 163, pp.416-428, 2019.
4. Garza-Fabre, M., Handl, J. and Knowles, J., An improved and more scalable evolutionary approach to multiobjective clustering. IEEE Transactions on Evolutionary Computation, 22(4), pp.515-535, 2018.
5. N. Moustafa, G. Creech, E. Sitnikova and M. Keshk, "Collaborative anomaly detection framework for handling big data of cloud computing," Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, pp. 1-6, 2017.
6. Jain, A., Bhatnagar, V. and Sharma, P., Collaborative and clustering based strategy in big data. In Web Services: Concepts, Methodologies, Tools, and Applications, IGI Global, pp. 221-239, 2019.
7. Azizpour, S., Giesecke, K. and Schwenkler, G., Exploring the sources of default clustering. Journal of Financial Economics, 129(1), pp.154-183, 2018.
8. Bangui, H., Ge, M. and Buhnova, B., A Research Roadmap of Big Data Clustering Algorithms for Future Internet of Things. International Journal of Organizational and Collective Intelligence (IJOCI), 9(2), pp.16-30, 2019.
9. Yu, Z., Big Data Clustering Analysis Algorithm for Internet of Things Based on K-Means. International Journal of Distributed Systems and Technologies (IJDST), 10(1), pp.1-12, 2019.
10. Tsai, C.W., Liu, S.J. and Wang, Y.C., A parallel metaheuristic data clustering framework for cloud. Journal of Parallel and Distributed Computing, 116, pp.39-49, 2018.
11. Irani, J., Pise, N. and Phatak, M., Clustering techniques and the similarity measures used in clustering: A survey. International Journal of Computer Applications, 134(7), pp.9-14, 2016.
12. Rathore, P., Kumar, D., Bezdek, J.C., Rajasegarar, S. and Palaniswami, M., A rapid hybrid clustering algorithm for large volumes of high dimensional data. IEEE Transactions on Knowledge and Data Engineering, 31(4), pp.641-654, 2019.
13. Sidorov, G., Gelbukh, A., Gómez-Adorno, H. and Pinto, D., Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación Systems, 18(3), pp.491-504, 2014.

AUTHORS PROFILE

Kapil sharma is an Assistant Professor in Guru Nanak Dev Engineering Collage, Ludhiana, India and Pursuing Ph.D. in CSE from RIMT University. He is having 6 years and a month of teaching experience. Published 16 papers in International Journals. Email: kapil4ravit@gmail.com

Dr. Satish Saini is Professor in the Department of Electronics and Communication, RIMT University, Punjab India and completed his Ph.D. in 2018 in computer science and engineering. He is having 16 teaching and industrial experience in the same field. Published 25 paper in Image processing and Artificial Neural Network.