International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Higher Dimensional Data Access and Management with Improved Distance Metric Access for Higher Dimensional Non-Linear Data

Sakshi Jolly, Neha Gupt

Abstract: ***Distance metrics for different kinds of data we daily use is the common approach for indentifying the insights of information and identifying the noisy and resolving the information with different scenarios and rules. The methodology imposed here is different in kinds of rules and the information we provide to the knowledge machine is the most important and considerable thing in designing and implementing distance metrics. The contradictory data is mixed information which the dataset includes is having the irrelevant information with the relevant information and identifying the novel thing from the information gathered. The information and the data gathered will be in the form of different formats of data and the most frequent thing we use is to make the clusters. In this article machine learning applications and the different data mining distance metric algorithms will be discussed and the information passed to the machine will be the ultimate and the dataset making is the quite challenging. In machine learning implementation the path of identifying the concept behind the everything to be predicted. The prediction works when the data is accurate and the information we get from the different repositories. All the data captured is not genuine and the combination of different such kind repositories make the contradicting data. The usage the additional distance metrics to manipulate and calculate the relation between the variables or the features we are considering. The machine learning is the finite mechanism which can help the researchers to identify the relationships between the variables or there is chance to find the inner relations among this kind of contradictory information. Contradictory data can help to identify the inner relations which can't be identified with normal distance metrics and here used a little advanced and succeeded in reaching the optimal result.***

Keywords: Machine Learning, Prediction, Contradictory Data, Result, Noisy Data

I. INTRODUCTION

Machine learning mechanisms using the distance metrics provide the semantic methodology for identifying better relation between the data points or among the multiple data points in the n- dimensional data plane. We plot the data point in the working space and if the error rate is high and that means the relation between the points and the hyper plane is not suitable for the better prediction and the clause here is to identify the contradictory data and the kind of irrelevance we use in managing and identifying the noisy information in our data is the main scenario we achieved and presented with the past and present work. There is much kind of distance metrics and we follow few to get some knowledge on the concept. The concept here is to manage the contradictory data which can be further used to identify the noisy information in an accurate manner[1-5]. The main reason behind using the contradictory information is as follow. We considered some real time examples to understand the real importance of the contradictory information.

Detecting the fake review:

Fake review system is the machine learning approach which work on natural language processing methodology to understand the human emotions on giving a review on anything over the internet. For example is there are mixed kind of review on a product then we can identify the average rating of that product. If there is a problem like giving the positive response on product with the negative star rating (lower rating) which feels like making some soup with the rotten tomatoes. The basic need is we need a wrong information to understand the right[6-8]. If there is any negative thing mixed with the positive thing we will give complete effort to understand the real truth from the information. As a data scientist we have to understand the trustworthiness of the information we provided and also which was provided to us. The information we got from the repository will be without pre-processing. Even though we perform the pre-processing step we need to indentify the other side of information we gathered.

The reason why we are focusing on the information is because "information is wealth". The right information lead us to the right path and we choose machine learning for such purpose and the information we capture from the repository may or may not contain the noisy information but there must be a trust worthy plot to trust the information. Distance metrics are used identify the similar data objects which can form the efficient prediction or decision making algorithms. Decision making the quite interesting part in contradicting data. We don't know how to make a decision when we have a contradicting data. We can't judge the information[9]. Lets take a simple example of negative conditions and positive conditions on an algorithm whose task is to identify the dolphin from a set of images. The machine was trained by all positive instances and unfortunately it came across a negative instance which gives the positive result[10].

Manuscript published on November 30, 2019.
* Correspondence Author
Sakshi Jolly*, Research Scholar, Faculty of Computer Applications Department, MRIIRS, Faridabad (India)
Dr. Neha Gupta, Associate Professor, Faculty of Computer Applications Department, MRIIRS, Faridabad (India)

For this prediction model we have to use few features like Beak size as 3,4, or 5 , whether the animal have teeth or not, whether it has gills or not, and also is it in black colour or not. Figure 1 explains the procedure of the DBScan and here if observe, there is no relation among the clusters and there will be high dissimilarity when this approach used for the contradictory information. This kind of contradictory information leads to the misplace of the knowledge to the machine or the model.

For this purpose of higher dimensional data visualization, DBScan is not much used and in the case we considered as example like fake review system, the machine has to identify whether the comment or the review given by human or a machine.

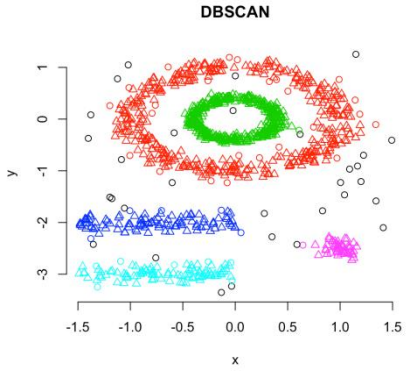

Figure 1: DBScan Result without considering the similarity with negative approaches.

Positive Instances:

We have set of positive instances which suits our decision

a. Beak = yes , Gills = No , Size = 4, Teeth= Yes
b. Beak = Yes , Gills= No , Size = 3, Teeth = Yes

For suppose we have the following negative instances.

a. Beak = No, Gills = no, Size= 5, teeth = yes
b. Beak = Yes, Gills = no, Size = 4, Teeth = Many

In the above example we have two positive instances which can be classifying the image as dolphin or not. Then the below case, we have negative instances which also need to be considered for training the algorithm. The more we train the algorithm, the best result we get. In the previous case if we observe the negative instances we can see that there is no beak. And also there are no gills, size is 5 and teeth- yes it has. Only one case we have the contradictory data and there might be cases where dolphin lost its beak because of some accident and we can't make it as an excuse to not selecting that as dolphin. This is the case where we get the contradicting data which was the combination of positive and negative as well. The information we use to train the model is the key thing for the less error rate. In this article we use two kinds of approach to explain the better usage of the contradictory data for the better utilization of decision making[11].

DBScan and Minkowski Distance are the two things we discuss in this article which makes difference in understanding the real purpose of usage of negative instances with the positives instances to maintain semantic ness in the information. The meaningfulness for the information will achieve when it has the proof of accepting the all the real time scenarios in its prediction model. The rule based generation will help the condition to understand the platform of machine learning.

II. EXISTING SYSTEM

Unsupervised learning mechanism provide the evidence based learning mechanism which can be helpful for learning on own. The model which we design will be helpful for understanding the current scenarios of the input and act according to it. The main scenario here is we need to form the cluster. The cluster we need to form is for the likelihood data which have the most amount of similarity without any variance. If there is more variance then we get the more error rate. Least squared error is the mechanism which can be used for identifying the most error rate in the model and solve the model for the less error or even avoiding the error. The existing system of this kind of concept is to use DBScan mechanism which is a flop mechanism for handling the multiple contradictory data. The contradictory data consists of numeric and the categorical variables. The categorical variables are the things which has to be assigned to a unique variable to represent the problem. For example if there is any gender in our data, then it can be a categorical variable and that has to be modified from an object to the numeric variable to process the information.

This kind of information cannot be handled by the DBScan mechanism. This kind of mechanism will be most used in making the clusters with the highest similarity among the data points. Sometimes we plot the spherical data which leads to confusion of information retrieval and the confusion leads to collapse of the model. That model needs some modifications like accepting the clusters with negative instances as well. DBScan most used on K-Means clustering which consists of the centroid, mean, median etc. The distance between the neighbour data point must be as less as possible to make a cluster. The following algorithm we followed to make better understanding of the existing approach[12].

Step 1: DBSCAN (D, eps, minpts)
Step 2: C=0
Step 3: For each point p in dataset D
Step 4: If p is visited
Step 5: Continue next point
Step 6: Mark p as visited
Step 7: NP = regionquery(p,eps)
Step 8: If sizeof(NP)<Minpts
Step 9: Mark p as noise
Step 10: End if
Step 11: else
Step 12: c= nextCluster
Step 13: expandCluster(p, NP, c, eps, Minpts)
Step 14: end else
Step 15: expandCluster(p, NP, c, eps, Minpts)
Step 16: add p to Cluster C
Step 17: Distance $d(i,j) = \frac{1}{m}\sum_{i=1}^{m} dij^{(f)}$
Step 18: for each point pp in NP
Step 19: NPP=regionQuery (pp, eps)
Step 20: NP= NP joined with NPP
Step 21: end while
Step 22: if pp is not yet member

Step 23: add pp Cluster c
Step 24: end class
Step 25: }
Step 26: regionQuery(p,eps)
Step 27: return all points within pp eps-neighbour hood (including p)

$$(dist(d_x - d_y))^2 = (d_{x,1} - d_{y,1}) + \cdots + (d_{x,1} - d_{y,1})$$

Condition base clustering

$$(dist(d_x - d_y) < EPS)$$

The main problem here is there is no possibility for understanding the existence of negative instances in this scenario of DBScan. With K-Means clustering mechanism, we have the standard distance vector algorithms like Euclidian. Which cannot handle this kind of information with the high dimension. It fails when we train the model with the Euclidian distance with the high dimensional information. To solve this we started understanding the usage of Minkowski distance metrics for better usage of high dimensional data. This was explained in the proposed system section.

III. PROPOSED SYSTEM

In this proposed system we started using Minkowski distance metrics with ward's. This is a very generalized method from Euclidian and Manhattan distance. This can be considered as the better approach for understanding the contradicting data. In the minkowski distance the order of p will be relay on two points X and Y. X is the independent variable and Y is the dependent variable. The independent variable can be used to predict something from the set and the Y is the predicted variable. If there is more amount of difference between predicted and actual value then there will be a high chance of error rate. The Minkowski distance metric used for our concept is explained the sample example as follows below[13].

Step 1: Start the Program

Step 2: X = { X_1,X_2,X_3,....X_n }

Step 3: L(0)=0 // Level Number

Step 4: M=0 // Sequence Number

Step 5: d[(r),(s)]=min d[(i).(j)]

Step 6: A← [ ]

Step 7: m=m+1

Step 8: L(m)=d[(r),(s)]

Step 9: While d.size > 1 do

Step 10: d[(k),(r,s)]=min(d[(k),(r)],d[(k),(s)])

Step 11: Remove dmin1 and dmin2 from d

Step 12: Add {dmin1,dmin2} to d

Step 13: L=L+1

Step 14: End while

$$\boldsymbol{dist} = \left(\sum_{k=1}^{n} | \boldsymbol{p_k} - \boldsymbol{q_k} |^{r}\right)^{\frac{1}{r}}$$

Step 15: for i = 1 to n

Step 16: ss=rij – sij

Step 17: return (root(sum(pow(abs(r-s), d)))

Step 18: if x!==null

Step 19: repeat step 2

Step 20: do c[i][j] ← sim(i,m,j)

Step 21: c[j][i] ← sim(i,m,j)

Step 22: I[m] ← 0

Step 23: Return X

Step 24: Stop the program

The main scenario is here we have the levels of understanding the information and the scrutiny we follow to understand the distances between the variables is high compared to any other kind of distance metrics. Minkowski provides the higher dimensional data evaluation and we try to provide as much of information we can for the data evaluation. The results are explained in the next section for better understanding[14-15]. The novelty in the proposed approach is like identifying the dolphin even with the negative instances. All the negative instances we consider with the data be handled easily by the Ward's algorithm in the higher dimensional work space. The main scenario we need to consider is to check whether the algorithm we are considering is working on the higher dimensions or not. The main plot here is to make the people sense that there is an unknown entity in the data which rules the result of the model when there is higher negativity in the features or the values. More that DBScan , ward's algorithm work more on identifying the insights of information and relations among the variables.

IV. RESULTS

We designed one application which can run on the wards algorithm based on the Minkowski distance metrics and the results we achieved from the algorithm very efficient to understand the methodology of the contradicting data. First we need to assign the dataset as the input to the machine. We need to train the machine with as much of data we can for the better understanding the situation. Then we need to select the DB scan methodology for analysing the information. Then we can get the clusters from the data. For example we achieved to get two clusters and we need to plot the similarities and dissimilarities between the clusters and the data points in the cluster. We achieve the accuracy with DBScan around 73% and now we try to achieve the same thing with Wards algorithm. We try to impose the ward algorithm on the same data and the clusters and we finally achieved to identify the contradictory data in the cluster with highest of 95% accuracy. The results are mentioned below. Figure 2 explains the DBScan result and Figure 3 explains the Ward's algorithm result.

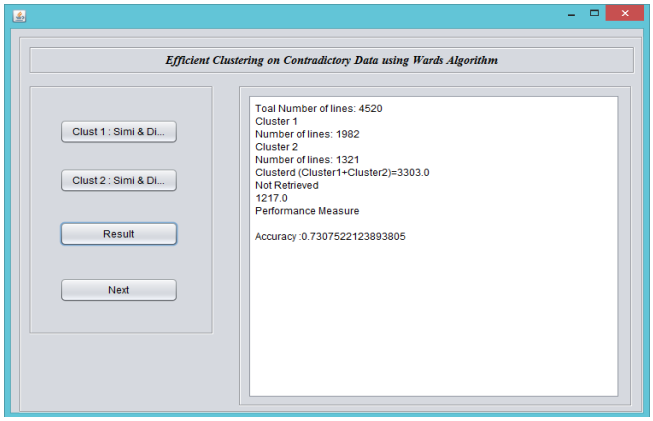

Figure 2: Result of DBScan

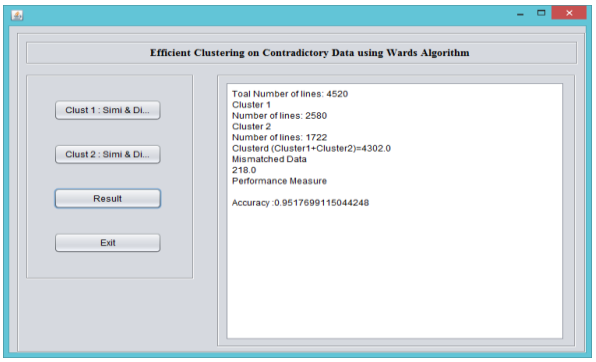

Figure 3: Result of Ward's Algorithm

There is a sample comparison between the same two models among two different data sets. These datasets are accepted with the sample references and all the information in those datasets is having contradictory information. Table 1 defines the comparison of the two datasets with two different algorithms. The table is as follows.

Table 1: Comparison of two different datasets with different distance metrics.

Data Set	WOCIL	OCIL	WKM	EWKM	WK-Modes	Proposed method	Difference
Soybeans	86	83	76	76	85	90.7	4±5%
voting	87	86	77	78	83	91.6	3±4%

There is a sample graph representation which would be our best result to compare the algorithms and the datasets. That was mentioned in figure 4 as follows.

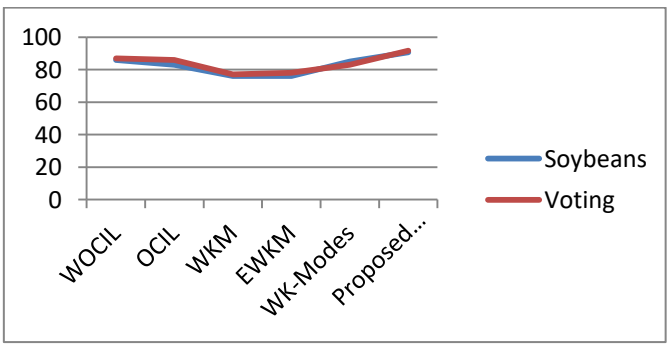

Figure 4: Comparison metrics.

V. CONCLUSION

Contradictory data is the most important thing to indentify the genuine information among the dissimilar information. The data points we get from the information from the new clusters with the highest dissimilarity with the DBScan and the Wards algorithm done well with handing the higher dimensional data than DBScan. DBScan works on the concept of Euclidian distance and this cannot handle the larger dimension information to form the cluster. Whereas the Wards algorithm with Minkowski distance metrics performs well and we achieved 95% accuracy in the result than the DBScan which achieved only 75% accuracy of identifying the similarity and dissimilarity in the data points.

REFERENCES

1. K. B. To and L. M. Napolitano, ``Common complications in the critically ill patient," *Surgical Clinics North Amer.*, vol. 92, no. 6, pp. 1519_1557, 2012.
2. C. M. Wollschlager and A. R. Conrad, ``Common complications in critically ill patients," *Disease-a-Month*, vol. 34, no. 5, pp. 225_293, 1988.
3. S. V. Desai, T. J. Law, and D. M. Needham, ``Long-term complications of critical care," *Critical Care Med.*, vol. 39, no. 2, pp. 371_379, 2011.
4. N. A. Halpern, S. M. Pastores, J. M. Oropello, and V. Kvetan, ``Critical care medicine in the United States: Addressing the intensivist shortage and image of the specialty," *Critical Care Med.*, vol. 41, no. 12, pp. 2754_2761, 2013.
5. A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, ``Machine learning and decision support in critical care," *Proc. IEEE*, vol. 104, no. 2, pp. 444_466, Feb. 2016.
6. O. Badawi *et al.*, ``Making big data useful for health care: A summary of the inaugural MIT critical data conference," *JMIR Med. Informat.*, vol. 2, no. 2, p. e22, 2014.
7. C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.
8. D. Gotz, H. Stavropoulos, J. Sun, and F. Wang, ``ICDA: A platform for intelligent care delivery analytics," in *Proc. AMIA Annu. Symp.*, 2012, pp. 264_273.
9. A. Perer and J. Sun, ``Matrix_ow: Temporal network visual analytics to track symptom evolution during disease progression," in *Proc. AMIA Annu. Symp.*, 2012, pp. 716_725.
10. Y. Mao,W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, ``An integrated data mining approach to real-time clinical monitoring and deterioration warning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. 2012, pp. 1140_1148.
11. J. Wiens, E. Horvitz, and J. V. Guttag, ``Patient risk strati_cation for hospital-associated C. Diff as a time-series classi_cation task," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 467_475.
12. S. Saria, D. Koller, and A. Penn, ``Learning individual and population level traits from clinical temporal data," in *Neural Inf. Process. Syst. (NIPS), Predictive Models Personalized Med. Workshop*, 2010.
13. R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, ``Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1,pp. 314_322, Jan. 2015.
14. M. Ghassemi *et al.*, ``Amultivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 446_453.
15. I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, ``A pattern mining approach for classifying multivariate temporal data," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 2011, pp. 358_365.
16. An overview on evocations of data quality at ETL stage, March 15, Available at:https://www.researchgate.net/publication/276922204, An overview on evocations of data quality at ETL stage.
17. Handling Mislaid/Missing Data to attain data trait, published in IJITEE, Available at ISSN: 2278–3075 , Volume-8 Issue-12, October 2019, Page No. 4308-4311.

AUTHORS PROFILE

Ms. Sakshi Jolly, Research Scholar in Faculty of Computer Applications Department, MRIU, Faridabad, (India)**,**total 4 years of teaching experience in the field of computers. She has authored 5 research papers in journals/conferences in the area of DQ.

Dr. Neha Gupta, Associate Professor, Faculty of Computer Applications Department, MRIU, Faridabad (India) has total of 14+ year of experience in teaching and research. She is a Life Member of ACM CSTA, Tech Republic and Professional Member of IEEE. She has authored and co-authored 34 research papers in SCI/SCOPUS/Peer Reviewed Journals (Scopus indexed) and IEEE/IET Conference proceedings in areas of Web Content Mining, Mobile Computing, and Cloud Computing. She has published books with publishers like IGI Global & Pacific Book International and has also authored book chapters with Elsevier, CRC Press and IGI global USA. Her research interests include ICT in Rural Development, Web Content Mining, Cloud Computing, Data Mining and NoSQL Databases. She is a technical programme committee (TPC) member in various conferences across globe. She is an active reviewer for International Journal of Computer and Information Technology and in various IEEE Conferences around the world.