International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Handwritten Digit Recognition of MNIST Data using Consensus Clustering

Monica Rexy. F, Lavanya. K

Abstract: ***In Big Data, Pattern Recognition and Consensus Clustering techniques have growing importance to the academic and professional communities. Today there is a great concern for categorizing the data, as data in inappropriate category means inaccurate information, which in turn results wastage of resources and harming the organisation. Pattern recognition (PR) helps in avoiding poor categorization of data by identifying the correct structure of data in dataset. Recognizing a pattern is the automated process of finding the exact match and regularities of data, which is closely related to Artificial Intelligence and Machine Learning. PR acts as a primary step to provide clustering since it analyses the structure and vector value of each characters in dataset. Consensus Clustering (CC) also called as clustering ensembles, plays a significant role in categorizing and maintaining any type of data. This is a technique that combines multiple clustering solutions to obtain stable, accurate and novel results. In this paper, to implement PR and CC techniques, we use MNIST dataset which is a large database of handwritten digits that is commonly used for training various systems in the field of Machine Learning.***

Keywords: Consensus Clustering, Pattern recognition, MNIST Dataset, Handwritten digit recognition.

I. INTRODUCTION

A. Patter Recognition (PR)

Today's digital world is filled with Patterns and these patterns can either be observed physically or it can be derived mathematically with the help of certain algorithms. Some of the examples of pattern are colours on clothes, speech pattern, alphanumerics in data, etc. In computer science, pattern is referred by the value of its vector features. In a simple PR application [11], initially the raw data is processed then the resultant data is converted into a machine understandable format. This involves two major phases of operation i.e., classification of patterns and clustering the patterns.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
Monica Rexy. F, Computer Science & Engineering, VIT University, Vellore, India.
Lavanya. K, Computer Science & Engineering, VIT University, Vellore, India.

In classification phase, each pattern is assigned an appropriate label based on the abstraction and is controlled by supervised learning. In clustering, the pattern with the same label is partitioned and categorized under one group and is controlled by unsupervised learning.

B. MNIST Dataset

The MNIST dataset (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training systems. The MNIST database is a subset of a much larger dataset known as the NIST Special Database [9]. This dataset contains both handwritten numerals and letters. It represents a much larger and more extensive classification task, along with the possibility of adding more complex tasks such as semantic interpretations through words interpretation. The accessibility of this MNIST dataset has certainly contributed to its widespread usage. The whole dataset is comparatively small (when compared to many recent benchmarking dataset), free to access and usage, and is encoded and stored in entirely straightforward manner. The encoding does not depend on complex storage structure, compressions, or any data format. For this reason, it is made very easy to access and this dataset can be included from any platform or through any programming languages. The NIST dataset, by contrast to the MNIST, has remained difficult to access and use. Resulting to the higher cost and the availability of storage when it was collected, the NIST dataset was originally stored in a an efficient and compact manner. Though source code to access the data is provided, it was very challenging to use on some recent computing platforms. For this reason, the NIST have recently released a second edition of the NIST dataset. However, the encoding of that dataset remains in the original format from which MNIST was extracted. The representation of the MNIST dataset is shown in figure

Fig. 1. MNIST Dataset

C. Consensus Clustering (CC)

As we all know, Clustering defines the process of gathering a specific set of data related to some category and aggregating them with respect to their characteristics. Whereas Consensus Clustering (CC) also known as aggregated clustering, defines the more robust kind of approach that depends on multiple iterations of the selected clustering method on subset of the dataset.

The main reason behind recommending CC [12] over simple clustering approach in pattern recognition is, CC can identify the pattern with greater accuracy, since it undergoes several iterations. Furthermore, MNIST dataset [10] consists of alphanumerics repeatedly in different fonts and styles. Here, the simple clustering approach could not understand this situation and it categorizes those repeated alphanumerics with different fonts in a separate individual category. But, CC works dramatically by subcategorizing the fonts under the category of alphanumerics. By this way, CC would be the best pairing to work with pattern recognition.

II. LITERATURE SURVEY

Zomaya et al. [2] proposes a survey about different categories of clustering algorithms. Some of the categories of clustering include Partitioning based clustering, Hierarchical clustering, Density based clustering, Grid based clustering and Model based clustering technique. In their article, they present a comparative definition between all these five categories with its most suitable algorithms. Their major were to find out the best suitable algorithm to cluster any big data. In [1], the author mainly focused on most popular and widely used algorithm, called K-means algorithm. Moreover they represented a detailed study on the functions and working process of K-means algorithm. Furthermore, another recent research [3] produces a simplified view on data mining algorithms. It also suggested some of the recent platforms that could have been used in Big data with its merits and demerits. This paper [8] discusses some of the effective algorithms which are applied in data mining. Their work is mainly to find out the most appropriate algorithm among all by applying comprehensive comparison technique. Nagpal and Mann's [6] research work does not include many recent clustering techniques. Despite, it only deal with Density based clustering algorithms, like DBSCAN and DENCLUE. They have also discussed about the advantages of density based clustering algorithms and its challenges. All the people in [7] were so interested in doing research about classification of algorithms that are used in the field of statistics and to apply then in specific traditional databases. Researchers in [5] presented an approach about the analysis and functioning of some very old algorithms that handled very large datasets. For example, Nearest neighbour search, Decision Tree and Neural Network. Herawan et al [4] discussed about various clustering techniques, that includes MapReduce and Parallel classification using MapReduce. They represented an overview of several categories in data mining and the pages of its clustering algorithms. Salma et al. [5] presented that clustering techniques in big data could be categorised under two major phases. They are single machine clustering techniques and Multiple machine clustering techniques. Later, their work draws more attention, because their proposed approach were faster and more adaptable to many challenges in big data. Their study covers most of the recent techniques and widely dealt with several different categories of data mining clustering algorithms. Henceforth, they additionally proposed few techniques that are used to fix new requirements in big data context. Example, Dimension reduction, parallel classification and Map Reduce framework.

III. PROPOSED MODEL

In Pattern Recognition, the selection of attributes and the representation of characters are said to be more important. Initially the MNIST dataset is put into the training phase which comprises of random projection of data, feature extraction and learning. When the input is fed into the training phase, every characters in MNIST dataset is subjected to be displayed in random manner, so that all the characters can be visualised clearly. After visualisation, each and every characters in the dataset is extracted individually to analyse its structure and vector value. The extracted structure then goes through a very important process called 'Learning'. Learning is a phenomena through which a system gets trained and becomes adaptable to give result in an accurate manner. It decides how well the system performs on the data provided to the system depends on the algorithm used on the data. These three steps in training phase are iterated until all the characters in the MNIST dataset get trained completely. In testing categorizing part, the characters with similar structure and vector value are grouped together under a single label, therefore some 'n' number of groupings or clusters are formed. From this formation of basic clusters, every character's feature is again analysed to find out it's styles and formats by applying CC algorithm. CC algorithm subcategorizes all the clusters by analysing the structure and it takes the best fit of all results with greater accuracy. By processing all the above mentioned steps the pattern recognition of MNIST dataset is done using consensus clustering. The diagrammatic representing of the processes in PR is shown in figure 2.

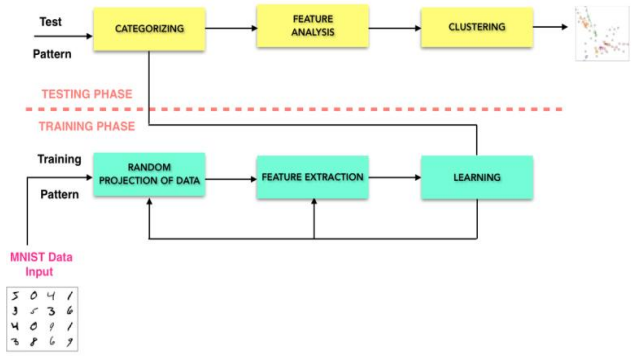

Fig. 2. Block diagram of Pattern Recognition

A. Algorithm for Consensus Clustering (CC)

Consensus Clustering (D, Resample, H, P, A, C)

Input: D is the MNIST Dataset in matrix format Mi [X0 , …, X19 , Y0,…,Y19] E N

Resample is the re-sampling scheme used for extracting a subset from dataset

H is the number of resamples

P is the percentage of rows extracted each time in the sub-sampling procedure

A represents Consensus clustering algorithm

Fig. 4. Projection of data

Output: C is the updated set of cluster, C = {Ci,..Ck}, k E N

Step 1: for each Mi do

Step 2: initialize empty connectivity matrices

Step 3: set createNewCluster = True;

Step 4: for 1>h>H do

perform Resample on D and assign to D(h)

Step 5: group elements of D(h) in C clusters using algorithm A

Step 6: build a connectivity matrix based on A's results for C

end for

Step 7: using connectivity matrices build a consensus matrix Mx,y for C

end foreach

Step 8: return C(Mx,y)

IV. AN ILLUSTRATIVE EXAMPLE

For demonstration purpose, we use 4X4 MNIST dataset with different styles of numerical values. Representation of the example dataset is shown in figure 3.

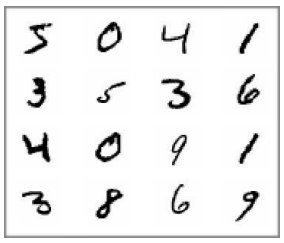

Fig. 3. Example Dataset

Step 1: The input dataset is taken in matrix format (Mi). All the data in Mi is analysed inside a for loop and projected separately to visualise and identify pattern. The diagrammatic representation for the projection of data is shown in figure 4.

Step 2: After analyzing all the data in dataset, empty connectivity matrix is initialized. This empty connectivity matrix is a matrix that consists of numerals with different structures. For example, we have numeral '5' with two different structures. Diagrammatic representation for different types of data's structure is shown in figure 5. From the diagram, 'structure A' of data is stored in one matrix and 'structure B' of data is stored in empty connectivity matrix which is just initialized, since structure B is different from structure A.

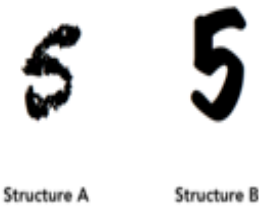

Fig. 5. Different structures of numeral 5

Step 3: By following step 2, the segregated types of data is categorised under a new cluster by setting createNewCluster = True. Until step 3, traditional K-means method of clustering is performed and is shown in figure 6.

Fig. 6. Generalised method of clustering

Step 4: In this step, resampling (analyzing data's structure) of each data inside various clusters is done and the result of resampling is assigned to D(h). D(h) is considered as a matrix that contains the numerals of different structures. Resampling of data is shown in figure 7.

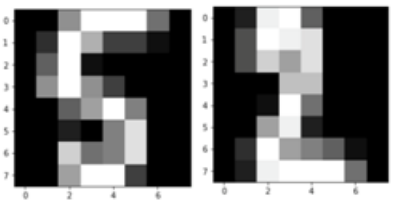

Fig. 7. Resampling of data

Step 5: All the elements in D(h) are grouped together to form a consensus clusters (C). In figure 8, the cluster with

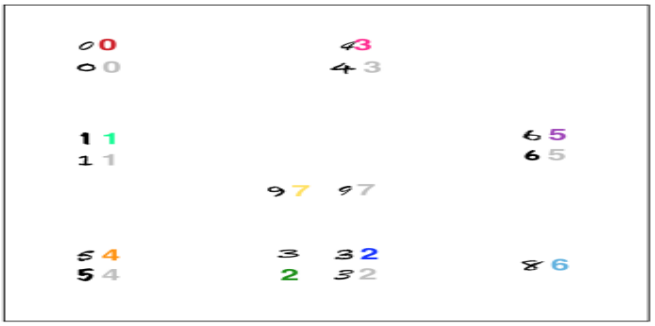

Fig. 8. Consensus clustering

grey colour label indicates consensus clustering's.

V. RESULT ANALYSIS

A. Dataset

In this paper, we use MNIST dataset to compute consensus clustering. The MNIST dataset is a large collection of handwritten digits which is commonly used for training various image processing systems. It was created by mixing the samples from NIST's original datasets. [13] The MNIST database contains 60,000 training images and 10,000 testing images. It was constructed from NIST's Special Database 3 (SD-3) and Special Database 1(SD-1) which contain binary images of handwritten digits. NIST originally uses SD-3 as their training set and SD-1 as its test set, moreover SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be recognized by the facts that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Extracting meaningful results from learning experiments requires, the result should be independent of the decision of training set and test set among the complete set of samples. Therefore it meant be necessary to build a new database by mixing NIST's datasets. The MNIST training set is acquired by combining 30,000 patterns from SD-3 and 30,000 patterns from SD-1. In this paper we use MNIST dataset that contains 20 rows and 20 columns, where each rows consist of numerals 0 to 5.

B. Performance analysis of Consensus Clustering

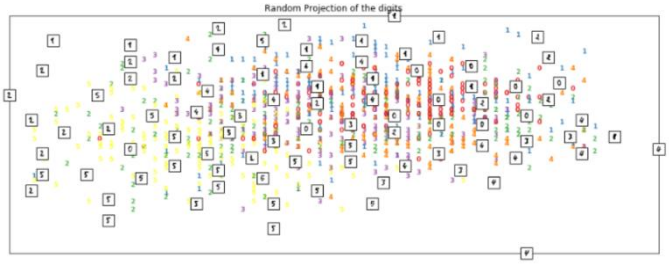

Fig. 9. Random projection of data

i. Projection of Data

Initially, the 20X20 MNIST dataset (Mi) is loaded into the program in matrix format and all the data in dataset is analysed inside a for loop to categorise the numerals. After analysing, each data is projected to the viewer by some random manner to visualise and recognize the structure of data. The purpose of this step is to make sure and verify the types and format of numerals in the dataset, also this random projection of data acts as a primary step to perform traditional clustering process and is represented in figure 9.

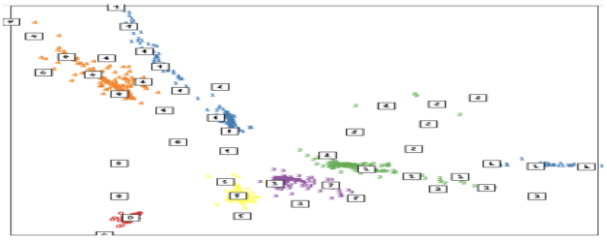

Fig. 9. Generalized method of clustering

ii. Generalized Method of Clustering

After projecting and analysing the data in step (i), all the numerals are categorized into different clusters with respect to its numerical value. For this generalized model of clustering, Traditional K - means methodology is applied, so that the results are expected to be very approximate with average level of accuracy. Here the data is recognized only by its numerical value but not by it's format and structure. Diagrammatic representation for generalized method of clustering is shown in fi
gure 9.

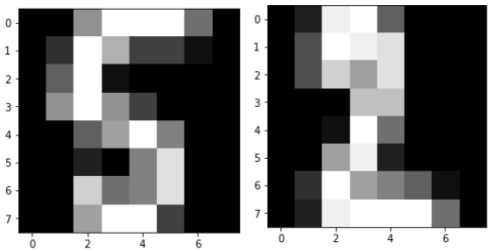

Fig. 10. Resampling of

iii. Pattern Recognition (PR

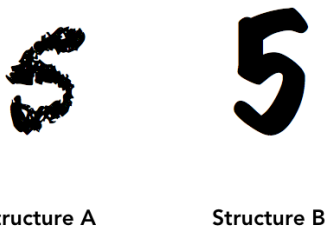

Fig. 11. Different structures of numeral 5

PR defines the process of identifying the structure of data by its vector value. This is done by resampling (analyzing different data structures for single numeral) all the data inside every clusters that are formed by traditional

method and the result of resampling is assigned to a matrix D(h). D(h) matrix consist of one specific structure of numeral, then another structure of a numeral is assigned to the base cluster which is formed before. For example, we have numeral '5'
with two different structures. 'Structure A' of the numeral is stored in base cluster and 'structure B' of data is stored in D(h) matrix. Diagrammatic representation for resampling of data and different structures of numeral 5 are shown in figure 10 and in figure 11 respectively.

iv. Consensus Clustering (CC)

After assigning the data to base cluster and D(h) matrix, all the elements in D(h) matrix are grouped together to form consensus clusters. Diagrammatic representation for consensus clustering is shown in figure 12. From this figure, red colour label under every base cluster indicates the consensus cluster. By overcoming the drawbacks in generalized K-means clustering, Consensus Clustering approach not only segregates the category of data but also identifies its structures and formatting styles with greater accuracy.

C. Comparison between various clustering technique.

With reference to the below graph, it is clearly understood that the dataset used in this paper delivers 95% accurate result by computing Consensus Clustering technique, while DBScan covers 70% of accuracy, Gaussian Mixture Model produces the result with 75% of accuracy, while Bisecting K-Means produces accuracy with 90%. From this we can conclude that consensus clustering achieved excellence in dividing our dataset into clusters.

VI. CONCLUSION

Pattern Recognition normally aims to provide a very reasonable answer to all possible inputs by taking its statistical variation into account. In this paper, the implementation of consensus clustering is discussed by taking MNIST dataset as the sample application. These implementation is done in the language Python with Anaconda package and Consensus Clustering package by using the tool Jupyter Notebook. And the comparative study with plot map is done between generalized clustering approach and consensus clustering approach to understand the accuracy of results in CC. In addition to that, CC is also used to determine the optimal number of clusters for a dataset and clustering algorithm. In Future, this CC approach can be applied in any stream and is very helpful to determine the best fit of datapoint among various comparisons of data in same dataset.

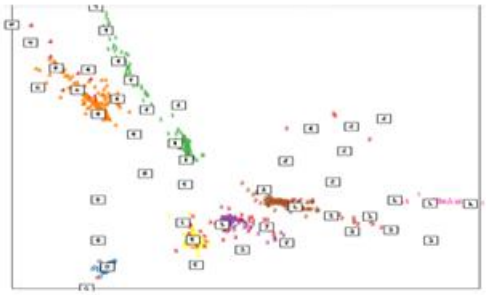

Fig. 12. Consensus clustering

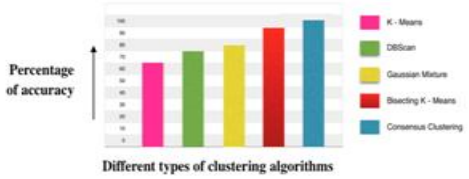

Fig. 13. Comparison between different clustering techniques

REFERENCES

1. https://www.journals.elsevier.com/pattern-recognition
2. https://en.wikipedia.org/wiki/Pattern_recognition
3. https://www.geeksforgeeks.org/pattern-recognition-introduction/
4. http://rasbt.github.io/mlxtend/user_guide/data/mnist_data/
5. A. Fahad, N. Alshatri, Z. Tari, A. ALAmri, A. Y. Zomaya, I. Khalil, F. Sebti, and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis," IEEE transactions on emerging topics in computing, 2014
6. A.BEN AYED, M.BEN HALIMA and M. ALIMI, "Survey on clustering methods: Towards fuzzy clustering for Big Data," In Soft Computing and Pattern Recognition (SoCPaR), 6th International Conference of. IEEE, p. 331-336, 2014.
7. A. Sherin, S. Uma, K.Saranya and M. Saranya Vani "Survey On Big Data Mining Platforms, Algorithms And Challenges". International Journal of Computer Science & Engineering Technology,Vol. 5 No, 2014.
8. S.ARORA, I.CHANA, "A survey of clustering techniques for Big Data analysis," in Confluence The Next Generation Information Technology Summit (Confluence), 5th International Conference-. IEEE.
9. P. Batra NAGPAL, and P. Ahlawat MANN, "Survey of Density Based Clustering Algorithms," International journal of Computer Science and its Applications.
10. R. Xu and D. WUNSCH, "Survey of clustering algorithms," Neural Networks, IEEE Transactions.
11. C. YADAV, S. WANG, et M. KUMAR, "Algorithm and approaches to handle large Data-A Survey," International Journal of computer science and network, vol 2, issue 3, 2013.
12. A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big Data Clustering: A Review," In Computational Science and Its Applications–ICCSA 2014. Springer International Publishing, p. 707-720. 2014.
13. http://bigdata-madesimple.com/what-is-clustering-in-data-mining/
14. https://towardsdatascience.com/consensus-clustering-f5d25c98eaf2
15. https://knowledgent.com/whitepaper/building-successful-data-quality-management-program/

AUTHORS PROFILE

Monica Rexy. F is currently doing M.Tech in School of Computer Science and Engineering(SCOPE) in VIT, Vellore. She also received BE degree in Computer Science and Engineering from the Pondicherry University, India, in 2014. Her current research interests include Information Security, Big Data Analytics, Artificial Intelligence.

Dr. K.Lavanya is currently working as an Associate Professor in the School of Computer Science and Engineering(SCOPE) in VIT, Vellore. She received her Ph.D. degree in Computer Science and Engineering from VIT University, Vellore, on August 2015 [July 2011 - August 2015]. She completed ME in Computer Science and Engineering from VIT University, Vellore, in the year 2011. She also received BE degree in Computer Science and Engineering from the Anna University, India,