A New Clustering Method Using an Augmentation to the Self Organizing Maps

Divyansh Kumar Roy
Department of Computer Science & Engineering
ASET, Amity University Uttar Pradesh
Noida, India
divyanshkumarroy@gmail.com

Hari Mohan Pandey
Department of Computer Science
Middlesex University
London, U.K.
profharimohanpandey@gmail.com

Abstract*—A technique is developed using Self Organizing Maps (SOM) to efficiently cluster the data and it is compared with existing clustering Techniques such as K-Means clustering, Hierarchical clustering and SOM Clustering. The proposed technique is used to cluster an Earthquake dataset and the performance is compared with the other existing clustering technique. The experimental results show that the proposed clustering method demonstrated better results as compared to other clustering methods.

Keywords: Clustering, Self Organizing Map (SOM), Hierarchical Clustering and K-means clustering.

I. Introduction

Clustering is the method used to group data into sets having similar characteristics. It can be used to observe common patterns in the data. Well formed clusters are those which are properly segregated and represent an order. Labeled data is easier to cluster as a penalty and reward system can be put into place to facilitate the efficient clustering of the data. However, it is difficult to cluster the unlabelled data since, there is no specific standard against which the clustering can be tested and the data is large enough to be properly clustered by human intervention. The need for Clustering was observed since long ago and several different clustering algorithms have been developed. However, lately there has been an increase in the use of these concepts to properly identify clusters. This has shown the various pros and cons of the clustering algorithms in place.

Over the years many algorithms have been developed to cluster the datasets. K Means clustering is a widely used technique which initializes by randomly assigning the cluster centroids for the K Clusters. Each iteration of the algorithm involves assigning data points to their nearest clusters and then recalculating the cluster centroid. It is a very simple and effective technique. However, it produces different cluster outputs based on the initial clusters. This drawback was removed by the K Means++ optimization where the cluster centroids were initialized according to a specific method rather than a random initialisation. The initial centroids were chosen in such a manner that they are distant from each other.. This technique provides better clustering outputs compared to its predecessor. Apart from that K Means also produces unsatisfactory outputs when the clusters don't possess the same variance.

Hierarchical clustering is a technique which produces an output in the form of a tree called a dendogram which is used to divide the dataset into any number of clusters [11]. However, it is a computationally expensive algorithm when the number of features is high. An improvement to the algorithm can be made when the number of clusters to be formed is predefined. The Top Down approach should be used when the value of K is small and the Bottom Up approach should be used when the value of K is large. It works well with both Gaussian Clusters and irregularly shaped clusters due to the several linkage options available. The Gaussian Clusters are formed properly due to the complete linkage option which allows every point in the cluster to be linked to the newly added point. The representation looks similar to a complete graph representation. The irregularly shaped clusters are formed due to the Single linkage option in Hierarchical clustering which enables linking of only one data point with the other which looks similar to an Euler Tour in a graph. Apart from these two linkages there are several other options available for different usages.

DBSCAN is one algorithm which has been highly appreciated for its performance. It is able to identify clusters on the basis of their densities. It works well with both Gaussian clusters and irregularly shaped ones. The main advantage being that the algorithm brings with itself the ability to filter out noise in the dataset. However, choosing the value of its initial parameters is a tough job which requires an expert. Otherwise, it will output clusters which are improperly formed. Additionally, the algorithm doesn't work well on datasets where the densities of the clusters have a large variance.

Self Organizing Map or Kohonen's Map (SOM) is a competitive learning algorithm based on Neural Networks [10]. It uses the concept of neurons which represent the clusters. The algorithm is highly useful as it can reduce an N dimensional feature representation to a 2D or 1D representation. It can be used for retraining the neural network without much modification to the algorithm. The initialisation phase starts with assigning random weights or coordinates to the neurons which represent the cluster centroids. Each iteration involves picking a data point and finding out the neuron closest to it. This is a competitive approach where neurons compete against each other to be chosen the closest neuron. The chosen neuron's weights are adjusted in such a manner that it comes closer to the data points whereas there is no change in the other neurons. The amount by which the neuron's weights are modified decreases gradually and is called the learning rate. The performance of the Self Organizing Map is reduced by the random initialisation of Neuron weights similar to K Means. An improvement in the initialisation of these weights can increase the performance of the algorithm significantly.

The aforementioned methods work well but there is no single best method amongst them. Usually the methods are used interchangeably based on the type of data which requires inputs from an expert. This human dependency should be gradually removed by developing an algorithm which outperforms the others irrespective of the type of clusters present in the dataset. The K Means Algorithm works well only if the actual data has clusters of similar sizes. Whereas, the Hierarchical Clustering seems a little impractical as it gives all the possible choices of cluster sets and not the actual answer to the problem [12]. The Self Organizing Map on the other hand fails to outperform these algorithms.

The rest of the paper is organized as follows: Section II presents background studies on clustering algorithms. We have presented the proposed algorithm in Section III. Section IV gives the experimental setup, results and analysis. Finally, we present concluding remark of this paper in Section V.

II. Background Study

The random initialization of these neuron's weights is a source of diminished results. The enhanced performance of the Kohonen's map can be extracted only if the neuron's are correctly initialized. The SOM converges to a local minimum rather than a global minimum hence if the neuron weights are randomly assigned there is no specific control on the clustering. This form of clustering will produce good results but it could be better if the initialisation was based on the pattern in the dataset. This can only be done in the case of batch processing of the dataset. There are several methods to initialize the neuron's weights which have been already developed.

An alternative to the random small valued weights for the neurons is the random initialisation of data points to the neurons. This technique serves better than the former one due to its accurate scaling dependent on the data. However, it is independent of the patterns in the dataset and is hence not much of an improvement to the original random initialisation.

A technique which combines K Means and SOM is occasionally used where the output of the K Means technique is used to initialize the neuron weights. It is even more computationally expensive that both K Means and SOM and it don't always produce superior outputs.

A technique named SOM++ was developed by Dogan. The neuron's weights are initialized by the help of the K-Means++ Algorithm. These weights are used as the starting point in the Kohonen SOM. It is faster to converge to the result than the traditional algorithms and it produces better results when compared to them.

An initialisation algorithm proposed by *"Ehsan Mohebi and Adil M. Bagirov in A New Modification of Kohonen Neural Network for VQ and Clustering Problems"*[6] is based on their split and merge procedure which is efficient in identifying areas with high density in the dataset. This helps in converging to a better local minimum than the original method.

Another initialisation method was proposed by "*Madhusmita Mishra and H.S. Behera in Kohonen Self Organizing Map with Modified K-means clustering"* [8] for High Dimensional Data Set where they use SOM to get the number of clusters. This output is used by a Genetic Algorithm which generates new initial centroids which are used by the K Means Algorithm. This technique is helpful as it is able to find out an appropriate value of K and the Genetic Algorithm finds out good initial cluster centroids. However, the use of SOM followed by K Means is computationally expensive even if we account the time saved due to faster convergence because of better initialization of centroids.

There are several initialisation techniques which have been formulated up till now but they're generally computationally expensive since they often use some other algorithm along with the original SOM which increases the time complexity of the respective approaches.

There is a requirement for a technique which is computationally inexpensive and is able to outperform the original SOM technique. It is noteworthy that a technique so developed could be used by other Clustering methods as well which require an initial cluster representation such as K Means. Hence a technique which has a time complexity that is O(N) or O(N log N) is required. A complexity this low would ensure that the initialisation algorithm takes almost no time when compared to the rest of the SOM procedure.

III. Proposed Clustering Algorithm

The technique proposed by us assigns specific weights to the neurons instead of a random assignment. The aforementioned clustering algorithms require a value of the number of clusters. K Means and SOM require the value of K to initialize their cluster centroids and neurons respectively. Hierarchical Clustering on the other hand doesn't require a value of K to produce it's Dendogram but it does require the value to give a result of the clustering algorithm. Similar to these methods, our approach requires the number of clusters to be formed from the data.

Algorithm-1: Cluster Centroid Initializations

Input: Point Set (P)

Output: Cluster Centroids for Initialisation

```
1.  Begin
2.      For p in P:
3.          Total Distance(p) ← $\sum$ Distance(p, a) for all a in P
4.      Store Extreme values of Total-Distance is Steps 2-3.
        Maximum Distance = max(Total Distance(p))
        Minimum Distance = min(Total Distance(p))
5.      Calculate Difference from Extreme values in Step 4.
        Difference = (Maximum Distance - Minimum Distance) / K
6.      For i in 1-K
7.          Cluster Distance(i) = Minimum Distance + Difference * (i-0.5)
8.          Cluster Centroid(i) = Coordinate(p) where
            |Total Distance(p) – Cluster Distance(i)| =
            min( |Total Distance(a) – Cluster Distance(i) |
            for all a in P.
9.      Initialize Cluster Centroids with the values obtained in Step 8.
10. End
```

Algorithm-1 is given in this paper to determine the initial cluster centroid positions. We consider every point and compute their distance from every other point. We note the extremum values of these distances in Step 3. We use these values in Step 5 to calculate the average distance between the distance metric of the cluster centroids. Then, we allocate the corresponding distance metrics to the cluster centroids in Step 7. Finally, we find the data points which have their distance metric closest to each centroid and assign their weight to the centroid's weight in Step 8.

The existing version of SOM that have been utilized in the earlier scientific literatures assigns random weights to the neurons. This arbitrary assignment seems ideal due to the randomness present but it isn't a good fit according to the data. The final result is affected since the assignment leads to a local minimum which is not as optimal as the global minimum. Attaining the global minimum is a rather hard problem. The proposed Algorithm-1 has the ability to overcome from the aforementioned issue. This is because it initializes the cluster depending on the actual dataset. We have observed that points that are nearer to each other tend to have similar values of the distance metric used by us. This hint is useful for classifying the data into K clusters.

IV. Experimental Setup, Results And Analysis

Extensive experiments have been conducted using Python 3.6.1 in an i7 7th Gen Processor clocked at 2.70 GHz and an 8 GB Memory. We have taken the earthquake data from National Center for Environmental Information, National Oceanic and Atmospheric Administration (NOAA) to perform our experiments [10]. The 1400 tuple data contains the latitudes and longitudes of the Earthquakes held from around 1000 AD to Present. The data is provided to the algorithms in the format of one whole training set and not in an online manner. The algorithms in comparison are used to cluster the data into 100 clusters.

In order to compare the performance we collected the results of K-Means, Hierarchical and SOM Clustering on the data. They were compared with the proposed algorithm on the basis of the following Performance Metrics:

1. Minkowski Euclidean Distance: It is the distance of the cluster centroid from the data points in the cluster. A lower value of the metric is desirable and represents clusters whose data points are closer to each other.

$$d(x,y)=\sqrt[q]{\sum_{i=1}^{n}(x_i-y_i)^q} \qquad (1)$$

Where, d (x, y) is the distance between two objects x and y. The number of features is represented by n. The value of q remains 2 for the Euclidean Distance.

2. Silhouette Value: It compares the distances of every point in a cluster to other points in the same cluster with its distance from every point in the neighbouring cluster. A higher value is desirable as it's represents well formed clusters. [9].

$$a(o_i)=\frac{1}{|C_A|-1}\sum_{o_j\in C_A, o_j\neq o_i} d(o_i,o_j) \qquad (2)$$

$$b(o_i)=\min_{C_B\neq C_A}\frac{1}{|C_B|}\sum_{o_j\in C_B} d(o_i,o_j) \qquad (3)$$

$$sil(o_i)=\frac{b(o_i)-a(o_i)}{\max\{a(o_i),b(o_i)\}} \qquad (4)$$

Where, sil(o_i) is the silhouette value for an object o_i. The cluster to which a point belongs is represented by C_A and the nearest cluster to the point is represented by C_B. The number of points in a cluster C_A is represented by $|C_A|$.

3. Average Error: It is the average distance between a data point and it's respective cluster centroid. A lower value is desirable as it represents proper assignment of the clusters. [9].

$$E(C)=\left[\sum_{i=1}^{K}\sum_{o\in C_i}\text{distance}(o_i,cen_i)\right]\Big/N \qquad (5)$$

Where, E(C) is the Average Error in the clustered output. The number of clusters is represented by K and the number of objects is represented by N. The object o belongs to a cluster C_i and cen_i represents the centroid coordinates of cluster C_i.

4. Cluster Utilisation: It is the ratio of the number of clusters utilised from the K clusters given. A higher value represents higher productivity since lesser space is wasted for unassigned clusters.

$$u = \frac{Number\ of\ \ Clusters\ Formed}{K} \tag{6}$$

TABLE I
RESULTS MATRIX OF VARIOUS CLUSTERING METHODS

Methods/Metrics	Metric 1	Metric 2	Metric 3	Metric 4
K Means	824.749	0.589336	1.97898	0.91
Hierarchical	1291.71	0.599106	2.23892	0.85
SOM	817.337	0.379404	2.78728	0.97
Proposed Algorithm	814.085	0.392424	2.71739	0.98

The following observations have been made using the results presented in TABLE I.

Metric 1 (Average Minkowski Euclidean Distance): A lower value of this metric is necessary since it represents closely shaped cluster members.

Metric 2 (Average Silhouette Value): A high value of this metric is necessary and it represents that the cluster members are closer to the centroid as compared to other neighboring centroids.

Metric 3 (Average Error): A lower error value is required. It represents correct centroid assignment.

Metric 4 (Cluster Utilization): A higher value of cluster utilization represents that the maximum number of clusters are utilized.

1. It has the Lowest Minkowski Euclidean Metric Value amongst all the Algorithms in comparison.
2. It has a Higher Average Silhouette Value as compared to standard SOM.
3. It has a Lower Average Error than the standard SOM.
4. Its cluster utilisation is the highest among all the techniques in comparison.

V. CONCLUSIONS

In this paper, we have presented a novel algorithm (Algorithm-1) to determine the initial cluster centroids for a Kohonen Map. The experimental results revealed that the proposed algorithm outperformed the other algorithms as it had better values of the performance metrics. It has a worse Average Silhouette Value and Average Squared Error than K-Means but an overall better result implies that it is a more practical option. The proposed approach produced better results than the SOM in all the criterions. Hence, it is recommended to use the proposed method for initializing the clusters centroids before using the Kohonen Map. The proposed algorithm evenly divided the Euclidean Distances among the clusters. Due to this, the average difference between distances assigned to Cluster$_i$ and Cluster$_{i+1}$ became the same for all values of i. However, in some datasets the clusters could be present in an uneven order. The above mentioned regular order would hinder the possible accuracy of the technique in those cases. The current technique enables online clustering by calculating the cluster centroids from the intermediate results. These centroids can be provided as input whenever more data points are available. However, with this technique the number of clusters wouldn't increase from the first step. This limitation can be diminished by assigning a little more neurons from the amount already being used and initializing them with Algorithm 1. Although, this too won't be the best solution to it. A technique needs to be developed that properly incorporates the modifications made in the dataset by adjusting the cluster centroids and the number of these clusters.

REFERENCES

1. Kohonen, Teuvo. (2012). Essentials of the self-organizing map. Neural networks : the official journal of the International Neural Network Society. 37. 10.1016/j.neunet.2012.09.018.
2. David Arthur , Sergei Vassilvitskii, k-means++: the advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, p.1027-1035, January 07-09, 2007, New Orleans, Louisiana
3. Martin Ester , Hans-Peter Kriegel , Jörg Sander , Xiaowei Xu, A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, August 02-04, 1996, Portland, Oregon
4. Mihael Ankerst , Markus M. Breunig , Hans-Peter Kriegel , Jörg Sander, OPTICS: ordering points to identify the clustering structure, Proceedings of the 1999 ACM SIGMOD international conference on Management of data, p.49-60, May 31-June 03, 1999, Philadelphia, Pennsylvania, USA [doi>10.1145/304182.304187]
5. (Dogan Y., Birant D., Kut A. (2013) SOM++: Integration of Self-Organizing Map and K-Means++ Algorithms. In: Perner P. (eds) Machine Learning and Data Mining in Pattern Recognition. MLDM 2013. Lecture Notes in Computer Science, vol 7988. Springer, Berlin, Heidelberg)
6. Mohebi, Adil M. Bagirov (2013) A New Modification of Kohonen Neural Network for VQ and Clustering Problems Ehsan. Australasian Data Mining Conference, Canberra, Australia
7. Akinduko, Ayodeji & Mirkes, Evgeny. (2012). Initialization of Self-Organizing Maps: Principal Components Versus Random Initialization. A Case Study.
8. Mishra, M & Behera, H. (2012). Kohonen self organizing map with modified k-means clustering for high dimensional data set. International Journal of Applied Information Systems. 2. 34-39.
9. Sabine Schulte im Walde, Experiments on the Automatic Induction of German Semantic Verb Classes, Computational Linguistics, v.32 n.2, p.159-194, June 2006 [doi>10.1162/coli.2006.32.2.159]
   National Geophysical Data Center / World Data Service (NGDC/WDS): Significant Earthquake Database. National Geophysical Data Center, NOAA. [doi>10.7289/V5TD9V7K]
10. D. E. Rumelhart , D. Zipser, Feature discovery by competitive learning, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations, MIT Press, Cambridge, MA, 1986
11. Chen, L.F., Jiang, Q.S. and Wang, S.R., A hierarchical method for determining the number of clusters. Journal of Software. v19 i1. 62-72.
12. Rui Xu , D. Wunsch, II, Survey of clustering algorithms, IEEE Transactions on Neural Networks, v.16 n.3, p.645-678, May 2005 [doi>10.1109/TNN.2005.845141]