International Journal of Research in Signal Processing, Computing & Communication System Design
4 (2), 2018, 28-32
http://www.publishingindia.com/ijrsp

Clustering of Hand Written Digits Using K-Means Algorithm and Self Organizing Maps

Maddimsetti Srinivas[1*], M. Venkata Srinu[2] and G. L. P. Ashok[3]
[1]Koneru Lakshmaiah Education Foundation, Vijayawada, Guntur(Dt.), Andhra Pradesh, India.
Email: maddimsetti34@kluniversity.in
[2]Koneru Lakshmaiah Education Foundation, Vijayawada, Guntur(Dt.), Andhra Pradesh, India.
Email: venkat486@kluniversity.in
[3]Koneru Lakshmaiah Education Foundation, Vijayawada, Guntur(Dt.), Andhra Pradesh, India.
Email: glpashok@kluniversity.in
*Corresponding Author

Abstract: Present work focuses on clustering of MNIST dataset using K-means clustering and Self-Organizing Maps (SOM). Histograms of Oriented Gradients (HOG) descriptors are used to extract the feature vectors and Principal Component Analysis (PCA) is applied on feature vectors to reduce the dimensionality. First two principal components are taken for cluster formation. Purity of cluster metric is used to evaluate the clusters. External criteria with prior information of true class is chosen to validate cluster. The performance of SOM is better than K-means in forming clusters. Out of 10 clusters K-means algorithm missed clusters of 3 digits (0, 7 and 9) whereas SOM missed clusters of 2 digits (5, 9).

Keywords: Clustering, Histograms of Oriented Gradients (HOG), K-means clustering, MNIST, Principal component analysis, Self organizing maps, Unsupervised learning.

I. Introduction

Hand written digit recognition is used in many applications in computer vision. Classification of hand written digits using standard machine learning algorithms by [1] used a database of 1200 samples of 12 writers. Based on the LeNet5 convolutional neural network architecture [2] a trainable feature extractor for hand written recognition is introduced using standard MNIST dataset [18]. The other comparative study on hand written recognition using Back Propagation, K-Nearest Neighbour (KNN) and Radial Basis Function done by [3]. The tested databases used by [4] are CENPARMI, CEDAR, and MNIST using gradient features and KNN, Support Vector Classifier (SVC) with Polynomial and Gaussian kernals, Multi Layers Perceptron (MLP) and Linear Vector Quantization (LVQ). Street View House Numbers (SVHN) Dataset is one more dataset similar to MNIST presented by [5] used unsupervised techniques for classification of hand written digits. Present work focuses on the clustering of hand written digits using standard MNIST dataset [18] using K-means clustering and Self Organizing Map.

Clustering algorithms are discussed in section 2, HOG feature extraction and dimensionality reduction is elaborated in section 3, MNIST dataset [18] details are given in section 4, cluster formation and evaluation is analysed in section 5, conclusion and future direction is given in section 6.

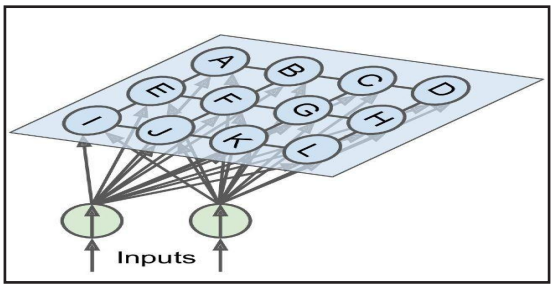

Fig. 1: Self Organizing Maps

II. Clustering Techniques

Clustering is a technique to grouping similar patterns, sets, or objects. Some of the applications of cluster analysis are text categorization [6], computer vision [7], information retrieval [8], marketing [9] and biology [10]. Clustering algorithms can be divided based on structure and operation. Different clustering methods are discussed in [14] [12]. In the present work K-means clustering [21] and Self Organising Map [19] are used to form clusters.

A. K-Means Clustering

In k-means clustering, k represents the number of clusters selected for partition of input instances. The k-means clustering

algorithm job is to assign each input instance to any of clusters selected. The algorithm can be explained as follows:

1. Out of all input instances select k instances as initial centroids using k-means ++ algorithm [21].
2. Calculate the distances of all input instances to all centroids.
3. The input instances which are close to the cluster centroid are joined to that cluster.
4. The average of all input instances in a cluster gives the new centroid. The new centroid for all k clusters is calculated in the similar way.
5. The above steps from 2 to 4 are repeated until the centroid value not deviated.

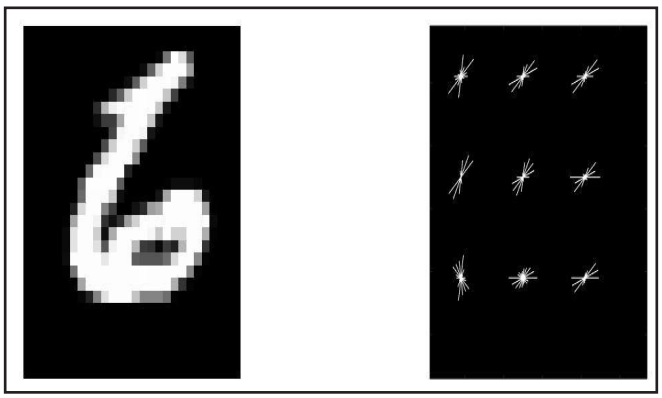

Fig. 2: HOG Visualization

B. Self Organizing Maps

Self Organizing Maps (SOM) [19] are used to produce a low-dimensional representation of a high-dimensional dataset, generally for visualization, clustering, or classification. The neurons are spread across a map as shown in Fig. 1 [22]. Each neuron has a weighted connection to every input. In the dataset a new instance will activate only one neuron whose weight vector is closest to the input vector. In general, instances that are nearby in the original input space will activate neurons that are nearby on the map. This makes SOMs useful for visuaslization.

Table I: HOG Parameters

HOG Feature Extractor Parameter List				
Cell Size	Block Size	Block Over Lab	Number of Bins	Orientation
[8, 8]	[2, 2]	Blocksize/2	9	[0, 180]

III. HOG Feature Extraction and Dimensionality Reduction

Histograms of Oriented Gradients (HOG) [17] are popular features for detecting objects. HOG can be applied in face recognition, pedestrian detection and other computer vision problems. HOG give excellent performance over wavelets, Scale Invariant Feature Transform (SIFT) [15]. To reduce the dimensionality of the input feature space Principal Component Analysis (PCA) [16] used [13]. The parameters of HOG are listed in Table I and HOG visualization for digit 6 is shown in Fig. 2.

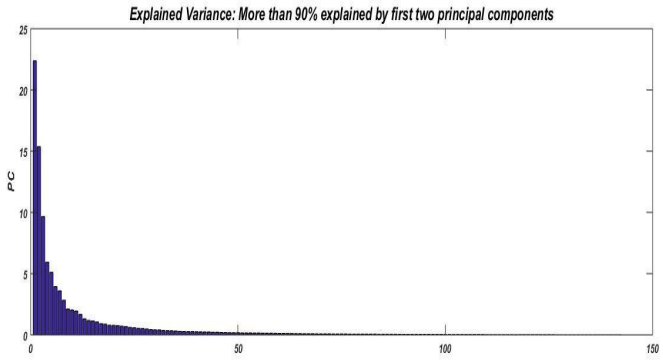

Fig. 3: Principal Components

IV. Hand Written Digit Dataset - MNIST

MNIST is most widely used hand written digit dataset in computer vision and it is available in mat file format, contains training instances, testing instances, training labels and testing labels. The dataset is divided as 60000 training instances and 10000 testing instances. The size of each input instance is 28*28 [4], [11], [2], [1] used MNIST dataset for their experiments.

Table II: Evaluation of K-Means Clustering

Purity Table - K-Means Clustering			
Cluster Number	Digit (Label)	Cluster	Members Purity(%)
1	6	1345	28.55
2	5	1005	28.55
3	4	890	30.33
4	1	442	92.53
5	8	1640	22.31
6	1	372	95.69
7	6	1138	23.55
8	2	1305	26.36
9	1	546	61.35
10	3	1317	35.99

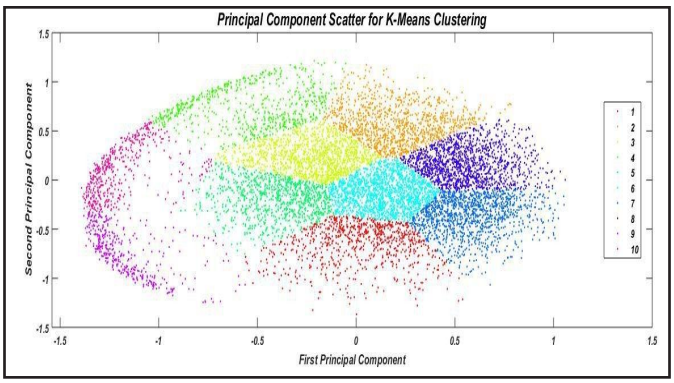

Fig. 4: K-Means Clusters

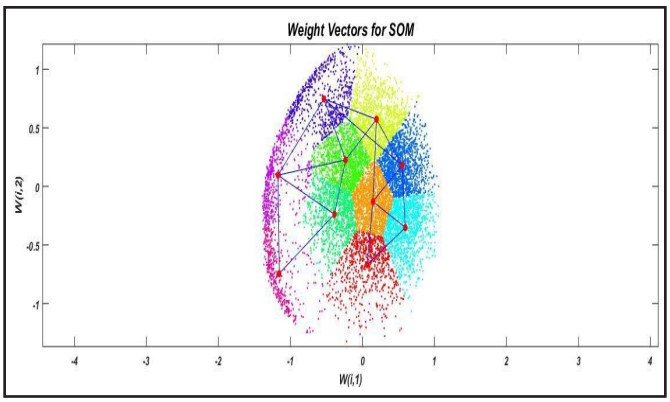

Fig. 5: SOM Clusters

V. Experimental Results

Comparison of clusters formed by K-means and SOM is analysed using purity and silhouette values of a cluster. Widely used MNIST dataset is taken input instances which are available as mat file in MATLAB. The dataset is divided as 60000 training instances and 10000 testing instances. The size of each input instance is 28*28 of gray scale. HOG features are extracted for each testing instances and the length of the feature vector is 144. The dimension of the feature vector is reduced using PCA for better visualisation. The first two principal components are less correlated as shown in Fig. 3, selected for further analysis. Purity of a cluster is defined as:

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_{k} \max_j |\omega_k \cap C_j| \quad (1)$$

where, $\Omega = \omega 1, \omega 2, \ldots \omega k$ is set of clusters and $C = c1, c2, \ldots ck$ is set of classes. The total number of clusters are selected are 10 for hand written digit dataset.

Table III: Cluster Evaluation of SOM

Purity Table - Self Organising Maps			
Cluster Number	Digit (Label)	Cluster	Members Purity(%)
1	7	823	18.00
2	8	1577	21.69
3	0	1159	21.72
4	6	1287	25.98
5	4	977	29.28
6	3	1277	34.16
7	2	1388	25.63
8	1	635	49.05
9	1	476	90.06
10	1	401	94.04

For K-means clustering, first two principal components are given as input. The cluster are shown in Fig. 4 for K-means clustering gives the cluster index with colour using first two principal components. SOM calculates the distance from an input instance from weight vectors for 10 neurons as shown on gray color in Fig. 6. The input instances which have minimum distances are activates the neuron and categorized as a separate cluster. Scatter plot gives an idea of how the weight vectors are changed for a particular input instance as shown in Fig. 5. Color patch size in Fig. 6 is based on the number of cluster members.

Table II shows the cluster index and corresponding digit classified based on voting of labels (classes). It can be observed from the table that 0, 7 and 9 digits not categorized into any of the clusters and these digits are added to cluster 5 i.e., digit 8. Compared to K-means clustering, the performance of SOM is better in the sense that only two digits 5 and 9 not classified to any cluster as highest in number shown in Table III. By investigation of labels 5 and 9, It is observed that these two digits are categorized to cluster 2 i.e., digit 8. The similarity between K-means and SOM is the missing digits clusters are categorised to digit 8. Further, investigation has to made how missing digits are joining the same digit 8 in the both clustering techniques.

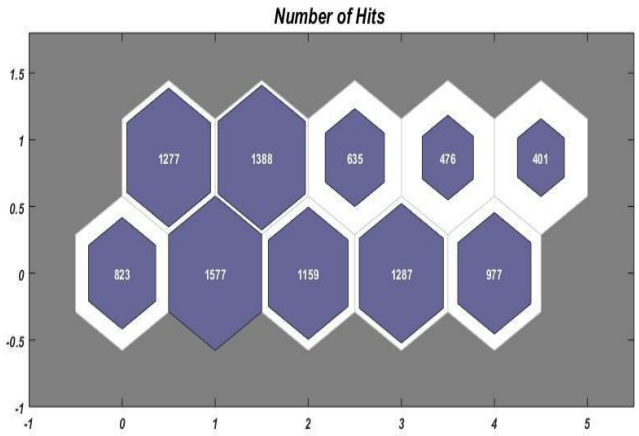

Fig. 6: Number of Hits to 10 Neurons from Input Vectors Using SOM Topology

VI. Conclusion and Future Direction

Classification of hand written digits is a problem for years using different datasets. Many deep learning techniques are used to classify MNIST dataset. Present work focuses on unsupervised learning using K-means Clustering and SOM. Purity of clusters is analysed for both techniques. It can be concluded from tables SOM is performed relatively better than K-means clustering in forming clusters.

In future this work can be extended by putting some internal criteria on clustering algorithm without prior information of classes i.e., ground truth to validate a cluster.

References

[1] Y. LeCun, L. D. Jackel, L. Bottou, C. Cortes, J. S. Denker, …. V. Vapnik, "Learning algorithms for classification: A comparison on handwritten digit recognition," in J. H. Oh, C. Kwon, and S. Cho (eds.), *Neural Networks: The Statistical Mechanics Perspective*, pp. 261-276, World Scientific, 1995.

[2] F. Lauer, C. Y. Suen, and G. Bloch, "A trainable feature extractor for handwritten digit recognition," *Pattern Recognition*, vol. 40, no. 6, pp. 1816-1824, 2007.

[3] Y. Lee, "Handwritten digit recognition using K nearest-neighbor, radial-basis function, and backpropagation neural networks," *Neural Computation,* vol. 3, no. 3, pp. 440-449, 1991.

[4] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: Benchmarking of state-of-the-art techniques," *Pattern Recognition*, vol. 36, no. 10, pp. 2271-2285, 2003.

[5] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, vol. 2011, no. 2, 2011.

[6] M. Iwayama, and T. Takenobu, "Cluster-based text categorization: A comparison of category search strategies," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 1995.

[7] H. Frigui, and R. Krishnapuram, "A robust competitive clustering algorithm with applications in computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 450-465, 1999.

[8] S. K. Bhatia, and J. S. Deogun, "Conceptual clustering in information retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 427-436, 1998.

[9] J. Hu, B. K. Ray, and M. Singh, "Statistical methods for automated generation of service engagement staffing plans," *IBM Journal of Research and Development*, vol. 51, no. 3.4, pp. 281-293, 2007.

[10] S. P. Hung, P. Baldi, and G. W. Hatfield, "Global gene expression profiling in Escherichia coli K12. The effects of leucine-responsive regulatory protein," *Journal of Biological Chemistry*, vol. 277, no. 43, pp. 40309-40323, 2002.

[11] D. C. Cirean, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep big, simple neural nets for handwritten digit recognition," *Neural Computation*, vol. 22, no. 12, pp. 3207-3220, 2010.

[12] A. K. Jain, "Data Clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010.

[13] C. Ding, and H. Xiaofeng, "K-means clustering via principal component analysis," *Proceedings of the Twenty-First International Conference on Machine Learning*, ACM, 2004.

[14] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264-323, 1999.

[15] M. Hassaballah, A. A. Abdelmgeid, and H. A. Alshazly, "Image features detection, description and matching," *Image Feature Detectors and Descriptors*, pp. 11-45, Springer, Cham, 2016.

[16] I. T. Jolliffe, "Principal component analysis and factor analysis," *Principal Component Analysis*, Springer, New York, NY, pp. 115-128, 1986.

[17] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," *International Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886-893, IEEE Computer Society, 2005.

[18] MNIST. Available: http://yann.lecun.com/exdb/mnist/

[19] T. Kohonen, "The self-organizing map," *Neurocomputing,* vol. 21, no. 1-3, pp. 1-6, 1998.

[20] J. Vesanto, and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586-600, 2000.

[21] D. Arthur, and S. Vassilvitskii, "k-means++: The advantages of careful seeding," *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, 2007.

[22] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Inc., 2017.