Supervised Deep Machine Learning Methods of Floral Data Image Processing

R. Lakshmi Priya  
Assistant Professor,  
Department of Statistics,  
Dr. Ambedkar Government Arts College, Vyasarpadi, Chennai

M. Salomi  
Assistant Professor,  
Department of Statistics,  
Madras Christian College, Chennai

N. Manjula Devi  
Bio Statistician,  
Department of Community Medicine,  
Karpaga Vinayaga Institute of Medical Sciences, Chegalpet

Manimannan G.  
Assistant Professor,  
Department of Statistics,  
TMG College of Arts and Science, Chennai

ABSTRACT
This research paper attempts to identify the pattern of three types of images using deep machine learning methods for cluster analysis. These three different images were collected on different web domains with different pixels and under the floral head. The flowers have a basic RGB colour and Black and white colour with different Kilo Byte (KB) sizes. Python data-based software creates image Width, Height and Size. The machine-readable image embedding widget image generates a vector base database from \( n_0, n_1, n_2, \ldots n_{2047} \) using inception v3 embedded. It generates the vectors from deep learning server or individual system. The cosine distance matrix proves the same images with the help of their links for these three types of images. These measures of the distance vector identify the image distance of three types of images using a variety of clustering techniques. In addition, clustered images are visualized using the image viewer widget. The images associated with each collection are displayed separately from other categories.

Keywords
Image Processing, Image Embedded, Data Mining, Hierarchical Clustering, Cosine Distance and Image Visualization.

1. INTRODUCTION
Image processing is very interesting and is an informative analytical tool for pattern recognition and data such as medical bio images, text and many other fields. The image is nothing more than a two-dimensional signal. The mathematical function \( f(x, y) \) is defined when \( x \) and \( y \) are two co-coordinates horizontally and vertically. The value of \( f(x, y) \) at any time gives the pixel value at that point of image. The separation of an image can be defined as separation of all the elements of the image or pixels in the image into different categories that reflects same features. Separation involves dividing an image into groups of identical pixels according to a specific procedure [1]. Different groups should not be united and adjacent groups should be different. Groups are called segments. Image classification is considered the most important function of meaningful analysis and interpretation of an acquired image. It is a critical and important part of image analysis system or pattern recognition system and is one of the most difficult tasks in application of images, determining the quality of final separation. Researchers have worked extensively on this basic issue and have come up with incredible options.

2. REVIEW OF LITERATURE
Many researchers used images to identify weight loss, similarities between other images, etc. High-level data can be converted into low-level codes by training a multilayer neural network with a small medium layer to reconstruct high input vectors. G. E. Hinton, et.al. [2] used Gradient dropper to fine-tune the instruments in such networks as ‘auto encoder’, but this only works best when the first instruments are close to a good solution. The author describes an effective tool for implementing tools that allow deep auto encoder networks to read low-dimensional codes that work much better than Principal Components Analysis (PCA) as a data reduction tool.

Separation is one of the most important steps in the analysis of digital images [3, 4]. It consists of separating an image into unbroken regions according to the similarity of their visual and / or space features (texture, size, shape, etc.). The most common method of classifying satellite imagery is based on data clustering algorithms [5]. Methods of integration can be divided into two main groups: consecutive and non-sequential. Non-sequential algorithms provide consistent data integration and hierarchical algorithms produce a system of embedded clusters corresponding to different sequence levels. Hierarchical representation makes it easy to interpret results in cases where details are needed at various levels of the structure of collection, as well as in cases where maximum number of desirable collections is unknown.

Traditional methods of collecting positions have some drawbacks. For example, a single connecting process is triggered by a so-called chain effect, and complete and standard coupling methods usually work best only with round clusters. Moreover, these methods do not allow for the separation of scattered clusters [6]. Another serious drawback of these methods is their high complexity of calculation, which does not allow it to be used for data processing as many images. The recent integration approach has been widely used to improve stability and performance of integration [7-11]. However, methods based on collection of congregational congregations have been the subject of only a few papers [12]. In addition, the algorithms used in it also take time to calculate. The main objective of this research paper is to identify a flower group that belongs to same group or different group using different group analysis by cosine range scale.

3. DATABASE
The database of floral data were collected on a variety of World Wide Web with three categories of flowers like RGB and BW (Black and White), images are in .jpeg formats. Image processing deep learning methods automatically generates vector database using Inception V3 embedded. The machine learning methods automatically generates the flowers height, width and size (KB). The embedded vectors of
are used for subsequent analysis of deep machine learning methods of supervised cluster analysis.

4. METHODOLOGIES
4.1 Hierarchical Clustering Methods
The hierarchical collection techniques are carried out by a series of consecutive encounters or a series of subsequent phases. The next step is the hierarchical clustering algorithm for collecting N objects or images or variables. [13]

Step 1: It starts with groups of N, each containing one business and N * N symmetric grades and defined by D = \{d_{ik}\}.

Step 2: Identify the matrix range of adjacent clusters. Allow distance between identical sets.

Step 3: Combine X collections with the newly created collection label(\(XY\)). Update everything in distance matrix (a) by removing rows and columns corresponding to group X and Y and (b) by adding a row and column to give the distances between group (\(XY\)) and remaining groups.

Step 4: Repeat steps 2 and 3 to a total of N − 1. (Figure 1.)

4.2 Average Linkage Method
Step 1: Intermediate communication treats the distance between two sets as a normal distance between all pairs of images where one number of pairs is another.

Step 2: Also, the input in average link algorithm can be distances or similarities, and the model can be used to collect images or objects or variables.

Step 3: The above general algorithm in the distance between (\(XY\)) and any other clusterW are computed by

\[d_{(xy)w} = \sum_{i} \sum_{k} d_{ik} \frac{N_{xy}N_{w}}{N_{y}}\]

Where \(d_{ik}\) is the distance i in the cluster between (\(XY\)) and the object k in cluster W and \(N_{xy}\) and \(N_{w}\) are the number of items in clusters (\(XY\)) and W respectively.

4.3 Weighted Average Linkage Method
Step 1: This method is also known as Weighted Pair of Group Method Average (WPCMA).

Step 2: The distance between two clusters is defined as the average distance between all points of data points, each of which comes from a different group.

Step 3: The difference is that the distances between the newly formed and others are weighed according to the number of data points in each collection. When two clusters \(C_i\) and \(C_j\) are merged, the distance to a third cluster \(C_l\) can be recomputed as:

\[D(C_l(C_i,C_j)) = \frac{n_i}{n_i+n_j}D(C_i + C_j) + \frac{n_j}{n_i+n_j}D(C_j + C_i)\]

4.4 Complete Linkage
Complete linkage clustering proceeds in the much same manner as single linkage clustering’s, with one important exception.

Step 1: At each stage the distance between clusters is determined by the two elements, one from each cluster, that are most distant.

Step 2: This method ensure that all items in a cluster are with in some maximum distance of each other.

Step 3: The general agglomerative algorithm repeat by finding the minimum entry in \(D_{ik}\) and merging the corresponding objects, such as \(U\) and \(V\), to get cluster \((U+V)\).

Step 4: The distances between \((U+V)\) and any other cluster W are computed by

\[d_{(uv)w} = \max\{d_{uw}, d_{vw}\}\]

Where, \(d_{uw}\) and \(d_{uw}\) are the distances between the most distant members of clusters \(U\) and W and clusters V and W respectively.

4.4 The Ward Method
Ward’s [14] is considered as a process of clustering positions based on minimizing the loss of information in joining two groups.

Step 1: This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion. ESS, first for a given cluster \(k\), let ESS, be the sum of squared deviations of every item in cluster from the cluster mean (centroid).

Step 2: If there are currently \(k\) clusters defines \(ESS_k\) as the sum \(ESS_k = ESS_1 + ESS_2 + ESS_3 + \cdots + ESS_k\).

Step 3: At each step in analysis, union of every possible pair of clusters is considered, and the two clusters whose combination results in a smallest increase in \(ESS\) are joined.
Step 4: Initially, each cluster consists of a single item, and if there are \( N \) items, \( ESS = 0 \),\( k = 1, 2, \ldots, N \), so \( ESS = 0 \).

Step 5: At the other extreme, when all the clusters are combined in a single group of \( N \) terms, the value of \( ESS \) is given by

\[
ESS = \sum_{j=1}^{N} (x_j - \bar{x})(x_j - \bar{x})
\]

Where, \( x_j \) is the multivariate measurement associated with the \( j \)th item and \( \bar{x} \) is the mean of all the items. The results of Ward’s method can be displayed as a dendrogram.

### 4.6 Cosine Distance Similarity

Cosine similarity is computed using the following formula:

\[
\text{Similarity}(X, Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2}}
\]

Values range between -1 and 1, where -1 is perfectly dissimilar and 1 is perfectly similar. The above formula contains both procedures and functions to calculate similarity between sets of data. The function is best used when calculating similarity between small numbers of sets. The procedures parallelize the computation and are therefore more appropriate for computing similarities on bigger datasets.

### 4.7 Image Processing and Model Developing Algorithm

Step 1: Three types of images (floral) were collected from a different URL and assigned to the input data matrix.

Step 2: Selected images in .jpeg format, with various sizes of images in KB (Kilo Bytes).

Step 3: Images are placed in a single folder and are provided as input data to insert the widget.

Step 4: A file widget connected to data table with image embedding widget. Image embedding calculates their vectors using Inception V3 embedded. Subsequently the distances are identified using Cosine Distance and identifies the similarities of images using image embedding widget.

Step 5: The distance widget connects to the hierarchical cluster analysis and it is visualized in dendrogram.

Step 6: The image view widget shows the clustered image, image name and cluster numbers. (Figure 2)

Image embedding widget is very important for all image analytics. A classification and regressions task requires data in the form of numbers and there isn’t a good way to perform such tasks with images unless, the researcher represents it in the form of numbers. This is where Embedding widget works by converting it to vectors of numbers. This widget reads images and uploads them to a remote server or evaluates them locally (folder).

### 5. RESULT AND DISCUSSION

#### 5.1 Proposed Image Processing Algorithm

Step 1: Input images generate data automatically using deep machine learning methods for data mining.

Step 2: Initially, it generate five parameters such as image name, image URL, height, width and size in Table 1 data.

Step 3: The image embedding widget generates the vectors from \( n_0, n_1, n_2, \ldots, n_{2047} \). (Table 1). These vectors are used for supervised clustering methods.

#### Table 1. Sample Image Data and their image, image name, Size, Width and Height

| Image Name | Height | Width | Size |
|------------|--------|-------|------|
| Image1     | 1200   | 1600  | 300K |
| Image2     | 1800   | 2400  | 500K |
| Image3     | 2400   | 3200  | 700K |
| Image4     | 3000   | 4000  | 1000K|
| Image5     | 3600   | 4800  | 1500K|

#### Table 2. Sample Image embedded (Inception V2) Data and their image, image name, Size, Width, Height and \( n_0, n_1, n_2, \ldots, n_{2047} \)

| Image Name | Image Data | Height | Width | Size |
|------------|------------|--------|-------|------|
| Image1     | 1200       | 1600   | 300K  |      |
| Image2     | 1800       | 2400   | 500K  |      |
| Image3     | 2400       | 3200   | 700K  |      |
| Image4     | 3000       | 4000   | 1000K |      |
| Image5     | 3600       | 4800   | 1500K |      |

Step 4: Embedding a widget connected to cosine distance method. The cosine distance method is the most common and
widely used in image analysis and will determine distance from the similarity of various images and is cluster together.

Step 5: Cluster is made up of a variety of statistical methods such as Average, Complete, Weighted Average and Ward’s method.

Step 6: All methods form a natural cluster and are represented in dendrogram program.

Step 7: Finally, selected clusters are displayed and labelled as $c_1, c_2, c_3, c_4, c_5$ and $c_6$ and are shown in picture with the help of image view widget.

The results of image minimization are shown in Figure 3. The separated elements are highlighted in blue and the values are zero. The cosine distance matrixes of the three categories of pictures are in table 3. The concentrations of the cosines are closed in those of blue and dark yellow. Small numbers are inserted in bright pink and blue colours, the diagonal matrix are highlighted in blue colour, the following results are interpreted separately in the following sections.

Image processing of data mining tools has given exciting results. All methods are well separated in dendrogram view and image view. Initially, in this research paper only (19, 19) images were used when looking at pictures of three categories and corresponding to their collection, the cosine distance gives good result between the ranges 0 to 1; the similarities within the images are well. These three images are divided into three colours with different types of distance model in dendrogram. These three types of images are well integrated and labelled as C1, C2, C3, C4, C5 and C6.

This method shows the result in the following dendrogram (Figure 4 and 5) and C3 cluster shows the results of RGB and BW images are visualized in Figure 5. The middle link method is divided into three sections with few overlap. The test image was re-integrated into their groups.

5.2 Average Linkage Method

Table 3. Distance Matrix of the Images

This method shows the result in the following dendrogram (Figure 4 and 5) and C3 cluster shows the results of RGB and BW images are visualized in Figure 5. The middle link method is divided into three sections with few overlap. The test image was re-integrated into their groups.

5.3 Weighted Linkage Method

The weighted linkage method shows the results in figure 6 and their images are shown in figure 7. This method is divided into two phases without overlap. The test image was re-integrated into their groups.
5.4 Complete Linkage Method
The complete connection method shows the results in the following dendogram (Figure 8) and their images are shown in figure 9. The complete communication method is divided into two phases without overlap. The test image was re-integrated into their groups.

5.5 Wards Method
Initially the Ward method shows the results in the following dendograms (Figure.10) and their images are shown in Figure 11. The Ward Road is divided into two sections without overlap. The test image was re-integrated into their groups.

6. CONCLUSION
This research paper begins with the introduction of image analysis tools that give a few new results. Then the, floral details upload images via import images widget. Image embedding widget calculates image vectors matrix. The cosine distance calculates the similarity measurement and achieved in pairs of the image range effect using a variety of data mining methods and visualized the results in dendogram. Clustered images are viewed via the image viewer widget. Also, the researchers learned to use image embedding widget to convert images into a vector of numbers. The Cosine Distances widget calculates the various methods of positioning methods and is displayed in collection dendogram. Finally, test it with a sample of flora images of RGB and BW colours. All methods of integration techniques have achieved C1, C2, C3, C4, C5 and C6 natural collections except few images.

Data does not always come in a nice tabular form. It can also be a collection of text, audio recordings, video materials or even images. However, computers can only work with numbers, so for any data mining, we need to transform such unstructured data into a vector representation.

For retrieving numbers from unstructured data, Orange can use deep network embedders. Currently, they are available for text and images.

7. REFERENCES
[1] Krishna Kant Singh, Akansha Singh (2010) , A Study Of Image Segmentation Algorithms For Different Types Of Images IJCSI International Journal of Computer Science.
[2] G. E. Hinton and R. R. Salakhutdinov (2206), American Association for Advancement of Science, Volume..313
[3] R. C. Gonzalez and R. E. Woods, Digital Image Processing (Tekhnosphera, Moscow, 2006) [Russian translation].
[4] P. A. Chochia (2014), “Image Segmentation Based on the Analysis of Distances in an Attribute Space,” Avtometriya 50 (6), 97–110 (2014) [Optoelektron. Instrum. Data Process. 50 (6), 613–624.]
[5] I. A. Pestunov and Yu. N. Sinyavskii (2012) “Clustering Algorithms in Problems of Segmentation of Satellite Images,” Vestn. KemGU 2 (4(52)), 110–125 (2012).
[6] R. Xu and D. J. Wunsch (2005) “Survey of Clustering Algorithms,” IEEE Trans. Neural Networks 16 (3), 645–678.
[7] A. K. Jain (2010) “Data Clustering: 50 years Beyond K-Means,” Pattern Recognition. Lett. 31 (8), 651–666.

[8] P. Hope, L. Hall, and D. Goldgof (2009), “A Scalable Framework for Cluster Ensembles,” Pattern Recognition. 42 (5), 676–688.

[9] R. Kashef and M. Kamel (2011) “Cooperative Clustering,” Patt. Recogn. 43 (7), 2315–2329 (2010).

[10] J. Jia, B. Liu, and L. Jiao, “Soft Spectral Clustering Ensemble Applied to Image Segmentation,” Front. Comput. Sci. China. 5 (1), 66–78.

[11] A. Mirzaei and M. Rahmati (2010) “A Novel Hierarchical-Clustering-Combination Scheme Based on Fuzzy-Similarity Relations,” IEEE Trans. Fuzzy Syst. 18 (1), 27–39.

[12] R. A. Johnson D. W. Wichern (2009), Applied multivariate Statistical Analysis, Fifth Edition, Published by PHI Learning Private Lts. India.

[13] Everit B. S. (1993), Cluster Analysis, Third Edition, London Edward Arnold.

[14] Manimannan G, et al. (2020), Machine Learning Method of Image Processing and Embedding of Flowers using Pattern Recognition, International Journal of Innovative and Mathematical Research, pp.12-21, India.