An Improved Hierarchical Clustering Technique for Character Recognition

Ritu Yadav[1], Sunila Godara[2]
[1](*Department of Computer Science and Engineering, Guru Jambheshwar University of Science & Technology, India)*
[2](*Department of Computer Science and Engineering, Guru Jambheshwar University of Science & Technology, India)*

Abstract

A Character Recognition is an activity, which covers all types of machine recognition of characters in various application domains. Character recognition systems can contribute tremendously to the advancement of the automation process and can improve the interaction between man and machine in many applications, including office automation, image processing, cheque verification and a large variety of banking, business and data entry applications. The purpose was to correctly recognize the letters from their images. In this Paper, we propose Weighted Euclidean Distance method for correctly classifying the characters from the Letter Image Recognition dataset using Hierarchical clustering and results are compared with Euclidean distance method in the Weka tool.

1. Introduction

A Character recognition System is a step towards the automation process, which requires in many fields, including image processing, office automation and data entry applications. The various techniques covered under the general term character recognition [1] fall into either the on-line or off-line category, each having its own hardware and recognition algorithms.
In on-line character recognition systems, the computer recognizes the symbols as they are drawn. The most common writing surface is the digitizing tablet, which typically has a resolution of 200 points per inch and a sampling rate of 100 points per second, and deals with one dimensional data. Some digitizers use pressure-sensitive tablets, which have layers of conductive and resistive material with a mechanical spacing between the layers. There are also, other technologies including laser beams and optical sensing of a light pen. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of the on-line character recognition techniques [2].
Off-line recognition is performed after the writing or printing is completed. Optical Character Recognition, OCR [3] deals with the recognition of optically processed characters rather than magnetically processed ones. Any OCR system goes through numerous phases including: data acquisition, pre-processing, feature extraction, classification and post-processing where the most crucial aspect is the pre-processing which is necessary to modify the data either to correct deficiencies in the data acquisition process due to limitations of the capturing device sensor.
Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. "The science of extracting useful information from large data sets or databases" [4].Classification is one of the techniques which classify the given data based on many attribute given in the data base. This paper includes the result of unsupervised classification that is clustering which implement weighted Euclidean distance method in Weka [5]. The technique is applied on the Letter Image Recognition data sets and the results are compared with existing technique which uses Euclidean distance formula.

2. Hierarchical Clustering

All Hierarchical clustering combine and divide existing groups, creating a hierarchical structure that reflects the order in which groups are merged or divided. The Algorithm is given below

1. Compute the proximity matrix containing the distance between each pair of patterns. Treat each pattern as a cluster.
2. Find the most similar pair of clusters using the proximity matrix. Merge these two clusters into one cluster. Update the proximity matrix to reflect this merge operation.
3. If all patterns are in one cluster, stop.

The tree is often called a "dendogram". The method is summarized below:

1. Place all points into their own cluster ,While there is more than one cluster, do
2. Merge the closest pair of clusters. The behavior of the algorithm depends on how "closest pair of clusters" is defined.

The merging criteria of clusters for hierarchical clustering are single link, average link and complete link [7].

3. Proposed Methodology

Hierarchical clustering is implemented using Euclidean distance formula in the Weka. The Euclidean distance between one vector x=(x1 ,x2,…xn) and another vector y=(y1 ,y2 ,…yn) is Euclidean distance d(xi, yi) .The Euclidean distance formula is given below

$$d = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$$

In this section, we define the proposed methodology which is weighted Euclidean distance formula [8] for clustering approach based on the assumption that each cluster corresponds to a class in character recognition system (see Fig 1). Firstly, the data is collected and preprocessed. Then, Segmentation is done if required to refine the data. The last and most important step is Recognition. This step is done with the help of Hierarchical clustering using the Weighted Euclidean distance Formula to recognize the inputted character. The Weighted Euclidean distance between one vector x=(x1,x2,…xn) and another vector y=(y1 ,y2 ,…yn) is the Euclidean distance d(xi, yi) which can be obtained as follow The proposed Weighted Euclidean distance formula is given as

$$d = \text{ulp}\left((\text{var})^{-1} \sum_{j=1}^{n} (x_j - y_j)^2 \right)$$

Where, var is the variance of the attributes of instances. Variance is calculated using the given formula

$$\text{var} = \frac{\sum_{j=1}^{n} (x_j - \overline{x})^2}{\mathbf{n - 1}}$$

Where, $\overline{x}$ is the mean of the attributes of instances.

The ulp() method returns the distance from a number to its nearest neighbors. This distance is called an *ULP* for unit of least precision or unit in the last place [9].

A proposed character Recognition System [10] is shown below in Fig. 1. This system has four stages as follows:

- Pre-Processing
- Segmentation
- Feature extraction
- Recognition

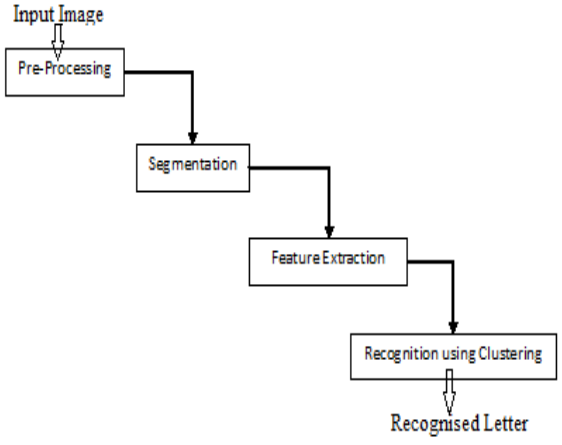

Fig.1 Proposed Character Recognition System

4. Dataset Description

To compare the result of proposed model to existing model, we used Letter Image Recognition dataset which is downloaded from UCI repository. The dataset is generated to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts), which were then scaled to fit into a range of integer values from 0 through 15 and 1 class attribute. The 16 integer attributes extracted from the raster scan images of the letters [11]. We tested our model with 530 instances of 5 letters (A, B, C, D, E, and F).

Attribute Information:

1. x-box horizontal position of box (integer)
2. y-box vertical position of box (integer)
3. width width of box (integer)
4. high height of box (integer)

5. onpix total # on pixels (integer)
6. x-bar mean x of on pixels in box (integer)
7. y-bar mean y of on pixels in box (integer)
8. x2bar mean x variance (integer)
9. y2bar mean y variance (integer)
10. xybar mean x y correlation (integer)
11. x2ybr mean of x * x * y (integer)
12. xy2br mean of x * y * y (integer)
13. x-ege mean edge count left to right (integer)
14. xegvy correlation of x-ege with y (integer)
15. y-ege mean edge count bottom to top (integer)
16. yegvx correlation of y-ege with x (integer)

5. Result Analysis

The experimentation detailed in this section was carried out within the Waikato Environment for Knowledge Analysis (Weka) [12] suite for machine learning. This software, developed in the Java programming language, offers a powerful testing harness for analysis of various data mining concepts and implementations. In addition, it offers the ability to extend the suite with additional modules, and in the case of the presented work, clustering methods.

We tested our model with 530 instances of Letter Image Recognition Dataset that has 16 numeric and 1 class attributes using Hierarchical clustering in Weka 6.6. In order to test the accuracy of obtained classification models we used the classes to cluster evaluation method. The model is built using just the instances in the training fold. In first experiment, we executed the Hierarchical clustering algorithms provided by Weka using all the available attributes using Euclidean distance formula. The result of classes to cluster evaluation is shown in Fig.2.

In further Experiment, we executed the Hierarchical clustering algorithm using Weighted Euclidean Distance formula in Weka with 530 instances of the Letter Image Recognition Dataset with 17 attributes. The result of classes to cluster evaluation method is shown in Fig. 3.

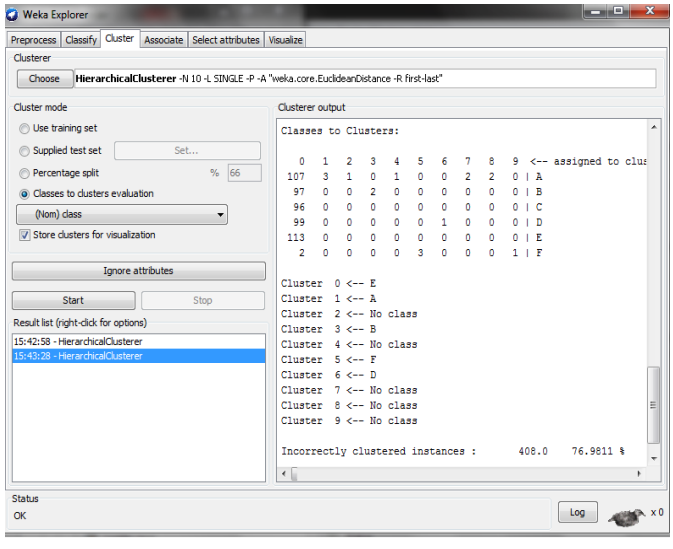

Fig 2: Result of clustering of Letter Image Recognition data with Hierarchical clustering

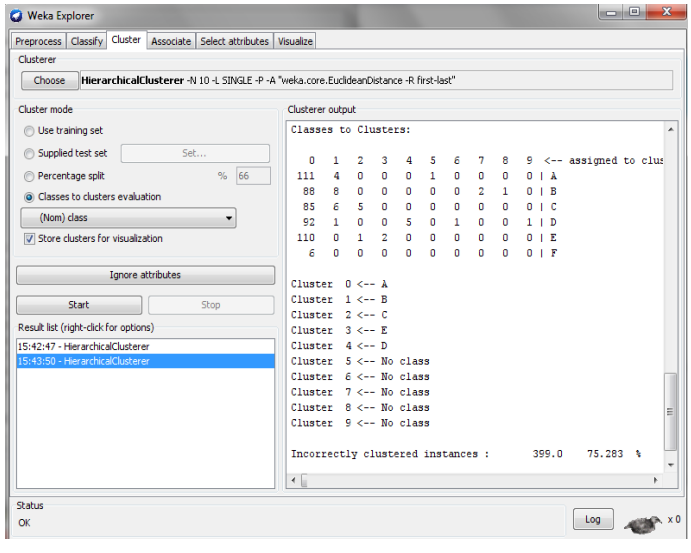

Fig 3: Result of clustering of Letter Image Recognition data with Hierarchical clustering with Weighted Euclidean Distance formula

The result of classes to cluster evaluation for Hierarchical clustering and improved Hierarchical clustering is shown in Table 1 and Table 2. The term nc used in tables represented no class assigned to the cluster. After comparing results from Table 1 and Table 2, we found that the improved Hierarchical clustering recognize Letter 'A' with the accuracy 95.69% as compared to Hierarchical clustering with accuracy 2.59%. Letter B, C, D, E, F are recognized with higher accuracy rate using improved Hierarchical clustering than Hierarchical clustering.

Table 1: Result of classes to cluster evaluation with Hierarchical clustering using Euclidean Distance for K=10

E	A	nc	B	nc	F	D	nc	nc	nc	
107	3	1	0	1	0	0	2	2	0	A
97	0	0	2	0	0	0	0	0	0	B
96	0	0	0	0	0	0	0	0	0	C
99	0	0	0	0	0	1	0	0	0	D
113	0	0	0	0	0	0	0	0	0	E
2	0	0	0	0	3	0	0	0	1	F

Table 2: Result of classes to cluster evaluation with Hierarchical clustering using Weighted Euclidean Distance for K=10

A	B	C	E	D	nc	nc	nc	nc	nc	
111	4	0	0	0	1	0	0	0	0	A
88	8	0	0	0	0	0	2	1	0	B
85	6	5	0	0	0	0	0	0	0	C
92	1	0	0	5	0	1	0	0	1	D
110	0	1	2	0	0	0	0	0	0	E
6	0	0	0	0	0	0	0	0	0	F

With the comparison of the above results obtained from Hierarchical and improved Hierarchical clustering algorithm for the Letter Image Recognition dataset in Table 1 and Table 2, the correctly identified number of letters is increased in improved Hierarchical clustering as compared to the Hierarchical clustering using the test option classes to cluster evaluation for a given number of cluster, i.e. K, equal to 10. An analysis of the results shown in Table 3 reveals that Hierarchical clustering when implemented with Weighted Euclidean Distance Formula has good level of accuracy as compared to Euclidean Distance Formula.

Table 3: Hierarchical and Improved Hierarchical Results on Letter Image Recognition data set for correctly classified instances

Clustering Name	Total Number of Instances	Correctly Classified Instances	Accuracy
Hierarchical Clustering	530	122	23.02
Improved Hierarchical Clustering	530	131	24.72

The graphical reprentation of result data given in Table 3 for Letter Image Recognition dataset is represented in Fig.4.

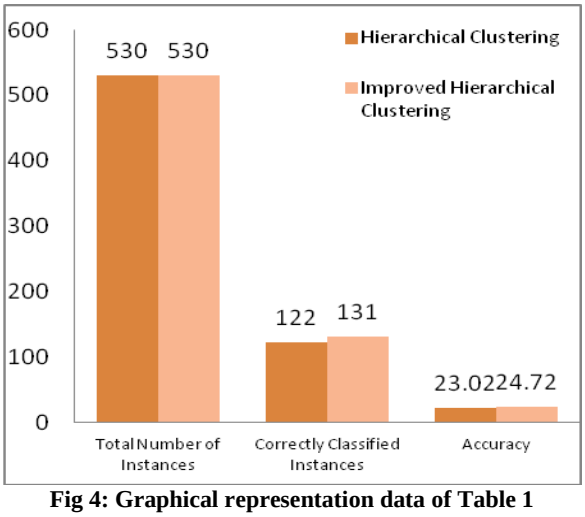

Fig 4: Graphical representation data of Table 1

With the comparison of the above results obtained from Hierarchical and improved Hierarchical clustering algorithm in Table 3, It is shown that Hierarchical clustering algorithm when implemented with Weighted Euclidean Distance formula has good accuracy and best classification of instances for Letter Image Recognition data sets .It is shown that the Improved Hierarchical Clustering has higher recognition rate for letters with 24.72% accuracy as compared to the Hierarchical clustering with 23.02% accuracy with the deployment of Letter Image Recognition data in Weka tool.

6. Conclusion

Character Recognition is a wide area of research where various techniques are applied for recognizing the character. In this paper Hierarchical and its improved version are employed for the purpose of recognition of character. Results of these clustering are compared on the basis of accuracy and correctly classified instances on Letter Image Recognition Dataset. The result showed that improved Hierarchical clustering has

obtained higher accuracy rate than Hierarchical clustering. It is also found that improved Hierarchical clustering has better performance than Hierarchical clustering for identifying the character in Character Recognition System.

7. References

[1] Arica N., Vural F.T.Y., "An Overview Of Character Recognition Focused On Off-line Handwriting" IEEE (1999), C99-06-C-203.
[2] R. Plamondon, S.N. Srihari, "On-Line And Off-Line Handwriting Recognition: A Comprehensive Survey", IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 63–84.
[3] Fujisawa H., "Forty Years Of Research In Character And Document Recognition-An Industrial perspective", Pattern Recognition 41 (2008) 2435 – 2446.
[4] Hand D., Mannila H., and Smyth P., "Principles of Data Mining", MIT Press, Cambridge, MA.(2001), ISBN 0-262- 08290-X.
[5] Markov Z., Russell I.. "An Introduction To Weka Data Mining System Tutorial", http://www.ccsu.edu.
[6] Wang J., Su X.; "An Improved K-means Clustering Algorithm", Communication Software and Networks (ICCSN), (2011), IEEE 3rd International Conference.
[7] Arai K., Barakbah A.R., "Hierarchical K-Means: An Algorithm For Centroids Initialization For K-Means", Reports of the Faculty of Science and Engineering, Saga University,(2007), Vol. 36, No.1, 25-31.
[8] Chim Y.C. , Kassim A. A., Ibrahim Y., "Character Recognition Using Statistical Moments", Image and Vision Computing 17 (1999) 299–307.
[9] http://www.ibm.com/developerworks/java/library/j-math2/index.html.
[10] Waard W.P. D., "An Optimized Distance Method For Character Recognition", Pattern Recognition Letters 16, (1995), pp 499-506.
[11] archive.ics.uci.edu/ml/datasets.html.
[12] Rammimmagadda S., Kanka P.,Yaramala V.B., "Implementation of Clustering Through Machine Learning Tool", IJCSI International Journal of Computer Science Issues,(2011), Vol. 8, Issue 1,pp: 395-401