ClusLine: A New Unsupervised Learning Algorithm

M.Fatih Amasyalı*, Okan Ersoy+
+School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA
phone: + (1-765) 494-6162, fax: + (1-765)494-3358, email: ersoy@purdue.edu
**Department of Computer Engineering, Yıldız Technical University, Beşiktaş 34349, Istanbul, TURKEY*
phone:+(90 212) 259 70 70/2801, email: mfatih@ce.yildiz.edu.tr
web:www.ce.yildiz.edu.tr/myindex.php?id=14

Abstract

Unsupervised learning (clustering) algorithms are generally used in discovering data structure and data compression by generating clusters of data. In this work a new such algorithm -ClusLine- is proposed. ClusLine is based on determining cluster centers by constructing a tree which divides data space into subspaces that have smaller standard deviation than all data's standard deviation. The algorithm is running on 14 real world dataset and compared popular clustering algorithms according to three cluster quality measures (Davies-Bouldin Index, Silhouette width and classification accuracy). According to quality measures, The Clusline algorithm can be used in all types of clustering problems because of its simplicity and acceptable performance.

1. Introduction

Clustering is defined as the process of organizing samples into clusters whose members are similar in some sense. Therefore, a cluster consists of samples which are "similar" within the cluster, and are simultaneously "dissimilar" to the samples of the other clusters. Data can be compressed by reducing the number of different samples by clustering. The structure of data can also be discovered by clustering. If the number of samples or the number of sample features is big. Many clustering algorithms have been developed for automated clustering by computers. The application area is very large. For example, finding customer groups having similar characteristics (shopping behaviors, credit risk factors, array of web sites links clicked in succession etc.), finding high risky earthquake areas, analyzing and clustering past earthquake records, and grouping huge web collections are some clustering applications.

In this work, a new clustering algorithm -ClusLine- was developed and compared with other popular methods. A detailed explanation of the Clusline algorithm is given in the second section. In the third section, the measures of clustering quality are described. In the fourth section, the comparison of Clusline and other methods on 14 different benchmark datasets is discussed. The conclusions are given in the last section.

2. ClusLine algorithm

The aim of Clusline is to find best cluster centers and sample groups. For this purpose, the division of the feature space into two regions with a hyperplane is utilized as in a tree. The distinct approach for the determination of the hyperplanes with Clusline is discussed in the next sections. The division process is repeated recursively until each region has samples having X times smaller standard deviation than the previous data. In the two-dimensional space, the boundary which separates classes is a line, in three dimensional spaces, it is a plane, and in higher dimensional spaces, it is a hyper plane. At each node of the tree, there is a hyperplane, and the samples are directed within the tree according to whether the sample points are on one side or the other of the hyperplane. The Clusline is summarized as Algorithm 1 below. In Algorithm 1, STD is the standard deviation of all samples.

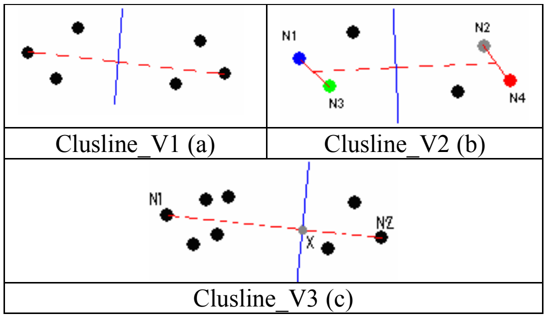
Algorithm 1. Clusline Algorithm

Start: Is the standard deviation of samples X times smaller than STD?	
Yes	**No**
Determine the cluster centre as the mean of samples. Return	Find the boundary hyperplane and create a root node for this hyperplane. Create two branches.
	Assign samples to branches according to whether the point is on one side or the other of the hyperplane.
	Return to start.

The algorithm has only one parameter, *X*. The number of clusters created by the algorithm is determined by the value of *X*. In the experiments discussed in the fourth section, the value of *X* has a range between 1.2 and 4. The number of clusters increases as *X* increases.

2.1. Determination of hyperplanes

Three methods are developed to determine the boundary hyperplanes discussed above. The methods will be explained in two dimensional (2-D) space for the purpose of easy visualization. In 2-D space, classes are separated with a line. Figure 1 illustrates the methods.



Clusline_V1 (a)

Clusline_V2 (b)

Clusline_V3 (c)

Figure 1. Determining boundary lines.

Method I:(Clusline_v1) Referring to Figure 1.a, the farthest two points are ‘A’ and ‘B’. There is only one line (dashed line) between these points. There is only one line which passes through the midpoint of this line and perpendicular to it. This line (straight line) is referred to as boundary line between clusters used in Clusline_v1 algorithm.

Method II:(Clusline_v2): ‘N2’ point is the second nearest point to ‘N1’ point after the ‘N4’ point. Similarly, ‘N3’ point is the second nearest point to ‘N4’ point after the ‘N1’ point. There is only one line which passes through the midpoint of the N1-N3 line and the midpoint of the N2-N4 line, say, Line 2 (dashed line). Then, the line perpendicular to Line 2 and passing through its midpoint as shown in Figure 1.b will be called boundary line (straight line) and used in Clusline_v2 algorithm.

Method III:(Clusline_v3) : Figure 1.c, the farthest two points are ‘N1’ and ‘N2’. The number of samples which are closer to N1 than N2, and The number of samples which are closer to N2 to N1 are found. In Figure 1.c, there are 4 samples which are closer to N1, and 2 samples which are closer to N2. The dashed line is divided proportionally according to these numbers. In Figure 1.c, N2-X distance is half of N1-X distance.

In Figure 2, The steps of Clusline_v1 algorithm and the constructed tree is shown in two dimensions.

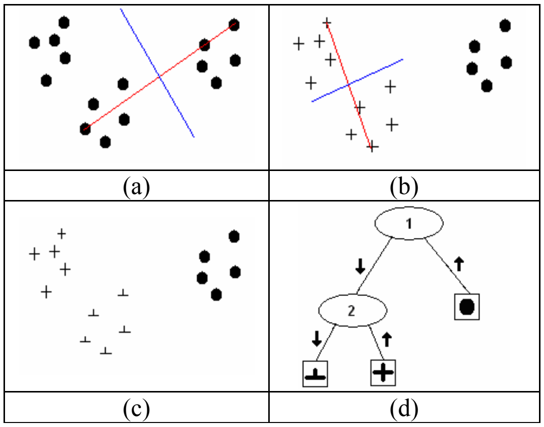


Figure 2. The steps of Clusline algorithm (a, b, c) and the constructed tree (d).

2.2. Determining parameters of hyperplanes

The equation of a hyperplane in the N-dimensional space is given by Eq. (1). The parameters to be found are $a_1,a_2,....a_n,a_0$.

$$a_1x_1 + a_2x_2 + ... + a_nx_n + a_0 = 0 \qquad (1)$$

Analytical geometry is used to obtain the parameters of the hyperplane (which is used in Clusline_v1) obtained from $A(x_1,x_2,\ldots,x_n)$ and $B(x_1,x_2,\ldots,x_n)$ points as shown in Figure 1.a. For example, in the three dimensional space, the equation of a plane which is perpendicular to $\overset{\rho}{U}=[a,b,c]$ and passes through $M(x_1,x_2,x_3)$ point is given by Eq. (2) [1].:

$$a(x-x_1)+b(y-x_2)+c(z-x_3)=0\ .(2)$$

More clearly,

$$\begin{array}{l} -ax_1 - bx_2 - cx_3 + ax + by + cz = 0 \\ a_1 = -a \\ a_2 = -b \\ a_3 = -c \\ a_0 = ax + by + cz \\ a_1x_1 + a_2x_2 + a_3x_3 + a_0 = 0 \end{array} \quad .\ (3)$$

The formula which is obtained at the end of Eq. (3) is the special case of Eq. (1) for the three dimensional space. In general, to get the parameters of the hyperplane, the equations for a vector perpendicular to the hyperplane and a point which is on the hyperplane are needed. The perpendicular vector is obtained with difference of ‘A’ and ‘B’ points:

$$\overset{\rho}{U}(x_1,x_2,...x_n)=\vec{A}(x_1,x_2,...x_n)-\vec{B}(x_1,x_2,...x_n)\,. \qquad (4)$$

The point which is on the hyperplane is the midpoint of ‘A’ and ‘B’, given by

$$M(x_1,x_2,...x_n)=(A(x_1,x_2,...x_n)+B(x_1,x_2,...x_n))/2\,. \quad (5)$$

The parameters can then be found with Eqs. 3, 4 and 5.

In Clusline_v2, there are four points N1, N2, N3 and N4. To obtain the parameters of the hyperplane, the midpoint of 'N1' and 'N3' (T_1), and the midpoint of 'N2' and 'N4' (T_2) are found. 'T1' and 'T2' point are used instead of 'A' and 'B' points in Clusline_v1. All other design equations are the same.

In Clusline_v3, the only difference from Clusline_v1 is in Eq. 5. M point is determined in terms of the proportion of the closer number of samples to N1 and N2 points.

3. Cluster quality measures

In the literature, there are many cluster quality measures. In this work, three methods (Davies-Bouldin Index [2], Silhouette Width [3] and Classification Accuracy) are used to measure the clustering qualities of the algorithms. While the samples and cluster centers (determined by clustering algorithm) are used in the first two methods, the third method uses sample classes.

3.1. Davies-Bouldin (D-B) index

After the division of dataset into N cluster $C = \{C_1, C_2, ..., C_n\}$, the Davies-Bouldin index of each cluster is calculated. In Eqs.6 and 7, the calculation of the D-B index is defined for the ith cluster:

$$DB_i = \max_{\substack{j=1,...,n \\ i<>j}} (DB_{ij}) \tag{6}$$

$$DB_{ij} = \frac{\{sc(C_i) + sc(C_j)\}}{cd(C_i, C_j)} \tag{7}$$

Where

$sc(x)$: Average distances of samples (belonging to x cluster) to center of x cluster,

$cd(x, y)$: The distance between the centers of x and y clusters.

The average value of the cluster indices is defined as D-B index of all clusters in Eq. (8):

$$DB = \frac{1}{n} \sum_{i=1}^{n} DB_i \tag{8}$$

The value of the D-B index and clustering quality are considered directly proportional.

3.2. Silhouette width (S-W)

To calculate S-W, first the S-W of each sample is found by Eq. (9). Then, the average S-W for each cluster and overall average S-W for all samples are calculated.

$$sw_i = \frac{sc(i) + sd(i)}{\max(sc(i), sd(i))} \tag{9}$$

where

$sc(i)$: Average distances between the ith sample and the other samples in the same cluster,

$sd(i)$: Average distance between the ith sample and the other samples which are nearest cluster to the ith sample's cluster.

If the value of S-W is close to 1, it means the sample is in an appropriate cluster. If it is close to 0, it means the sample can also be in the nearest cluster to the ith sample's cluster. If it is close to -1, it means this sample is not in an appropriate cluster.

3.3. Classification accuracy

To find classification accuracy, samples' class labels are used. Each sample's cluster is determined by the closest cluster center. A cluster's class is defined to be the majority class among this cluster's existing classes.

The numbers of samples for which the sample class and the sample's cluster's class are the same are found. The classification accuracy is the proportion of this number to the total number of samples.

For a dataset with M samples and N clusters, classification accuracy is found by Eq. (10):

$$cd = \frac{1}{M} \sum_{i=1}^{M} [S_i == sm(\arg\min_{j=1,...N} \| X_i, ce_j \|)] \tag{10}$$

Where

$sm(x)$: xth cluster's class,

$\|x, y\|$: The distance between x and y,

ce_j : jth cluster center, S_i : ith sample's class.

4. Experimental results

Experiments were made with the 14 UCI datasets shown in Table 1 with their names and characteristics [4]. WEKA Data Mining Tool [5] was used for the clustering algorithms except for Clusline, which was programmed in Matlab. The Clusline was compared to Farthest First [6], K-Means, and Self Organizing Map (SOM) [7] algorithms according to 3 clustering quality measures described in the previous section.

The algorithms were run on 14 benchmark datasets for at least two different cluster numbers.

Table 1. Datasets and properties.

Dataset ID	Dataset Name	# of Attributes	# of Classes	Size of training dataset
1	Glass	9	6	170
2	derma	34	6	286
3	ecoli	7	8	266
4	breast-cancer	30	2	456
5	weather	34	2	281
6	iris	4	3	120
7	New-thyroid	5	3	143
8	segmentation	19	7	210
9	bupa	6	2	175
10	wine	13	3	118
11	waveform	21	3	2460
12	Monks1	6	2	124
13	Monks2	6	2	169
14	Monks3	6	2	122

So each algorithm was applied to 37 clustering problems. In Figure 3, the algorithms' success values according to classification accuracy are given.

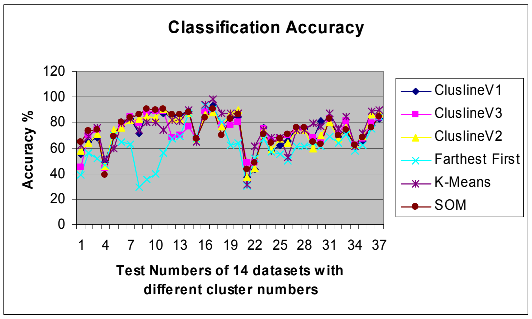


Figure 3. Classification Accuracies.

The average classification accuracies are given in Table 2. As seen, there is little difference between the algorithms except Farthest First algorithm.

Table 2. Average Classification Accuracies

Clusline_V 1	Clusline_V 2	Clusline_V 3	Farthest First	K-Means	SOM
73,4	72,7	73.3	61,5	75,1	74,2

In Figure 4, experimental results for Davies-Bouldin Index and Silhouette width are given. The algorithms have again 37 different problems.

According to Silhoutte width, Clusline_v1 has 11 (out of 37) times highest score. SOM is the best and Clusline_v1 is the second best clustering algorithm according to Davies-Bouldin Index.

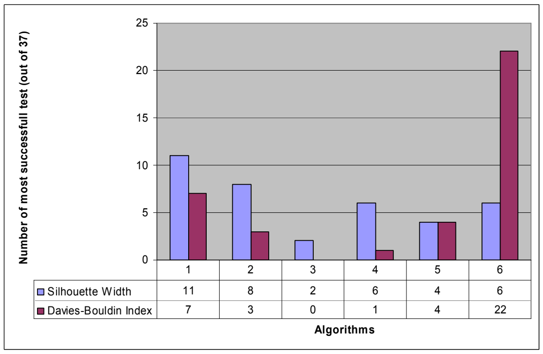


Figure 4. Results of Cluster Quality Measures

5. Conclusions

In this paper, a new clustering algorithm -Clusline- is proposed. The Clusline algorithm constructs binary, multivariate trees that can be used with numerical attributes. At each node of the tree, there is a hyperplane which separates the clusters. Three different methods were proposed to determine the hyperplanes. 14 different real world datasets were used in performance tests. The Clusline was compared with popular clustering algorithms according to three clustering quality measures. In classification accuracy tests, Clusline has comparible accuracy with other algorithms. According to Silhouette width, Clusline is the best; according to Davies-Bouldin, Clusline is the second best clustering algorithm. In Clusline versions, highest performance was achieved by Method I on determining hyperplanes. Further development of the Clusline algorithm is expected to yield better results. In conclusion, the Clusline algorithm can be used in clustering applications because of its simplicity and acceptable performance.

References

[1] http://www.deu.edu.tr/userweb/mustafa.ozel/dosyalar/Uzay%20Analitik%20Geometri.ppt

[2] D.L. Davies, D.W. Bouldin. "A cluster separation measure", *IEEE Trans. Pattern Anal. Machine Intell.* , pp.224-227, 1979

[3] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, 20, pp. 53-65, 1987.

[4] Blake, C., Merz, C., *UCI Repository of machine learning databases*, available at: http://www.ics.uci.edu/mlearn/MLRepository.html, 1998

[5] WEKA 3: Data Mining Software in Java. available at: http://www.cs.waikato.ac.nz/~ml/weka

[6] Hochbaum and Shmoys, "Farthest First Traversal Algorithm: A best possible heuristic for the k-center problem", *Mathematics of Operations Research*, 10(2): pp.180-184, 1985.

[7] Kohonen,T., "The self-organizing map". *Proceedings of the IEEE*, 78(9), pp.1464-1480, 1990.