Clustering based on Dissimilarity First Derivatives

Ana Fred

Telecommunications Institute
Instituto Superior Tecnico, Lisbon, Portugal,
afred@lx.it.pt

Abstract. A hierarchical agglomerative clustering algorithm based on the analysis of dissimilarity increments between neighboring patterns is presented. The first derivative of dissimilarity between neighboring patterns inside a natural cluster is modelled by an exponential distribution, this statistic characterizing the cluster. A cluster isolation criterion is defined based on estimates of each cluster dissimilarity increments mean value, continuously updated along the clusters formation process, under a hierarchical agglomerative framework. Unreliable estimates, mainly occurring when cluster cardinality is low, can lead to over-fragmentation of the data into spurious, small sized clusters. In order to prevent this situation, a regularizing function is proposed to widen the estimates of the exponential distribution mean, when the number of samples is small. Analysis of the method is performed in a comparative study with the well known single-link and k-means algorithms. Application examples using both syntectic and real data show the ability of the method to identify arbitrary shaped clusters.

1 Introduction

Clustering - the partitioning of a set of objects into groups or clusters - is very important in exploratory pattern analysis and data mining. Various clustering algorithms and techniques have been reported in the literature [1, 2], from model-based [3–5], non-parametric density estimation based methods [6], central clustering [7] and square-error clustering [8], graph theoretical based [9, 10], to empirical and hybrid approaches.

Two main strategies are used for clustering: hierarchical and partitional methods [11, 1]. Hierarchical methods, mostly inspired by graph theory, consist of a sequence of nested data partitions in a hierarchical structure, graphically represented as a dendrogram; a partition may be obtained by setting a threshold on the dendrogram. The most popular and well known algorithm in this class is the single-link method [1, 2]. Partitional methods organize patterns into a small number of clusters, either as a result of an optimization process over some cost function, or based on some heuristic criterion. The k-means is probably the best known and widely used algorithm in this category, being a computationally efficient clustering technique based on the square-error criterion.

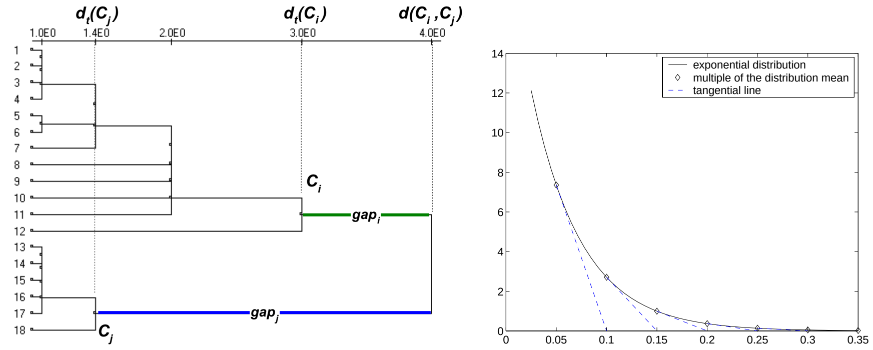
Underlying each clustering algorithm is a concept about data similarity. Clustering techniques based on intra-cluster compactness criteria, such as the k-means, tend to organize the data into hyper-spherical clusters. Graph-based approaches supported on the minimum spanning tree concept, such as the single-link method, are able to handle elongated clusters, but have difficulties in addressing situations of uneven density clusters. Another undesirable characteristic is the "chaining effect", meaning the gathering of distinct clusters whenever there is a chain of data points bridging the gap.

Recently, a new clustering algorithm based on the analysis of dissimilarity increments between neighboring patterns was proposed [3]. Its ability to identify clusters that have arbitrary shape and size, intrinsically finding the number of clusters, was illustrated in a set of application examples. Assuming a parametrical model for cluster representation – an exponential distribution summarizing dissimilarity increments statistics, it is essentially an agglomerative type hierarchical method supported on a new cluster isolation criterion. Estimates of the mean value of dissimilarity increments between neighboring patterns within a cluster are produced and updated along the clustering process, being crucial to the cluster isolation step. It has been pointed out that unreliable estimates of distribution means, particularly occurring when cluster sizes are very small, may lead to over-fragmentation of data, due to premature isolation of clusters. The procedure adopted in [3] to prevent this situation consisted in inhibition of cluster isolation when clusters candidate for merging had both a very small cardinality. In this paper we address the problem of unreliable estimates of distribution means proposing a smooth widening function of the isolation parameter, thus overcoming the difficulties reported above.

Section 2 introduces the cluster isolation criterion and outlines the algorithm's steps. The regularizing function for estimates of the distribution mean of dissimilarity increments, based on a small number of samples, is presented in section 3. Evaluation of the clustering algorithm in comparison with the single link method and the k-means algorithm is performed in section 4 through a set of examples.

2 The Clustering Algorithm

The algorithm described in [3] is a pair-wise clustering method based on the analysis of the statistics of dissimilarity increments between neighboring patterns. It proposes an iterative cluster merging procedure, starting with one sample per cluster, a dendrogram type graph being built along the process. Figure 1(a), plotting a dendrogram, illustrates the basic concepts. Most similar patterns are joined first in a cluster, corresponding to leftmost links on the graph. Vertical lines represent dissimilarity values between samples (when joining single element clusters) or between clusters, in the later case dissimilarity being computed between the most resembling patterns from each cluster (nearest-neighbor rule for computing the distance between clusters). Dissimilarity increments between neighboring patterns are represented on the graph as *gaps*.



(a) Dedrogram and definition of *gap*.

(b) Exponential disttribution.

Fig. 1. Dendrogram and *gap* statistics. The patterns are linked according to the nearest neighbor rule.

The underlying assumption is that dissimilarity increments, or *gaps*, statistics are characteristic of each cluster, an exponential distribution being used to model cluster structure. Taking the one sided view of a cluster C_i, the gap to the other cluster (candidate for merging), gap_i, is computed and compared with the statistic of the first cluster – exponential distribution mean, μ_i. If it is not consistent with this statistic (gap values located far on the tail of the exponential distribution - see fig. 1(b)), the first cluster is isolated in the dendrogram, the algorithm proceeding with the remaining data.

The cluster isolation criterion can be stated as follows:

– *Let C_i, C_j be two clusters which are candidates for merging, and let μ_i, μ_j be the respective mean values of the dissimilarity increments in each cluster. Compute the increments for each cluster, gap_i and gap_j. If $gap_i \geq \alpha\mu_i$ ($gap_j \geq \alpha\mu_j$), isolate cluster C_i (C_j) and proceed the clustering strategy with the remaining patterns. If neither cluster exceeds the gap limit, merge them.*

The term $\alpha\mu_i$ corresponds to a threshold to which gap values are compared, with α being typically set to 3 when dealing with well separated clusters (for a discussion on the selection of this parameter see [3]), and μ_i being replaced by the current mean value estimate based on data present in the cluster.

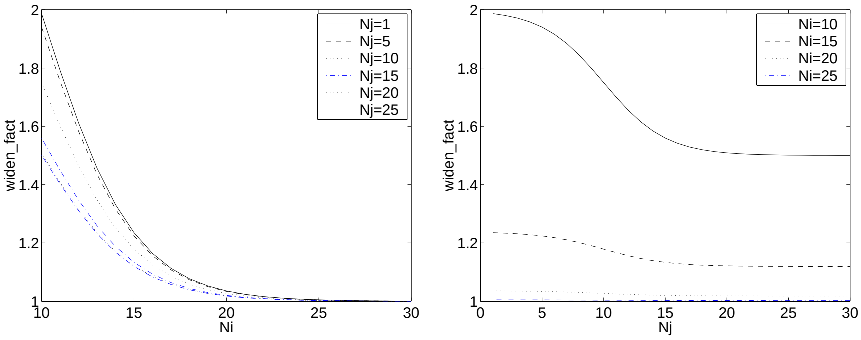
In this paper we propose to replace the threshold $\alpha\mu_i$ by a more adequate dynamic threshold, $th_{dyn\_C_i}$, compensating the effect of under-estimation of *gaps* statistics in the early stages of the clustering algorithm, when a small number of samples is present in each cluster. The revised algorithm with the enhanced dynamic threshold is outlined in table 1.

Table 1: Enhanced clustering algorithm based on dissimilarity increments between neighboring patterns.

```
Input: $N$ samples; $\alpha$ (default value is 3).
Output: Data partitioning.
Steps:
  1. Set: $Final\_clusters = \phi$; $n = N$;
     Put the $i$th sample in cluster $C_i$, $i = 1, \ldots, n$;
     $Clusters = \bigcup_i C_i$, $i = 1, \ldots, n$;
     $d_t[i] = \mu[i] = jumps[i] = 0$, $i = 1, \ldots, n$;
  2. If ($Clusters == \phi$) or ($n == 1$)
     then stop, returning the clusters found in $Final\_clusters \bigcup Clusters$;
     else continue.
  3. Choose the most similar pair of clusters $(C_i, C_j)$ from $Clusters$. Let
     $gap_i = d(C_i, C_j) - d_t[i]$        ni=jumps[i]
     $gap_j = d(C_i, C_j) - d_t[j]$        nj=jumps[j]
  4. If $((gap_i < th_{dyn\_C_i}(\mu[i], ni, nj))$ and $(gap_j < th_{dyn\_C_j}(\mu[j], nj, ni)))$
     then
       join the clusters $C_i$, $C_j$ into cluster $C_{i,j} : C_{i,j} = C_i \bigcup C_j$
       Let $I$ be the index for the merged cluster;
       Replace $C_i$, $C_j$ by $C_{i,j}$ in $Clusters$;
       $d_t[I] = d(C_i, C_j)$;
       $jumps[I] = jumps[i] + jumps[j] + 2$;
       $\mu[I] = \mu[i]\frac{jumps[i]}{jumps[I]} + \mu[j]\frac{jumps[j]}{jumps[I]} + \frac{gap_i + gap_j}{jumps[I]}$;
       Go to step 2.
     else continue.
  5. If $(gap_i \geq \alpha\mu[i])$
     then set $Final\_clusters = Final\_clusters \bigcup C_i$;
       Remove $C_i$ from $Clusters$;
       $n = n - 1$.
     end if
     If $(gap_j \geq \alpha\mu[j])$
     then set $Final\_clusters = Final\_clusters \bigcup C_j$;
       Remove $C_j$ from $Clusters$;
       $n = n - 1$.
     end if
    Go to step 2.
```

3 Adaptive Threshold for Reduced Number of Samples

When a reduced number of samples are gathered in a cluster, estimates of dissimilarity increments statistics are not reliable. It is important to prevent premature cluster isolation in these situations due to low estimates of the distribution mean, $\hat{\mu}$. In order to overcome this difficulty we propose to increase the value of the estimate $\hat{\mu}$ by multiplying it by a term $widen_{fact}(ni, nj)$ greater or equal to 1.



(a) Amplification factor as a function of the number of terms used in the computation of the gaps distribution mean for cluster C_i ($\beta = 1$).

(b) Reinforcement of the amplifying term as a function of the number of elements in cluster C_j ($\beta = 1$).

Fig. 2. Amplification term $widen_{fact}$ associated with the estimate $\hat{\mu}$ for cluster C_i.

We define the amplifying term $widen_{fact}(ni, nj)$ as a monotonous decreasing function of ni, nj - the number of elements available for the computation of the distribution means for cluster C_i and C_j respectively:

$$\begin{aligned} & widen_{fact}(ni, nj) = \\ & \quad 1+\beta \times \underbrace{\left(1 - \frac{1}{1 + e^{-.4(ni-10)}}\right)}_{f_1(ni)} \\ & \quad \times \underbrace{\left(2 - \frac{1}{1 + e^{-.4(nj-10)}}\right)}_{f_2(nj)} \end{aligned} \tag{1}$$

The reasoning underlying expression 1 is the following (see figure 2). If cluster C_i has few samples, the estimate $\hat{\mu}(C_i)$ should be enlarged to compensate for possible underestimation of the true distribution mean; this widening effect smoothly vanishes as the number of terms ni used in the computation of the estimate $\hat{\mu}(C_i)$ increases (fig. 2(a)), which is modelled by the term $f_1(ni)$, a sigmoid-like function. The term $f_2(nj)$ expresses the reinforcement of the widening effect when the number of elements in the competing cluster C_j is also low (fig. 2(b)), taking values greater or equal to 1. When both clusters have low cardinality the combined action of f_1 and f_2 favors clusters merging. When cluster C_i has already a sufficiently large number of elements, the estimate of $\hat{\mu}(C_i)$ is considered to be reliable and term $f_1(ni)$ tends to zero, thus annihilating the influence of term f_2 (the size of cluster C_j becomes irrelevant – see fig. 2(a), $ni \geq 25$). In expression 1, β is a scaling parameter (default value: 3).

When the number of elements available for the estimation of the dissimilarity increments statistic, n_i, is extremely low (such as when the number of cluster's samples is less than 10) the estimate for the μ parameter is very poor. Applying a multiplicative factor to the threshold term may not solve the under-estimation problem in this situation, in particular when $\hat{\mu}$ is near zero. To cope with this situation the proposed dynamic threshold takes a new additive term with a high value and short domain (vanishes for $n_i = 10$)

$$delta_{fact}(ni) = big_{val} \times \left(1 - \frac{1}{1 + e^{-10(ni-5)}}\right), \tag{2}$$

where big_{val} is a large positive number. The final expression of the dynamic threshold to which gaps, seen from cluster C_i perspective, are to be compared is given by:

$$th_{dyn\_C_i}(\hat{\mu}_i, ni, nj) = delta_{fact}(ni) + \alpha\hat{\mu}_i \times widen_{fact}(ni, nj) \tag{3}$$

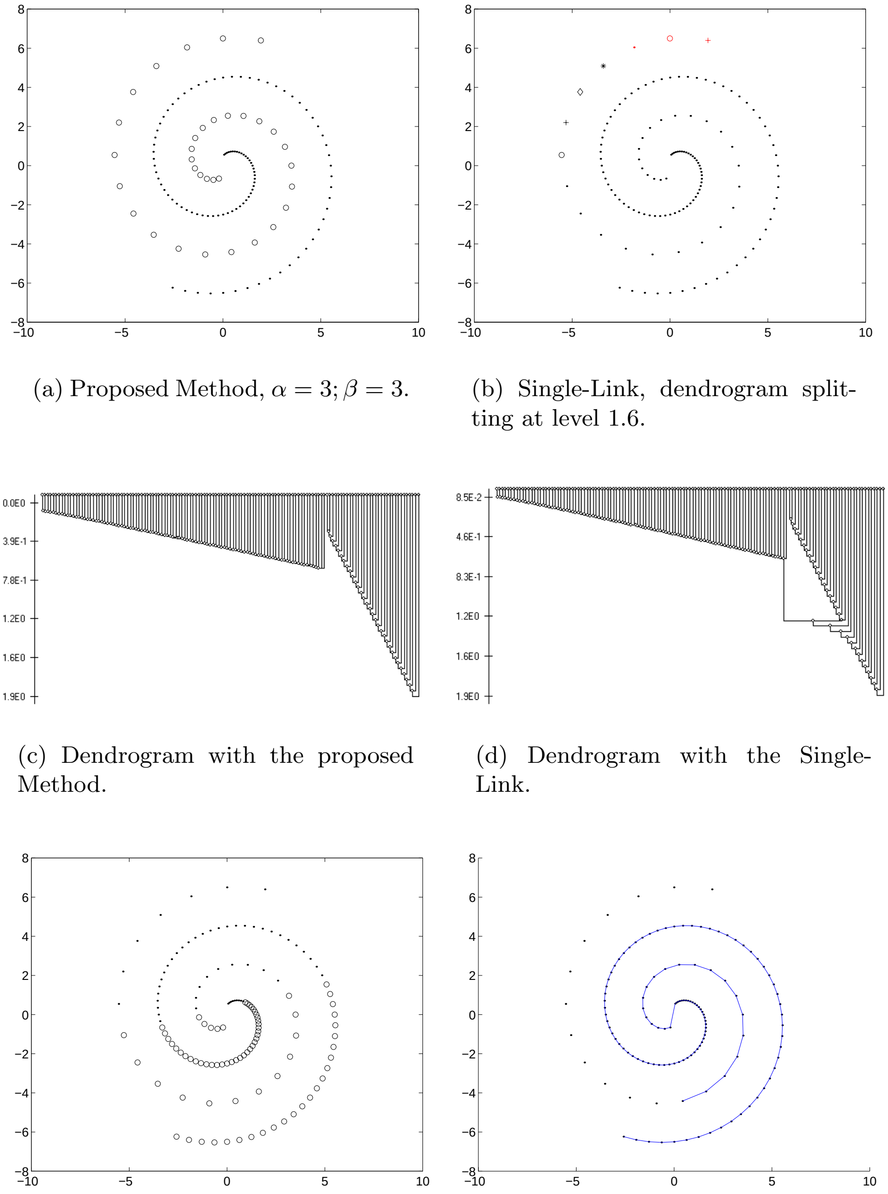
4 Examples

In this section results of application of the proposed method are discussed in comparison with two well known clustering algorithms: a hierarchical agglomerative technique – the single-link method; a partitional algorithm – k-means.

4.1 Spiral Data

The two spiral arms with uneven data sparseness plotted in figure 3(a) constitute an example of a challenging cluster structure for most clustering algorithms reported in the literature.

Clustering results are plotted in figure 3. As shown (figures 3(a) and 3(c), the proposed algorithm correctly identifies the clusters for $2 \leq \alpha < 100$ (β was set to 3). The single-link method and the k-means algorithm, however, are unable to handle this type of data structure. The focus on data compactness provided by the k-means algorithm fails to capture essential properties of the data, the clustering results being represented in fig. 3(e). The single-link method produces the dendrogram depicted in figure 3(d). Thresholding on this graph is equivalent to cutting weak edges in the minimum spanning tree (fig. 3(f)), resulting in a large cluster gathering points from both spiral arms, and additional single element clusters (fig 3(b)). The ability of the proposed method to distinguish between different gaps statistics leads to isolation of the cluster with a higher number of points (denser data), the corresponding sub-graph on the dendrogram being frozen, thus enabling correct formation of the second cluster, as shown in fig. 3(c).



(a) Proposed Method, $\alpha = 3; \beta = 3$.

(b) Single-Link, dendrogram splitting at level 1.6.

(c) Dendrogram with the proposed Method.

(d) Dendrogram with the Single-Link.

(e) K-means, $k = 2$.

(f) Thresholding on the minimum spanning tree, th=1.3

Fig. 3. Spiral data (133 samples.)

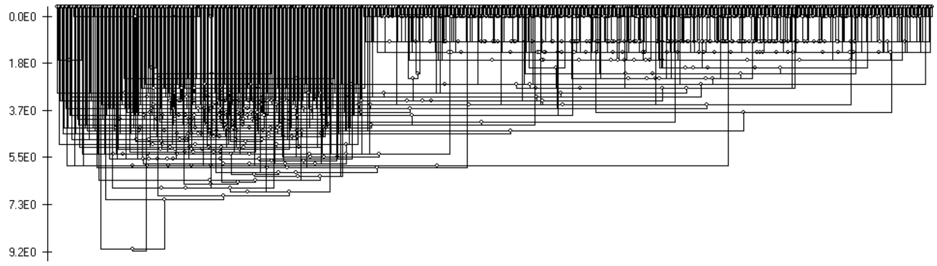


Fig. 4. Dendrogram produced by the single-link method for the breast cancer data.

4.2 Breast-Cancer Data

The Wisconsin Breast Cancer Data set (available at the UCI Machine Learning Repository [12]) consists of two classes (benign and malignant, 444 and 239 samples, respectively), represented by 9 features. Class labels are ignored for clustering.

References [13] and [14] present clustering results using cluster center based methods, according to which performances obtained on this data were 94.28% and 95.5% correct classifications, respectively. Using the k-means algorithm, with k=2, results are dependent on the initial cluster centers. After several experiments, the best accuracy achieved was 96.49% (24 samples misclassified).

Figure 4 represents the dendrogram produced by the single-link method, by ordering data according to their class labels: benign patterns are on the right side of the graph. As shown, this method is not able to differentiate between the two types of data: simple thresholding on this graph leads to a cluster with most of the samples, and spurious single pattern clusters being formed. Analysis of this graph shows that the two classes are not well separated but exhibit different structures of inter-pattern distances.

With the proposed method a single cluster is obtained for $\alpha = 3$ (which assumes good cluster separation). By setting this threshold to 1 two clusters are identified (with β in expression 1 taking an arbitrary value) corresponding to a recognition rate of 96.63% (23 samples were misclassified). This result compares favorably to the results reported above.

Table 2. Clustering results obtained with the proposed method on the breast cancer data.

$\alpha = 1$	#Clusters	Recognition rate
$th = \alpha\hat{\mu}_i$	3	96.49 %
$th_{dyn\_C_i}$	2	96.63 %

Table 2 presents the clustering results when using the dynamic threshold as defined in expression 3 or the threshold $th = \alpha\hat{\mu}_i$. As shown, samples in the spurious cluster produced in the later situation are distributed to the correct clusters by using the dynamic threshold, thus improving the overall performance.

5 Conclusions

This paper presented and enhanced version of a clustering algorithm [3] based on dissimilarity increments between neighboring patterns. The novelty of this contribution consisted of addressing the problem of unreliable estimates of distribution means by proposing a smooth widening function to replace the threshold parameter given in the cluster isolation criterion underlying the work in [3].

The ability of the enhanced clustering algorithm to produce correct data partitioning has been demonstrated on an artificially created data set as well for a real world application. Examples included clusters with arbitrary shapes and sizes, the method correctly identifying the intrinsic data structure. For the breast-cancer data set, the influence of the new dynamic threshold function was crucial to increase the percentage of correct classifications, the final result outperforming both the single-link and k-means algorithms, as well as other clustering results reported in the literature.

Acknowledgments

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Portuguese Ministry of Science and Technology, and FEDER, under grant POSI/33143/SRI/2000.

The breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The database is available at the UCI repository of Machine Learning Databases [http://www.ics.ici.edu/ mlearn/ MLRepository.html].

References

1. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988)
2. Jain, A., Murty, M.N., Flynn, P.: Data clustering: A review. ACM Computing Surveys **31** (1999) 264–323
3. Fred, A.L., Leitão, J.: Clustering under a hypothesis of smooth dissimilarity increments. In: Proc. of the 15th Int'l Conference on Pattern Recognition. Volume 2., Barcelona (2000) 190–194
4. McLachlan, G., Basford, K.: Mixture Models: Inference and Application to Clustering. Marcel Dekker, New York (1988)

5. Roberts, S., Husmeier, D., Rezek, I., Penny, W.: Bayesian approaches to gaussian mixture modelling. IEEE Trans. Pattern Analysis and Machine Intelligence **20** (1998)
6. Pauwels, E.J., Frederix, G.: Fiding regions of interest for content-extraction. In: Proc. of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII. Volume SPIE Vol. 3656., San Jose (1999) 501–510
7. Buhmann, J., Held, M.: Unsupervised learning without overfitting: Empirical risk approximation as an induction principle for reliable clustering. In Singh, S., ed.: International Conference on Advances in Pattern Recognition, Springer Verlag (1999) 167–176
8. Mirkin, B.: Concept learning and feature selection based on square-error clustering. Machine Learning **35** (1999) 25–39
9. El-Sonbaty, Y., Ismail, M.A.: On-line hierarchical clustering. Pattern Recognition Letters (1998) 1285–1291
10. Zahn, C.: Graph-theoretical methods for detecting and describing gestalt structures. IEEE Trans. Computers **C-20** (1971) 68–86
11. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1973)
12. Merz, C.J., Murphy, P.M.: Uci repository of machine learning databases. [http://www.ics.uci.edu/ mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science (1996)
13. Kothari, R., Pitts, D.: On finding the number of clusters. Pattern Recognition Letters **20** (1999) 405–416
14. Chakravarthy, S.V., Ghosh, J.: Scale-based clustering using the radial basis function network. IEEE Trans. Neural Networks **7** (1996) 1250–1261