Non Parametric Local Density-Based Clustering for Multimodal Overlapping Distributions*

Damaris Pascual[1], Filiberto Pla[2], and J. Salvador Sánchez[2]

[1] Dept de Ciencia de la Computación, Universidad de Oriente,
Av. Patricio Lumunba s/n, Santiago de Cuba, CP 90100, Cuba
dpascual@csd.uo.edu.cu
[2] Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I,
12071 Castelló, Spain
{pla, sanchez}@lsi.uji.es

Abstract. In this work, we present a clustering algorithm to find clusters of different sizes, shapes and densities, to deal with overlapping cluster distributions and background noise. The algorithm is divided in two stages. In a first step, local density is estimated at each data point. In a second stage, a hierarchical approach is used by merging clusters according to the introduced cluster distance, based on heuristic measures about how modes overlap in a distribution. Experimental results on synthetic and real databases show the validity of the method.

1 Introduction

Many application problems require tools aimed at discover relevant information and relationships in databases. These techniques are mainly based on unsupervised pattern recognition methods like clustering. The problem of clustering can be defined as: Given n points belonging to a d-dimensional space, and provided some measure of similarity or dissimilarity, the aim is to divide these points into a set of clusters so that the similarity between patterns belonging to the same cluster is maximized whereas the similarity between patterns of different clusters is minimized.

There are two main approaches in clustering techniques: the partitioning approach and the hierarchical approach [8]. The partitioning methods build a partition splitting a set of n objects into k clusters. These algorithms usually assume a priori knowledge about the number of classes in which the database must be divided. The K-means is one of the best known partitioning algorithms.

Other clustering algorithms are based on parametric mixture models [3]. However, this work focuses on non parametric approaches, since they can be applied in a more general way to metric and non metric feature spaces, just defining a dissimilarity measure in the feature space.

Hierarchical methods consist of a sequence of nested data partitions in a hierarchical structure, which can be represented as a dendogram. There exist two hierarchical

* This work has been partially supported by projects ESP2005-07724-C05-05 and TIC2003-08496 from the Spanish CICYT.

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 671 – 678, 2006.

approaches: agglomerative and divisive. The first one can be described in the following way: initially, each point of the database form a single cluster, and in each level, the two most similar clusters are joined, until either a single cluster is reached containing all the data points, or some stopping condition is defined, for instance, when the distance between the clusters is smaller than certain threshold. In the divisive approach, the process is the other way around.

The Single Link (SL) and the Complete Link (CL) methods are the most well known hierarchical strategies [4]. Some hierarchical algorithms are based on prototypes selection, as CURE [5]. On the other hand, in density–based algorithms, the clusters are defined as dense regions, where clusters are separated by low density areas [6]. Some of the most representative works of the density-based approach are DBSCAN [1], KNNCLUST [8] and SSN [2] algorithms.

The main problems of these algorithms are the fact that clusters are not completely separable, due to the overlapping of cluster distributions, and the presence of noisy samples. The main contribution of the work presented here is the use of a hybrid strategy between the hierarchical and density-based approaches, and the cluster dissimilarity measure introduced, both aimed at dealing with overlapped clusters and noisy samples, in order to discover the most significant density based distributions in databases with high degree of cluster overlapping and clusters with multiple modes.

2 Clustering Process

The objective of the algorithm here presented is to detect clusters of different shapes, sizes and densities even in the presence of noise and overlapping cluster distributions. The algorithm here presented is a mixture of a density-based and a hierarchical-based approach, and it is divided in two stages. In the first stage, the initial clusters are constructed using a density-based approach. In a second stage, a hierarchical approach is used, based on a cluster similarity function defined in terms of cluster density measures and distances, joining clusters until either arriving to a pre-defined number or reaching a given stopping criterion.

2.1 Point Density Estimation

Let X be a set of patterns provided with a similarity measure between patterns d. Let x be an arbitrary element in the dataset X , and $R>0$. The neighbourhood V_R of radius R of x is defined as the set $V_R(x)=\{y/d(x,y)\leq R\}$, and the local density $p(x)$ of the non-normalized probability distribution at point x as:

$$p(x)=\sum_{x_i\in V_R(x)}\exp\left(-\frac{de^2(x,x_i)}{R^2}\right) \tag{1}$$

where x_i are the points that belong to the neighbourhood of radius R of x, V_R, and de, the Euclidean distance.

In the algorithm presented here, we will differentiate between two concepts: *core cluster* and *cluster*. We will refer to *core clusters* to the sets that are obtained after

applying the first stage of the algorithm, and we will refer to *clusters* to the groups of core clusters that will be grouped into clusters in a further stage.

2.2 Di-similarities Between Clusters

As part of the hierarchical approach, we need to define a di-similarity measure that takes into account two possible facts, when clusters are overlapped or completely separated. Let us define the following di-similarty function d between two clusters K_i and K_j ,

$$d(K_i, K_j) = do(K_i, K_j)\ \ (1 + ds(K_i, K_j)) \tag{2}$$

where $do(K_i,K_j)$ is a measure of overlapping between clusters K_i and K_j , and $ds(K_i,K_j)$ is a measure of separability between those clusters.

The separability measure can be defined as

$$ds(K_i, K_j) = \min \{dsc(C_m, C_n)\},\ \forall\ C_m,\ C_n / C_m \in K_i \text{ and } C_n \in K_j$$

where C_m , C_n are two core clusters, one from each cluster, and let us define the distance between two core clusters as:

$$dsc(C_m, C_n) = \min \{de(x_m, x_n)\};\ \forall\ x_m, x_n / x_m \in C_m \text{ and } x_n \in C_n$$

That is, the distance or di-similarity measure of separability between two clusters is the shortest distance between any pair of points, one point from each cluster. Therefore, for overlapped clusters, $ds=0$.

On the other hand, about the cluster overlapping measure in equation (2), $do(K_i,K_j)$, let us suppose that each cluster corresponds to one mode in Figure 1. A non parametric measure of the degree of overlapping of such modes can be defined referring to the density value of the border point x_b between both modes.

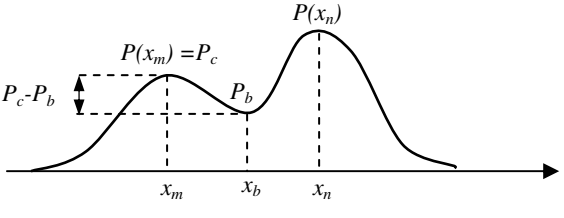

Fig. 1. Overlapping measures between two distribution modes

Therefore, let us define the overlapping degree of the two modes in Figure 1, $doc(C_m ,C_n)$, as the relative difference between the density of the modes centres, x_m and x_n, with respect to the density at the border x_b between both modes. We can express this relative measures as

$$d(C_m, C_n) = \frac{P_c - P_b}{P_c} \tag{3}$$

Given a core cluster C_m , the centre of the core cluster x_m is defined as the point whose density is maximal within the core cluster. Let x_m and x_n be the centres of C_m and C_n respectively. Therefore, P_c in equation (3) is defined as the minimum density of the core cluster centres x_m and x_n, that is, $Pc=\min(p(x_m),p(x_n))$. Note that the di-similarity measure of overlapping in equation (3) is normalized in the range [0,1].

In equation (3), P_b is the density at the midpoint of the border between both core clusters, which is defined as the midpoint between the nearest points xb_m and xb_n, one from each core cluster, C_m.and C_n . Finally, the measure of the degree of overlapping between two clusters $do(K_i,K_j)$, is be defined as

$$do(K_i,K_j) = \min \{doc(C_m,C_n)\};\ \forall\ C_m, C_n / C_m \in K_i \text{ and } C_n \in K_j$$

In a few words, the di-similarity measure defined in (2) is aimed at considering that clusters are more similar when their probability distributions are either nearer in the feature space, measured by means of the separability measure $ds()$, or when their probability distributions are more overlapped. When the probability distributions are overlapped ($ds=0$), the measure of similarity becomes the overlapping degree of the probability density term $do()$, which is a heuristic local estimate of the mixed probability distributions at the border between clusters (Figure 1).

2.3 Clustering Algorithm

The clustering algorithm here presented consists of a hierarchical agglomerative strategy based on a Single Link approach, using the di-similarity measures defined in the previous Section. The use of such di-similarity measures defines the behaviour of the clustering process and the response to the overlapping of the local distributions of patterns in the data set.

Therefore, the proposed algorithm can be summarized in two stages as follows:

First stage:

```
Input:  radius R, data points and density noise threshold
Output: data points grouped into N core clusters
```

1. Initially, each point of the database is assigned to a single core cluster.
2. For each point x, calculate its neighbourhood of radius R, $V_R(x)$
3. For each point x in the database, estimate its probability density p(x) according to expression (1).
4. Assign each point x to the same core cluster of the point xc in its neighbourhood, being xc the point with maximal density in the neighbourhood of x.
5. Mark all core clusters with density less than the density noise threshold as noise core clusters. The rest of the core clusters are the resulting N core clusters.

Second stage

```
Input:  N core clusters
Output: K clusters
```

1. Initially, assign each one the N core clusters from the first stage to a single cluster. Therefore, there are initially N clusters with one core cluster.
2. Repeat until obtaining K clusters,
   - 2.1 Calculate the distance between each pair of clusters using expression (2)
   - 2.2 Join the two clusters in step 2.1 that their distance is minimum
3. Eventually, assign the noise core clusters to a nearest.

3 Experimental Results

In this section, some experimental results are presented aimed at evaluating the proposed algorithm, hereafter named *H-density*, and to compare it with some other similar algorithms referred in the introduction, DBSCAN, CURE and K-means. In order to test the algorithm, three groups of experiments are performed. The first one uses synthetic databases based on overlapped Gaussian distributions, in order to see the response of the proposed algorithm in these controlled conditions. The second experiment uses two synthetic databases from [7], for comparison purposes, and to test the problem of the presence of noise, overlapping, and clusters of different sizes and shapes. Finally, some experiments are performed on three real databases.

3.1 Gaussian Databases

Several databases using Gaussian distributions were generated with different number Gaussians, sizes and overlapping degrees. The results obtained in one of these databases are shown in Figure 2, where we can notice how the algorithm has been able to correctly detect each one of the existing Gaussian distributions, even in the presence of significant overlapping.

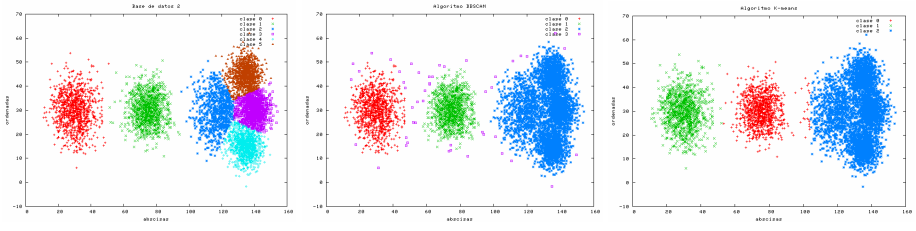

Fig. 2. Results on a Gaussian database of (left to right) H-density, DBSCAN and k-means

The DBSCAN algorithm did not correctly detect all the Gaussians in different data-bases because it is not able to separate the overlapped distributions. The CURE and K-means algorithms correctly detected the three main clusters. However, in the case of trying to find six clusters, they could not detect the 4 Gaussians highly overlapped.

3.2 Synthetic Databases

In [7], some experiments were presented for the DBSCAN and CURE algorithms using the databases of Figure 3 (see [7] for comparison results with those algorithms and note the satisfactory results of the proposed H-density algorithm). Notice the presence of clusters of different shapes, sizes, noise and overlapping. Figure 4 shows the result of applying the proposed H-density algorithm on these databases. Note how the algorithm has correctly grouped the main clusters present in the data set. Figure 4 shows the result of the K-means algorithm for 6 clusters (left) and 9 clusters (right) of the corresponding databases. The errors in the grouping are noticeable.

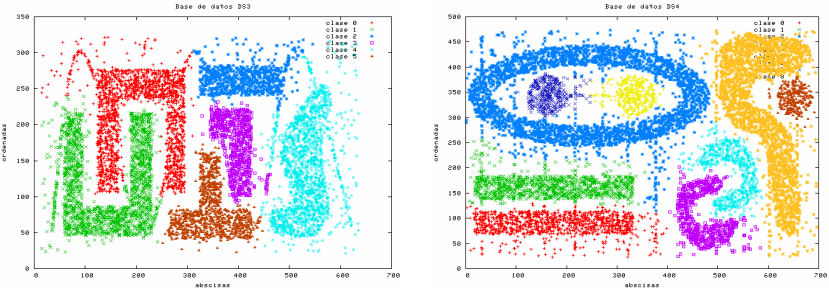

Fig. 3. Results of the H-density algorithm on databases from [7]

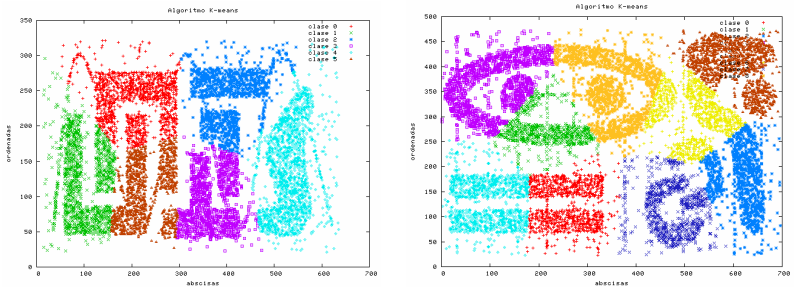

Fig. 4. Results of the K-means algorithm on databases from [7]. Left: for 6 clusters. Right for 9 clusters.

3.3 Real Databases

Two real databases were used in this experiment, Iris and Cancer. These databases were used for comparison purposes with the results presented in [4]. The first one is a database of Iris plants containing 3 known class labels, with a total of 150 elements, 50 each of the three classes: Iris Setosa, Iris Versicolour, Iris Virginica. The number of attributes is 4, all numeric. The first class, Iris Setosa, is linearly separable from the other two classes.

In order to compare the clustering results with the ones presented in [4], there was provided an error classification measure using the NN classifier, taking as training set the resulting clusters of the clustering algorithms, and as a test set the original labelled data set. The class assigned to each cluster was the class of the majority of patterns with the same class from the original dataset.

In the first experiment, all the algorithms were run to obtain two classes, and all of them obtained 100% of correct grouping or classification, that is, all the tested algorithms were able to correctly separate the Setosa class from the other ones.

In a second experiment, the algorithms were run to find three clusters. The results are shown in Table 1. Notice how, due to the overlapping between Versicolour and Virginica classes, the proposed H-density algorithm outperforms the other ones reaching a 94% correct classification. In the case of the Cancer database, it has 2 classes. The proposed H-density algorithm obtained a 95.461% of correct classification, the same as CURE (Table 2).

Table 1. Classfification rate of the clustering algorithm on Iris database

Algorithm	% in two classes	% in three classes
DBSCAN	100	71.33
CURE	100	83.33
K-means	100	88.33
H-Density	100	94.00

Table 2. Classification rate of the clustering algorithms in Cancer database (two classes)

Database	DBSCAN	CURE	K-means	H-Density
Cancer	94.28	95.461	95.04	95.461

Finally, the H-Density algorithm was run on a dataset consisting of the chroma values of the *Lab* representation of the "house" image (Figure 5 left). This image has 256x256 pixels, and the clustering was performed in the *ab* space to find 5 different colour classes. Note how the algorithm has been able to correctly identify 5 different clusters with a high degree of overlapping and different shapes and sizes (Figure 5 right). To see the goodness of the clusters found Figure 3 (middle) shows the labelled pixels with the corresponding assigned clusters.

5 Conclusions and Further Work

A hierarchical algorithm based on local probability density information has been presented. The way the density of the probability distribution is estimated, and the use of this information in the introduced dissimilarity measure between clusters, provides to the algorithm a mechanism to deal with overlapping distributions and the presence of noise in the data set. The experiments carried out show satisfactory and promising results to tackle these problems usually present in real databases. The experiments

also show the proposed algorithm outperforms some existing algorithms. Future work is directed to unify the treatment of noise and overlapping in the process, and to introduce a measure to assess the "natural" number of clusters in the hierarchy.

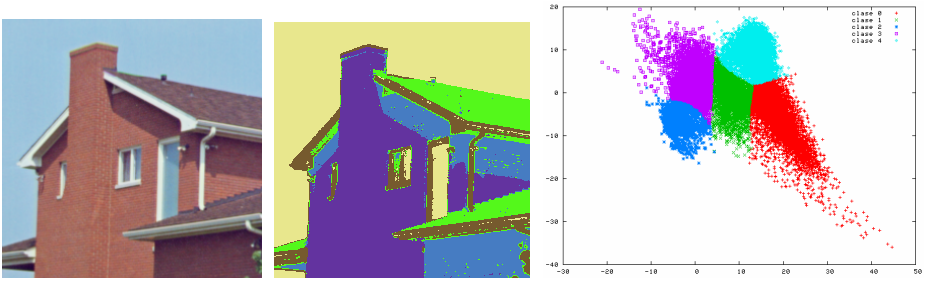

Fig. 5. Result of the H-density algorithm on the "house" image. Left: original image. Middle: labelled image. Right: 5 colour clusters found of pixels in the *ab* space.

References

1. Ester, M.; Kriegel, H. P.; Sander, J. and Xu, X.; A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of the second International Conference on Knowledge Discovery and Data Mining, Portland, (1996) 226-231.
2. Ertöz, L.; Steinbach, M. and Kumar V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In Proceedings of Third SIAM International Conference on Data Mining, (2003).
3. Figueiredo, M. and Jain, A.K.; Unsupervised Learning of Mixture Models, IEEE Trans. on PAMI, Vol 24, No 3 (2002) 381-396.
4. Fred A. L. and Leitao J.: A New Cluster Isolation Criterion Based on Dissimilarity Increments. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol 25, No 8, (2003) 944-958.
5. Guha, S.; Rastogi, R. and Shim, K.; CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, . ACM, New York, (1998) 73-84.
6. Hinneburg A. and Keim D.A.: An efficient Approach to Clustering in Large Multimedia Databases with Noise. In Proc. of the ACM SIGKDD, (1998).
7. Karypis, G.; Han, E.H. and Kumar, V.; Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling. In the IEEE Computer Society. Vol 32, No 8 (1999) 68-75.
8. Tran T. N., Wehrens R. and Buydens L.M.C.: Knn Density-Based Clustering for High Dimensional Multispectral Images. Analytica Chimica Acta 490 (2003) 303–312.