SCALE-BASED CLUSTERING WITH LATENT VARIABLES

F.M. Frattale Mascioli, M. Panella, A. Rizzi, and G. Martinelli
INFO-COM Dpt., University of Rome “La Sapienza”,
Via Eudossiana 18, 00184 Roma, ITALY
Tel: +39 06 44585488; fax: +39 06 4873300
e-mail: panella@infocom.uniroma1.it; mascioli@infocom.uniroma1.it

ABSTRACT

The use of clustering systems is very important in those real-world applications where an efficient, both accurate and economical, representation of the data to be processed is necessary. When dealing with statistical models, such a problem is usually related to the estimate of their parameters in the Maximum Likelihood context. At this regard, we propose an EM-based algorithm that uses a hierarchical growing approach, based on a given splitting procedure, to determine in an efficient way the parameters of a mixture of Gaussian clusters. The splitting procedure and the determination of the correct number of clusters are based on a scale-based approach, which imitates the human perception of images. Moreover, each cluster is modelled by means of latent variables, which also ensure a local linear dimension reduction of the data being processed.

1 STATISTICAL CLUSTERING MODELS

The main purpose of clustering systems is the unsupervised detection of statistical regularities of a data set. This is achieved by separating the data in subsets contained in suitable regions of the space, where they are represented as points. The success of the clustering is mandatory to the success of other operations of data elaboration as: encoding, compression, function approximation, statistical analysis and combinatorial optimisation. Basic topics toward a successful clustering procedure are:

- The construction of a model of the probabilistic distribution, given a finite sample drawn from that distribution.
- The specific strategy followed and its motivation.

We will briefly discuss these two points, in order to establish the premises of the algorithm to be proposed.

With regard to the first point, the goal is to model the unknown process which yields the patterns we are analysing. Since we can properly estimate only the moments of the first and second order of the data, the Gaussian model is the most reasonable. In fact, the model should be in agreement with the known characteristics of the process without adding arbitrary information with regards to what is unknown. In other words, the probabilistic description of the data source should maximise its entropy defined in the sense of information theory [1]. This result is just achieved with Gaussian model, since it is characterised by the largest entropy among the processes with same first and second order moments [2]. In conclusion, a valid model of the data density is a probabilistic mixture of normal distributions, each representing the model of a cluster with its own independent mean $\mu$ and covariance matrix $\Sigma$, i.e.

$$p(x) = \sum_{i=1}^{C} \pi_i p(x|i) \quad (1)$$

where: $x$ is a point of the d-dimensional space where the data are represented; $p(x|i)$ is the conditional distribution of the $i$-th component process (the cluster model); $\pi_i$ is the mixing coefficient which satisfies the requirements $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{C} \pi_i = 1$; $C$ is the number of processes (clusters) of the model. The mixing coefficient $\pi_i$ can be interpreted as the prior probability for the $i$-th cluster.

The second point to be discussed regards the strategy to be followed. It is inspired by two biological remarks, which suggest the use of scale-based principle and latent variables.

1.1 Scale-based Principle

The first remark concerns with how the brain perceives important structures of an image [3]. Only structures that survive over a large range of scale are perceived and therefore taken into consideration.

When applied to clustering, the imitation of this mechanism requires the introduction of a scale parameter for simulating the sight of the data set, represented in a suitable space (the data space), by an imaginary observer located at different distances [4]. Varying the scale parameter is similar to modify the distance from the data and hence, the resolution level at which they are observed. The structures of the data relevant to clustering are the locations of high-density regions. They are natural candidates to the role of clusters. Taking account of the mechanism to be imitated, not all these locations are retained but only those surviving the largest variations of scale [5]. The implementation of the scale-based principle should require that: 1) the algorithm is constructive and hierarchical; 2) there is a control parameter (the scale); 3) the optimal partition is chosen on the basis of the scale range size where it is present.
1.2 Latent Variables Models

The second remark concerns with how the brain stores in an economical way the information. There is evidence that the representation of knowledge in the brain assumes the form of a feature map geometrically organised over appropriate regions. This was the motivation of the development of the SOM’s by Kohonen [6].

These neural nets imitate the brain in producing a mapping between the data space and the bidimensional neuron array with important property of preserving the topological relations while performing a dimensionality reduction of the data space. The reduction from a d-dimensional to a bidimensional data space, accomplished in the brain maps and imitated by the SOM’s, is very easy to show that p(x) will be reduced to (2). The simplest one is linear, i.e.

\[ p(x) = \pi \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)) \]

where \( \pi \) allows the conditional distribution to have a non-zero mean. The latent variable distribution has been chosen to be a zero-mean unit covariance Gaussian, and the noise model for \( u \) is also a zero-mean Gaussian with isotropic covariance matrix \( \sigma_u^2 I \). Under these assumptions, it is easy to show that p(x|i) remains Gaussian with mean \( \mu_i \) and covariance matrix

\[ \Sigma_i = (\sigma_u^2 I + W_i W_i^T) \]

This model belongs to the class of the linear models known as factor analysis models. From (4) it is evident that the number of parameters in p(x|i) will be reduced from d(d+1)/2 in \( \Sigma_i \) to (d+q) in \( \sigma^2_i \) and \( W_i \).

In conclusion, the data density model (1) will be represented by a mixture of C clusters each related to a Gaussian distribution p(x|i), with the parameters to be determined, in addition to the prior probability \( \pi_i \), are \( \mu_i \), \( W_i \), and \( \sigma^2_i \).

2 DETERMINATION OF THE MIXTURE PARAMETERS VIA EM ALGORITHM

The mixture parameters, for a fixed number C of clusters, can be obtained by applying a specific version of the well-known Expectation-Maximisation (EM) algorithm. It alternates between two steps (E and M), whose iteration is guaranteed to find a local maximum of the log-likelihood \( L \) of the data set to be clustered [9]. The algorithm stops when the variation of \( L \) is less than a suitable threshold.

As clustering result, each pattern \( x_n, n=1…N \), can be assigned to the i-th cluster having the maximum posterior responsibility for generating \( x_n \). This responsibility \( R_{ni} \) can be reasonably evaluated by Bayes theorem via

\[ R_{ni} = \pi_i p(x_n | i); \quad i=1…C, \quad n=1…N \]

A specific version of the EM algorithm, developed for latent variables mixtures models [8], is considered:

- In the E-step, starting with ‘old’ values of the parameters, \( L \) and \( R_{ni} \) are evaluated.
- In the M-step \( L \) is maximised with respect to the parameters \( \pi_i, \mu_i, W_i, \) and \( \sigma^2_i \), i=1…C, to give a revised set of them. Then, the mixing coefficients and components means are re-estimated using

\[ \pi_i = \frac{1}{N} \sum_{n=1}^{N} R_{ni} \]

\[ \mu_i = \frac{\sum_{n=1}^{N} x_n R_{ni}}{\sum_{n=1}^{N} R_{ni}} \]

The parameters \( W_i \) and \( \sigma^2_i \) are obtained by first evaluating the weighted covariance matrix given by

\[ S_i = \frac{\sum_{n=1}^{N} R_{ni} (x_n - \mu_i) (x_n - \mu_i)^T}{\sum_{n=1}^{N} R_{ni}} \]

and then applying

\[ \sigma^2_i = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j \]

\[ W_i = U_i \left( A_i - \sigma^2_i I \right)^{1/2} Q_i \]

where \( \lambda_j \), j=1…d, are the eigenvalues of \( S_i \), ordered in decreasing magnitude, the q largest of which are stored in the diagonal matrix \( A_i \), the columns in the dxq matrix \( U_i \) are the corresponding q normalised eigenvectors; and \( Q_i \) is a qxq orthogonal rotation matrix that can be omitted since it does not affect (4).
The relation between latent variables, factor analysis and principal components is shown in (8). Each cluster is modelled by means of a local linear principal component analysis of the data. Namely, a local linear dimension reduction from a d-dimensional to a q-dimensional data space is obtained for each cluster, so that the mixture (1), called in this case Mixture of Probabilistic Principal Component Analysers (MPPCA), yields to a global non-linear dimension reduction of the data space [10].

3 THE PROPOSED ALGORITHM: HMPPCA

The EM algorithm, which determines the clusters, is not sufficient for deriving the overall model (1) of the data set: the key problem remains the determination of the number of clusters C. The optimisation of this number, very critical in clustering procedures, could be achieved by minimising cost functional depending either on C or on a parameter strictly related to it. That would require several EM algorithm applications for each value of C to be considered, so that the optimisation procedure may result very expensive.

In this paper, we propose a hierarchical approach for increasing C progressively and thus for exploiting the information about previous clustering result. The basic operations of the algorithm, which will be denoted as Hierarchical MPPCA (HMPPCA), are listed below:

- **Initialisation.** At the beginning, there is a unique cluster modelling the entire data set. The determination of its parameters is straightforward by using (5)-(8) with C=1.

- **Generic step.** There are C clusters obtained in the previous step with their initial values of parameters. The EM algorithm determines their final values. Then, according to the scale-based criterion explained in the following, one cluster is selected to be split. From this cluster, the splitting procedure generates a pair of new ones and computes their initial parameters values. They are necessary for the EM algorithm in the successive generic step.

- **Stop criterion.** In order to avoid an excessive complexity of the resulting clustering model, the algorithm terminates if C exceeds a suitable fraction N≤N of the training set cardinality.

The splitting procedure of a cluster is based on a suitable choice among its principal directions (i.e. the eigenvectors of its covariance matrix). A scale parameter is determined by simulating this procedure for each principal direction of each cluster that could be split. Let us consider the h-th cluster, h=1…C, and its j-th principal direction, j=1…d. The patterns belonging to the cluster are divided in two groups (offspring clusters) on the basis of their projections \( \xi_{\text{hj}}^{(m)} \) onto the chosen direction, i.e.

\[
\xi_{\text{hj}}^{(m)} = u_{\text{hj}}^T x_m
\]

where \( u_{\text{hj}} \) is the j-th eigenvector of the cluster covariance \( S_{\text{h}} \), and \( x_m \) is the m-th pattern belonging to the h-th cluster on the basis of (5). The pattern is assigned to one group if \( \xi_{\text{hj}}^{(m)} \leq \xi_{\text{hj}}^{(\mu)} \), where \( \xi_{\text{hj}}^{(\mu)} \) is the projection of the mean \( \mu_{\text{h}} \); else it is assigned to the other group. We define as corresponding scale the Euclidean distance between the means of the resulting offspring clusters.

By using this procedure, it is evident that the largest value of the scale can be reasonably associated to the coarsest resolution level of the data. Therefore, according to the scale-based principle, the cluster and its principal direction that yielded the largest value of the scale will be selected for the splitting procedure in the current generic step of the HMPPCA algorithm. It will be also the scale associated with this generic step.

After the hierarchy is generated, the optimal clusters must be selected. As suggested by the scale-based principle, the most persistent clusters over the variation of the scale are chosen. We outline that, as a cluster splits, the resolution level at which its data are observed increases and the scale reduces. Good results are obtained by taking into account the intervals of existence of clusters present at different levels of the hierarchy (local stability) [4].
A bidimensional example is useful for illustrating the previous optimality criterion. In Fig. 1, the data set to be clustered is shown together with the optimal partition obtained by the local lifetime criterion. There are 5 clusters, each represented by the mean of the mixture process and by an ellipse related to the unit Mahalanobis distance from the cluster centre. In the case of the data set in Fig. 1, we obtain the dendrogram of Fig. 2: the dotted lines correspond to the clusters having the largest lifetimes and thus determining the optimal partition in Fig. 1.

4 SIMULATION TESTS AND CONCLUSIONS

Exhaustive simulation tests were carried out for ascertaining the performance of the proposed algorithm. The performance evaluation of a clustering algorithm is difficult and depends on the specific application. In this section we report the results concerning the use of clustering as an unsupervised classification system. In particular, the efficient performance of the SOM’s suggests the use of \( q=2 \) as default value for a large reduction of the model parameters.

The first test concerns with the classification of the IRIS data set [11]. In Tab. 1 is reported the resulting confusion matrix. With only 2 errors, the algorithm yields a 98.67% classification accuracy, which is the performance measure adopted in this case.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class I</td>
<td>Class II</td>
</tr>
<tr>
<td>Cluster A</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Cluster B</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>Cluster C</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Confusion matrix of IRIS data set.

The second test regards the classification of the Cancer data set [12]. The confusion matrix reported in Tab. 2 shows that the Benignant class is decomposed in two clusters (A and B), while the Malignant one is modelled by a single cluster (C). The algorithm yields in this case a 96.14% classification accuracy, with 27 misclassified patterns.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benignant</td>
</tr>
<tr>
<td>Cluster A</td>
<td>201</td>
</tr>
<tr>
<td>Cluster B</td>
<td>250</td>
</tr>
<tr>
<td>Cluster C</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of Cancer data set.

The results obtained on real data sets by the actual implementation favourably compare with the best performances obtained by different approaches reported in the technical literature.

In conclusion, the clustering algorithm proposed in the present paper is based on the imitation of two biological mechanisms: the scale-based principle underlying human vision and the formation of brain maps for storing information from sensory stimuli. The latter is strictly connected to the dimensionality reduction of data, by projecting them from the input space onto a latent space. The introduction of a scale is obtained by following a hierarchical approach, which starts from a unique cluster. This strategy directly saves the problem of initialisation and improves the computational efficiency of the algorithm. It could be further improved by investigating appropriate aspects only touched in the paper.

References