UNSUPERVISED NON PARAMETRIC DATA CLUSTERING BY MEANS OF BAYESIAN INFERENCE AND INFORMATION THEORY

Gilles Bougenière, Claude Cariou, Kacem Chehdi
TSI2M Laboratory, University of Rennes 1 / ENSSAT, 6 rue de Kerampont, 22300 Lannion, France

Alan Gay
Institute of Grassland and Environmental Research (IGER), Plas Gogerddan, Aberystwyth, Ceredigion, SY23 3EB, UK

Keywords: Clustering, Classification, Bayesian method, Maximum A Posteriori, Information theory, Remote sensing, Multispectral images.

Abstract: In this communication, we propose a novel approach to perform the unsupervised and non parametric clustering of *n*-D data upon a Bayesian framework. The iterative approach developed is derived from the Classification Expectation-Maximization (CEM) algorithm, in which the parametric modelling of the mixture density is replaced by a non parametric modelling using local kernels, and the posterior probabilities account for the coherence of current clusters through the measure of class-conditional entropies. Applications of this method to synthetic and real data including multispectral images are presented. The classification issues are compared with other recent unsupervised approaches, and we show that our method reaches a more reliable estimation of the number of clusters while providing slightly better rates of correct classification in average.

1 INTRODUCTION

Merging objects having similar characteristics is a very important problem in various contrasting research fields such as medicine, genetics, chemistry, computer vision, etc. Despite several decades of research in this area, the task is still difficult because of the continual improvement of the technology and the increase of the size of the data to be analyzed. Without any prior information, the grouping of objects has to be done in an unsupervised way. This processing is called *clustering*, in contrast to the *classification* which is the grouping of samples in a supervised way. The different groups are then called clusters, and they are formed of the closest individuals, according to a similarity measure. In the particular case of clustering of multispectral images, the individuals are the pixels which are grouped on their spectral information characteristics. To help the clustering of image pixels, one can also use the spatial information and the fact that two neighboring pixels are more likely to belong to the same cluster (Cariou et al., 2005).

Clustering methods can be distinguished by the similarity function used to realize the clustering (Tran et al., 2005). The similarity functions fall in two categories: the deterministic similarity functions and the probabilistic similarity functions.

In the deterministic case, a distance function is often used. This is the case of the well known *k*-means algorithm (MacQueen, 1967) which associates each object with the cluster label for which the corresponding representative object (typically the centroid of the objects in that cluster) is the closest according to the distance function used. At each iteration the centroid is computed again. This algorithm is very simple and has been improved since its initial development until recently (Huang and Ng, 2005; Laszlo and Mukherjee, 2006). For instance, a modified version which can automatically associate a weight to each feature during the clustering process has been developed, leading to a more accurate result (Huang and Ng, 2005). Genetic algorithms have also been proposed as a reliable approach of determining centers of clusters (Laszlo and Mukherjee, 2006).

The *k*-means algorithm provides a hard partitioning of the individuals which involves a lack of precision, particularly in case of overlapping between clusters. The fuzzy *c*-means (FCM) algorithm (Dunn, 1973; Bezdek, 1981) is the fuzzy equivalent of the *k*-means algorithm. Each object is potentially asso-

Bougenière G., Cariou C., Chehdi K. and Gay A. (2007).
UNSUPERVISED NON PARAMETRIC DATA CLUSTERING BY MEANS OF BAYESIAN INFERENCE AND INFORMATION THEORY.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 101-108
DOI: 10.5220/0002141301010108

ciated to different clusters, the degree of membership to each cluster being determined according to the distance function. This algorithm is known to yield better results than the *k*-means algorithm in most cases. The FCM-GK algorithm (Gustafson and Kessel, 1979) uses an adaptive distance and thus it can more efficiently fit the different cluster sizes and shapes.

In the probabilistic case, one makes use of the Bayesian paradigm, which in most cases requires a parametric modelling of class-conditional probability density functions (pdf). A parametric modelling of class-conditional pdfs is often difficult to obtain because of some non trivial cluster shapes which can occur as in multispectral and hyperspectral image processing. It is the case of the mixture modelling methods based on a statistical approach. Each cluster is modelled by a multivariate distribution f with parameters θ_c and the dataset is described by a linear combination of those conditional distributions. A maximization of the likelihood is often used to find the best parameters of each cluster. This maximization is often performed by using the iterative EM algorithm (Dempster et al., 1977). However the SEM algorithm, which is a stochastic version of the EM algorithm, can avoid some drawbacks of the EM algorithm such as its slow convergence (Celeux and Diebolt, 1987). Using one of these parameters estimation methods, a classification can be obtained for instance by associating to each individual the class label with the highest posterior probability.

In order to avoid the use of parametric (e.g. Gaussian) conditional distributions, a recent approach using a Fourier-based description of those distributions has been proposed (Zribi and Ghorbel, 2003). This approach guarantees that the conditional distributions are smooth enough to correctly model the variability of each cluster without any parametric modeling assumption, despite the fact that "negative" probabilities may artificially occur in the course of the iterations.

Another approach to clustering is density-based clustering. Its principle is to estimate the conditional densities using the data samples. The high density areas are characteristic of a cluster whereas the low density areas correspond to the boundaries. A density threshold and a volume are necessary to compute the local densities, and then the number of clusters follows automatically. However, density based clustering methods often have difficulty to handling high dimensional data because of the very odd-shaped cluster densities. In (Tran et al., 2006), a new algorithm named KNNClust dealing with this problem is presented.

We present in this paper a new clustering algorithm, based on the SEM algorithm called the Non Parametric SEM algorithm (NPSEM). It is a non parametric and unsupervised clustering algorithm which has the ability to estimate the number of clusters during the clustering process. The originality of the work is in the extension of the SEM algorithm to the estimation of non parametric conditional distributions and the weighting of the posterior probabilities by a coherence function which is based on the conditional entropy of each cluster. It allows to regularize the estimation and to stabilize the decision step result.

The second section is devoted to the presentation of our algorithm and its links to and inspirations from the SEM and the *k*-means algorithm. In the third section we present some results on different datasets. Comparisons with other state of the art algorithms are also given. Finally, a conclusion is given in the fourth section.

2 PROPOSED CLUSTERING METHOD

In this section we present the NPSEM clustering method and show its similarities with the *k*-means and SEM algorithms.

The SEM algorithm, as for the algorithm EM from which it rises, aims to maximize, in an iterative way, the likelihood of a parametric model when this model depends on incomplete data. In the case of a mixture density, the goal of the EM and SEM algorithms is to estimate the mixture parameters of K distributions:

$$f(\mathbf{X}) = \sum_{k=1}^{K} f(\mathbf{X}|\theta_k)p_k \quad , \qquad (1)$$

where $\{f(\mathbf{X}|\theta_k)\}, k = 1 \ldots K$ are the conditional distributions of parameters θ_k and p_k are the clusters prior probabilities. Although this algorithm is basically dedicated to parameter estimation, its use in classification is also possible, in particular via the Classification EM algorithm (CEM) (Celeux and Govaert, 1992; Masson and Pieczynski, 1993). The difference between the algorithms EM and SEM comes from the introduction into the latter of a stochastic step aiming to produce a current partition of the data (pseudo-sample), at each iteration, using a random sampling according to the posterior distribution computed thanks to the current parameter estimates. The CEM algorithm was recognized as a generalization of the *k*-means algorithm (Same et al., 2005). The SEM is also close to it, and particularly at two points: (i) the maximization step is mostly very similar, and consists of parameter estimation of the clusters formed; (ii) the

construction of a posterior pseudo-sample is carried out by updating the estimated parameters. However, the major difference between the two approaches is in the purely *deterministic* feature of the *k*-means and CEM algorithms: at each iteration, the label of an individual is given according to a decision criterion of minimal distance to the current cluster representative in the case of the *k*-means, or according to the MAP criterion for the CEM. This deterministic aspect has a major disadvantage, namely the convergence to a local likelihood maximum, whereas the SEM algorithm makes it possible to avoid this problem. In order to carry out a compromise between the SEM and CEM approaches, we first propose to re-examine the E (*Estimation*) step of the SEM algorithm, by computing membership posterior *pseudo*-probabilities of the individuals $\mathbf{x}_m, 1 \leq m \leq M$ to each cluster k in the following way:

$$p_\alpha(C=k|\mathbf{X}=\mathbf{x}_m) = \frac{[p_k\ f(\mathbf{X}=\mathbf{x}_m|\theta_k)]^\alpha}{\sum_{k=1}^K [p_k\ f(\mathbf{X}=\mathbf{x}_m|\theta_k)]^\alpha} \quad (2)$$

where C is the (random) cluster label of an individual, $\alpha \in [1,+\infty[$ is a parameter controlling the degree of determinism in the construction of the pseudo-sample: $\alpha = 1$ corresponds to the SEM (stochastic) algorithm, while $\alpha \to +\infty$ corresponds to the CEM (deterministic) algorithm.

In the above form, the algorithm only allows the use of parameterized conditional distributions (for example normal distributions), which can sometimes be insufficient to manage complex shaped clusters, as for instance in multispectral imagery. Consequently, we have taken into account this constraint by replacing at each iteration the parameterized conditional distributions in (2) by non parametric conditional distributions $f(\mathbf{X}|C)$, estimated from the pseudo-sample by the use of a Gaussian isotropic kernel $g_\gamma(\mathbf{x})$ with aperture γ. This aperture can be fixed automatically with respect to the dimensionality of the data, as soon as it is centered and reduced. The joint distribution, estimated by:

$$f(\mathbf{X}=\mathbf{x}_m, C=k) = \frac{\sum_{l=1}^M g_\gamma(\mathbf{x}_l-\mathbf{x}_m)\ \mathbf{1}_{C(m)=k}}{\sum_{m=1}^M \sum_{l=1}^M g_\gamma(\mathbf{x}_l-\mathbf{x}_m)} \quad , \quad (3)$$

$$\forall\ 1 \leq k \leq K,\ \forall\ 1 \leq m \leq M$$

where $C(m)$ represents cluster label affected to the individual with index m in the current iteration, makes it possible to estimate the prior probabilities p_k and the conditional distributions $f(\mathbf{X}=\mathbf{x}_m|C=k)$. Those conditional distributions cannot be directly used in the above algorithm because the mixture distribution is no longer identifiable. We then propose a further modification of the posterior distribution computation by introducing a regularizing heuristic as follows:

$$p_\alpha(C=k|\mathbf{X}=\mathbf{x}_m) = \frac{\left[p_k\ f(\mathbf{X}=\mathbf{x}_m|C=k)\ e^{-H(\mathbf{X}|k)}\right]^\alpha}{\sum_{k=1}^K \left[p_k\ f(\mathbf{X}=\mathbf{x}_m|C=k)\ e^{-H(\mathbf{X}|k)}\right]^\alpha} \quad , \quad (4)$$

where $H(\mathbf{X}|k)$ measures the conditional entropy of the current k-th cluster. Its effect on the posterior probabilities is as follows: a low entropy conditional distribution will support the membership of an individual $\mathbf{x}_m$ to the corresponding cluster if this individual strongly contributes to the coherence of this cluster. This heuristic thus tends to agglomerate the individuals according to coherent and low entropy clusters (and conditional distributions). Finally, the clustering itself is carried out by using the MAP criterion, i.e. one chooses for each individual m the cluster k which maximizes Equation (4).

An important consequence of the proposed algorithm is to allow the estimation of the number of clusters. Indeed, starting from an upper bound of the number of clusters, the algorithm reduces the number of clusters as soon as a cluster proportion is lower than a previously specified threshold of representativeness. In this case, the individuals which belong to the cluster which disappears are redistributed into the remaining clusters.

3 EXPERIMENTS AND RESULTS

In this section, we evaluate the efficiency of NPSEM on the following three real datasets for which a ground truth is available: (1) Fisher's iris dataset (150 individuals with 4 variables partitioned in 3 classes) (Fisher, 1936); (2) The wine dataset, composed of 178 individuals of 3 types (clusters) and 13 variables per individual (Aeberhard et al., 1992); (3) the Morfa dataset is a portion of a CASI hyperspectral image, composed of 747 pixels with 48 spectral radiance measurements each (equally spaced from 405nm to 947nm). This dataset was acquired in 2006 by the IGER (Institute of Grassland and Environmental Research) in Morfa Mawr, Wales, UK, during the survey of a barley crop field containing two different species which are infected or not by the mildew (4 classes). In addition we have used an additional synthetic dataset used to assess our algorithm on non Gaussian data. This 2D dataset is composed by two classes as shown by the ground truth in Figure 5-a.

Comparisons have been carried out with some other partitioning algorithms from the state of the art: *k*-means, FCM, FCM-GK, EM-GM, witch is a clustering algorithm based upon a Gaussian Mixture ob-

tained thanks to the EM algorithm, and the KNNClust, a non parametric, non supervised version of the KNN (k nearest neighbors) algorithm which can also estimate the number of clusters. This algorithm also shares with the NPSEM the property that it is not deterministic, i.e. it can provide a different clustering result at each run. The ground truth is available for each dataset which makes it possible to compute the correct classification rate obtained by the different algorithms in our experiments.

For the wine and Morfa datasets, the clustering has been performed after a data reduction consisting of keeping the first three components resulting from the principal component analysis (PCA).

A correct classification rate is computed, as well as the κ index which is a classification rate weighted to compensate the effect of chance on the clustering results. The κ index is computed as follows :

$$\kappa = (P_o - P_e)/(1 - P_e) \quad (5)$$

where P_o is the correct classification rate and

$$P_e = \frac{1}{M^2 \sum_{k=1}^{K} n_{clust_k} \sum_{k=1}^{K} n_{truth_k}}$$

with n_{clust_k} the number of individuals associated to cluster k during the clustering process and n_{truth_k} the number of samples that are in cluster k according to the ground truth. K is the number of clusters and M the number of individuals to cluster.

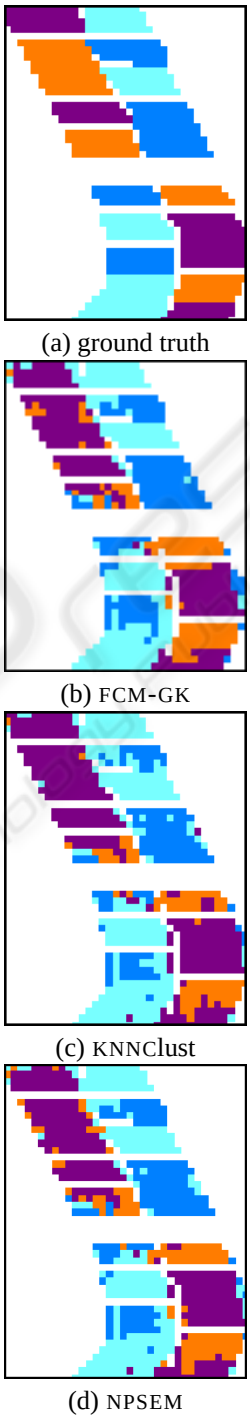
Table 1: Correct classification rates (in %) obtained by the clustering algorithms on the different datasets and their average.

	Synth	Wine	Iris	Morfa	**avg.**
k-means	62.3	88.9	78.0	64.3	**73.4**
EM-GM	61	92.7	94	72.6	**80.1**
FCM	57	97.1	88.0	73.8	**79.0**
FCM-GK	57.4	95.5	91.3	75.9	**80.0**
KNNClust	61.4	95.5	83.3	73.0	**78.3**
NPSEM	99.4	95.4	83.0	73.9	**87.9**

Table 2: Kappa index of agreement (in %).

	Synth	Wine	Iris	Morfa	**avg.**
k-means	24.6	83.6	66.0	52.4	**56.6**
EM-GM	21	88.9	91.3	63.5	**66.2**
FCM	13.9	95.8	82.0	65.1	**64.2**
FCM-GK	14.7	93.3	87.0	67.9	**65.7**
KNNClust	22.3	91.3	75.0	63.9	**63.1**
NPSEM	98.8	93.1	76.9	65.2	**83.5**

For the KNNClust algorithm, different values for the number of nearest neighbors have been tried. Also, in all experiments with the NPSEM algorithm, the upper bound for the number of clusters was fixed to $K = 5$, the Gaussian kernel aperture in Eq. (3) to $\gamma = 0.2$ and the pseudo-probabilities reinforcement coefficient in Eq. (4) to $\alpha = 1.2$. For each algo-



(a) ground truth

(b) FCM-GK

(c) KNNClust

(d) NPSEM

Figure 1: Ground truth of Morfa hyperspectral image (48 bands, 4 classes) and clustering results by the FCM-GK, the KNNClust and the NPSEM algorithm.

Table 3: Rate of correct estimation (in %) of the number of clusters for fully unsupervised methods.

	Wine	Iris	Morfa	**average**
KNNClust	95	45	65	**68.3**
NPSEM	100	80	80	**86.7**

rithm, only the best results have been kept. To compute the correct classification rate for the KNNClust and the NPSEM algorithms, which both can estimate the number of clusters, we have taken into account the only results where the correct number of clusters has been found. The correct classification rates are shown in Table 1, and the κ index results are given in Table 2. Table 3 shows the behavior of those two algorithms regarding the estimation of the number of clusters. For the other algorithms, the correct number of clusters K was given a priori.

The overall correct classification and kappa rates show better results for the NPSEM algorithm. More precisely we can see that the results obtained on real datasets are equivalent to the GK version of the FCM, the KNNClust and the NPSEM with a little advantage to FCM-GK. But on our synthetic dataset all the algorithms have failed to recover the two clusters, except our NPSEM algorithm.

Moreover, as is shown in Table 3, the NPSEM gives more reliable estimates of the number of clusters than the KNNClust. This reliability is also confirmed in the iris dataset, where the correct number of clusters has been obtained in 80% of the experiments for the NPSEM against only 45% for the KNNClust, both reaching nearly the same correct classification rate when the correct number of clusters was found. Moreover, the overall κ index is slightly better for the NPSEM method compared to the KNNClust method which reveals a better agreement of the classification with the ground truth.

Figure 1 shows typical classification results given by the FCM-GK, NPSEM and KNNClust on the Morfa dataset. On this example, the correct classification rate is 75.9% for the FCM-GK, 73.9% for the NPSEM and 73% for the KNNClust. Figure 2 depicts the same clustering results in the feature space, through a projection onto the first two principal axis resulting from the PCA of the Morfa dataset. Figure 3 shows the ground truth and clustering results which were obtained for the wine dataset, Figure 4 the ground truth and clustering results obtained for the iris dataset and Figure 5 the ground truth and clustering results obtained for our synthetic dataset.

4 CONCLUSION

In this communication, we have described the behaviour of a new clustering algorithm, the Non-Parametric Stochastic Expectation Maximisation (NPSEM) algorithm. This algorithm, inspired from the SEM algorithm and based on the use of a kernel function and an entropy based weighting, has the advantage to deal with non parametric conditional pdfs. This enables the algorithm to best fit different shapes of cluster. This feature is very important in the case of multispectral image clustering where the shape of clusters may be very different. Our algorithm can also estimate the number of clusters during the clustering process. The only parameter which is needed is an upper bound estimate of the number of clusters.

We have tested this algorithm on three different datasets, and we have compared the results with five other clustering algorithms. four of them were classical algorithms (*k*-means, EM-GM, FCM, FCM-GK) which are well known for their efficiency and/or simplicity. Their main drawback is that they are not fully unsupervised in the sense that the number of clusters must be given. The fifth one is the KNNClust algorithm which can also estimate the number of clusters automatically. This method also requires one parameter, i.e. the number of neighbors which most of the time is not easier to determine than the number of clusters.

The results of our first experiments are promising: NPSEM has shown to be more efficient in terms of estimation of the number of clusters while giving in average better classification rates than other comparable approaches on datasets with clusters of different shapes.

In further works, we plan to consider especially the case of multispectral and hyperspectral image segmentation by adding spatial information to the spectral information for each pixel. By doing so, we hope to be able to improve the clustering results, whilst keeping the advantage of reliable estimation of numbers of clusters.

ACKNOWLEDGEMENTS

This work is supported by the European Union and co-financed by the ERDF and the Regional Council of Brittany through the Interreg3B project number 190 PIMHAI.

REFERENCES

Aeberhard, S., Coomans, D., and deVel, O. (1992). The classification performance of RDA. Technical report, 92-01, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University, North Queensland, Australia.

Bezdek, J. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Cariou, C., Chehdi, K., and Nagle, A. (2005). Gravitational transform for data clustering - application to multicomponent image classification. In *Proc. IEEE ICASSP 2005*, volume 2, pages 105–108, Philadelphia, USA.

Celeux, G. and Diebolt, J. (1987). A probabilistic teacher algorithm for iterative maximum likelihood estimation. In *Classification and Related Methods of Data Analysis*, pages 617–623. Amsterdam: Elsevier, North-Holland.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. In *Computational Statistics and Data Analysis*, number 3, pages 315–332.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Gustafson, D. and Kessel, W. (1979). Fuzzy clustering with a covariance matrix. *IEEE Conference on Decision and Control*, pages 761–766.

Huang, J. and Ng, M. (2005). Automated variable weighting in *k*-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668.

Laszlo, M. and Mukherjee, S. (2006). A genetic algorithm using hyper-quadtrees for low-dimensional *k*-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):533–543.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

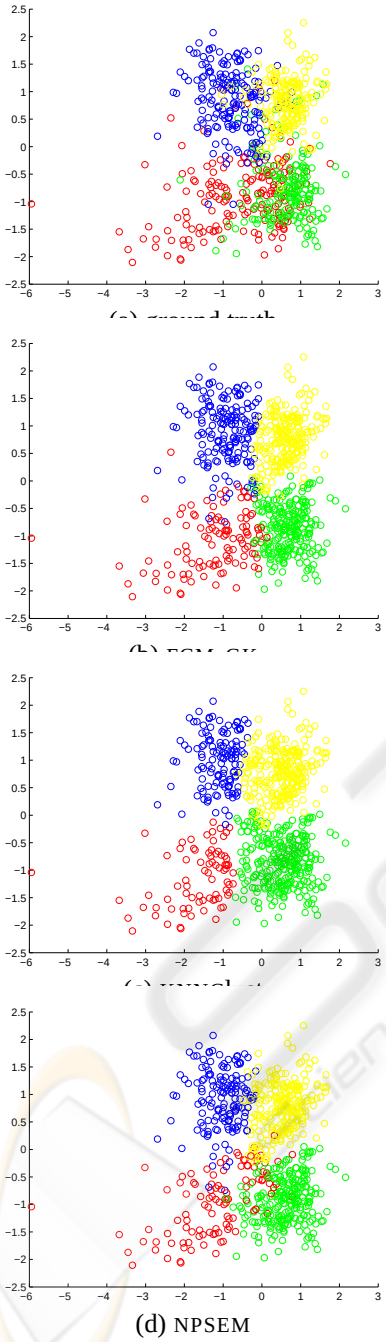
Masson, P. and Pieczynski, W. (1993). SEM algorithm and unsupervised statistical segmentation of satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 31(3):618–633.

Same, A., Govaert, G., and Ambroise, C. (2005). A mixture model-based on-line cem algorithm. In *Advances in Intelligent Data Analysis, 6th International Symposium on Data Analysis, IDA 2005, 8-10 Oct. 2005, Madrid, Spain*.

Tran, T., Wehrens, R., and Buydens, L. (2005). Clustering multispectral images: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 77:1–2.

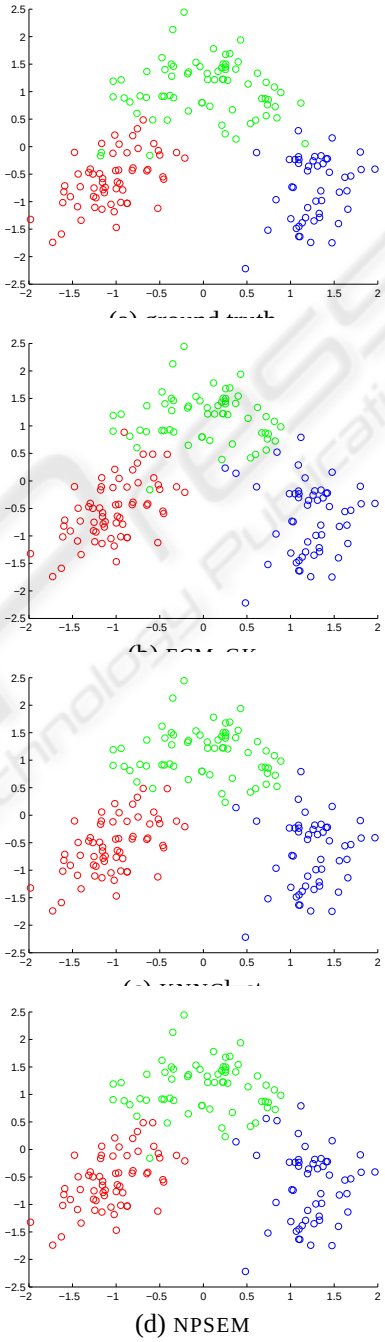
Tran, T., Wehrens, R., and Buydens, L. (2006). KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics and Data Analysis*, 51:513–525.

Zribi, M. and Ghorbel, F. (2003). An unsupervised and non-parametric bayesian classifier. *Pattern Recognition Letters*, 24(1):97 – 112.



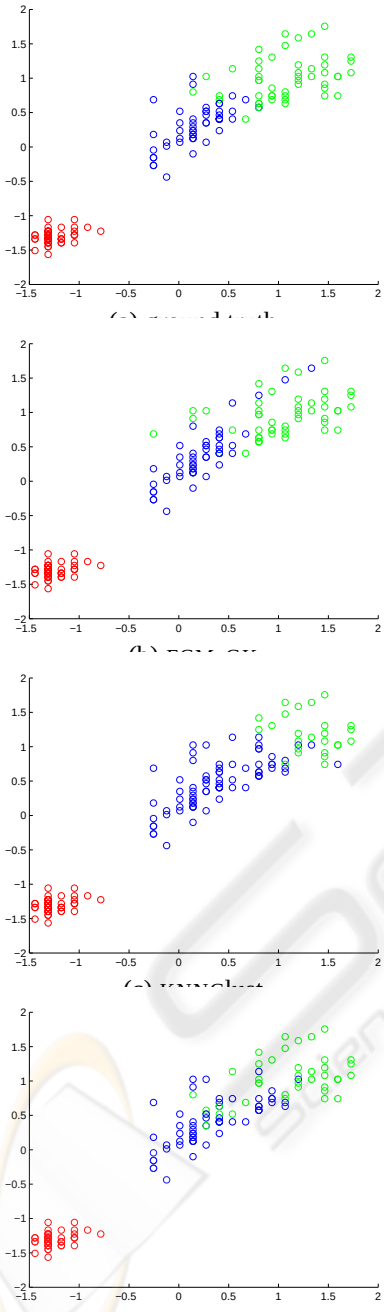
(d) NPSEM

Figure 2: Ground truth and clustering results on Morfa dataset after selection of the first three principal components. The data is projected onto the first two principal axis.



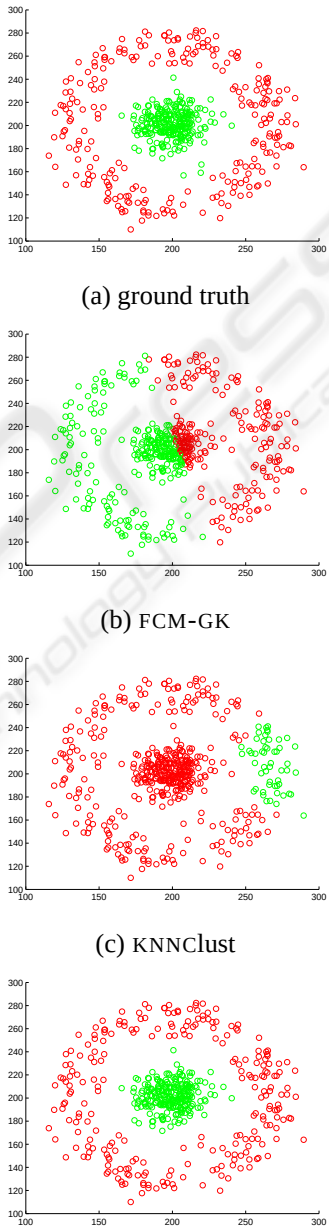
(d) NPSEM

Figure 3: Ground truth and clustering results on wine dataset after selection of the first three principal components. The data is projected onto the first two principal axis.



(d) NPSEM

Figure 4: Ground truth and clustering results on iris dataset. The data is projected onto the first two principal axis.



(a) ground truth

(b) FCM-GK

(c) KNNClust

(d) NPSEM

Figure 5: Ground truth and clustering results on a 2D synthetic dataset.