Assessing the Number of Clusters in a Mixture Model with Side-information

Edith Grall-Maes and Duc Tung Dao
ICD - LM2S - UMR 6281 CNRS - Troyes University of Technology, Troyes, France

Keywords: Clustering, Model Selection, Mixture Model, Side-information, Criteria.

Abstract: This paper deals with the selection of cluster number in a clustering problem taking into account the side-information that some points of a chunklet arise from a same cluster. An Expectation-Maximization algorithm is used to estimate the parameters of a mixture model and determine the data partition. To select the number of clusters, usual criteria are not suitable because they do not consider the side-information in the data. Thus we propose suitable criteria which are modified version of three usual criteria, the bayesian information criterion (BIC), the Akaike information criterion (AIC), and the entropy criterion (NEC). The proposed criteria are used to select the number of clusters in the case of two simulated problems and one real problem. Their performances are compared and the influence of the chunklet size is discussed.

1 INTRODUCTION

Clustering is used in many fields with an increasing interest. It aims to determine a partition rule of data such that observations in the same cluster are similar to each other. The estimation of mixture models has been proposed for quite some time as an approach for clustering. It assumes that data are from a mixture of clusters in some proportions and that the probability density function is a weighted sum of parameterized probability density functions. When the number of clusters is known, the problem consists in determining the parameters of the density functions and the proportions of each cluster. However the number of clusters is generally unknown and it has to be assessed.

In this paper, we consider the problem of data with side-information, which gives the constraint that some data originate from the same source. In particular, when different measures are realizations of the same random variable, these points belong a same chunklet. It is the case when some spatiotemporal measures are available and it is known that for example the random variable does not depend on the time; this means that all the measures originating from the same position in space belong a same cluster. An example is the temperature in a given month in different towns and in different years. The temperature is a random variable and the values for a same town and for all the years make a chunklet. The clustering problem consists in grouping similar towns, considering the values of the different years as a chunklet. It has to be noticed that the number of samples in a chunklet is not fixed. Another application is provided by time series. In a series, all points arise from the same system then they have to belong to the same cluster.

The mixture models using side-information have already been studied. For a given number of clusters, an algorithm for determining the parameters of the probability density functions and the proportions has been introduced in (Shental et al., 2003). A modified version has been proposed in (Grall-Maës, 2014) when partitioning the data is the main concern. However as in classical clustering problems, the number of clusters is generally unknown.

This paper addresses the problem of assessing the number of clusters in a mixture model for data with the constraint that some points arise from the same source. To select the number of clusters, usual criteria are not suitable because they do not consider the side-information in the data. Thus we propose suitable criteria which are based on usual criteria.

This paper is organized as follows. Section 2 describes the method for determining jointly the parameters of the mixture model and the cluster labels, with the constraint that some points arise from the same source, in the case of a known number of clusters. At the same time it introduces notations. In section 3, three criteria based on usual criteria are proposed. The Bayesian information criterion (BIC), the entropy criterion (NEC), and the Akaike information criterion

Grall-Maes, E. and Dao, D.
Assessing the Number of Clusters in a Mixture Model with Side-information.
DOI: 10.5220/0005682000410047
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 41-47
ISBN: 978-989-758-173-1

(AIC) are modified in order to be adapted to the problem of clustering with constraint. The results using two examples on simulated data and one example on real data are reported in section 4. The criteria are compared and the influence of the chunklet size is discussed. We conclude the paper in section 5.

2 CLUSTERING WITH SIDE-INFORMATION

The data we consider is a set of N observations $\mathcal{X} = \{s_n\}_{n=1..N}$. Each observation s_n is assumed to be a chunklet, which is a set of $|s_n|$ independent points that originate from the same source: $s_n = \{x_i^n\}_{i=1..|s_n|}$.

The observation set is assumed to be a sample composed of K sub-populations which are all models of the same family. Each model corresponds to a statistical law parameterized by θ_k. The latent cluster labels directly related to the parameters θ_k are described by $Z = \{z_n\}_{n=1..N}$ where $z_n = k$ means that the n^{th} realization originates from the k^{th} cluster. This means that all points x_i^n are within the k^{th} cluster due to the side-information.

Then in the case of data with side-information, the observation set and the cluster label set are respectively given by:

$$\mathcal{X} = \{s_n\}_{n=1..N} \text{ with } s_n = \{x_i^n\}_{i=1..|s_n|} \quad (1)$$

and

$$Z = \{z_n\}_{n=1..N}. \quad (2)$$

In order to compare this problem to an equivalent case without side-information, we define the observation set $\mathcal{X}'$ which is composed of N' points with $N' = \sum_{n=1}^{N} |s_n|$ and the cluster label set Z' respectively by:

$$\mathcal{X}' = \{x_i^n\}_{i=1..|s_n|,n=1..N} \quad (3)$$

and

$$Z' = \{z_i^n\}_{i=1..|s_n|,n=1..N}. \quad (4)$$

The mixture model approach to clustering (McLachlan and Basford, 1988) assumes that data are from a mixture of a number K of clusters in some proportions. The model is parameterized by $\boldsymbol{\theta}_K = \{\theta_k, \alpha_k\}_{k=1..K}$ where α_k is the probability that a sample belongs to class k, $\alpha_k = P(Z = k)$, and θ_k is the model parameter value for the class k. Then the density function of a sample s given $\boldsymbol{\theta}_K$ writes as :

$$f(s|\boldsymbol{\theta}_K) = \sum_{k=1}^{K} \alpha_k f_k(s|\theta_k)$$

where $f_k(s|\theta_k)$ is the density function of the component k.

The maximum likelihood approach to the mixture problem for a data set $\mathcal{X}$ and a given value K consists of determining $\boldsymbol{\theta}_K$ that maximizes the log-likelihood.

The log-likelihood is given by:

$$L_{\mathcal{X}}(\boldsymbol{\theta}_K) = \sum_{n=1}^{N} \log f(s_n|\boldsymbol{\theta}_K).$$

Due to the constraint given by the side-information, and the independence of points within a chunklet, we get

$$f(s_n|\boldsymbol{\theta}_K) = \sum_{k=1}^{K} \alpha_k f_k(s_n|\theta_k) = \sum_{k=1}^{K} \alpha_k \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k) \quad (5)$$

Then

$$L_{\mathcal{X}}(\boldsymbol{\theta}_K) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \alpha_k \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k) \right) \quad (6)$$

The log-likelihood in the equivalent case without side-information, for a data set $\mathcal{X}'$ and a parameter set $\boldsymbol{\theta'}_K = \{\theta'_k, \alpha'_k\}_{k=1..K}$ is:

$$\begin{aligned} L_{\mathcal{X}'}(\boldsymbol{\theta'}_K) &= \sum_{n=1}^{N} \sum_{i=1}^{|s_n|} \log f(x_i^n|\boldsymbol{\theta'}_K) \\ &= \sum_{n=1}^{N} \sum_{i=1}^{|s_n|} \log \left(\sum_{k=1}^{K} \alpha'_k f_k(x_i^n|\theta'_k) \right) \end{aligned} \quad (7)$$

A common approach for optimizing the parameters of mixture models is the expectation-maximization (EM) algorithm (Celeux and Govaert, 1992). This is an iterative method that produces a set of parameters that locally maximizes the log-likelihood of a given sample, starting from an arbitrary set of parameters.

A modified EM algorithm has been proposed in (Grall-Maës, 2014) for taking into account the side-information, as in (Shental et al., 2003), and with the aim of getting a hard partition, as in (Celeux and Govaert, 1995). It repeats an estimation step (E step), a classification step, and a maximization step (M step). The E step at iteration m requires to compute the posteriori probability $c_{nk}^{(m)}$ that the n^{th} chunklet originates from the k^{th} cluster. It is given by:

$$\begin{aligned} c_{nk}^{(m)} &= p(Z_n = k|s_n, \boldsymbol{\theta}_K^{(m-1)}) \\ &= \frac{\alpha_k^{(m-1)} \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k^{(m-1)})}{\sum_{r=1}^{K} \alpha_r^{(m-1)} \prod_{i=1}^{|s_n|} f_r(x_i^n|\theta_r^{(m-1)})} \end{aligned} \quad (8)$$

The M step consists in estimating the parameters that maximize the expected value of log-likelihood determined on the E step.

Let denote $L^*(K)$ the maximized log-likelihood for a given number K

$$L^*(K) = \max_{\boldsymbol{\theta}_K} L_{\mathcal{X}}(\boldsymbol{\theta}_K) = L_{\mathcal{X}}(\boldsymbol{\theta}_K^*) \quad (9)$$

where $\boldsymbol{\theta}_K^*$ is the optimal parameter set. One can compare $L^*(K)$ with the usual log-likelihood without side-information $L'^*(K)$ which is defined similarly for the set $\mathcal{X}'$.

The model complexity increases with K and consequently the maximized log-likelihood $L'^*(K)$ is generally an increasing function of K. Then in the classical case (without side-information) the maximized log-likelihood cannot be used as a selection criterion for choosing the number K. In the case with side-information, this is the same then $L^*(K)$ cannot be used as a selection criterion for choosing the number K.

3 CRITERIA

In order to choose the number of clusters K, a criterion for measuring the model's suitability which balances the model fit and the model complexity has to be used. Various criteria have been previously proposed for data without side-information. In this paper we modify three criteria to adapt them to data with side-information.

Generally a criterion allows to select a model in a set. The complexity of a model depends on the parameter dimension r of the model. For instance for d-dimensional K components Gaussian mixture, $r = K - 1 + dK + \frac{Kd(d+1)}{2}$.

Let $\{M_m\}_{m=1,\dots,M}$ denote the set of candidate models, where M_m corresponds to a model of dimension r_m parameterized by ϕ_m in the space Φ_m.

3.1 Criterion BIC

One of the most used information criterion is BIC (Schwarz, 1978) which is a likelihood criterion penalized by the number of parameters in the model. The idea of BIC is to select a model from a set of candidate models by maximizing the posterior probability:

$$P(M_m|\mathcal{X}) = \frac{P(\mathcal{X}|M_m)P(M_m)}{P(\mathcal{X})}$$

With hypothesis that $P(M_1) = P(M_m), \forall m = (1,\dots,M)$, the maximization of $P(M_m|\mathcal{X})$ is equivalent to the maximization of $P(\mathcal{X}|M_m)$. It can be obtained from the integration of the joint distribution :

$$\begin{aligned} P(\mathcal{X}|M_m) &= \int_{\Phi_m} P(\mathcal{X}, \phi_m|M_m)\, \mathrm{d}\phi_m \\ &= \int_{\Phi_m} P(\mathcal{X}|\phi_m, M_m) P(\phi_m|M_m)\, \mathrm{d}\phi_m \end{aligned}$$

The exact calculation of this integral can be approached by using the Laplace approximation (Lebarbier and Mary-Huard, 2006). The maximization of $P(\mathcal{X}|M_m)$ is equivalent to the maximization of $\log P(\mathcal{X}|M_m)$. Neglecting error terms, it is shown that

$$\log P(\mathcal{X}|M_m) \approx \log P(\mathcal{X}|\widehat{\phi_m}, M_m) - \frac{r_m}{2}\log(N)$$

When the model depends directly to the number of clusters, the criterion BIC for the set $\mathcal{X}'$ defined by relation 3 is given by:

$$BIC(K) \quad = \quad -2L'^*(K) + r\log(N')$$

This is the value of the criterion in the case of no side-information, for a model with r free parameters.

In order to take into account the side-information, the criterion has to be adapted. For the set $\mathcal{X}$ given by relation 1 the number of observations is N and the maximum log-likelihood $L^*(K)$, which takes into account the positive constraints is computed differently. Then the BIC criterion is given by:

$$BIC(K) \quad = \quad -2L^*(K) + r\log(N) \qquad (10)$$

Then the criterion does not depend directly on the total number of points $\sum_{n=1}^{N}|s_n|$. It depends on the number of chunklets N.

3.2 Criterion AIC

The criterion AIC proposed in (Akaike, 1974) is another largely used information criterion to select a model from a set. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth M_0.

$$d_{KL}(M_0, M_i) = \int_{-\infty}^{+\infty} P(\mathcal{X}|M_0)\log(\mathcal{X}|M_0)$$

$$-\int_{-\infty}^{+\infty} P(\mathcal{X}|M_0)\log P(\mathcal{X}|M_i)$$

It is equivalent to select the model giving the maximized value of

$$\int_{-\infty}^{+\infty} P(\mathcal{X}|M_0)P(\mathcal{X}|M_i)$$

The criterion AIC without side-information takes the form:

$$AIC(K) \quad = \quad -2L'^*(K) + 2r$$

For taking into account that some points arise from the same source, we propose to replace $L'^*(K)$ by $L^*(K)$. Then the modified AIC is given by:

$$\begin{aligned} AIC(K) &= -2L^*(K) + 2r \qquad (11) \\ &= -2\sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\alpha_k\prod_{i=1}^{|s_n|}f_k(x_i^n|\theta_k)\right) + 2r \end{aligned}$$

3.3 Criterion NEC

The normalized entropy criterion (NEC) proposed in (Celeux and Soromenho, 1996) is derived from a relation underscoring the differences between the maximum likelihood approach and the classification maximum likelihood approach to the mixture problem.

In the case of a data set without side-information $\mathcal{X}'$, the criterion is defined as :

$$NEC(K) = \frac{E'^*(K)}{L'^*(K) - L'^*(1)} \tag{12}$$

where $E'^*(K)$ denotes the entropy term which measures the overlap of the mixture components.

This criterion is expected to be minimized in order to assess the number of clusters of the mixture components. Because NEC(1) leads to an indeterminate form a new procedure has been proposed in (Biernacki et al., 1999) to retain the number K. This procedure is equivalent to setting NEC(1)=1 and to retain the value K leading to the minimal NEC value.

Considering a data set with side-information $\mathcal{X}$, we need to modify the computation of the terms of entropy and of log-likelihood. Since $\sum_{k=1}^{K} c_{nk} = 1$, we can rewrite $L_{\mathcal{X}}(\boldsymbol{\theta}_K)$ given by relation (6) as :

$$L_{\mathcal{X}}(\boldsymbol{\theta}_K) = \sum_{n=1}^{N}\sum_{k=1}^{K} c_{nk} \log \sum_{r=1}^{K} \alpha_r \prod_{i=1}^{|s_n|} f_r(x_i^n|\theta_r).$$

Using the value of c_{nk} adapted to data with side-information:

$$c_{nk} = \frac{\alpha_k \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k)}{\sum_{r=1}^{K} \alpha_r \prod_{i=1}^{|s_n|} f_r(x_i^n|\theta_r)}$$

we obtain

$$L_{\mathcal{X}}(\boldsymbol{\theta}_K) = \sum_{k=1}^{K}\sum_{n=1}^{N} \log \frac{\alpha_k \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k)}{c_{nk}}$$

which can be rewritten as

$$L_{\mathcal{X}}(\boldsymbol{\theta}_K) = C_{\mathcal{X}}(\boldsymbol{\theta}_K) + E_{\mathcal{X}}(\boldsymbol{\theta}_K)$$

with

$$C_{\mathcal{X}}(\boldsymbol{\theta}_K) = \sum_{k=1}^{K}\sum_{n=1}^{N} c_{nk} \log \left(\alpha_k \prod_{i=1}^{|s_n|} f_k(x_i^n|\theta_k) \right)$$

and

$$E_{\mathcal{X}}(\boldsymbol{\theta}_K) = -\sum_{k=1}^{K}\sum_{n=1}^{N} c_{nk} \log c_{nk}.$$

Thus we propose to use the criterion given by:

$$NEC(K) = \frac{E^*(K)}{L^*(K) - L^*(1)} \tag{13}$$

in which $E^*(K) = E_{\mathcal{X}}(\boldsymbol{\theta}_K^*)$ and $L^*(K)$ is given by (9).

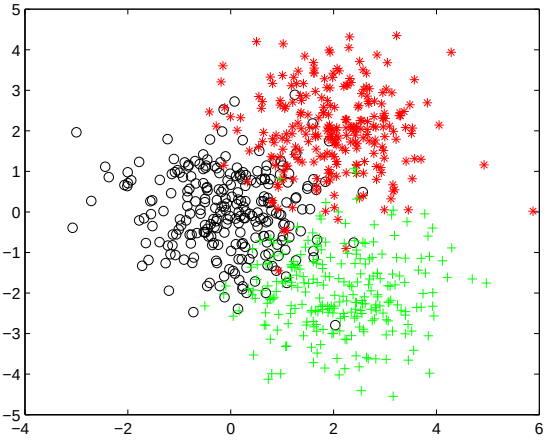

Figure 1: An example of a mixture of three Gaussian components.

4 RESULTS

The performances of the three criteria have been assessed using two simulated problems, a Gaussian mixture and a Gamma process, and one real problem using climatic data.

4.1 Gaussian Mixture

We considered a mixture of three two dimensional Gaussian components. The observation data have been generated with the following parameter values

$$m_1 = [0,0], m_2 = [2,2], m_3 = [2,-2],$$
$$\Sigma_1 = \Sigma_2 = \Sigma_3 = I,$$

$N = 150$ (50 chunklets per cluster) and $|s_n| = 5\ \forall n$. Consequently the total number of points is equal to 750. An example of data is given on figure 1. Let note that on this figure we are not able to show the chunklets and then we can only see the points. The number of clusters was determined using each of the three criteria. This experiment has been repeated 200 times in order to estimate the percent frequency of choosing a K-component mixture for each criterion.

Three other cases have been tested changing the number of chunklets and the parameter values of the Gaussian components. The value of N and $|s_n|$ have been changed to have the same total number of points but a larger number of points in each chunklet. We have used $N = 15$ (5 chunklets per cluster) and $|s_n| = 50$. Thus the "side-information" is more important than in the reference case. Then the value of m_2 and m_3 have been modified to have less separated clusters. We have used $m_2 = [1,1]$ and $m_3 = [1,-1]$.

The results for the four cases and for each of the criteria BIC, AIC and NEC are given in table 1.

Table 1: Percent frequencies of choosing K clusters.

		K	BIC	AIC	NEC
$m_1 = [0,0]$ $m_2 = [2,2]$ $m_3 = [2,-2]$	$N = 150$ $\lvert s_n \rvert = 5$	1	0	0	0
		2	0	0	40
		3	**97**	**79**	**59**
		4	3	17	1
		5	0	4	0
$m_1 = [0,0]$ $m_2 = [2,2]$ $m_3 = [2,-2]$	$N = 15$ $\lvert s_n \rvert = 50$	1	0	0	0
		2	0	0	32
		3	**99**	**95**	**68**
		4	1	5	0
		5	0	0	0
$m_1 = [0,0]$ $m_2 = [1,1]$ $m_3 = [1,-1]$	$N = 150$ $\lvert s_n \rvert = 5$	1	0	0	0
		2	0	0	**78**
		3	**96**	**78**	17
		4	4	15	0
		5	0	7	5
$m_1 = [0,0]$ $m_2 = [1,1]$ $m_3 = [1,-1]$	$N = 15$ $\lvert s_n \rvert = 50$	1	0	0	0
		2	0	0	33
		3	**97**	**94**	**67**
		4	3	6	0
		5	0	0	0

The best criterion to evaluate the number of clusters for data with side-information is the BIC; over 95% success is obtained in all cases. AIC has a slight tendency to overestimate the number of clusters, while NEC has a tendency to underestimate. This conclusion is in accordance with the results given in (Fonseca and Cardoso, 2007) obtained in case of the classical criteria for mixture-model clustering without side-information. It is also mentioned that for normal distributions mixtures, BIC performs very well. The NEC criterion shows the worst behavior with a rate success under 80%. In (Celeux and Soromenho, 1996), it is mentioned that this criterion is efficient when some cluster structure exists. Since the cluster structure is not obvious in this experiment, this criterion has low performances.

Comparing the four cases, the results are not surprising. When the side-information increases, or when the cluster overlapping decreases, it is easier to determine the right number of clusters.

4.2 Gamma Process

The Gamma process is widely used for the modeling of monotonic and gradual deterioration (Van Noortwijk, 2009). This process is defined by parameters that describe the deterioration evolution in time. The parameters depend on the properties or operational conditions of the observed system. Their values are usually estimated using data obtained on the observed systems. In the case of data coming from different systems, within an unknown finite number of operational conditions, it is required to group similar systems to estimate the parameters of the Gamma process model. Then it is necessary to determine the number of components (i.e. process models), the parameters of each component, and the component used to model each system.

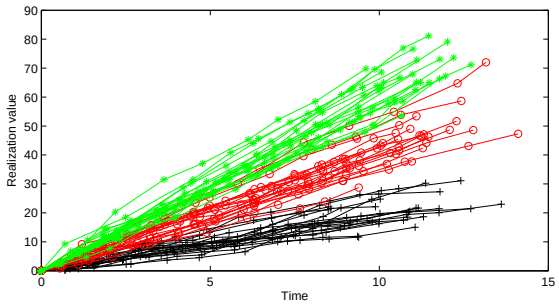

Figure 2: An example of simulated data with three Gamma processes, 10 realizations for each process, and 10 points for each realization.

An homogeneous Gamma process is parameterized by a shape and a scale parameters respectively noted a and b. Each observed increment of deterioration Δx observed for a time increment Δt, is a random variable which follows a Gamma distribution $\Gamma(a\Delta t, b)$. Instead of defining the Gamma process by the shape and the scale parameters, it is possible to define it using the mean m and variance v of the distribution $\Gamma(a,b)$. It is more convenient since m (resp. v) corresponds to the mean value (resp. the variance value) of the degradation per unit of time. m and v are given by:

$$m = \frac{a}{b} \quad v = \frac{a}{b^2}.$$

We used Monte-Carlo simulation to generate data with mean $m_k = 2k$, variance $\sigma_k^2 = 2$ for $k = 1 \ldots K$, where K is the number of clusters, i.e. the number of processes. An example of simulated data is given on figure 2, in the case of three Gamma processes (clusters), a number of systems (chunklets) equal to 10 for each process, and a number of points per chunklet equal to 10, i.e. there are 10 measures for each system. It means that $K = 3$, $\lvert s_n \rvert = 10\ \forall n$ and $N = 30$ and the total sample size is equal to 300.

Two experiments have been done: one with constant values for N and $\lvert s_n \rvert$ and varying the value of K, and one with constant value for K and varying the value of N and $\lvert s_n \rvert$.

In the first experiment, we have used 20 realizations for each process and $\lvert s_n \rvert = 10$. We considered 4 cases for the value of the cluster number: $K = 1$, 2, 3, and 4. Then the sample size was equal to $200K$ points. A Gamma mixture model was used in the clustering algorithm. The number of clusters was selected using each of the three proposed criteria (mod-

Table 2: Percent frequencies of choosing K clusters in the case of $N = 20K$ and $|s_n| = 10$ and different cluster numbers.

theoretical K	chosen K	BIC	AIC	NEC
1	**1**	**96**	**96**	**44**
	2	4	4	42
	3	0	0	8
	4	0	0	6
2	1	0	0	0
	2	**100**	**98**	**78**
	3	0	2	22
	4	0	0	0
3	1	0	0	0
	2	0	0	46
	3	**96**	**94**	**52**
	4	4	6	2
4	1	0	0	0
	2	0	0	**68**
	3	0	0	8
	4	**98**	**94**	24
	5	2	4	0
	6	0	2	0

ified BIC, AIC, and NEC). The experiment has been repeated 200 times for each value of K for estimating the percent frequency of choosing K clusters. The results are reported in table 2.

As in the case of the Gaussian mixture experiment, the best results are obtained with the criterion BIC whatever the value of K. The results with the criterion AIC are close to that with BIC. The results obtained with the criterion NEC are not good. This is due to the fact that the cluster overlapping is quite important.

In the second experiment, we have used $K = 3$. We considered 6 cases for the couple $(N, |s_n|)$. The experiment has been repeated 200 times for each case for estimating the percent frequency of choosing 3 clusters. The results are given in table 3.

As expected, for a given value of N, the efficiency of the clustering algorithm and the efficiency of the criteria are increasing with the value of $|s_n|$. For a given value of $|s_n|$, they are increasing with the value of N. However the influence of $|s_n|$ is more important than the value of N. This is due to the fact that $|s_n|$ is related to the amount of side-information, while N modifies only the size of the learning data base.

4.3 Climatic Data

The public website *donneespubliques.meteofrance.fr* provides climatic data in France. We have used the average temperature and the average rainfall for the months of January and July in 109 available towns. Then the dimension of a point is equal to 4. The mea-

Table 3: Percent frequencies of choosing 3 clusters in the case of 3 clusters and different couples $(N, |s_n|)$.

N	$\|s_n\|$	BIC	AIC	NEC
$20K$	3	68	62	20
$20K$	5	92	78	20
$20K$	20	98	92	70
$6K$	10	96	94	24
$10K$	10	98	96	52
$40K$	10	98	96	78

sure for each town is a random variable in dimension 4. We have used the data for the years 2012 to 2014. It is assumed that the random variable for a given town has the same distribution within all years. Then we can consider that 3 realizations are available for each random variable. It also means that the number of points for each town (chunklet) is equal to 3. And we assume that we can group random variables which follow the same distribution. We have used a Gaussian mixture model for the clustering algorithm.

The selected number of clusters with the criterion BIC is 5. The results are reported on figure 3. We can see different towns with climates which are rather mediterranean, maritime, mountainous, continental, and semi-continental. In a future work these results will be compared with results obtained by increasing the data dimension (adding data from other months and from other parameters such as the wind), and by increasing the numbers of points for each town (adding data from other years).

The proposed approach for the clustering with side-information allows to deal with a variable number of points within each chunklet (i.e. town). Thus it can deal with missing data for some years in some towns. It is very convenient since it occurs that sensors are off.

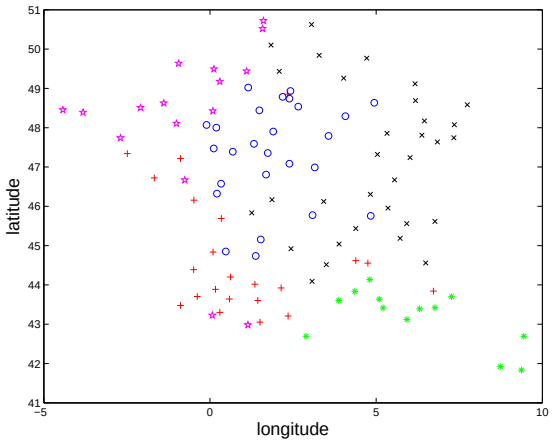

Figure 3: Classification of climates in France.

5 CONCLUSION

In this paper, criteria are proposed for assessing the number of clusters in a mixture-model clustering approach with side-information. The side-information defines constraints, grouping some points in the same cluster.

Three criteria used for assessing the number of clusters in a mixture-model clustering approach without side-information have been modified. These criteria are the Bayesian information criterion (BIC), the Akaike information criterion (AIC) and the entropy criterion (NEC). For adapting the criteria, the computation of the log-likelihood has been modified. It takes into account that some points arise from the same source. In addition the criteria depend on the number of chunklets but not on the total number of points.

Experiments have been done with simulated problems: Gaussian mixtures and Gamma processes, and with a real problem. The simulations allowed to compare the efficiency of the criteria to determine the right number of clusters. The climatic data problem has given an application example.

The side-information helps to determine the clusters mainly when the clusters overlap. Thus the criteria fitted to such situations are the most efficient. The experiments have shown the best behavior of BIC compared with the two other criteria. AIC presents a slight tendency to overestimate the correct number of clusters while NEC has an underestimating tendency. Because NEC is strongly efficient when the mixture components are well separated, its performance is quite poor for the considered experimental cases.

The influence of point number per chunklet on the performance of the proposed criteria has also been studied. The larger the chunklet size is, the better the clustering algorithm performances are, and the better the estimated number of clusters is.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parcimonious clustering models. *Pattern Recognition*, 28:781–793.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13(2):195–212.

Fonseca, J. R. and Cardoso, M. G. (2007). Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(1):155–173.

Grall-Maës, E. (2014). Spatial stochastic process clustering using a local a posteriori probability. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014)*, Reims, France.

Lebarbier, E. and Mary-Huard, T. (2006). Le critère BIC: fondements théoriques et interprétation. Research report, INRIA.

McLachlan, G. and Basford, K. (1988). Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, pages 461–464.

Shental, N., Bar-Hillel, A., Hertz, T., and Weinshall, D. (2003). Computing gaussian mixture models with EM using side-information. In *Proc. of the 20th International Conference on Machine Learning*. Citeseer.

Van Noortwijk, J. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21.