Belgian Journal of Operations Research, Statistics and Computer Science Vol. 36 (4) 96

A likelihood ratio test based on the Poisson point model in clustering problems

Tite KUBUSHISHI and Jean-Paul RASSON [1]

[1] F.U.N.D.P., Département de Mathématique,
Rempart de la Vierge, 8 B-5000 Namur, Belgium.
Tél: (32)81 72 49 33 Fax: (32) 81 72 49 14
Email: tku@math.fundp.ac.be

Summary: When testing the presence of $(k+1)$ clusters versus the presence of k clusters, Hardy (1983) considers a stationary Poisson point process in some domain $D \subset I\!R^d$ which is the union of k disjoint convex compact domains D_i $(i = 1, 2, \cdots, k)$ (k fixed). In order to derive a stopping rule for determining the 'optimal' number of clusters present in a given set of data, Hardy (1983) [5] proposed the likelihood ratio test for $H_0 : v = k$ versus $H_1 : v = k+1$. However, one can see that k (the number of components) is not a parameter of the model. The goal of this small note is to give a more accurate formulation of this test, which is based on the concept of finite mixture models (see Redner and Walker (1984) [9], Izenman and Sommer (1988)[3]).

Keywords: Poisson point process, likelihood ratio test, hypervolume criterion, finite mixture model.

1 Introduction

Let us consider that the observations $X = (X_1, X_2, \cdots, X_n)$ are a realization of a stationary Poisson point process in the domain D of $\mathbb{R}^d$.
We assume that the set D is the union of k disjoint convex compact domains D_i $(i = 1, 2, \cdots, k)$ (with k fixed).
Let $P_k = \{D_1, D_2, \cdots, D_k\}$ be a partition of the domain D into k clusters: $D_j \neq \emptyset, D_i \cap D_j = \emptyset$ for $i, j = 1, 2, \cdots, k; i \neq j$ and $\cup_{j=1}^{k} D_j = D$.
Then to measure the quality of the partition P_k, we will consider the hypervolume criterion $W(P_k)$ defined as follows:

$$W : \mathcal{P}_k \longrightarrow \mathbb{R}^+ : \mathcal{P}_k \longrightarrow W(P_k) = \sum_{i=1}^{k} m(D_i)$$

where $m(D_i)$ is the Lebesgue measure of the domain D_i;
$\mathcal{P}_k$ is the set of all partitions of the domain D in k nonempty clusters.
Hence, the clustering problem consists of finding the partition P^* which minimizes $W(P_k)$, i.e.

$$W(P^*) = \min_{P_k \in \mathcal{P}_k} \sum_{i=1}^{k} m(D_i)$$

The mapping W is called the hypervolume criterion. Thus the hypervolume criterion minimizes the sum of the Lebesgue measures of these classes.
In practice the domains $D_i, i = 1, \cdots, k$ are unknown.
Let's consider X_i as the restriction of the sample $X = (X_1, X_2, \cdots, X_n)$ over the convex domain D_i, $1 \leq i \leq k$. The convex hull of the sample X_i inside the convex domain D_i, $1 \leq i \leq k$, is denoted by $H(X_i)$. The maximum likelihood estimates of these unknown convex domains D_i are their convex hulls $H(X_i)$, see Ripley and Rasson (1977) [10]. We see that taking the homogeneous point process as a model in cluster analysis, the hypervolume criterion is reduced to

$$W : P_k \longrightarrow \mathbb{R}^+ : \forall P_k \in \mathcal{P}_k : W(P_k) = \sum_{i=1}^{k} m(H(X_i))$$

and minimizes the sum of the Lebesgue measures of the convex hulls of the classes. Finally our clustering problem consists to find the partition P^* which minimizes $W(P_k)$ (Hardy (1993) [7]), i.e.

$$W(P^*) = \min_{P_k \in \mathcal{P}_k} \sum_{i=1}^{k} m(H(X_i))$$

Now, considering the likelihood ratio test proposed by Hardy (1983)[5], Hardy (1992)[6], we have the hypotheses $H_0 : v = k$ against $H_1 : v = k + 1$.
Then using the hypervolume criterion, it becomes possible to apply this test in clustering problems to characterize genuine clusters present in a given data set.
However, one can remark in this test that v, which represents in fact the unknown number of clusters present in the data, is not a parameter of the statistical model. Then in the following sections, we want to propose a more accurate formulation of this test. A formulation which must take care of the real parameters of the underlying model and use the well-known general concept of finite mixture model, see Redner and Walker (1984) [9].

2 The finite mixture model

The finite mixture model is a natural concept for many problems occuring in applied statistics especially in pattern recognition, discriminant analysis and cluster analysis. In this section we briefly recall the basic elements of this general concept, see for instance Redner and Walker (1984) [9], Izenman and Sommer (1988)[3].
Assume that X is a random variable with a probability density function [1] $f(x;\phi)$, where $\phi \in \Phi$ is the parameter of the model. Then let consider a parametric family of finite mixture densities, i.e. a family of probability density functions of the form

$$f(x,\phi) = \sum_{i=1}^{L} p_i \, g(x,\theta_i) \tag{1}$$

where we have:

$$\Gamma = \{ \, p_1, p_2, \cdots, p_L : \sum_{j=1}^{L} p_j = 1; \;\; \text{with} \;\; p_j > 0, j = 1,2,\cdots,L \, \} \tag{2}$$

The parameters $p_1, p_2, \cdots, p_L$ are called the mixture proportions or the mixing weights and each $g(x,\theta_i); i = 1,2,\cdots,L$, parameterized by $\theta_i \in \Theta$, is the component density of the mixture. Each $g(x,\theta_i); i = 1,2,\cdots,L$ is also a probability density function, see Izenman and Sommer (1988). In such a case, we say that X has a finite mixture distribution and that that $f(x;\phi)$, defined in (1), is a finite mixture density function.
Then it becomes clear that the parameter space is

$$\Phi = \Gamma \times \Theta^L$$

and the parameters of the mixture model are

$$\phi = (p_1, p_2, \cdots, p_L, \theta_1, \theta_2, \cdots, \theta_L)$$

In fact, from (2), we conclude that the general mixture model has $(L-1)$ independent mixing weights $p_1, p_2, \cdots, p_{L-1}$. Then,

$$p_L = 1 - \sum_{j=1}^{L-1} p_j$$

In cluster analysis, many situations are modeled by a mixture model. In these problems it is supposed that the statistical population (or the sample) $X = (X_1, X_2, \cdots, X_n)$ is composed of k homogeneous classes $D_1, D_2, \cdots, D_k$ with a distribution density $f_i(x), i = 1,2,\cdots,k$ in each cluster. Hence the mixing density of X has the form

$$f(x) = p_1 f_1(x) + p_1 f_2(x) + \cdots + p_k f_k(x)$$

When determining the 'optimal' clusters present in some data, stopping rules based on the mixture model can be used. A stopping rule that has received the most attention in clustering problems is the test procedure of Wolfe (1970). Let us give the formulation of the Wolfe's test (Izenman and Sommer (1988)).
Let $k \geq 1$ be a given integer and let the parameter space Φ be partioned into two disjoint

[1] or mass function in the discrete case, however we prefer to use the term probability density function in both the continuous and discrete cases.

sets Φ_0^k and Φ_1^k, as follows. Then Wolfe's test for k components in a mixture distribution entails specifying a null hypothesis,

$$H_0^k : \phi \in \Phi_0^k$$

confirming the existence of k components,
and an alternative hypothesis,

$$H_1^k : \phi \in \Phi_1^k$$

corresponding to $k+1$ components.
Hence the following likelihood ratio can be computed:

$$\Lambda_k = \frac{\sup_{\phi \in \Phi_0^k} \mathcal{L}(x;\phi)}{\sup_{\phi \in \Phi_1^k} \mathcal{L}(x;\phi)}$$

where $\mathcal{L}(x;\phi)$ is the likelihood function for ϕ, given the sample $X = (X_1, X_2, \cdots, X_n)$. Then $-2log\ \Lambda_k$ is compared to an appropriate critical value of the χ^2 distribution with degrees of freedom equal to twice the difference in the number of parameters estimated for the two models, not including the mixing proportions. This test is repeated for a succession of increasing values of k. If H_0^r $(r \geq 1)$ is the first null hypothesis not rejected, then the number of mixture components is r and the testing procedure is terminated.
Unfortunately, the asymptotic null distribution of the Wolfe's statistic is not χ^2 (see McLachlan and Basford (1988) [8]), so the test is not statistically valid. Remember that Wolfe's test is based on the normality assumption. In this paper, this assumption is relaxed and we consider that the points in the clusters are distributed according to a homogeneous Poisson point process.

3 Application to the Poisson point model

3.1 The mixture model based on the Poisson point model

The aim of this section is to recall that the Poisson point model can be considered as a mixture model.
Let us consider $X = (X_1, X_2, \cdots, X_n)$ a realization of the stationary Poisson point process N on the domain $D \subset \mathbb{R}^d$. Then, from this model and its conditional uniformity, we have that X is a random variable uniformly distributed on the domain D, with density:

$$f(x, D) = \frac{1}{m(D)} I_D(x) \tag{3}$$

where $m(D)$ is the Lebesgue measure of the domain D and I_D is the indicator function defined as:

$$I_D(x) = \begin{cases} 1 & \text{if } x \in D \\ 0 & \text{if } x \notin D \end{cases}$$

As the domain D is formed by L disjoint convex domains $D_j, j = 1, 2, \cdots, L$, the probability density function takes the form:

$$f(x, D) = \frac{1}{\sum_{j=1}^{L} m(D_j)} \sum_{j=1}^{L} I_{D_j}$$

Now take

$$p_j = \frac{m(D_j)}{m(D)} \tag{4}$$

as mixing proportions and

$$f(x, D_j) = \frac{I_{D_j}(x)}{m(D_j)} \tag{5}$$

as the mixture components, $j = 1, 2, \cdots, L$. Then, considering (4) and (5) we see that (3) has the general form of a finite mixture model. The mixture density function has the form

$$f(x, D) = \sum_{j=1}^{L} p_j f(x, D_j)$$

With this general formulation given above, one can conclude that the Poisson point model can be interpreted as a mixture model, an approach which is very popular in clustering problems. Remember that the likelihood function of the derived mixture model becomes, see Hardy (1983)[5]:

$$\begin{aligned}
\mathcal{L}_D(X) &= \prod_{i=1}^{n} \{\sum_{j=1}^{L} p_j f(x_i, D_j)\} \\
&= \prod_{i=1}^{n} \{\sum_{j=1}^{L} \frac{m(D_j)}{m(D)} \times \frac{1}{m(D_j)} \times I_{D_j}(x_i)\} \\
&= \prod_{i=1}^{n} \{\sum_{j=1}^{L} \frac{1}{m(D)} \times I_{D_j}(x_i)\} \\
&= \prod_{i=1}^{n} [\frac{1}{m(D)} I_D(x_i)] \\
&= \frac{1}{[m(D)]^n} \prod_{i=1}^{n} I_D(x_i) \\
&= \frac{1}{[m(D)]^n} I_D(H(X))
\end{aligned}$$

where $H(X)$ is the convex hull of X. Also note that $H(X)$ is a maximum likelihood estimator and a minimal sufficient statistic of the domain D.

3.2 A likelihood ratio test

In this case, we consider again the parameter space $\Phi = \Gamma \times \Theta^L$ where Γ is defined in (2),

$$p_i = \frac{m(D_i)}{m(D)} \quad i = 1, 2, \cdots, L$$

and Θ is a subset of the Euclidean space $\mathbb{R}^d$. The parameter ϕ (which in fact represents the mixing weights) of the Poisson point model is the
vector $\phi = (p_1, p_2, \cdots, p_L, D_1, D_2, \cdots, D_L)$. With these notations, a likelihood ratio test will be formulated. We attempt to formulate criteria for the number of components in a parametric mixture model.

H_0^k: $\phi \in \Phi_0^k$; corresponding to k components in the mixture distribution
versus
H_1^k : $\phi \in \Phi_1^k$; corresponding to $k + 1$ components in the mixture distribution

Let define some useful notations:

- $P_{k+1} = \{C_1, C_2, \cdots, C_k, C_{k+1}\}$= optimal partition of the domain D into $(k+1)$ clusters;
- $P_k = \{D_1, D_2, \cdots, D_k\}$= optimal partition of the domain D into k clusters.

Then, we can compute the following likelihood ratio:

$$Q_k(x) = \frac{\sup_{\phi \in \Phi_0^k} \mathcal{L}(x;\phi)}{\sup_{\phi \in \Phi_1^k} \mathcal{L}(x;\phi)}$$

Using the notations above, the likelihood ratio becomes

$$Q_k(x) = \frac{\sup \mathcal{L}_D(x; v = k)}{\sup \mathcal{L}_D(x; v = k+1)}$$

By the hypervolume criterion, we obtain that

$$\begin{aligned} Q_k(x) &= \frac{\frac{1}{\left(\sum_{i=1}^{k} m(H(D_i))\right)^n}}{\frac{1}{\left(\sum_{j=1}^{k+1} m(H(C_j))\right)^n}} \\ &= \left(\frac{\sum_{j=1}^{k+1} m(H(C_j))}{\sum_{i=1}^{k} m(H(D_i))}\right)^n \end{aligned}$$

The last expression takes the form:

$$Q(x) = (\frac{W(P_{k+1})}{W(P_k)})^n$$

where we have

$$0 \leq Q_k(x) \leq 1$$

and the critical region becomes

$$\begin{aligned} RC &= \{x | Q_k(x) > K\} \\ &= \{\frac{W(P_{k+1})}{W(P_k)} > K\} \\ &= \{S > K\} \end{aligned}$$

where $0 \leq S \leq 1$ and K is a constant.
As we do not know the distribution of the test statistic, it becomes difficult to study the properties of the test. However, in practice, one can use a naive approach which rejects the null hypothesis H_0 when the value of the statistic

$$S = \frac{W(P_{k+1})}{W(P_k)}$$

is near 1.
The test is then repeated for a succession of increasing values of k. If H_0^r $(r \geq 1)$ is the first null hypothesis not to be rejected, then the number of mixture components is r and the testing procedure is terminated.

3.3 Remark

The reader can easily verify that this test and the gap test are closely related since

$$W(P_k) = W(P_{k+1}) + m_{k,k+1}$$

where

$$m_{k,k+1} = m\left(H(D_k \cup D_{k+1})\right) - m(H(D_k)) - m(H(D_{k+1}))$$

In fact, $m_{k,k+1}$ is the gap space between the clusters D_{k+1} and D_k. If we consider again the statistic of the proposed test, we have:

$$\begin{aligned} Q(x) &= \left(\frac{W(P_{k+1})}{W(P_{k+1}) + m_{k,k+1}}\right)^n \\ &= \left(\frac{1}{1 + \frac{m_{k,k+1}}{W(P_{k+1})}}\right)^n \end{aligned}$$

Finally, we have that:

$$\begin{aligned} S &= \frac{W(P_{k+1})}{W(P_k)} \\ &= \frac{1}{1 + \frac{m_{k,k+1}}{W(P_{k+1})}} \end{aligned}$$

and we conclude easily that $0 \leq S \leq 1$.
The elbow technique measures the gap space between the clusters D_{k+1} and D_k. The presence of a significant knee means that we have a big variation of the quantity $\frac{m_{k,k+1}}{W(P_{k+1})}$ between the disjoint clusters D_k and D_{k+1}. But again we need a threshold to decide whether the quantity $\frac{m_{k,k+1}}{W(P_{k+1})}$ is large or small enough to be unusual. This question can be answered if the distribution of the statistic

$$S = \frac{1}{1 + \frac{m_{k,k+1}}{W(P_{k+1})}}$$

is known. Unfortunately, we are now unable to determine the exact or the asymptotic distribution of this statistic.

4 Example: Govaert data

The test has been applied on many examples, see Hardy [6]. In this paper we will apply the test on the Govaert data, a data set which is very difficult to classify.

4.1 Description of the data

The Govaert data are a set of 106 points of $\mathbb{R}^2$ distributed in 7 classes. They were first used in clustering problems by Govaert [2]. Nowadays, this data set, portrayed in figure

1, is also considered as data-test for new clustering methods.

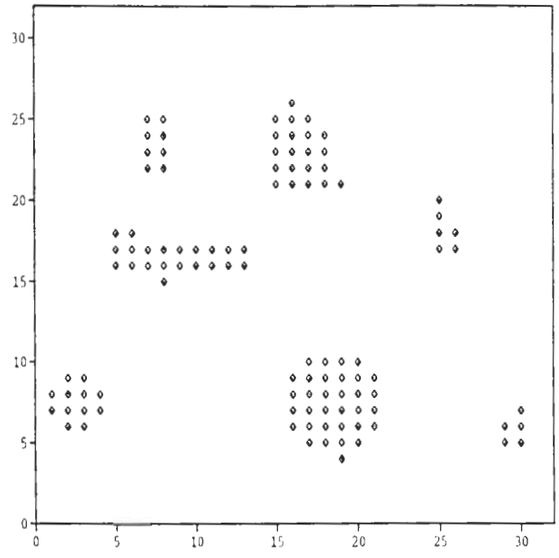

Figure 1: Govaert data.

4.2 Applications

Let's first give the value of the hypervolume associated with the optimal partitions of the Govaert data into k clusters.

k	$W(P_k)$
1	463.50
2	308
3	217
4	160.50
5	114
6	85
7	68.50
8	61.50
9	53.50
10	52

When we apply the elbow technique [2]to the data above, we obtain the following graphic:

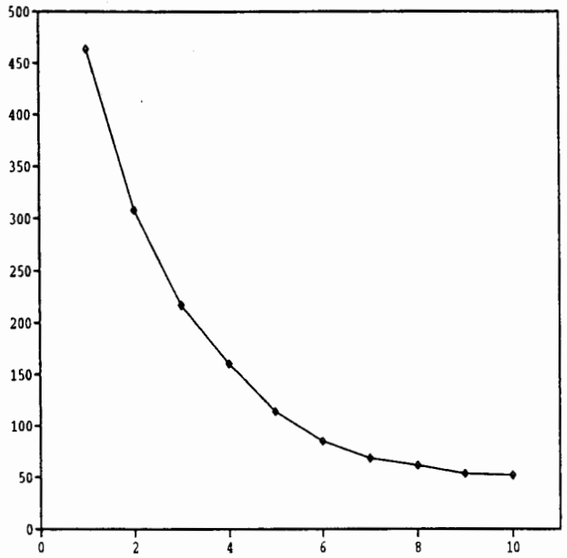

Figure 2: Results obtained when the elbow criterion is applied on the Govaert data.

Then, calculating the ratio

$$V(k) = \frac{W(P_{k+1})}{W(P_k)} \quad k = 1, 2, \cdots, 9$$

yields the following table:

k	$V(k)$
1	0.66
2	0.70
3	0.74
4	0.71
5	0.75
6	0.81
7	0.90
8	0.87
9	0.97

[2]the elbow criterion recommends the value k which yields a marked decrease of the hypervolume criterion $W(P_k)$

The reader who analyses carefully the table above can see for himself that it is not easy to decide whether the ratio $\frac{W(P_{k+1})}{W(P_k)}$ $k = 1, 2, \cdots, L$ is close to 1 or not, in other words when $k+1$ rather k classes are present in the data. In this example, the situation is very gloomy. However, based on the a priori knowledge of the Govaert data, we can state arbitrary that $V(k) = 0.81$ is near 1 and say that the data contain 6 optimal clusters. A conclusion which does not reveal the reality!!!

5 Conclusions

The test of the number of components in finite mixture models considered in this paper provides us with a very useful tool in classification problems. It seems to be original since it is based on the homogeneous Poisson point process model. However, the test can only be performed in practice if we can obtain the distribution of the test statistic; which is a problem for further research.

Acknowledgements

The authors thank Dr Michel Hermans who first pointed out that the formulation of the test was not correct, and after some discussions invited us to find an appropriate formulation of this test. We also thank the referee for insightful comments.

References

[1] Cox, D.R. and Isham, V. (1980): *Point processes.* Chapman and Hall, London.

[2] Govaert, G. (1975): *Classification automatique et distances adaptatives.* Thèse de 3ième cycle, Université Paris *VI*, Paris, France.

[3] Izenman, A.J. and Sommer, C.J. (1988): Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, **83**, 404, 941–953.

[4] Jain, A.K. and Dubes, R.C. (1988): *Algorithms for clustering data.* Prentice–Hall, Englewood Cliffs, New Jersey.

[5] Hardy, A. (1983): *Statistique et classification automatique, un modèle, un nouveau critère, des algorithmes, des applications.* Dissertation doctorale, FUNDP, Namur, Belgium.

[6] Hardy, A. (1992): On tests concerning the existence of a classification. *Jorbel*, **31**, 3–4, 111–126.

[7] Hardy, A. (1993): An examination of procedures for determining the number of clusters in a data set. In *New approaches in classification and data analysis*, editors Diday, E. et al., Springer Verlag, 178–185.

[8] Mclachlan, G.J. and Basford, K.E. (1988): *Mixture models: inference and applications to clustering.* Marcel Dekker, New York.

[9] Redner, R. A. and Walker, H.F. (1984): Mixture densities, maximum likelihood and the EM algorithm. *SIAM review* , **26**, 2, 195–239.

[10] Ripley, B.D. and Rasson, J-P. (1977): Finding the edge of a Poisson forest. *Journal of Applied probability*, 14, 483–491.