Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

A MEASUREMENT OF OVERLAP RATE BETWEEN GAUSSIANCOMPONENTS

HAO-JUN SUN[1], MEI SUN[1], SHENG-RUI WANG[2]

[1]Faculty of Mathematics and Computer Science, Hebei University, Baoding 071002, China
[2]D′epartment d'informatique, Universit′e de Sherbrooke, Sherbrooke, QC, Canada J1K 2R1
E-MAIL: haojun.sun, mei.sun@mail.hbu.edu.cn, shengrui.wang@usherbrooke.ca

Abstract:

Overlapping clusters often appear in cluster analysis in the data mining. However, the phenomenon of cluster overlapping is still not mathematically well characterized, especially in multivariate cases. In this paper, we are interested in the overlap phenomenon between Gaussian clusters, since the Gaussian mixture is a fundamental data distribution model suitable for many clustering algorithms. We introduce the novel concept of the ridge curve and establish a theory on the degree of overlap between two components. Based on this theory, we develop an algorithm for calculating the overlap rate. We investigate factors that affect the value of the overlap rate, and show how the theory can be used to generate "truthed data" as well as to measure the overlap rate between a given pair of clusters or components in a mixture.

Keywords:

Mixture model; Ridge curve; Overlap rate; Cluster analysis

1. Introduction

The Gaussian mixture plays an important role in cluster analysis. More recently, it has been recognized that these models can provide a principled statistical approach to practical questions that arise in applying clustering methods. The simplified Gaussian mixture in which each component has a spherical shape is frequently used as the data model for popular algorithms such as K-Means and the Fuzzy C-means, though it is not the only data model suitable to these algorithms. All these clustering algorithms or their improved versions are widely used in practice [1-3]. However, their performance often depends on whether the data set contains well separated clusters, or in other words, whether and how the components of the Gaussian mixture overlap each other.

Given a data set satisfying the distribution of a mixture of Gaussians, the degree of overlap between components affects the number of clusters "perceived" by a human operator or detected by a clustering algorithm. The component overlapping phenomenon is illustrated in Fig. 1. Fig. 1-a to 1-c show the 1-D case, with two components that are (almost) non-overlapping, partially overlapping and strongly overlapping. Fig. 1-d to 1-f show their counterparts in the 2-D case. Non-overlapping clusters are relatively easy to be discovered by clustering algorithms. Partially overlapped clusters are more difficult to separate and strongly overlapping clusters are in general very difficult to separate whichever clustering algorithm is used. We are interested in establishing analytically the relationship between the degree of overlap and the parameters of each component. This relationship will have an important impact in many real applications. For example, generating truthed data sets with a prescribed degree of overlap between clusters provides a way of evaluating the capacity of existing clustering algorithms to identify overlapping clusters.

The main contributions of this paper are creating a theoretical framework for component overlap in a mixture, proposing a practical algorithm for measuring the degree of overlap between two components, and describing its application to hierarchical clustering. We do not make any assumption regarding the covariance structure of each component. For the sake of simplicity, the theory will be introduced in the 2-D case. All of the results hold in the multidimensional case and the main equations for the multidimensional case are discussed. The theory is based on a novel concept, that of the "ridge curve". A series of theorems will be established to show the main characteristics of component overlapping. This allows us to give a feasible definition of the overlap rate, as well as developing algorithms for measuring it and generating truthed data sets.

1-4244-0973-X/07/$25.00 ©2007 IEEE

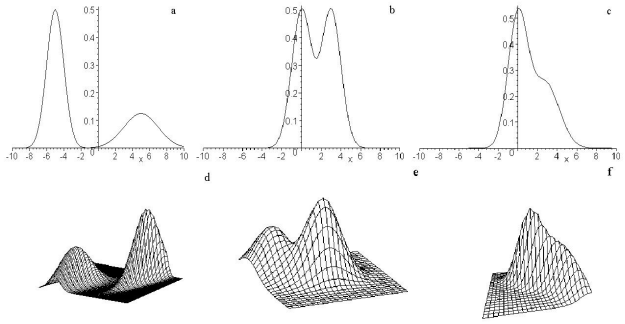

Figure 1. Two components that are non-, partially and strongly overlapping, in the 1-D and 2-D cases

This paper is organized as follows. In the next section, we give a theoretical framework for measuring the similarity between two components in a Gaussian mixture, introduce the concept of the "ridge curve", and prove a series of theorems to establish the theoretical framework for describing the phenomenon of overlapping components. In Section 3, we define the overlap rate between two components in a mixture and design an algorithm for calculating this overlap rate. In Section 4, we consider the factors affecting the value of the overlap rate in order to generate truthed data sets with prescribed overlap rates. At last, we conclude our paper with a discussion of some extensions of the research.

2. Theoretical framework

2.1. Mixture models

Mixture models satisfy some intuitive definitions of cluster structure. Two key properties of a cluster are internal cohesion, which requires that entities within the same cluster should be similar to each other, and external isolation, which requires that entities in one cluster should be separated from entities in another cluster by fairly empty areas of space [4]. Internal cohesion is an inherent property of mixture models, for a given cluster, data are generated from the same distribution. External isolation concerns the degree of overlap between the components of the mixture model [5].

A set of n entities forming a k-mode two-way array can be presented as $X = \{X_1, ..., X_n\}$, where X_i is a vector of dimension d. In the finite mixture models dealt with here, each X_i can be viewed as arising from a mixture of k Gaussian distributions and the probabilistic density function (pdf) is given (in the d-dimensional data space) by:

$$\begin{cases} p(X) = \sum_{i=1}^{k} \alpha_i G_i(X, \mu_i, \Sigma_i) \\ x \in R^d \end{cases} \quad (1)$$

with the restrictions $\alpha_i > 0$ for $i = 1, ..., k$ and $\sum_{i=1}^{k} \alpha_i = 1$.

(α_i, Σ_i) denotes, respectively, the mean and the covariance matrix for the i^{th} distribution G_i. G_i is the i^{th} component, given by:

$$G_i(X, \mu_i, \Sigma_i) = \frac{\exp(-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i))}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \quad (2)$$

In this paper, our investigation of the overlap phenomenon follows a geometrical approach which is intuitive and practical in real applications. Because of complex non-linear equations involved and of the real need in practice, the current theoretical study is restrained to the overlap of two components. The overlap phenomenon of three or more components could be dealt with using our theory in a pair-wise way. Without loss of generality, we assume, in the development of our theory, that k = 2 in Eq. 1. If the two components are (almost) non-overlapping or only partially overlapping, then the *pdf* of the mixture has two divided local peaks. In this case, the data arising from the mixture model are viewed as two clusters (see Fig. 1 a, b, d, e). On the other hand, if the *pdf* has only one peak, then the data are viewed as having two strongly overlapping components. Visually, it is very difficult to distinguish between the two components in this case (see Fig. 1-c and f).

2.2. Related work

Most of the methods used to measure the degree of overlap between two components are probabilistic in nature. A few others are geometric. Probabilistic methods are based on a hypothesis that the data are drawn from one of several probabilistic distributions. The mathematical model is a mixture of these probabilistic distributions. The most commonly used model is the Gaussian mixture. The classification error rate, $\int min\{\alpha_1 G_1, \alpha_2 G_2\} dX$ R, as well as various distances such as the Mahalanobis distance [6], [7], and an extension of it, the Bhattacharyya distance [8], are used as measures of the similarity between two Gaussian distributions [9]. Because of the difficulty of computing the classification error rate, it is often replaced by its upper bounds (for example, the Bhattacharyya bound [8] in practical applications.

Distance measures fail to characterize component overlap for various reasons. For instance, the distance between the centers of components, $|\mu_1-\mu_2|$, ignores the covariance matrix and other probability parameters. The Mahalanobis distance measures the similarity between two normal distributions, assuming that they have the same covariance matrix. The Bhattacharyya distance takes covariance matrices into account, but ignores the influence of the coefficients of the components. However, its main problem is that it does take account explicitly of the geometrical properties of the *pdf*.

Geometric methods consider the geometrical properties of the *pdf* of a mixture. The typical approach is based on comparing between the local minimum and the local maximum of *pdf* or their locations. In [10], Tabbone studied the presence of a false edge between the steps of a staircase edge. He treated this as a mixture of two Gaussians with equal covariance, Σ, in the one-dimensional case. There is a false edge (the two Gaussians strongly overlap) if and only if $2\sigma < |\mu_1 - \mu_2|$. In [5], Aitnouri et al. extended the results in [10] to a mixture of Gaussians with non-equal covariance. The extension remained restricted to the one-dimensional case. In [5], the overlap rate was defined as the ratio of the height of the minimum between the two component centers (if such a minimum exists) to the height of the lower maximum also situated between the component centers. This definition does not have a natural extension to the multi-dimensional case since there is no local minimum between the two component centers.

In what follows, we try to characterize this phenomenon as a function of the mixture parameters. In particular, we derive an efficient procedure for verifying whether the two components strongly overlap and to compute an overlap rate when they partially overlap. For the sake of simplicity, we establish the theory for the two-dimensional case ($d=2$). All of the theorems are also valid in the multi-dimensional case with an appropriate definition of the ridge curve, the key concept of this theory, which will be discussed in the following subsection. There, we will describe that the peaks of the *pdf* in R^2 can be found by a search procedure which follows a curve linking the centers of the two components.

2.3. Peaks of the *pdf* and ridge curve

If the *pdf* of a mixture has two divided local peaks, then the overlap rate between the two components depends on the geometrical properties of the two peaks. In this case, there is a saddle between the two peaks, and the value of the saddle is less than the value of either peak.

We propose to use the ratio of the saddle to the lower peak to measure the degree of overlap between the two components. This idea relates to the results described in [5]. The challenge in the multi-dimensional case is how to compute the overlap rate.

As we know, the peaks of the *pdf* satisfy the following system of stationary equations:

$$\begin{cases} \dfrac{\partial p}{\partial x_1} = A_{x_1}\alpha_1 G_1 + B_{x_1}\alpha_2 G_2 \quad (I) \\ \dfrac{\partial p}{\partial x_2} = A_{x_2}\alpha_1 G_1 + B_{x_2}\alpha_2 G_2 \quad (II) \end{cases} \tag{3}$$

where

$$\begin{pmatrix} A_{x_1} \\ A_{x_2} \end{pmatrix} = \nabla \left\| X-\mu_1 \right\|^2_{\Sigma_1^{-1}} = -\Sigma_1^{-1}(X-\mu_1)$$
$$\begin{pmatrix} B_{x_1} \\ B_{x_2} \end{pmatrix} = \nabla \left\| X-\mu_2 \right\|^2_{\Sigma_2^{-1}} = -\Sigma_2^{-1}(X-\mu_2) \tag{4}$$

Because of the involvement of G_1 and G_2 in Eq. 3, this system does not have a closed-form solution. A naive numerical solution would imply searching a whole region of R^2 (R^d in the general case). The following theorems illustrate that the search procedure can be restricted to a curve.

Now we introduce a new concept, that of the ridge curve (RC) of a mixture model of two Gaussian distributions, as follows:

Definition 1. Given a mixture of two Gaussians (Eq. 1), the quadratic curve

$$A_{x_1}B_{x_2} - A_{x_2}B_{x_1} = 0 \tag{5}$$

is called the ridge curve (RC) of the mixture.

Theorem 1 The ridge curve is a hyperbola or a line.

Theorem 2 The means of the two components and the stationary points (peak and saddle points) of the pdf are on one branch of the ridge curve.

These properties guarantee that the search of the peak(s) and saddle point (if any) can be done along the ridge curve instead of a region in the feature space. Moreover, the following theorem confirms that the search can simply be done between the centers of the two components.

Theorem 3 The stationary points of the *pdf* fall on the segment between the two means of the components of the ridge curve.

Based on the result of *Theorem 3*, to determine the stationary points of Eq. 1, we need only search the segment of the RC, $A_{x_1}B_{x_2} - A_{x_2}B_{x_1} = 0$, between the two means. The search for the stationary points of a Gaussian mixture is thus reduced to a linear search procedure.

3. Degree of overlap

In this section, we will give the definition of the overlap rate (OLR) between two components of a mixture and design an algorithm to compute the OLR.

3.1. Definition of the overlap rate

We propose a geometric method for determining the overlap rate between two Gaussians. In general, a definition for the overlap rate between two Gaussian components implements the following principle: 1) the value of the overlap rate lies within a normalized interval such as [0, 1]; 2) the overlap rate tends to decrease (→ 0) as the two components become more widely separated; 3) the overlap rate increases (→1) as the two components become more strongly overlapped. As mentioned previously, the following definition of overlap rate relies on the ridge curve concept developed in the previous section.

Definition 2. The overlap rate of two Gaussian components in a mixture is defined by:

$$OLR(G_1, G_2) = \begin{cases} 1 & pdf \text{ has one peak} \\ \dfrac{p(X_{saddle})}{p(X_{lower\_peak})} & pdf \text{ has two peaks} \end{cases} \quad (6)$$

where X_{Saddle} is the saddle point of $p(X)$ which is on the ridge curve and $X_{Sub\_Max} = arg(Sub\_Max\ p(X))$ is the lower peak point of p(X). Fig. 2 illustrates different elements of this definition and provides an intuitive interpretation of the overlap rate.

The OLR describes the degree of overlap between two components (clusters). It is not the percentage of the data falling in the "overlapping region", nor is it linearly dependent on this percentage. The relation between OLR and parameters in the mixture will be further investigated in Section IV. In approximately speaking, as in the writing of this paper, we say that the two components (clusters) are (visually) well separated if the value of OLR is less than 0.6; if the OLR belongs to (0.6, 0.8], they are partially overlapping; and if the OLR is greater than 0.8, they are strongly overlapping (see also Fig. 6-10). These terms are utilized only for facilitating discussions.

3.2. Algorithm for calculating OLR

To compute *OLR*, the stationary points of $p(X)$ need to be determined. In fact, these points are the solution of the following equations:

$$\begin{cases} A_{x_1} G_1 + B_{x_1} G_2 = 0 \\ A_{x_1} B_{x_2} - A_{x_2} B_{x_1} = 0 \end{cases} \quad (7)$$

Expressing x_2 in terms of x_1 from (II) and substituting

x_2 in (I), we obtain the following equation:

$$M_1(x_1) e^{N_1(x_1)} + M_2(x_1) e^{N_2(x_1)} = 0 \quad (8)$$

where $M_1(x_1)$, $M_2(x_1)$, $N_1(x_1)$, $N_2(x_1)$ are rational functions of x_1. This equation does not have a closed-form solution. For this reason, we need a numerical algorithm for finding the stationary points of $p(X)$. The main idea of the algorithm is to search, based on *Theorem 3*, the segment of the ridge curve $A_{x_1} B_{x_2} - A_{x_2} B_{x_1} = 0$ between the means of the two components. A local maximum point is a peak of $p(X)$ and the minimum point, if it exists, is a saddle point. *Algorithm COLR* below computes the overlap rate of two mixture components. Using this algorithm, we can estimate the overlap rate of any two clusters in a given set of data by first estimating the mean, covariance matrix and prior probability of each cluster.

Algorithm COLR (for computing *OLR* of the mixture in Eq. 1)

1) Input the parameters of two distributions (μ_1, μ_2, Σ_1, Σ_2, α_1, α_2);
2) Compute the ridge curve $A_{x_1} B_{x_2} - A_{x_2} B_{x_1} = 0$;
3) Numerically search for the maximum and minimum values of $p(X)$ on the ridge curve between μ_1 and μ_2;
4) Compute *OLR* of the two components by Eq. 6

3.3. Extension to the high-dimensional case

The theory on component overlap presented above can naturally be extended to the high-dimensional case. In the d-dimensional case (d > 2), the two main concepts of the theory, ridge curve and overlap rate, can be described as the same form. Based on the concept of the ridge curve in the d-dimensional case, the overlap rate between G_1 and G_2, $OLR(G_1,G_2)$, can be defined using the same formula as in the 2-dimensional case (Eq. 6). The computation of the OLR can be carried out by *Algorithm COLR* .

4. Generating truthed data sets with prescribed OLRs

In this section, we propose a general framework for generating truthed data sets. This is the inverse operation to computing the OLR: given an OLR value, olr, what are the values of the parameters a mixture must have so that the maximum of the OLR between each pair of components is OLR. We restrict our discussion to the case in which there are two components in a mixture. The mixture with

multiple overlapped components can be made up of various separate pairs of components. The aim of this section is to provide some ideas on how to modify the parameters in the mixture to obtain data sets with different values of the OLR. A more systematic use of overlap theory in generating valid data sets and evaluating the performance of clustering techniques is reported in another paper [11].

4.1. Factors affecting the OLR

The problem would become much easier if the OLR could be expressed as an analytical function of the parameters of the mixture. In the 1-dimensional case, Aitnouri et al [5] give an approximate solution to the problem based on a piece-wise linear approximation to the Gaussian components. However, this approach cannot be effectively extended to the multidimensional case. For this reason, we consider the factors that affect the OLR. The aim of the following subsections is to show the influence of different parameters of the mixture on the OLR. We try to show some general trends in the dependence of the OLR on different parameters.

Let C_1 and C_2 be two Gaussian components of a mixture model. Without loss of generality, we suppose that the initial parameters of the two components are given by:

$$\begin{cases} G_1 \quad \alpha_1 = 0.5 \quad \mu_1 = (0,0)^T \qquad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ G_2 \quad \alpha_2 = 0.5 \quad \mu_2 = (3,0)^T \quad \Sigma_2 = \begin{pmatrix} 2.17 & 1.82 \\ 1.82 & 2.17 \end{pmatrix} \end{cases} \tag{9}$$

In what follows, we will show the evolution of the OLR when one of the parameters is varied.

1) Effects of Varying the Mixing Coefficient: Fig. 3 shows the effects of varying the mixing coefficient $\alpha_1 : 0 \rightarrow 1$. In this case, the OLR is a piece-wise function of α_1. Experimental results show that the OLR reaches its minimum (represented by r_{min}) when $\alpha_1 = 0.46$. The value of r_{min} depends on the components. The fact that the OLR reaches 1 at both ends means that the two components overlap completely when the coefficient crosses a threshold. The difference between the two covariance matrices is the main reason for the asymmetry of the OLR curve. Varying the value of the mixing coefficient is an easy way to obtain a mixture with an OLR value varying between r_{min} and 1.

2) Effects of Varying the Distance Between the Two Means: Fig. 3 shows also the relationship between the OLR and the distance between the two means. For this experiment, we varied μ_2 from $(0, 0)^T$ to $(8, 0)^T$. When the two means are very close to each other (distance < 2.16), the OLR is 1. The OLR decreases rapidly to zero once it falls below 1 (once two partially overlapped components appear). We notice that modifying the distance between the two means leads to values of the OLR varying in [0, 1]. So we conclude that, for a given OLR ($0 \leq OLR \leq 1$), we can find a value of the distance between the two means to match the OLR. The distance between the two means is the most convenient parameter to control in order to obtain all possible values of the OLR.

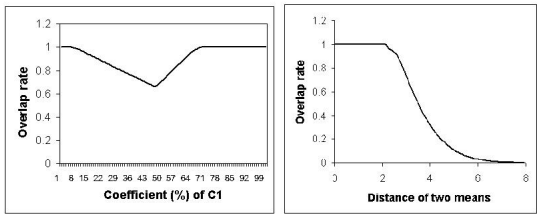

Figure 3. The effect of varying the mixing coefficient and the distance between the two means on the OLR

3) Effects of Varying the Covariance Matrix: Describing the relationship between the OLR and the difference between the two covariance structures is more complex. For simply, we consider the effects of varying the angle between the main axes of the two contours. For this factor (see Fig. 4), the angle θ varies from $-\pi/2$ to $\pi/2$. It is easily understood that OLR reaches its maximum (not necessarily 1; represented by r_{max}) around $\theta = 0$ and its minimum (not necessarily 0; represented by r_{min}) around $\theta = \pm\pi/2$, since the main axis of the ellipse is (almost) aligned with the center of the circle (or vertical). If the given OLR is in $[r_{min}, r_{max}]$, we can find two angle values (positive and negative) to match the OLR.

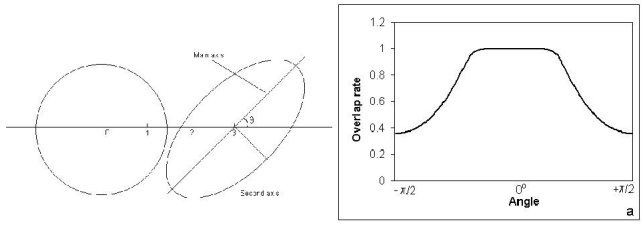

Figure 4. Contours of the two components and the effects of varying the two covariance matrices on the OLR

4.2. Examples of generating truthed data sets

In this subsection, we will show some truthed data sets generated to match certain overlap rates by modifying the different parameters mentioned above. We chose 0.99, 0.80, 0.60 and 0.40 as four OLR values.

The initial components are given by Eq. 9. The first example involves choosing the coefficients of the mixture described in Section 4.1 to match the OLRs. Fig. 5 shows

three data sets. Their OLRs are 0.99, 0.80 and 0.67. In this case, 0.67 is the minimum reachable OLR, i.e. r_{min}, attainable when only mixing coefficients are modified. Fig. 6 and 7 show the data sets generated by adjusting each of the other parameters.

The overlap rate is an integrative concept. It depends on all the parameters of the mixture and the combinations of these parameters. To generate controlled data sets with overlapped clusters, one can select many approaches for modifying the parameters to obtain various types of clusters, such as spheres and ellipses of different sizes. In another work [11], we use the technique developed in this paper to generate groups of data sets with set OLRs for use in evaluating the efficiencies of clustering validity indices.

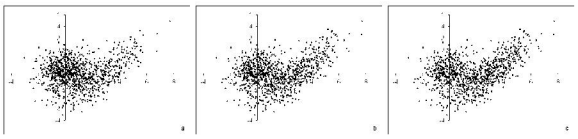

Figure 5. Generated data sets based on the coefficients: a) OLR is 0.99,coefficients are 0.67 and 0.33; b) OLR is 0.8, coefficients are 0.54 and 0.46; c) OLR is 0.67, coefficients are 0.46 and 0.54

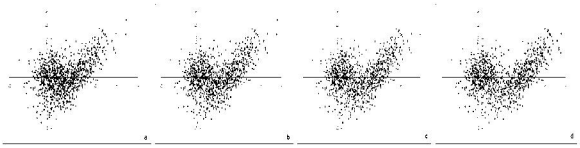

Figure 6. Generated data sets based on the distance between the two means: a) OLR is 0.99, distance is 2.16; b) OLR is 0.8, distance is 2.84; c) OLR is 0.60, the distance between two means is 3.28; d) OLR is 0.40, distance is 3.76.

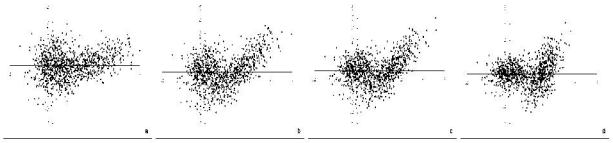

Figure 7. Generated data sets based on the angle between two main axes: a)OLR is 0.99, angle is 14.9o; b) OLR is 0.80, angle is 40.7o; c) OLR is 0.60, angle is 52.1o; d) OLR is 0.40, angle is 73.3o.

5. Discussion and conclusion

The main contribution of this paper is establishment of a theory to explain the phenomenon of overlap in mixtures of Gaussians. The significance of this theory is that it provides a mathematically rigorous way to explain the overlap phenomenon and a computationally feasible way to calculate the degrees of overlap between the components of a mixture. It provides a foundation for generating overlapped data sets for use in validating clustering and classification algorithms.

We are currently pursuing our research in several directions. These include investigations into whether the theory can be extended to other mixtures, how to adapt the theory to deal with overlapping phenomena in a sub-space only, and how to use the theory to develop (hierarchical) clustering algorithms capable of extracting clusters of arbitrary shape.

Acknowledgements

This work has been supported by Science Found from Department of Education, Hebei Province, China, and Found from Department of Human Resource, Hebei Province, China.

References

[1] V. Ramos and F. Muge, "Map segmentation by colour cube genetic k-mean clustering," in *ECDL 2000*, vol. 1923, (Lisbon, Portugal), pp. 319–323, 2000.

[2] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "Incremental genetic k-means algorithm and its application in gene expression data analysis,"*BMC Bioinformatics*, vol. 5, p. No. 172, Oct. 2004.

[3] C. Fraley, "Algorithm for model-based Gaussian hierachical clustering,"*SIAM J. Sci. Comput*, vol. 20, no. 1, pp. 270–281, 1998.

[4] G. Milligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrica*, vol. 45, No. 3, pp. 325–342, 1980.

[5] E. Aitnouri, F. Dubeau, S. Wang, and D. Ziou, "Controlling mixture component overlap for clustering algorithms evaluation," *J. of Pattern Recog. and Image Analysis*, vol. 12, no. 4, pp. 331–346, 2002.

[6] N. Day, "Estimating the components of a mixture of two normal distributions," *Biometrics*, vol. 56, pp. 463–474, 1969.

[7] G. McLachlan and K. Basford, "Mixture Models". Marcel Dekker, Inc. N.J., 1988.

[8] K. Fukunaga, "Introduction to Statistical Pattern Recognition" (2nd edition).Academic-Press, 1990.

[9] H. Chanl, A. Chung, A. N. S. Yu, and W. Wells, "Clustering web content for efficient replication," in *2003 Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. II, 2003.

[10] S. Tabbone, "Edge Detection, Subpixel and Junctions Using Multiple Scales." Ph.D. thesis, Institut National Polytechnique de Lorraine, France, In French, 1994.

[11] M. Bouguessa, S. Wang, and H. Sun, "An objective approach to cluster validation," *Submitted to Pattern Recognition Letters*, 2005.