OUTLIER DETECTION IN EXPERIMENTAL DATA USING A MODIFIED EXPECTATION MAXIMIZATION ALGORITHM

Vedran Novoselac* and Zlatko Pavić

Mechanical Engineering Faculty in SlavonskiBrod, J. J. Strossmayer University of Osijek, Croatia
Corresponding author e-mail: vnovosel@sfsb.hr

Abstract

The paper studies the problem of clustering data sets with the E-M (Expectation Maximization) algorithm. Within the E-M algorithm is implemented procedure omitting data that have a low probability of belonging to a Gaussian mixture model components. For this purpose, the threshold is determined by the rejection of data, which are considered as outliers. For this procedure is used Mahalanobis distance of the observed data to the expectations component model that describes a particular cluster. Mahalanobis distance in this situation proved to be a good choice for Gaussian mixture models that describe clusters.

Keywords: E-M algorithm, Mahalanobis distance, data clustering, outliers

1. Introduction

This paper considers the problem of clustering (grouping) of a data set S with $m > 2$ elements in the presence of outliers [3]. Data clustering finds its application in medicine, biology, psychology, robotic visualization and navigation, image segmentation etc. [2]. For this purpose, the E-M algorithm is presented. E-M algorithm is based on the principle of soft grouping, where the boundaries between clusters are not solid. Specifically, it is a probabilistic grouping that each element of the reference data set determines the probability of belonging to each cluster. E-M algorithm is generally based on the Gaussian mixture model. Gaussian mixture model approximates the data as a linear combination of k density

$$p(x) = \sum_{i=1}^{k} w_i \, f_i(x \mid \theta_i), \qquad (1)$$

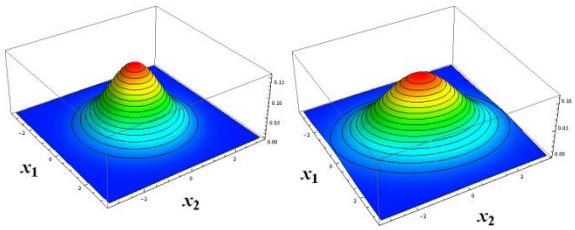
where x is l-dimensional vector, and weights w_i, $i = 1,2,\dots,k$ respectively represent the percentage of data belonging to a cluster π_i, $i = 1,2,\dots,k$, what imply $\sum_{i=1}^{k} w_i = 1$. Parameter θ_i of density function $f_i(x|\theta_i)$ in Gaussian mixture model is presented with expectation μ_i and covariance matrix Σ_i determining the density function for the normal (Gaussian) distribution, i.e. $\theta_i = (\mu_i, \Sigma_i)$. Into the E-step of E-M algorithm is implemented reduction of data which are not taken into the calculation. For that purpose the Mahalanobis distance is considered for every component of Gaussian mixture model during of algorithm execution [1],[4]. If observed data exceed appointed threshold σ from expectation μ_i for every cluster, it is not take into the calculation. In final partition such type of data are considered as an outliers. Measurement are conducted with well known Davis-Bouldin clustering index via threshold σ [5], taking into account percentage of reduced data considered as an outliers produced with modified E-M.

2. Standard E-M algorithm

As mentioned, in Gaussian mixture model clusters are presented with a linear combination of k density presented as (1). Each cluster is presented with corresponding Gaussian mixture model component, i.e. expectation μ_i and covariance matrix Σ_i. Presenting density function $f_i(x|\theta_i)$ is normal distributed, i.e.

$$f_i(x \mid \theta_i) = \frac{1}{\sqrt{(2\pi)^l \mid \Sigma_i \mid}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \right), \qquad (2)$$

where μ_i is l-dimensionaldata vector and Σ_i $l \times l$ covariance matrix. In Figure 1 is shown two illustrative example of (2) for two dimensional case, $l = 2$. In Figure 1(a) is present case with expectation $\mu = (0,0)$, and identity covariance matrix $\Sigma = I$. Figure 1(b) present case with $\mu = (0,0)$ and covariance matrix $\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$.



(a) (b)

Figure 1. PDF of Gaussian distribution

Quality of Gaussian mixture model presented with parameters $\Phi = \{(w_i, \mu_i, \Sigma_i) : i = 1,\dots,k\}$ is measured with log-likelihood

$$L(\Phi) = \sum_{x \in S} \log\left(\sum_{i=1}^{k} w_i \, f_i(x \mid \mu_i, \Sigma_i) \right). \qquad (3)$$

The process is repeated until the log-likelihood of the mixture model at the previous iteration is sufficiently close to the log-likelihood of the current model. The algorithm proceeds as follows for Gaussian mixture model:

ALGORITHM 1. (standard E-M)
STEP 0.
Initialization of parameters $\Phi^0 = \{(w_i^0, \mu_i^0, \Sigma_i^0) : i = 1,\ldots,k\}$ (zero partition), $s = 0$, and stopage criterion $\varepsilon > 0$ (set by user or at random).
STEP 1. (E step)
For every $x \in$ S calculate π_i cluster probability as

$$w_i^s(x) = \frac{w_i^s f_i(x \mid \mu_i^s, \Sigma_i^s)}{\sum_{n=1}^{k} w_n^s f_n(x \mid \mu_n^s, \Sigma_n^s)}, \quad i = 1,\ldots,k.$$

STEP 2. (M step)
Calculation of new parameters for Gaussian mixture model for every $i = 1,2,\ldots,k$:

$$w_i^{s+1} = \sum_{x \in S} w_i^s(x),$$

$$\mu_i^{s+1} = \frac{\sum_{x \in S} w_i^s(x) x}{\sum_{x \in S} w_i^s(x)},$$

$$\Sigma_i^{s+1} = \frac{\sum_{x \in S} w_i^s(x)(x - \mu_i^{s+1})(x - \mu_i^{s+1})^T}{\sum_{x \in S} w_i^s(x)}.$$

STEP 3.
If $|L(\Phi^s) - L(\Phi^{s+1})| \le \varepsilon$ then STOP, else $s = s + 1$ and go to **STEP 1.**

E-M, like many other iterative clustering algorithms, is known to be sensitive to initial parameter values. E-M computes a local solution to the problem of maximizing the log-likelihood of the data given the model. Since this is a local optimization procedure, the quality of the local solution is dependent upon the initial parameter values. Standard practice usually calls for running E-M from many different (possibly randomly) initial parameter values and choosing the mixture model solution with best quality.

3. Mahalanobis distance

Mahalanobis distance is defined as

$$d_M(x, y) = \sqrt{(x - y)\Sigma^{-1}(x - y)^T}. \quad (4)$$

In this case, the constant distance $d_M(x, \mu) = c$ curves are ellipses (hyperellipses in general case).

Indeed, the covariance matrix is symmetric and it can always be diagonalized by a unitary transform

$$\Sigma = U \Lambda U^T, \quad (5)$$

where $U^T = U^{-1}$ and $\Lambda = diag\{\lambda_1, \ldots, \lambda_l\}$ is the diagonal matrix whose elements are eigenvalues of Σ. U has as its columns the corresponding (orthonormal) eigenvectors

$$U = [v_1, v_2, \ldots, v_l]. \quad (6)$$

Combining (5) and (6) in $d_M(x, \mu) = c$, we obtain

$$(x - \mu) U \Lambda^{-1} U^T (x - \mu)^T = c^2. \quad (7)$$

Define $x' = U^T x^T$. The coordinates of x' are equal $v_k^T x$, $k = 1, \ldots, l$, that is, the projections of x onto the eigenvectors. In other words, they are the coordinates of x with respect to a new coordinate system whose axes are determined by v_k, $k = 1, \ldots, l$. Equation (8) can now be written as

$$\frac{(x_1' - \mu_1')^2}{\lambda_1} + \cdots + \frac{(x_l' - \mu_l')^2}{\lambda_l} = c^2. \quad (8)$$

This is the equation of a hyperellipsoid in the new coordinate system. The center of ellipse is μ', and the principal axes are aligned with the corresponding eigenvectors and have length $2\sqrt{\lambda_k}c$, respectively. Thus, all points having the same Mahalanobis distance from a specific point are located on an ellipse.

Example 1.
Mahalanobis distances with covariance matrix

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

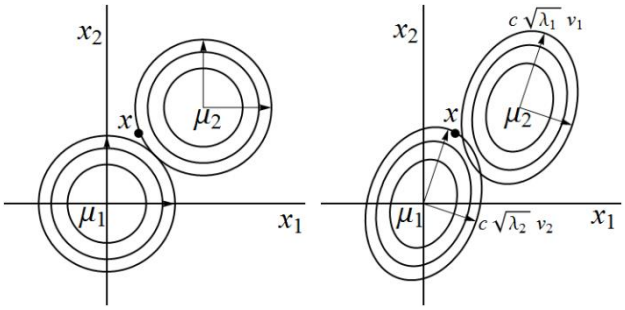
From $x = (1.0, 2.2)$ *to two mean vectors* $\mu_1 = (0, 0)$ *and* $\mu_2 = (3, 3)$ *is*

$$d_M^2(x, \mu_1) = (1.0, 2.2) \begin{bmatrix} 0{,}95 & -0{,}15 \\ -0{,}15 & 0{,}55 \end{bmatrix} (1.0, 2.2)^T = 2{,}952$$

and

$$d_M^2(x, \mu_2) = (-2.0, -0.8) \begin{bmatrix} 0{,}95 & -0{,}15 \\ -0{,}15 & 0{,}55 \end{bmatrix} (-2.0, -0.8)^T = 3.3672.$$

Given statement is presented with next figure.



(a) (b)

Figure 2. Mahalanobis distance

Notice that the given vector $x = (1.0, 2.2)$ *is closer to* $\mu_2 = (3, 3)$ *with respect to the Euclidean distance. Figure 2(a) shows curves of equal*

Euclidean distance from the mean. They are obviously circles of radius c (hyperspheres in the general case). Figure 2(b) shows that principal axes are aligned with the corresponding eigenvectors and have length $2\sqrt{\lambda_k}c$, respectively.

4. Modified E-M algorithm

Modification of E-M algorithm for the purpose of outlier detection is provided into the E-step. Mahalanobis distances generated with corresponding Gaussian mixture model components, i.e. expectation μ_i and covariance matrix Σ_i was used. With specified threshold σ the rejection of data is provided. The modified E-M algorithm proceeds as follows:

ALGORITHM 2. (modified E-M)

STEP 0.

Initialization of parameters $\Phi^0 = \{(w_i^0, \mu_i^0, \Sigma_i^0) : i = 1,\ldots,k\}$ (zero partition), $s = 0$, and stopage criterion $\varepsilon > 0$ (set by user or at random), and $\sigma > 0$.

STEP 1. (E step)

For every $x \in S$ calculate π_i, $i = 1,2,\ldots,k$ cluster probability in the following way.

If

$$d_M^i(x, \mu_i^s) \le \sigma,$$

then

$$w_i^s(x) = \frac{w_i^s f_i(x \mid \mu_i^s, \Sigma_i^s)}{\sum_{n=1}^{k} w_n^s f_n(x \mid \mu_n^s, \Sigma_n^s)},$$

else

$$w_i^s(x) = 0.$$

The distance

$$d_M^i(x, \mu_i^s) = \sqrt{(x - \mu_i^s)(\Sigma_i^s)^{-1}(x - \mu_i^s)^T}$$

is Mahalanobis distance generated with corresponding Gaussian mixture model components.

STEP 2. (M step)

Calculation of new parameters for Gaussian mixture model for every $i = 1,2,\ldots,k$:

$$w_i^{s+1} = \sum_{x \in S} w_i^s(x),$$

$$\mu_i^{s+1} = \frac{\sum_{x \in S} w_i^s(x) x}{\sum_{x \in S} w_i^s(x)},$$

$$\Sigma_i^{s+1} = \frac{\sum_{x \in S} w_i^s(x)(x - \mu_i^{s+1})(x - \mu_i^{s+1})^T}{\sum_{x \in S} w_i^s(x)}.$$

STEP 3.

If $|L(\Phi^s) - L(\Phi^{s+1})| \le \varepsilon$ then STOP, else $s = s + 1$ and go to **STEP 1.**

As standard E-M, the results from modified E-M depends of initial parameters values. For very large σ modified E-M act as standard E-M, taking into the calculation every data. Managing with threshold σ omitting data is provided which exceed from mean vector of corresponding Gaussian mixture model component. In that sense is provided better grouping characteristic and outlier detection.

5. Clustering Validity Criteria

Many different clustering validity measures exist that are very useful in practice as quantitative criteria for evaluating the quality of data partitions. Some of the most well-known validity measure, also referred to as relative validity (or quality) criteria, are possibly the Davis-Bouldin index, defined as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i, \tag{9}$$

where

$$R_i = \max_{i \ne j} R_{ij} \tag{10}$$

$$R_{ij} = \frac{r_i + r_j}{D_{ij}} \tag{11}$$

$$D_{ij} = d(\mu_i, \mu_j) \tag{12}$$

$$r_i = \frac{1}{m_i}\sum_{x \in \pi_i} d(x, \mu_i). \tag{13}$$

For a given assignment of clusters π_i, $i = 1,2,\ldots,k$, a lower DB index indicates better clustering. In (12), and (13) $d : R^l \times R^l \to R$ present Euclidian distance measurement, while m_i in (13) is the cardinal number of corresponding cluster π_i. It is a hard task for the user, however, to choose a specific measure when he or she faces such a variety of possibilities. To make things even worse, new measures have still been proposed from time to time. For this reason, a problem that has been of interest for more than two decades consists of comparing the performances of existing clustering validity measures and, eventually, that of a new measure to be proposed. In our case percentage of outlier provided with modified E-M is taken into the insight. In next example we present our method for detection of outliers in observed data set.

Example 2.
On given data the standard and modified E-M are tested, i.e. Algorithm 1 and Algorithm 2. The data are clustered in three clusters (finding the optimal k partition present separable problem in data clusters analysis). In Figure 3(a) is present result of standard E-M algorithm, i.e. Algorithm 1. The results are presented with contours of linear combination of $k(=3)$ density presented with (1), while the red points presented mean vector of corresponding Gaussian mixture model component. Blue coloured data belongs to first cluster, while green belongs to second, and yellow to third. Figure 3(b) presents results of modified E-M, Algorithm 2, where black data are detected as an outliers. For threshold is observed $\sigma = 2.75$.

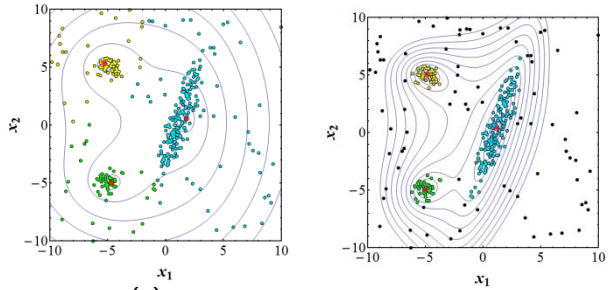


Figure 3. Clustering

Calculation of DB and outlier percentage via threshold σ of modified E-M is presented in next figure.

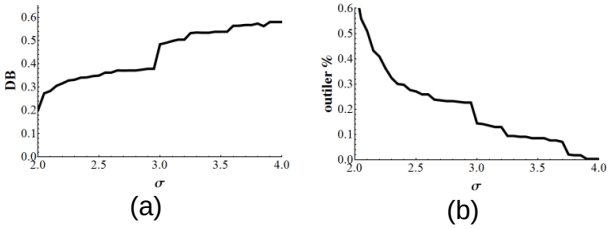


Figure 4. Clustering Validity Criteria

From measures, presented in Figure 4, we can see that the big skip at the threshold value $\sigma = 3$ is happened. This situation can be discussed as a good clustering until breakdown point at $\sigma = 3$ happened. Afterwards great account of an outliers are rejected. It can mean that the outliers are taken into the clusters, what has happened in our situation. In this situation good clustering properties are disturbed, what is presented with DB index presented in Figure 4(a). From that, a conclusion that a good outlier detection and clustering properties are happened for a values smaller then breakdown point. For a greater values of a threshold σ it can be seen that a great amount of an outliers exceeds, Figure 4(b), what implies perform of standard E-M for a large σ, and degradation of clustering quality of data with outlier presence.

6. Conclusion

Proposed modification of E-M algorithm into the E-step with Mahalanobis distance shows good properties of an outlier detection via threshold σ. Managing with σ is also established better grouping characteristic of clusters.

7. References

[1] P. S. Bradley, U. M. Fayyad, C. A. Reina, "Scaling EM (Expectation-Maximization) Clustering to Large Databases", *Microsoft Researc*h, 1999.
[2] G. Gan, C. Ma, J. Wu, "Data Clustering, Theory, Algorithms and Application", *SIAM*, 2007.
[3] P. J. Rousseeuw, A. M. Leroy, "Robust Regression and Outlier Detection", *Wiley*, 2003.
[4] S. Theodoridis, K. Koutroumbas, "Pattern Recognition", Fourth edition, *Elsevier,* 2009.
[5] Lucas Vendramin, Ricardo J. G. B. Campello, Eduardo R. Hruschka, "On the Comparison of Relative Clustering Validity Criteria", Proceedings of the SIAM International Conference on Data Mining, April 30-May 2, 2009, Sparks, Nevada, USA, *SIAM* 2009, 733-744.