An Interval Expectation Maximization Algorithm for Outlier Detection in Linear Regression

Daniel B. Barreiros[1], **Marco A.O. Domingues**[2], **Renata M.C.R. Souza**[1], **and Francisco J. A. Cysneiros**[3]
[1]Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brazil. dbb2@cin.ufpe.br, rmcrs@cin.ufpe.br
[2]CTADS, Instituto Federal de Pernambuco, Recife, PE, Brazil. marcodomingues@recife.ifpe.edu.br
[3]Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil. cysneiros@de.ufpe.br

Abstract— *Outlier detection has an important role in diverse fields of research and application domains including pattern recognition, exploratory data analysis and data mining. In classical regression analysis, these outliers are often removed from the data set, being usually regarded as errors of the process. However, in SDA domain, this procedure is unsuitable because a single symbolic data observation may represent the generalization of a subset of other classical observations. This paper introduces an expectation-maximization algorithm for interval data in order to detect atypical intervals concerned with regression analysis problems. The algorithm is evaluated regarding different simulated and real interval data sets.*

Keywords: Outlier detection, Expectation-Maximization, Interval Data, Symbolic Data Analysis

1. Introduction

Interval data have been considered in real world applications: analysis of census data [1], electricity load profiling [2], scientific production of researches [3]. This kind of data has been studied mainly in *Symbolic Data Analysis* (SDA) which is a research field related to multivariate analysis, pattern recognition and artificial intelligence [4]. SDA aims to provide a comprehensive way to summarize data sets by means of symbolic data resulting in a smaller and more manageable data set which preserves the essential information. In the literature of SDA, several approaches for interval data have been introduced: recommendation systems [5], classification [6], principal component analysis [7], regression [8]. Symbolic data allow multiple values for each variable. These new variables (set-valued, interval-valued, and histogram-valued) make it possible to hold data intrinsic variability and/or uncertainty from the original data set as shown in [4].

Interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Another common source of interval data is the aggregation of data into a reduced number of groups. In this case, SDA starts extracting knowledge from a data set in order to provide symbolic descriptions that are mathematically modeled by a generalization process applied to a set of individuals. An example is an amanita mushroom specie data set formed by 23 mushroom species. The intervals of this data set were obtained by aggregating individual mushrooms according to the kind of species. Each individual mushroom is described by three interval variables that are: stipe length, stipe thickness and pileus cap.

Figure 1 shows the amanita data set. In this figure, we can observe that there are two intervals which are substantially different from all other ones. They were obtained from the generalization process applied to the amanita data set.

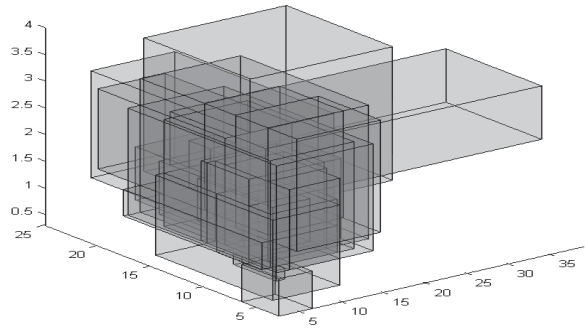

Fig. 1: Interval *Amanita* mushroom data

According to [4], overgeneralization problems can arise when these extreme values are actually outliers or when the set of individuals to generalize is in fact composed of subsets of different distributions. Indeed, in these situations interval outliers can be found, as it is highlighted in the amanita interval data set, and methods that identify them are essential. Investigation methods of outliers as primary analysis is an opened research topic.

In classical data analysis, point outliers are observations in a data set which do not follow the pattern of the other observations. Such data play important role in regression since they can lead to inaccurate regression estimates. It is a common practice to distinguish between two types of outliers: on the response variable, called outlier, represents a model failure and may indicate a sample peculiarity, a data entry error or another problem; and with respect to the predictors variables, called leverage points. This paper addresses outliers on response variables.

In SDA, interval outliers are also unusual observations and

interval regression is an extension of the classical regression for symbolic interval data [8]. In the amanita data set of the Figure 1, the regression problem concerns to estimate pileus cap (response interval variable) from stipe length, stipe thickness (predictor interval variable).

The main contribution of this work is to propose an EM-type algorithm regarding a multivariate gaussian mixture model for interval data to identify atypical intervals in regression analysis. The proposed algorithm is evaluated with different real and simulated interval data sets. For simulated interval data, the performance of the proposed algorithm is measured by the false negative and false positive rates in the framework of a Monte Carlo experiment.

The rest of the paper is organized in the following form: Section 2 presents the simulated and real data sets used in this work. Section 3 describes the EM-type algorithm for detecting atypical intervals. Section 4 presents a performance analysis. Finally, Section 5 gives the concluding remarks.

2. Interval data sets

Different simulated interval data sets which comprises two arrangements for interval outliers and the Amanita interval data set are presented in this section.

2.1 Simulated interval data sets containing outliers

Initially, each seed s_i^x $(i = 1, \ldots, n)$ on coordinate X arises from an uniform distribution $[a, b]$. A seed s_i^y on coordinate Y is related to the seed s_i^x as $s_i^y = \beta_0 + \beta_1 s_i^x + \varepsilon_i$ $(i = 1, \ldots, n)$ where β_0 and β_1 are simulated from an uniform distribution $[c, d]$ and ε_i is simulated from a standard normal distribution.

Thus, seed data sets are now formed by bivariate points (s_i^x, s_i^y) $(i = 1, \ldots, n)$. For each point i, a random sample of size 30 is drawn from a bivariate gaussian distribution with mean vector and the diagonal covariance matrix $\boldsymbol{\mu} = (s_i^x, s_i^y)$ and $\Sigma = \sigma \mathbf{I}$ where σ is a parameter of scale. From each sample, the rectangle i is defined by a vector of two intervals

$$\mathbf{v} = (x_i = [a_i, b_i], y_i = [\lambda, \gamma])'$$

where $a_i = Q_1^x$, $b_i = Q_3^x$, $\lambda_i = Q_1^y$ and $\gamma_i = Q_3^y$ are first and third quartiles of the samples on coordinates X and Y, respectively.

Interval outliers are created in the following way. First of all, the sets are sorted ascending by the dependent variable Y^c and a small cluster containing the m first points of the sorted set (y_i^c, x_i^c) is selected. The observations of this cluster are changed into outlier points by

$$x_i^c = x_i^c + f_x . S(X^c)$$

$$y_i^c = y_i^c + f_y . S(Y^c)$$

where $S(Y^c)$ and $S(X^c)$ are, respectively, the standard deviation of $(y_1^c, \ldots, y_n^c)$ and the standard deviation of $(x_1^c, \ldots, x_n^c)$, and, f_x and f_y are fixed values.

Two different configurations for rectangles containing remote intervals in terms of position (center of the intervals) are considered in this paper. Figures 2 and 3 display the interval data sets 1 and 2, respectively, with $s^x \sim U[a, b] = [10, 40]$, $\beta_0, \beta_1 \sim U[c, d] = [1, 10]$, $n = 50$ and $\sigma = 3$. Figure 2 ($f_x = 0$ and $f_y = 10$) shows a scenario in which there are intervals that are strongly outliers. Figure 3 ($f_x = 5$ and $f_y = 10$) considers a data set with a group of intervals that are slightly outliers.

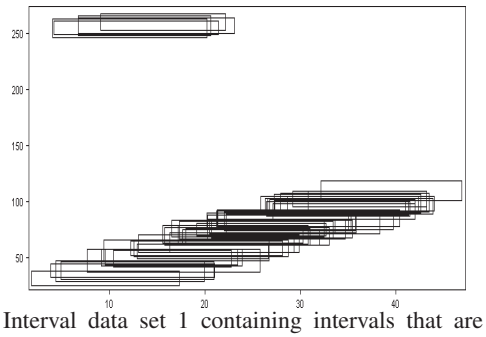

Fig. 2: Interval data set 1 containing intervals that are strongly outliers.

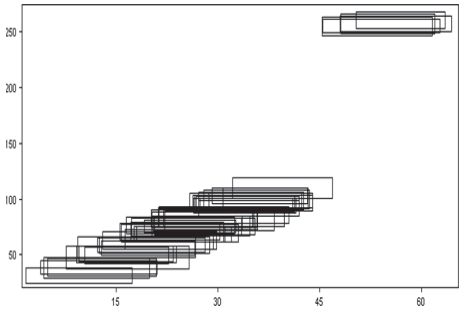

Fig. 3: Interval data set 2 containing intervals that are slightly outliers.

2.2 Amanita interval data set

Table 1 shows a mushroom specie data set. These mushroom species are members of the genus *Amanita* in which the values were collected from the *Fungi of California Species Index* $(http : //www.mykoweb.com/CAF/species\_index.html)$.

From the values in the table above, three species are candidate outliers on the response variable. They are: Lanei, Muscaria and Pachycolea. Regarding the pileus cap response variable, the Lanei and Muscaria species have atypical intervals.

Table 1: Ranges of pileus cap, stipe length and stipe thickness of the *Amanita* mushroom family.

Amanita species	Interval Variables		
	Pileus Cap	Stipe Length	Stipe Thickness
Lanei	[8.00 : 25.00]	[10.00 : 20.00]	[1.50 : 4.00]
Constricta	[6.00 : 12.00]	[9.00 : 17.00]	[1.00 : 2.00]
Franchetii	[4.00 : 12.00]	[5.00 : 15.00]	[1.00 : 2.00]
Novinupta	[5.00 : 14.00]	[6.00 : 12.00]	[1.50 : 3.50]
Muscaria	[6.00 : 39.00]	[7.00 : 16.00]	[2.00 : 3.00]
Ocreata	[5.00 : 13.00]	[10.00 : 22.00]	[1.50 : 3.00]
Pachycolea	[8.00 : 18.00]	[10.00 : 25.00]	[1.00 : 3.00]
Pantherina	[4.00 : 15.00]	[7.00 : 11.00]	[1.00 : 2.50]
Phalloides	[3.50 : 15.00]	[4.00 : 18.00]	[1.00 : 3.00]
Protecta	[4.00 : 14.00]	[5.00 : 15.00]	[1.00 : 3.00]
Vaginata	[5.50 : 10.00]	[6.00 : 13.00]	[1.20 : 2.00]
Velosa	[5.00 : 11.00]	[4.00 : 11.00]	[1.00 : 2.50]
Aprica	[5.00 : 15.00]	[3.30 : 9.10]	[1.40 : 3.50]
Bivolvata	[7.00 : 10.00]	[13.00 : 15.00]	[1.60 : 2.50]
Gemmata	[3.00 : 11.00]	[4.00 : 15.00]	[0.50 : 2.00]
Magniverrucata	[4.00 : 13.00]	[7.00 : 11.50]	[1.00 : 2.50]
Smithiana	[5.00 : 17.00]	[6.00 : 18.00]	[1.00 : 3.50]
Cokeri	[7.00 : 15.00]	[10.00 : 20.00]	[1.00 : 2.00]
Porphyria	[3.00 : 12.00]	[5.00 : 18.00]	[1.00 : 1.50]
Silvicola	[5.00 : 12.00]	[6.00 : 10.00]	[1.00 : 2.50]
Californica	[6.00 : 7.00]	[6.00 : 10.00]	[0.60 : 0.80]
Farinosa	[2.50 : 6.50]	[3.00 : 6.50]	[0.30 : 1.00]
Breckonii	[4.00 : 9.00]	[7.00 : 10.00]	[0.90 : 2.00]

3. EM-type algorithm for interval data

The Expectation Maximization(EM) algorithm [9] has been widely applied to estimation problems involving incomplete data, or in problems which can be modeled as mixture of distributions. In brief, the EM algorithm aims at finding maximum likelihood estimates of parameters in probabilistic models in the presence of missing or hidden data. Due to its simplicity, the EM for multivariate gaussian mixture model is by far the most employed mixture model with many applications in cluster analysis and statistical pattern recognition (see, for instance, [10]).

In the outlier framework, the EM algorithm can be employed as a tool for detecting atypical observations from the data sets . For this reason, a EM-type algorithm for interval data (EM-IVD) is introduced in this paper. EM-IVD extends the standard EM algorithm for multivariate gaussian mixture model to treat interval-valued data.

Consider $\mathbf{X}^*$ as a $n \times r$ input data matrix and whose each row is represented as an interval feature vector $\mathbf{x}_i^* = (x_{i1}^*, \ldots, x_{ir}^*)^{'}$ where $x_{ij}^* = [a_{ij}, b_{ij}]$, $(j = 1, \ldots, r) \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$. The interval Expectation-Maximization (iE-M) algorithm sets an initial partition and alternates two steps such an expected log likelihood-type function reaches a stationary value representing a local maximum.

Let $\{C_1, C_2\}$ be a partition of $\mathbf{X}^*$ in 2 clusters (outliers and inliers) and $\boldsymbol{\theta}_k = (\tau_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{'}$ $(k \in \{1, 2\})$ be a parameter vector of C_k where $\boldsymbol{\mu}_k = ([\mu_{kl}^1, \mu_{ku}^1], \ldots, [\mu_{kl}^r, \mu_{ku}^r])^{'}$ is an average interval vector, $\boldsymbol{\Sigma}_k$ be a covariance matrix and τ_k be a mixture coefficient of C_k. In the iE-M method, there is an average interval vector represented as $\boldsymbol{\mu}_k$ that correspond average values of boundaries of intervals and a single covariance matrix $\boldsymbol{\Sigma}_k$ whose the values measure the variability of the intervals related to this average interval vector.

3.1 Initialization step

Randomly choose 2 different objects $\mathbf{g}_1$ and $\mathbf{g}_2$ belonging to $\mathbf{X}^*$ and assign each objects i to a class C_m such that $m = \arg\min_{k=1,2} d(\mathbf{x}_i^*, \mathbf{g}_k)$ where d is the normalized Hausdorff distance [11] between two interval vectors.

Let $\mathbf{x}_i^*$ e $\mathbf{x}_h^*$ two interval vectors in $\Re^r$, the normalized Hausdorff distance between these vectors is given by:

$$d(\mathbf{x}_i^*, \mathbf{x}_h^*) = \left\{ \sum_{j=1}^{r} \left[\frac{Max[\ |a_i^j - a_h^j|, |b_i^j - b_h^j|\ ]}{H_j} \right]^2 \right\}^{1/2}, \tag{1}$$

with

$$H_j^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{h=1}^{n} \Big[Max[\ |a_i^j - a_h^j|, |b_i^j - b_h^j|\ ] \Big]^2.$$

Given a partition $\{C_1, C_2\}$, the initial values for the parameters of the class C_k $(k = 1, 2)$ are computed as:

- average interval vector

$$\hat{\boldsymbol{\mu}}_k = ([\hat{\mu}_{kl}^1, \hat{\mu}_{ku}^1], \ldots, [\hat{\mu}_{kl}^r, \hat{\mu}_{ku}^r])^{'} \tag{2}$$

  with

$$\hat{\mu}_{kl}^j = \frac{1}{|C_k|} \sum_{i \in C_k} a_i^j$$

  and

$$\hat{\mu}_{ku}^j = \frac{1}{|C_k|} \sum_{i \in C_k} b_i^j.$$

- covariance matrix $\hat{\boldsymbol{\Sigma}}_k = (\hat{\sigma}_k^{vj})$ with

$$\hat{\sigma}_k^{vj} = \frac{\sum_{i, v \in C_k} \Big[(a_i^v - \hat{\boldsymbol{\mu}}_{kl}^v)(a_i^j - \hat{\boldsymbol{\mu}}_{kl}^j) + (b_i^v - \hat{\boldsymbol{\mu}}_{ku}^v)(b_i^j - \hat{\boldsymbol{\mu}}_{ku}^j)) \Big]}{2|C_k|}. \tag{3}$$

- mixture coefficient

$$\hat{\tau}_c = \frac{|C_k|}{n}. \tag{4}$$

3.2 E step

Let $\hat{\boldsymbol{\mu}}_{kl} = (\hat{\mu}_{kl}^1, \ldots, \hat{\mu}_{kl}^r)'$ and $\hat{\boldsymbol{\mu}}_{ku} = (\hat{\mu}_{ku}^1, \ldots, \hat{\mu}_{ku}^r)'$ be vectors associated to lower and upper bounds of the intervals of $\hat{\boldsymbol{\mu}}_k$. Consider also $\mathbf{x}_{il}^* = (a_i^1, \ldots, a_i^p)'$ and $\mathbf{x}_{iu}^* = (b_i^1, \ldots, b_i^p)'$ as vectors associated to lower and upper bounds of the intervals of the pattern $\mathbf{x}_i^*$ $(i = 1, \ldots, n)$.

Given the parameter vector $\hat{\boldsymbol{\theta}}_k = (\hat{\tau}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)'$ $(k \in \{1, 2\})$, the probability of the object i belong to C_k is defined as:

$$\hat{Pr}(C_k|\mathbf{x}_i^*) = \frac{\hat{\tau}_k \hat{Pr}(\mathbf{x}_i^*|C_k)}{\sum_{k=1}^{2} \hat{\tau}_k \hat{Pr}(\mathbf{x}_i^*|C_k)}, \tag{5}$$

where

$$\hat{Pr}(\mathbf{x}_i^*|C_k) = \frac{\exp^{-\frac{1}{2}[A+B]}}{\sqrt{(2\pi)^p \times |\hat{\boldsymbol{\Sigma}}_k|}}, \tag{6}$$

$$A = (\mathbf{x}_{il}^* - \hat{\boldsymbol{\mu}}_{kl})^T \hat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x}_{il}^* - \hat{\boldsymbol{\mu}}_{kl})$$

$$B = (\mathbf{x}_{iu}^* - \hat{\boldsymbol{\mu}}_{ku})^T \hat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x}_{iu}^* - \hat{\boldsymbol{\mu}}_{ku})$$

3.3 M step

The parameter vector $\hat{\boldsymbol{\theta}}_k = (\hat{\tau}_k, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)'$ $(k \in \{1, 2\})$ is updated by:

$$\hat{\tau}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{Pr}(C_k|\mathbf{x}_i^*), \tag{7}$$

$$\hat{\boldsymbol{\mu}}_k = ([\hat{\mu}_{kl}^1, \hat{\mu}_{ku}^1], \ldots, [\hat{\mu}_{kl}^r, \hat{\mu}_{ku}^r])'$$

with

$$\hat{\mu}_{kl}^j = \frac{\sum_{i \in \Omega} a_{ij} \cdot \hat{Pr}(C_k|\mathbf{x}^*)}{\sum_{i \in \Omega} \hat{Pr}(C_k|\mathbf{x}^*)}, \tag{8}$$

$$\hat{\mu}_{ku}^j = \frac{\sum_{i \in \Omega} b_{ij} \cdot \hat{Pr}(C_k|\mathbf{x}^*)}{\sum_{i \in \Omega} \hat{Pr}(C_k|\mathbf{x}^*)} \tag{9}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i \in \Omega} \hat{Pr}(C_k|\mathbf{x}^*) \times (W + V)}{2 \cdot \sum_{i \in \Omega} \hat{Pr}(C_k|\mathbf{x}^*)}, \tag{10}$$

with

$$W = (\mathbf{x}_{il}^* - \hat{\boldsymbol{\mu}}_{ki})(\mathbf{x}_{il}^* - \hat{\boldsymbol{\mu}}_{kl})'$$

$$V = (\mathbf{x}_{iu}^* - \hat{\boldsymbol{\mu}}_{ku})(\mathbf{x}_{iu}^* - \hat{\boldsymbol{\mu}}_{ku})'.$$

3.3.1 Algorithm schema

The iE-M algorithm has the following steps:

Algorithm 1 A EM-type algorithm for interval data.

1: ***Initialization step***: Randomly choose a partition $(C_1^{(0)}, C_2^{(0)})$ of $\mathbf{X}^*$ or randomly choose 2 distinct objects $\mathbf{g}_1^{(0)}, \mathbf{g}_2^{(0)}$ belonging to $\mathbf{X}^*$ and assign each objects i to the closest prototype $\mathbf{g}_m$, where $m = \arg\min_{k=1,2} d(\mathbf{x}_i^*, \mathbf{g}_k)$ and d is the Hausdorff distance defined in Eq. (1). Obtain initial estimate for parameters $\hat{\tau}_k^0, \hat{\boldsymbol{\mu}}_k^0$ and $\hat{\boldsymbol{\Sigma}}_k^0$ $(k = 1, 2)$ according to the Eqs. (2), (3) and (4), respectively. Do $t = 1$.

2: **E-*step***: For $i = 1, \ldots, n$, compute the probability $\hat{Pr}(\mathbf{x}_i^*|C_k)^t$ $(k = 1, 2)$ using the Eq. (5).

3: **M-*step***: For $k = 1, 2$, compute the vector $\hat{\boldsymbol{\theta}}_k^t = (\hat{\tau}_k^t, \hat{\boldsymbol{\mu}}_k^t, \hat{\boldsymbol{\Sigma}}_k^t)$ according to the Eqs. (7), (8), (9) and (10).

4: ***Stopping criterion*** If $||\frac{\hat{\boldsymbol{\theta}}_k^t - \hat{\boldsymbol{\theta}}_k^{t-1}}{\hat{\boldsymbol{\theta}}_k^t}|| < \varepsilon$ for $k = 1, 2$ then go to step 5 else do $t = t + 1$ and go to 2.

5: ***Classification step***: For $i = 1, \ldots, n$ find the cluster C_{k^*} such that

$$k* = arg\ \max_{1 \leq k \leq 2} \hat{Pr}(\mathbf{x}_i|C_k).$$

Let K be the number of classes (here, $K = 2$). The time complexity of the **E-*step*** is $O(nKr^2t)$ and the time complexity of the **M-*step*** is $O(nKt + nKrt + nKr^2t)$. Therefore, the time complexity of the iE-M algorithm is $O(nKr^2t)$.

4. Performance Analysis

For simulated interval data sets 1 and 2, the performance is measured by the false positive and false negative rates (FNR and FPR) in the framework of a Monte Carlo experience with 100 replications for each interval data set. Here, FNR is the number of elements of the inlier class labeled as belonging to outlier class divided by the size of the inlier class and FPR is the number of elements of the outlier class labeled as belonging to inlier class divided by the size of the outlier class.

For each data set, four situations are considered taking into account the quantity (percentage of the data set) of outlying observations presents in each interval data set, that is, 2%, 6%, 10% and 20% of the interval data are indeed interval outliers. Moreover, values for the seed s^c are generated from an uniform $U[1, 10]$ and the values for the parameters β_0, β_1 are selected randomly from an uniform distribution $U[1, 10]$. Each interval data set has two clusters, one with regular intervals and the other with outlying intervals.

Tables 2 shows the the average of the false negative and false positive rates (FNR and FPR). The iE-M method performs well in terms of false positive rate for all cases. Moreover, this method based on the full covariance surpasses that based on the diagonal matrix for both scenarios. This is expected because the linear relation assumed for the interval variables. Regarding the false negative rate, the iE-M method improves when the number of outliers increases and it is important to observe that this method has the worst

Table 2: FPR(%) and FNR (%) for scenarios 1 and 2.

Outliers	Scenario 1				Scenario 2			
	FPR		FNR		FPR		FNR	
	Diag	Full	Diag	Full	Diag	Full	Diag	Full
2%	0.00	0.00	10.90	10.07	0.00	0.00	11.09	13.45
6%	10.34	0.00	1.18	0.83	2.67	0.40	1.62	1.03
10%	18.80	0.00	1.00	0.76	7.00	0.40	1.32	0.92
20%	32.50	0.00	0.65	0.55	20.50	0.10	0.78	0.70

Table 3: Average number of iterations for the iE-M algorithm.

Outliers	Scenario 1		Scenario 2	
	Diagonal Matrix	Full Matrix	Diagonal Matrix	Full Matrix
2%	3.20	3.22	3.53	3.18
6%	4.66	3.13	4.16	3.11
10%	4.48	3.05	4.03	3.04
20%	3.83	3.15	3.46	3.22

performance for the data sets containing a small group of outliers (2% of the data set).

Table 3 shows the average number of iterations for the iE-M algorithm and scenarios 1, 2 and 3. In general, the convergence of this method was achieved with less than five iterations. The algorithm based on full covariance matrix achieves the convergence faster than the algorithm based on diagonal covariance.

With respect the application of the iE-M algorithm to the amanita data set, two groups are obtained. The first group contains 20 species: Lanei, Constricta, Franchetii, Novinupta, Pantherina, Phalloides, Protecta, Vaginata, Velosa, Aprica, Bivolvata, Gemmata, Magniverrucata, Smithiana, Cokeri, Porphyria, Silvicola, Californica, Farinosa and Breckonii. The second group contains 3 species: Muscaria, Ocreata and Pachycolea. From these results and Figure 1 that points out two outliers belonging to the amanita data set, we can say that the Muscaria, Ocreata and Pachycolea species are candidate outliers.

5. Conclude remarks

In this paper, an interval Expectation-maximization for detecting outlier in the framework of regression analysis which is related to symbolic data analysis is presented. The method has as input data a set of predictor interval symbolic variables and a response interval symbolic variable. The experiments regarding different scenarios of simulated interval data sets containing interval outliers and an application with a mushroom interval data base showed the usefulness of this algorithm.

6. Acknowledgements

The authors would like to thank CNPq, CAPES and FACEPE (Brazilian Agencies) for their financial support.

References

[1] Cariou, V. Extension of multivariate regression trees to interval data. Application to electricity load profiling. Computational Statistics 21 325-341, (2006).

[2] Giusti, A. and Grassini L. Cluster analysis of census data using the symbolic data approach. Advances in Data Analysis and Classification, 2 (2) 163-176 DOI: 10.1007/s11634-008-0024-5, (2008).

[3] Pimentel, B.A. & Souza, R.M.C.R. Using Weighted Clustering and Symbolic Data to Evaluate Institutesś Scientific Production. IN Artificial Neural Networks and Machine Learning ICANN 2012. Lecture Notes in Computer Science Volume 7553, 435-442, 2012.

[4] Noirhomme-Fraiture, M. & Diday, E. Symbolic Data Analysis and the SODAS Software, Wiley, (2008).

[5] Bezerra, B.L.D and De Carvalho, F.A.T. Symbolic data analysis tools for recommendation systems Knowledge Information System 26, 3, 385-418, (2011)

[6] Malia, K. & Mitra, S. Symbolic classification, clustering and fuzzy radial basis function network. Fuzzy Sets and Systems 152 553-564, (2005).

[7] Gioia, F. & Lauro, C.N. Principal Component Analysis on Interval Data. Computational Statistics 21, 343-363, (2006).

[8] Lima Neto, E.A. & De Carvalho, F.A.T.. Centre and Range method for fitting a linear regression model to symbolic interval data. Computational Statistics & Data Analysis 52, 1500-1515, (2008).

[9] Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B 39, 1, 1-38, (1977).

[10] He, Yi, Pan, Wei and Lin, Jizhen Lin. Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. Computational Statistics & Data Analysis 51, 2, 641-658, (2006).

[11] Billard, L., Diday E., Symbolic Data Analysis: Conceptual Statistics and Data Mining, Wiley, West Sussex, England, (2006).