ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Automatic Group-Outlier Detection

Amine Chaibi and Mustapha Lebbah and Hanane Azzag

LIPN-UMR 7030
Université Paris 13 - CNRS
99, av. J-B Clément - F-93430 Villetaneuse
{firstname.secondname}@lipn.univ-paris13.fr

Abstract. We propose in this paper a new measure called GOF (Group Outlier Factor) to detect groups outliers. To validate this measure we integrated it in a clustering process using Self-organizing Map. The proposed approach is based on relative density of each group of data and simultaneously provides a partitioning of data and a quantitative indicator (GOF). The obtained results are very encouraging to continue in this direction.
Keywords: outliers, group outlier, clustering, self-organizing maps.

1 Introduction

An outlier is an observation or a point that is significantly different from the rest of the data. In data exploration, there is a need to discover not only outliers but also outlier groups. An outlier group is a small data set considerably isolated from the rest of the data. Outliers points can be problematic because they may bias the results, especially for methods based on distances between individuals. In this paper, we have look at density-based approaches for outlier detection. The main task is to define pairwise distances between data points and identify outliers by examining the distance or relative density of each data point to its local neighbors. The advantage of density-based approaches is that they do not make any assumption for the generative distribution of the data. Local Outlier Factor approach [1][6] still the most used density based models. For each object in the dataset, LOF indicates its degree of outlier-ness. LOF method compares the local density of an observation with the average density of its k-nearest neighbors (k-NN). The traditional local outlier factor [1] is meaningful and adequate under certain conditions, but not satisfactory for the general case when clusters of different densities exist. In this direction, the authors of [2] proposed a self-organizing map approach for spatial outlier detection. Spatial outliers are abnormal data points, which have significantly distinct non-spatial attribute values compared with their neighborhood. This method was improved by [3] in the aim of reducing the size of the data, maintaining topological map informations and reducing the influence of potential outliers.

In this paper, we develop a new formal definition of group outliers, which avoids the shortcomings present in traditional approach. The key difference between our notion and existing notions of outliers is that being outlying is not associated to only observation. Instead, we assign to each cluster an outlier factor (GOF: Group Outlier Factor), which is estimated during the learning phase. This factor is the degree of being outlying associated to cluster. For

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

validation, we include this measure in topological maps algorithm. This allows to learn the data structure while providing a new parameter GOF.

2 Simultaneous clustering and Group Outlier Detection

The proposed approach to perform clustering and cluster relevance is designed to search simultaneously for the outlier cluster, and the optimal clustering. A group outlier is a cluster that deviates so much from other clusters. It may be a group of interest, a group of novelty... etc.

In our approach we use self-organizing maps as clustering algorithm [7]. SOM is increasingly used as tools for visualization, as they allow projection in small spaces that are generally two dimensional. Our model consists on a discrete set $\mathcal{C}$ of cells called map. This map has a discrete topology defined by undirected graph, it is usually a regular grid in 2 dimensions. For each pair of cells (c,r) on the map, the distance $\delta(c, r)$ is defined as the length of the shortest chain linking cells r and c on the grid. For each cell c this distance defines a neighbor cell. Let $\Re^d$ be the euclidean data space and $\mathcal{D} = \{\mathbf{x}_i; i = 1, \ldots, N\}$ a set of observations, where each observation $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^d)$ is a vector in $\Re^d$. In our approach, each cell c of the grid $\mathcal{C}$ is associated with two parameters : a prototype vector $\mathbf{w}_c = (w_c^1, w_c^2..., w_c^j, ..., w_c^d)$ of dimension d and a new parameter $GOF_c \in \Re$ (Group Outlier Factor), which is the cluster degree is being outlying : more the value is large, more the cluster is probably outlier group. We denote thereafter $\mathcal{W} = \{\mathbf{w}_c, \mathbf{w}_c \in \mathcal{R}^d\}_{c=1}^C$ the set of prototypes and by $GOF = \{GOF_1, ..., GOF_C\}$ the set of outlier indicators, where C is the number of cells. Each prototype is associated for subset of data assigned to the cell c denoted by P_c. The set of the subsets form a partition $\mathcal{P} = \{P_1, ...P_c, ..., P_C\}$ of the data set $\mathcal{D}$.

In this paper, we emphasize on density-based observation $f_c(\mathbf{x})$ and new outlier factor. $f_c(\mathbf{x})$ is a function used to estimate data density assigned to cell c. Thus $f_c(\mathbf{x})$ is defined as follows: $f_c(\mathbf{x}_i) = \exp^{-\frac{\|\mathbf{w}_c - \mathbf{x}_i\|^2}{2\sigma^2}}$, where σ is the standard deviation. In the particular case of topological maps, we propose to minimize the following cost function:

$$\begin{aligned}\mathcal{R}(\mathcal{W}, \phi, GOF) &= \sum_{i=1}^{N}\sum_{c=1}^{C} K(\delta(\phi(\mathbf{x}_i), c))\|\mathbf{w}_c - \mathbf{x}_i\|^2 \\ &+ \sum_{i=1}^{N}\sum_{c=1}^{C} K(\delta(\phi(\mathbf{x}_i), c)) \left(GOF_c - \frac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{\frac{1}{f_c(\mathbf{x}_i)}} \right)^2 \quad (1)\end{aligned}$$

where ϕ assign each observation $\mathbf{x}_i$ to a single cell of the map. The first term depends on the parameters $\mathcal{W}$, and ϕ, which allow to estimate the prototypes. The second term depends on the cluster outlier factor (GOF) associated for each cell. The minimization of $\mathcal{R}(\mathcal{W}, \phi, GOF)$ is run by iteratively performing three steps until stabilization. After initialization step of prototype set $\mathcal{W}$ and the associated group outlier factor set GOF, at each training step (t + 1) a sample

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

observation $\mathbf{x}_i$ is randomly chosen from the input data set, and we carry out the following stage:

1. Competition phase: Assign data $\mathbf{x}_i$ by using the function

$$\phi(\mathbf{x}_i) = \arg \min_{1 \leq j \leq C} \parallel \mathbf{x}_i - \mathbf{w}_j \parallel^2$$

2. Adaptation phase:

   - Update prototypes $\mathbf{w}_c$ of each cell c

$$\mathbf{w}_c(t) = \mathbf{w}_c(t-1) - \varepsilon(t) K(\delta(\phi(\mathbf{x}_i, c))) \left(\mathbf{w}_c(t-1) - \mathbf{x}_i\right)$$

   - Update the values of GOF_c associated to each cell c: $GOF_c(t) = GOF_c(t-1) - \varepsilon(t) K(\delta(\phi(\mathbf{x}_i, c))) \left(GOF_c(t-1) - \frac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{\frac{1}{f_c(\mathbf{x}_i)}} \right)$
   Where $\varepsilon(t)$ is the step of learning.

3 Experimentation

In this section, with the proposed approach of taking some maximum GOF value within the range, we show that our idea can be used successfully to identify cluster outliers, which appear to be meaningful, but cannot be identified by other methods as LOF method [1]. We used some public dataset [5] modified with adding more difficulties (see figure 1) and synthetical dataset, for which we show the outlier factors for all cell of map, in order to provide an intuitive notion of the GOF values computed. Table 1 presents the description of the synthetic and public dataset and the map size used in the learning phase.

dataset	size	Map size	dataset	size	Map size
RingModif	1072	14×12	CircleModif	638	13×10
HeptaModif	212	9×8	LsunModif	400	11×9
TargetModif	951	13×12	GolfBallModif	4343	19×17
simulated base 1	160	5×13	simulated base 2	234	3×26
simulated base 3	569	8×15	simulated base 4	402	8×13

Table 1: Dataset description.

In order to select some GOF values, we used an heuristic method called "*Scree Acceleration Test*" to detect outliers groups [4]. The basic idea is to display the curve associated to $GOF = (GOF_1, GOF_2..., GOF_c, ..., GOF_C)$ parameters and find from which there is seems to behave randomly. The number of components to keep is the number of values preceding this "Scree". Often this "Scree" appears where the slope of the graphic change radically. Thus we have to process the following steps:

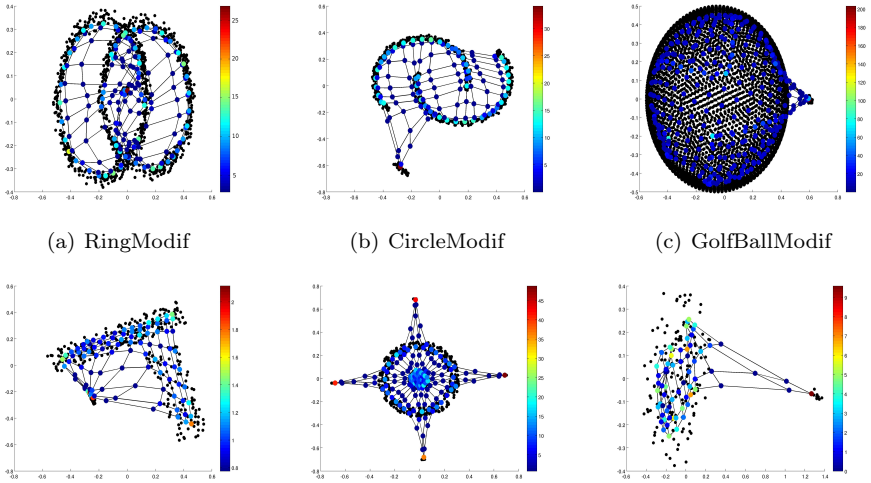
1. Sort the GOF in descending order. The new ordered is noted $GOF = (GOF^1, GOF^2, ..., GOF^j, ..., GOF^C)$;

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2. Compute the first difference $df_i = GOF^i - GOF^{i+1}$;
3. Compute the second difference (acceleration) $acc_i = df_i - df_{i+1}$
4. Find the scree : $\max_i (abs(acc_i) + abs(acc_{i+1}))$

This process allows to select all components located before the abrupt change.

To check visually the results, we project the dataset and the map using PCA. At this point, we get the same result with the traditional self-organizing maps, except that we estimate during learning phase a GOF parameter. In figure 1 the GOF value is indicated with color degree (More the color is red, more the group has a high value of GOF). It is clear for the presented dataset that the red color corresponds to the isolated cluster with high value of GOF.



(a) RingModif (b) CircleModif (c) GolfBallModif

(d) LsunModif (e) TargetModif (f) Simule1

Fig. 1: GOF-SOM

Clustering evaluation

Table 2 presents the results obtained after applying the "Scree Acceleration Test" on databases set. Each GOF value selected represents an outlier cell. We compare the cell selected and the right outlier known, thus for all dataset our algorithm allows to detect the outlier clusters. In the case of simulated database 3, "Scree Test" has selected five group outliers, where 2 groups are subsets of simulated outlier cluster. The other two groups belong to another cluster and the last group is the third outlier cluster. To show the interest of detect outliers groups, we compute the traditional LOF parameter for each database. Particularly for simulated dataset, we observe that the use of LOF can not detect outliers. This disadvantage is known because LOF is based on two principles: the distance between the observation and the density of each observation. Usually in each

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

cluster the distances between observation is small and local densities of data comparing with average densities of their *k-NN* are relatively equal. Thus, in the case of simulated bases, LOF values are almost equal for all observations.

Dataset	#Known GO	#GO selected with "Scree Test"	#GO selected without repetition	Dataset	#Known GO	#GO selected with "Scree Test"	#GO selected without repetition
RingModif	1	1	1	CircleModif	1	1	1
HeptaModif	1	1	1	LsunModif	1	1	1
TargetModif	4	4	4	GolfBallModif	1	1	1
simulated base 1	1	1	1	simulated base 2	2	2	2
simulated base 3	3	5	3	simulated base 4	4	6	4

Table 2: Automatic detection of outliers cells. GO: Cell selected as group outlier using scree Test. Real GO : Outlier known.

To evaluate the quality of map clustering, we adopt the approach of comparing the results to a "ground truth". We use the clustering accuracy, usually named purity measure and Rand index for measuring the clustering results. This is a common approach in the general area of data clustering. Table 3 lists the purity (accuracy) and rand index obtained with GOF-SOM approach and classical SOM. The analysis of the results confirm that the GOF parameter integrated on SOM provides equivalent performance in terms of purity index or rand index. Thus, the introduction of the GOF parameter in the learning phase does not disrupts the classical operation of SOM. We observe that we reach similar performance with an additional parameter GOF that allow to quantify automatically the outlying of groups or clusters.

Databases	Purity index		Rand index	
	GOF-SOM	SOM	GOF-SOM	SOM
RingModif	1	1	0.570	0.570
CircleModif	1	1	0.574	0.573
LsunModif	1	1	0.643	0.642
TargetModif	1	1	0.674	0.674
Hepta	1	1	0.899	0.904
GolfBallModif	1	1	0.150	0.150
simulated base 1	1	1	0.139	0.137
simulated base 2	1	1	0.262	0.256
simulated base 3	1	1	0.675	0.668
simulated base 4	1	1	0.783	0.768

Table 3: Comparison between GOF-SOM and SOM.

4 Conclusion and perspectives

Finding outliers is an important task for many applications. Existing proposals consider being an outlier only for observation. In this paper we show that it is meaningful to consider being an outlier as the degree to which the cluster is isolated from its neighborhood. Thus we introduce a new notion named GOF (Group Outlier Factor), which measure the cluster degree of being outlying. A

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

series of experiments were conducted to validate the proposed method by integrating GOF parameter on Self-Organizing Map. Experimental results demonstrate that our heuristic is promising and identify meaningful outlier-clusters that previous approaches cannot find as the well known LOF. There are many perspectives to study after this results. The first consists on further improving the performance of GOF computation. Second we will use GOF parameter more tightly with clustering in order to use it for novelty detection.

References

[1] M. Breunig, H. Kriege, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. *ACM SIGMOD 2000 International congerence on Management of Data*, 2000.

[2] Q. Cai, H. He, and H. Man. Somso: A self-organizing map appoach for spacial outlier detection with multiple attributes. *Proccedings of International Joint Conference on Neural Networks*, 2009.

[3] Q. Cai, H. He, H. Man, and J. Qiu. Iterativesomso: An iterative self-organizing map for spatial outlier detection. *Proccedings of International Joint Conference on Neural Networks*, 1:325–330, 2010.

[4] R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245–276, 1966.

[5] A. Frank and A. Asuncion. Uci machine learning repository. *Technical report, University of California, Irvine, School of Information and Computer Sciences, available at :http ://archive.ics.uci.edu/ml*, 2010.

[6] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J. Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recogn. Lett.*, 30:994–1002, August 2009.

[7] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.