# Self-Organizing map and Points of Interest

Hanane AZZAG* and Mustapha LEBBAH* and Julien LAVERGNE** and Nicoleta ROGOVSCHI*

**LIPN-UMR 7030 University of Paris 13*
*99, av. J-B Clément - F-93430 Villetaneuse, France*
*hanene.azzag@lipn.univ-paris13.fr and Mustapha.lebbah@lipn.univ-paris13.fr and Nicoleta.rogovschi@lipn.univ-paris13.fr*

***Laboratoire d'Informatique de l'université François-Rabelais de Tours, EA 2101, France*
*Julien.lavergne@univ-tours.fr*

***Abstract* — we propose in this paper a new manner to learn topographic clustering which uses points of interest. This new approach introduces in the learning phase a new concept of referent of interest. These referents are automatically detected during the learning phase or deducted from knowledge of the database. Referents of interest will have more active role in the topological map organization. At the end, we evaluate the performance of our new approach on several databases with different difficulties. The obtained results are encouraging and promising.**

***Index Terms* —Points of Interest, self organizing map, topographic clustering.**

## I. POINTS OF INTEREST AND DATA MINING

The points of interest (POIs) are specific data in database, they represent the points (In this paper we will use the words point and observation in the same way) which are more informative. They are detected automatically or provided by the expert. They represent a particular observations that may give us an idea on the whole information present in the database. When we know this information we can easily generalize some characteristics of these points to other points close to them. The methods based on points of interest are generally used in supervised learning, especially in methods of visualization. Among them are in 2D, we can cite systems VIBE [13], SQWID [16], Radial [3], Radviz [12] and POI2D [6] which has been used to displaying documents provided from search engine. Generally the selected POIS are keywords used in the request and data is the documents generated. We can also cite 3D visualization methods such as VR-VIBE system[4] and POI3D [7].

In Da Costa and Venturini approach [6], authors present an approach based similarity measure to cluster all types of data (numeric, symbolic images, Text, etc.) and which allow the expert, using a visualization based on points of interest (POIs), to label data.

In the supervised way POIS are selected so as to have for each class a representative data. We can also cite the POIS of H. Frigui [8] which proposes to use a pre-processing technique to select automatically and label the points of interest from non labeled data.

The methods based on points of interest are often used in the supervised and semi-supervised case and it is close to the active learning.

Active learning is based on learning methods which allow model to interact with its environment by selecting relevant information [5].

It allows interaction with a human expert to select data. It also permits to build all learning during training [8] [17]. So, active strategies can really accelerate learning by considering the most Information [9] [14].

The unsupervised learning is a complex optimization task that presents several minima local. The problem is more challenging when the data sets are large, noisy, and with limited knowledge. The self-organizing maps are often used for this type of unsupervised learning. The methods based on points of interest have proven their efficiency in several supervised and semi-supervised tasks; for these reasons, we are motivated to show the possibility of their use in the unsupervised case. We are inspired by these systems to detect and use the relevant points in learning of self-organizing map. This new notion of points of interest will be noted in the area of the map references of interest (ROIs: References Of Interest).

## II. SELF-ORGANIZING MAP AND REFERENT OF INTEREST

Self-organizing maps are increasingly used as tools for visualization, as they allow projection over small areas that are generally two dimensional.

The basic model proposed by Kohonen consists on a discrete set $C$ of cells called map. This map has a discrete topology defined by undirected graph; usually it is a regular grid in 2 dimensions. We denote $p$ the number of cells.

For each pair of cells $(c, r)$ on the map, the distance $d(c, r)$ 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; in order to control the neighborhood area, we introduce a kernel positive function $K(K>=0$ *and* $\lim K(x)=0\ x\text{->}\infty\}$ ). We define the mutual influence of two cells $c$ and $r$ by $K(d(c,r))$.

In practice, as for traditional topological map we use smooth function to control the size of the neighborhood as

$$\mathcal{K}(\delta(c,r)) = \exp(\frac{-\delta(c,r)}{T})$$

Using this kernel function, *T* becomes a parameter of the model. As in the Kohonen algorithm, we decrease *T* from an initial value *Tmax* to final value *Tmin*. Let $R^d$ be the Euclidean data space and *(zi; i=1,...N)* a set of observations, where each observation $z\_i=(z^1 i, z^2 i,..., z^d i)$ *is* a continuous vector in Rd. For each cell *c* of the grid, we associate a referent vector $wc=(w^1 c,w^2 c,...,w^j c,...,w^d c)$ of dimension *d*. We denote by *W* the set of the referent vectors. The set of parameter *W* has to be estimated from *A* iteratively by minimizing a cost function defined as follows:

$$\mathcal{J}(\phi,\mathcal{W}) = \sum_{\mathbf{z}_i\in\mathcal{A}}\sum_{r\in C}\mathcal{K}(\delta(\phi(\mathbf{z}_i),r))||\mathbf{z}_i-\mathbf{w}_r||^2$$

Where φ assign each observation *z* to a single cell in the map *C*. We will discuss in this work the problem of automatic detection of referent of interest (ROIs) and their use during the learning process. In this paper the modified topological algorithm will be named SOMROIs \footnote{Self-Organizing Map and Referents of Interest}. Indeed, during the learning process and the assignment phase, each observation is assigned to a cell *c* in *C* which is associated to referent *wc*. When database *A* is assigned, each cell of the map will be characterized by its size (the number of observations collected by cell). All referents of the map are not equivalent and therefore there will be referents more important than others. They will be also used in the learning process to better organize the map topology. All of the ROIs will be denoted by $W^{ROIs}$. It is obvious that the detection of ROIs involves the detection of POIs. Indeed each referent is associated to observation subset "cluster".

For the automatic detection of ROIs in the SOM map, we use the distortion measure (In the real case we have use the *Som-distortion* function of the SOM Toolbox). We take into account the size of cell; the referents which receive more observations are the most important.

In the following we present the algorithm (SOMROIs) we propose and which permits to produce in one step topological organization and ROIS detected automatically.

```
BEGIN SOMROIs
REQUIRE SOM map init, data set D, threshold of distortion,
ENSURE SOM map, set of ROIs
Step1 : ROIs search s
  - Compute distortion using the current neighborhood
  - Select referents which have quantification error lower than threshold distortion
Step 2 : Assignment
FOR{j=1 à N}
  - Z = =D[j]
  - Find the best match unit among a set of ROIs
        $\phi(\mathbf{z}) = \arg\min_{\mathbf{w}_c\in\mathbf{W}^{ROIs}}(\mathbf{z}-\mathbf{w}_c)$
  - Select the referents which are in neighborhood of ROI, V φ(z)
  - Select in this neighborhood V φ(z) the best match unit of referent
        $\Phi(\mathbf{z}) = \arg\min_{c\in V_{\phi(\mathbf{z})}}(\mathbf{z}-\mathbf{w}_c)$
ENDFOR
- Traditional quantization phase which consist to adapt the weight of referent vectors.
END
```

It is clear that in the proposed algorithm in addition to the automatic detection of ROIs, the principal difference with traditional topological maps still the assignment phase.

First, for each observation, *z*, we find the best match ROI which permit us to define an assignment area Vφ (z). This region firstly will be limited to only one referent ROI and grow in the learning process to include other referents.

From the point of view complexity, the assignment phase in traditional topological maps performs throughout the learning phase for each example *p* assignment. With our proposal studies the number of assignment change throughout the learning from $|W^{ROIs}|$ *to* $|W|$ = *p* where *p* is the number of cell. We can note that the referents of interest have a real active role in self-organizing map.

## III. EXPERIMENTATIONS

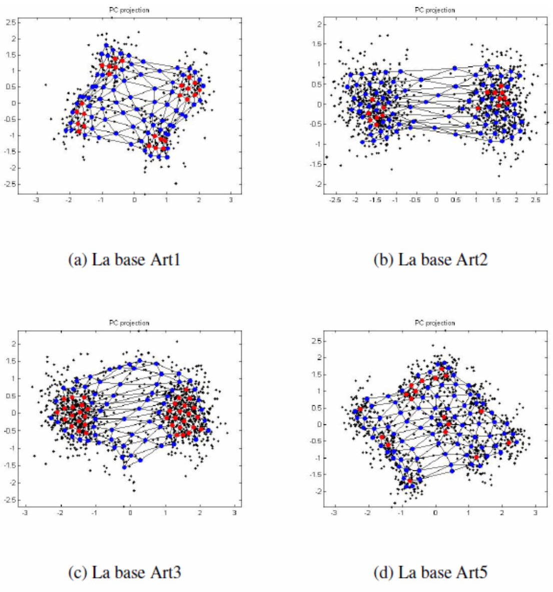
We have evaluated and compared our algorithms on a several databases. The databases ART1 to ART5 are artificial and have been generated with Gaussian and Uniform distributions. The others have been extracted from the machine learning repository [1] [2].

Table I presents for each data set, the actual number of real classes c_*r*, the data dimension *d* and the total number of data *N*.

TABLE I. DATASETS USED IN THE EXPERIMENTATION

| Bases | $C_R$ | $d$ | $N$ |
|---|---|---|---|
| iris | 3 | 4 | 150 |
| Hepta | 7 | 3 | 212 |
| Lsun | 3 | 2 | 400 |
| Tow diamonds | 2 | 2 | 800 |
| WingNut | 2 | 2 | 1016 |
| Art1 | 4 | 2 | 400 |
| Art2 | 2 | 2 | 1000 |
| Art3 | 4 | 2 | 1100 |
| Art4 | 2 | 2 | 200 |
| Art5 | 9 | 2 | 900 |
| Art6 | 4 | 8 | 400 |
| Glass | 7 | 9 | 214 |
| pima | 2 | 8 | 768 |
| Thyroïd | 3 | 5 | 215 |

Figure I show some map 2D projections. The red cells indicate the ROIS and the the remaining cells are in blue. Analyzing projections, we can detect the interesting areas on the map. Indeed in Figure I , the ROIs present connex referents. We observe that the number of clusters indicated by connex ROIs is similar with the real number of clusters.



(a) La base Art1

(b) La base Art2

(c) La base Art3

(d) La base Art5

We present in the table II a comparison between the traditional Kohonen algorithm (SOM) and the SOMROIs. We observe an increase of empty cells for the majority of data set compared to the standard version.

TABLE I. COMPARISON BETWEEN THE SIZES OF SUBSET

| Bases | SOM | SOMROI |
|---|---|---|
| Art1 | 10 | 9 |
| Art2 | 10 | 10 |
| Art3 | 3 | 6 |
| Art4 | 25 | 46 |
| Art5 | 5 | 4 |
| glass | 29 | 35 |
| iris | 38 | 39 |
| Lsun | 22 | 24 |
| pima | 6 | 7 |
| two diamonds | 11 | 13 |
| tyroid | 27 | 49 |
| wingnut | 10 | 10 |

We have also compared the quantization error of the SOMROIs with standard algorithm of SOM Toolbox (table III). We observe that we have the same results for both algorithms which can confirm that our approach works efficiently and provide more information's than a traditional SOM. Table VI provide a comparison of topological error between the standard algorithm (SOM) and the proposed method.

We observe an improvement in the topological error for some data set compared to the standard version. We also observe that some result of topological a bad result; this is due to a distribution measure and the use the ROIs to organize the map.

TABLE III. COMPARISON WITH QUANTIZATION ERROR

| Bases | SOM | ROIs |
|---|---|---|
| Anneaux | 0,027 | 0,065 |
| Art1 | 0,0075 | 0,0025 |
| Art2 | 0,034 | 0,024 |
| Art3 | 0,0327 | 0,04 |
| Art4 | 0,02 | 0,03 |
| Art5 | 0,0011 | 0,0022 |
| atom | 0,04 | 0,08375 |
| Engytime | 0,0183 | 0,0280 |
| glass | 0 | 0,0280 |
| golfball | 0,0582 | 0,0657 |
| iris | 0,0333 | 0 |
| Lsun | 0,0375 | 0,055 |
| pima | 0,0325 | 0,06380 |
| target | 0,0454 | 0,003 |
| tetra | 0,030 | 0,0375 |
| two diamonds | 0,0062 | 0,00625 |
| tyroid | 0,0186 | 0,0186 |
| wingnut | 0,0187 | 0,0137 |

TABLE VI. COMPARISON WITH TOPOLOGICAL ERROR

| Bases | SOM | ROIs |
|---|---|---|
| Anneaux | 0,1584 | 0,1639 |
| Art1 | 0,2423 | 0,24 |
| Art2 | 0,1659 | 0,1641 |
| Art3 | 0,1983 | 0,1995 |
| Art4 | 0,1458 | 0,1471 |
| Art5 | 0,2239 | 0,2306 |
| atom | 0,3374 | 0,3425 |
| Engytime | 0,1827 | 0,1808 |
| glass | 1,0668 | 1,0902 |
| golfball | 0,2763 | 0,2827 |
| iris | 0,3675 | 0,3841 |
| Lsun | 0,1389 | 0,1389 |
| pima | 1,5023 | 1,5216 |
| target | 0,1946 | 0,1918 |
| tetra | 0,4029 | 0,4131 |
| two diamonds | 0,1468 | 0,1482 |
| tyroid | 0,5736 | 0,5996 |
| wingnut | 0,1664 | 0,1699 |

## IV. CONCLUSION

In this paper we have presented a new approach for unsupervised learning of the topological map, which is based on points of interest or more precisely referents of interest (ROIs). We have presented the detailed algorithm and its process (SOMROIs). The obtained results have been compared to those found by the traditional Kohonen algorithm (SOM). This comparison confirmed that the proposed approach is promising and can be used in various applications of data mining. Our model presents in a single phase a hierarchical view of data (the ROIS, then the topological map with the associated prototypes and the data). There are several perspectives with this work. First we must find other criteria which allow us to better characterize and detect automatically the referents of interest. Secondly we need to improve the efficiency of computation time of our algorithm by optimizing the Matlab code.

## REFERENCES

[1] Asuncion, A. et D. Newman (2007). UCI machine learning repository.

[2] http ://www.ics.uci.edu/_mlearn/MLRepository.html.

[3] Au, P., M. Carey, S. Sewraz, Y. Guo, et S. M. Ruger (2000). New paradigms in information visualization. In ACM SIGIR2000, (ACM, pp. 307–309. Press.

[4] Benford, S., D. Snowdon, C. Greenhalgh, R. Ingram, L. Knox, et C. Brown (1995). VR-VIBE : A virtual environment for co-operative information retrieval. j-CGF 14(3), 349–360.

[5] Cebron, N. et M. R. Berthold (2007). An adaptive multi objective selection strategy for active learning. Technical report.

[6] COSTA, D. D. (2007). Visualisation et fouille interactive de données à base de points d'intérêt. Ph. D. thesis, Polytech'Tours, Tours, France.

[7] Costa, D. D. et G. Venturini (2006). Visualisation interactive de données avec des méthodes à base de points d'intérêt. In EGC, pp. 335–346.

[8] Frigui, H. (2005). Unsupervised identification of points of interest for semi-supervised learning. In Fuzzy Systems, 2005, pp. 91–96. IEEE Computer Society.

[9] Hasenjäger, M., H. Ritter, et K. Obermayer (1999). Active data selection for fuzzy topographic mapping of proximities. In G. Brewka, R. Der, S. Gottwald, et A. Schierwagen (Eds.), Fuzzy Neuro Systems 99, Leipzig, pp. 93–103. Leipziger Universitätsverlag.

[10] Hemmje, M. (1995). Lyberworld : a 3d graphical user interface for fulltext retrieval. In CHI 95 Conference Companion, pp. 417–418.

[11] Hemmje, M., C. Kunkel, et A. Willett (1994). Lyberworld - a visualization user interface supporting fulltext retrieval. In W. B. Croft et C. J. van Rijsbergen (Eds.), SIGIR, pp. 249– 259. ACM/Springer.

[12] Hoffman, P., G. G. Grinstein, et D. Pinkney (1999). Dimensional anchors : A graphic primitive for multidimensional multivariate information visualizations. In Workshop on New Paradigms in Information Visualization and Manipulation, pp. 9–16. ACM.

[13] Korfhage, R. (1991). To see, or not to see : Is that the query ? In A. Bookstein, Y. Chiaramella, G. Salton, et V. V. Raghavan (Eds.), SIGIR, pp. 134–141. ACM.

[14] Mazzoni, D., K. Wagstaff, et M. C. Burl (2006). Active learning with irrelevant examples. In

[15] J. Fürnkranz, T. Scheffer, et M. Spiliopoulou (Eds.), ECML, Volume 4212 of Lecture Notes in Computer Science, pp. 695–702. Springer.

[16] Mccrickard, D. S. et C. M. Kehoe (1997). Visualizing search results using sqwid. In In Proceedings of the Sixth International World Wide Web Conference.

[17] Nguyen, H. T. et A. W. M. Smeulders (2004). Active learning using pre-clustering. In ICML.

[18] Vesanto, J., M. Sulkava, et J. Hollmén (2003). On the decomposition of the self-organizing map distortion measure. In Proceedings of the Workshop on Self-Organizing Maps (WSOM'03,