Clustering through SOM Consistency

Nicolau Gonçalves and Ricardo Vigário

Department of Information and Computer Science,
Aalto University School of Science,
P.O. Box 15400, FI-00076 Aalto, Finland
{firstname.lastname}@aalto.fi

Abstract. Clustering is a classical tool in image analysis, with wide applications. Yet, most of its algorithmic solutions include a considerable amount of stochasticity, *e.g.* due to different initialisations. Here, we introduce a clustering method rooted on self organizing maps, that exploits the maps' intrinsic variability, to produce reliable clustering. Although only a subset of the data is consistently clustered, we show that this set is trustworthy, and can be used for posterior classification.

1 Introduction

Data clustering is widely used in statistical data analysis. Its applications include data mining, image analysis and bio-informatics. Clustering means partitioning the data into different groups, clusters, so that in each group similarity can be found, whereas differences exist across clusters. This shared similarity is typically measured by some defined distance.

Clustering methods are typically based either on competitive learning [1], or statistical model identification [9]. In competitive learning based clustering, first the parameters are adjusted through learning. After this step, the network is ready for generalisation. A common example of competitive learning is k-means, where the objective is to find cluster centres and assign the data to the nearest cluster centre. In the case of k-means, the learning occurs by minimizing the squared distances inside each cluster.

Statistical modeling methods assume that a mixture of underlying probability distributions generates the data. Parameter estimation can then be performed, *e.g.* through maximisation of the log-likelihood function. An advantage of using a statistical model is that the choice of the clustering criterion is less arbitrary than in competitive learning and the approach includes rigorous statistical tests. Yet, the definition of the used model is not always easy to set.

On most methods, noise often renders the identification of the true clusters difficult. Furthermore, the algorithms may approach the solution via different paths, depending on the algorithm's initial conditions. To solve this problem, one can run the algorithm multiple times, making sure that the starting conditions differ in each run. Then, using a measure of intra-cluster variance, the best clustering result can be identified.

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part I, LNCS 7324, pp. 61–68, 2012.

Another approach is to use the different results of each clustering run, together with their variability [13]. We propose a clustering method based on an analysis of consistency in self-organizing maps (SOM,[6]). Such analysis is particularly relevant for an algorithm such as SOM, with its intrinsic stochastic nature and dependence on the initialisations. Because of its ease of use, as well as ability to efficiently map high-dimensional data into a 2D lattice, SOM has been widely used in many applications, with over 10.000 published papers ([7]). In our approach, several maps are built, and the clustering consistency assessed to produce a set of overall reliable clusters.

2 Methods

Self-organising Maps

SOM may be formally described as a non-linear, ordered, smooth mapping of high-dimensional input data manifolds onto the elements of a regular, low-dimensional array. It performs a lattice projection that preserves similarity information in the input space, through competitive learning with an Hebbian learning rule [8]. After training, the result is a topographic map representing the input patterns. In this map, similar input patterns are represented by neurons that are close in the SOM space.

To produce quantitative descriptions of data properties, interesting groups of map units, *i.e.* clusters must be identified among the local minima of the SOM [11].

With different random initialisations, the SOM representations of the data may vary. Therefore, data points sharing the same cluster in a given run of SOM may be projected to different clusters in another. We define as consistent clusters those comprising elements that are grouped together in a large number of runs.

To better illustrate this concept, we show three possible runs using toy-data in Fig. 1. On the leftmost example, the numbered circles are all in the same cluster. In the other two cases, circle 3 appears in different clusters from the other two. Consistency in clustering membership, such as 1 and 2, is indicative of a good class estimation. Circle 3 has too much variability in its clustering grouping to represent the same class. Note that these changes in clustering assignments are typical for different runs of stochastic algorithms.

In our experiments, we trained one hundred SOMs, using the SOM Toolbox [2]. The algorithm was run in batch mode, with random initialisations for each run. To further increase the variability in the clustering conditions, the dimensions of the maps were also changed in every run.

Consistent Clusters

To analyze the consistency of the various clusters, one may use a variety of distance measures, and subsequently join redundant clusters [12]. We used two such measures, one based on the elements included in each cluster, and another on the value distributions of the elements belonging to different clusters.

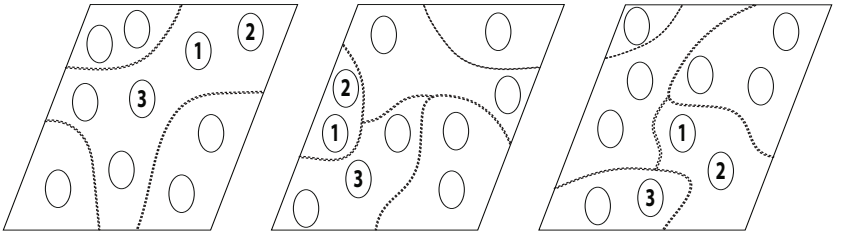

Fig. 1. Clustering results for three different SOM initialisations, using toy data. The circles represent different data points, while the dashed lines define the clusters. Circles 1 and 2 always appear together for different SOM runs, representing one same class. Circle 3 won't be picked to represent any class due to the variability of its clustering.

The first measure is defined as

$$d_{ij}^1 = \frac{1}{N} \frac{\sum_{n=1}^{N} c_i^n \wedge c_j^n}{\sum_{n=1}^{N} c_i^n \vee c_j^n}, \tag{1}$$

where **c** represents a cluster with N elements and {i,j} represent cluster indices; c_l^n, for $l = i, j$, is 1 if element n belongs to cluster l, and 0 otherwise. N is the total number of data elements. $\wedge$ and $\vee$ are the AND and OR logic operators, respectively.

We joined clusters that exceeded a heuristically selected threshold of $d_{ij}^1 > 0.8$. This value does not depend on the dataset, and its purpose is to remove redundant clusters.

After this first merging analysis, we proceed to a second stage of grouping. In this stage, the clusters are compared using their data distribution:

$$d_{ij}^2 = \sqrt{(\overline{\mathbf{g}}_i - \overline{\mathbf{g}}_j)^2}, \tag{2}$$

where

$$\overline{\mathbf{g}}_l = \begin{pmatrix} \mathsf{E}\{\mathbf{X}\} \\ \mathsf{Var}\{\mathbf{X}\} \end{pmatrix}.$$

$\mathbf{X}$ represents the values of the elements belonging to $\overline{\mathbf{g}}_l$.

During this second stage, clusters with $d_{ij}^2 > .9$ are merged. This value was heuristically selected to allow for some variability, which accounts for noise and other artefacts in the observed data. A higher value would increase the number of clusters obtained, while a smaller value would group together clusters that might represent different classes.

The two aforementioned measures can not be combined into one due to their different points of application. The first deals with element-cluster assignments, whereas the second uses distribution information.

Fig. 2. Two data-sets used for experimental illustrations of the proposed method. Only one height is shown for both data, although the experiments were done for the full brain volume.

3 Experiments

Simulated Data

Synthetic data from the BrainWeb database [3,4] was used to test the behaviour of the clustering algorithm in different settings, see Fig. 2(a). It comprised a normal phantom with labels for cerebrospinal fluid (CSF), grey matter (GM) and white matter (WM). The image slices, with $1mm$ of thickness, were generated from 3 different sequences (T2, PD and T1). The images were available with and without magnetic field inhomogeneity. Different levels of noise (0/1/3/5/9% of the maximum image intensity) were used for each homogeneity condition.

The ground truth for the simulated data was readily available from the BrainWeb database, see Fig. 3(a). This allowed for a thorough evaluation of the clustering results, although this information was not used for clustering.

For clarity, only the visual results for the BrainWeb set with no inhomogeneity and 5% noise level are shown in Fig. 3(b). All other results can be found in the summarizing tables.

The SOM map dimensions changed in every run according to a bi-dimensional normal distribution with a mean of 20 and a variance of 2.

The consistent results shown in Fig. 3(b) correspond to around 60% of the total amount of voxels present in the images, with minimal erroneously classified voxels (less than 2%), see Table. 1. Table. 2 gives a discriminated percentage of how many voxels are correctly identified, for each tissue, when compared to the ground truth. Consistent voxels for CSF and WM are usually more than 80% of the total number of voxels of that tissue. The addition of inhomogeneity and/or noise typically results in a concomitant decrease in the number of voxels found,

especially in the case of GM and WM. This effect is not so evident in CSF since this class is located mainly in areas not affected by inhomogeneity and has a high intensity uniformity.

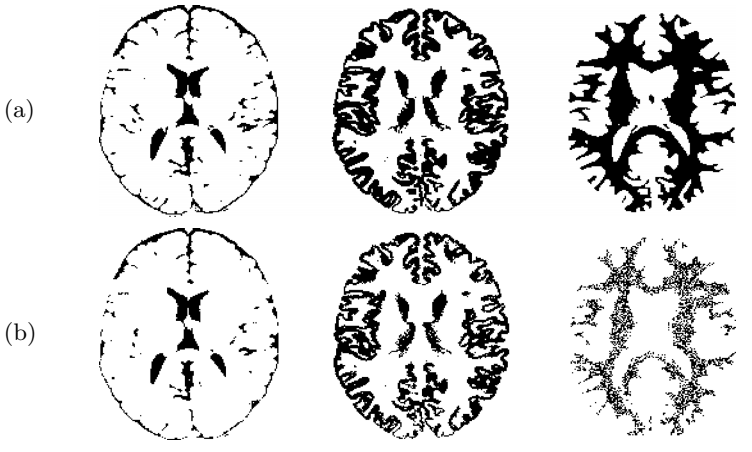

Fig. 3. Results for the segmentation using the simulated data, 5% noise level and no inhomogeneity. The first row shows the ground truth. The clusters found are shown in the second row. The classes are CSF, GM and WM, from left to right respectively.

Table 1. Percentage of true positives of the consistent clustering for the simulated data-set with different levels of noise, and presence of a 20% inhomogeneity field

Noise %	0		1		3		5		9	
Inhomogeneity	no	yes	no	yes	no	yes	no	yes	no	yes
CSF	100%	99%	99%	99%	97%	96%	97%	98%	98%	97%
GM	100%	100%	100%	100%	100%	100%	98%	98%	96%	96%
WM	100%	100%	100%	100%	99%	99%	99%	99%	98%	90%

Grand Challenge Data

We also tested our approach on real data, from the Grand Challenge II [10], illustrated in Fig. 2(b). The images used were acquired with a 3T Siemens scanner (scan parameters: axial plane, FOV $250mm$, matrix 512x512, slice thickness: $5mm$, interslice gap $0.5mm$) and included FLAIR, T2 and T1 sequences. Several foci of lesion can be clearly seen in the FLAIR sequence, as its brightest voxels, near the ventricles and inside the white matter. We only show results for the multiple sclerosis (MS) lesion class, since this is the only one with segmentation ground truth, defined by 2 different expert raters.

Table 2. Total number of ground truth voxels in parenthesis, and percentage of those found to be consistent in the same conditions as in Table. 1

Noise %	0		1		3		5		9	
Inhomogeneity	no	yes	no	yes	no	yes	no	yes	no	yes
CSF (154110)	93%	91%	92%	92%	95%	95%	80%	72%	83%	70%
GM (483145)	61%	15%	48%	30%	30%	20%	53%	61%	45%	72%
WM (493222)	86%	83%	84%	63%	84%	84%	58%	56%	57%	47%

Segmentation results for the Grand Challenge dataset are displayed in Fig. 4. One should note that there is a clear difference between the number of lesion voxels and that of other tissues. Such difference is likely to affect the clustering performance of SOM.

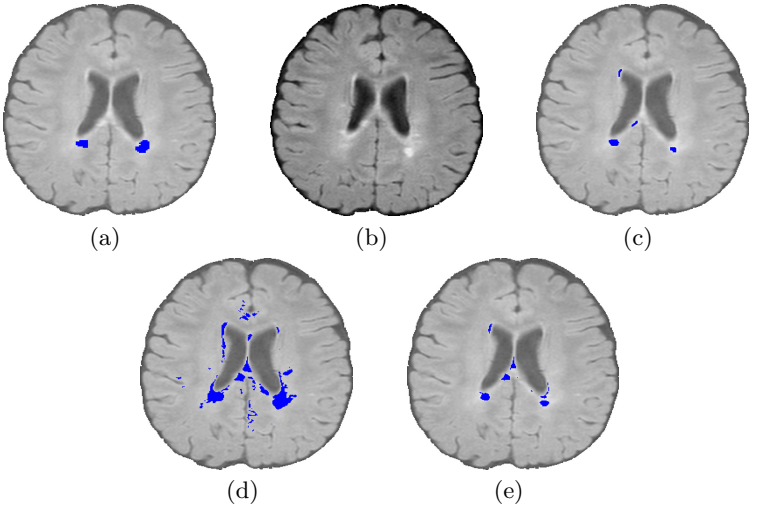

Fig. 4. Clustering found for MS lesion in the Grand Challenge data. Figures 4(a) and 4(c) correspond to the ground truth, as rated by two different annotators. Figure 4(b) shows the original FLAIR image. Figure 4(d) shows the result of our consistent clustering, while Fig. 4(e) represents the refined clustering obtained from the estimate shown in Fig. 4(d).

The process of annotating real structural MR images is often subjective. Comparing the sub-figures (a) and (c) in Fig. 4, corresponding to two independent expert annotators, it is clear that both diverge in what they consider to be lesion. If one defines as total ground truth all voxels identified by both annotators (0.6% out of 3188536 total voxels), the lesion selection shown in Fig. 4(a) missed 33%, while that of Fig. 4(c) missed 53%. Applying our method to the data,

see Fig. 4(d), the number of true positives only reached 9%. Many voxels that have similar intensity to lesion were considered lesion by the method.

As noted earlier, the number of lesion voxels is very small, rendering the clustering process difficult. To circumvent such limitation, we refined our study by applying a second time the proposed method, now only applied to our first estimate. This procedure, displayed in Fig. 4(e), resulted in missing only 44% of the lesion, which is in line with the annotators' mismatches.

4 Discussion

Our method aims at contributing to the robustness and consistency of self-organizing maps. It allows to obtain reliable clusters in the data, with high internal similarity.

To show the relevance of a consistency study, we can compare the results of one run with the ones from the consistent runs. For the simulated data with 5% noise case and no inhomogeneity, a typical result of one SOM is 64/84/78% of true positives for CSF, GM and WM respectively, with 5% misclassification. Our methodology improves both the misclassification and the number of true positives found.

The method can also be applied to other stochastic clustering algorithms, such as k-means. Using the aforementioned data, a typical k-means result is 64/84/96% of true positives for CSF, GM and WM respectively, with more than 10% erroneously clustered voxels. Using consistent k-means the misclassification is reduced to less than 1%, while the number of true positives stays the same. These results are better than those of a single run, but worse than the ones obtained with consistent SOMs.

Although promising, as observed from the results shown in this article, not all data points can be clustered. Therefore a subsequent classification method can be used to extend the results to the whole dataset. This classification can then use the clusters found through our method as reliable labels. One particular example, using the same simulated data set, can be found in [5].

When using our method in unbalanced data, like the Grand Challenge dataset, the clusters obtained are not as reliable. Clustering relies on clear differences between groups of data. If this distance is not significant, and the number of available sample of one of those classes is clearly insufficient to represent it, clustering will suffer. When compensating for this, by using a smaller number of voxels, the method performs remarkably well. In the Grand Challenge data, the results obtained in this smaller set are comparable with the ones from the annotators. Even the topographic locations and intensity values of the voxels detected are in line with the voxels selected manually.

We have observed in own experiments that small clustering errors can still be compensated during classification, if a sufficient amount of consistent labels exist.

Acknowledgements. Nicolau Gonçalves was funded by grant number SFRH/BD/36178/2007 from Fundação para a Ciência e Tecnologia.

References

1. Ahalt, S., Krishnamurthy, A., Chen, P., Melton, D.: Competitive learning algorithms for vector quantization. Neural Networks 3(3), 277–290 (1989)
2. Alhoniemi, E., Himberg, J., Parhankangas, J., Vesantoen, J.: SOM Toolbox, http://www.cis.hut.fi/projects/somtoolbox/ (visited October 2011)
3. Cocosco, C., Kollokian, V., Kwan, R.S., Evans, A.: BrainWeb: Online interface to a 3D MRI simulated brain database. In: NeuroImage (Proceedings of 3rd International Conference on Functional Mapping of the Human Brain), Copenhagen, vol. 5(4, part2/4), p. S425 (May 1997), http://www.bic.mni.mcgill.ca/brainweb/
4. Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C.: Design and construction of a realistic digital brain phantom. IEEE Trans. Med. Imaging 17(3), 463–468 (1998)
5. Gonçalves, N., Nikkilä, J., Vigário, R.: Partial Clustering for Tissue Segmentation in MRI. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008. LNCS, vol. 5507, pp. 559–566. Springer, Heidelberg (2009)
6. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer (2001), Huang, T.S., Kohonen, T., Schroeder, M.R. (eds.)
7. Laaksonen, J., Honkela, T. (eds.): WSOM 2011. LNCS, vol. 6731. Springer, Heidelberg (2011)
8. MacKay, D.J.C.: Information Theory, Inference & Learning Algorithms, 1st edn. Cambridge University Press (2002)
9. McLachlan, G.J., Basford, K.E.: Chapters 1 and 2. In: McLachlan, Basford (eds.) Mixture Models: Inference and Applications to Clustering, pp. 1–69. Marcel Dekker, Inc. (1988)
10. Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S.: 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. In: MICCAI 2008 Workshop, MIDAS Journal (September 2008), http://hdl.handle.net/10380/1509
11. Vesanto, J., Alhoniemi, E.: Clustering of the Self-Organizing Map. IEEE Neural Networks 11(3), 586 (2000)
12. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 1073–1080. ACM, New York (2009)
13. Ylipaavalniemi, J., Vigário, R.: Analyzing consistency of independent components: An fMRI illustration. NeuroImage 39(1), 169–180 (2008)