2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Human-like Perception Using Sequential Superparamagnetic Clustering

Thomas Ott and Ruedi Stoop

Institute of Neuroinformatics, ETH/UNIZH Zurich, Switzerland
Email: tott@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—For coping with the emergent properties of multi-component systems (embracing multi-circuit and multi-sensory systems), macroscopic states need to be identified. This can be achieved by means of clustering algorithms. Using a nontrivial two-dimensional toy system, we comparatively explain the advantages of our recently developed sequential superparamagnetic clustering approach. The ability to provide 'natural', i.e. intrinsically unbiased, system states, excels our algorithm above the standard methods, such as K-means or linkage methods. It is this ability that provides human-like perception.

1. Introduction

Clustering is a fundamental ingredient of cognition and particularly important to, nowadays, prominent fields of science dealing with multi-component, and multi-sensory, systems. Recent applications include the fields of robotics [1], chemoinformatics [2] and bioinformatics [3]. The ultimate goal of clustering is to find natural classes of similar items in a given set, where, usually, no information about number of classes or class sizes is available. This search for natural classes is complicated by a number of factors: (**a**) Most clustering approaches are based on a similarity measure, that discriminates between similar and dissimilar data points. The problem of how to choose an optimal similarity measure is highly nontrivial (it will, however, not be discussed here.) (**b**) Sometimes, some items will belong to several classes. This aspect can be handled with particular care by using fuzzy logic. (**c**) Results of clustering have an inherent branching nature. A priori, it is not clear which clustering resolution provides the most natural, or useful, classes.

Among the variety of clustering approaches, two methods are most widely used: The linkage methods family (championed by Ward clustering [2]) and the K-means method. Both methods, however, have obvious disadvantages. In K-means clustering, the number of classes has to be provided a priori. The results also depend on the initial placement of the cluster centers. Ward's method, on the other hand, provides a cluster hierarchy (dendrogram), leading to the problem of finding the appropriate resolution level. Both methods often fail to recognize clusters of complex shape, as they are biased to recognize hyper-sphere clusters [2]. Furthermore, nonuniform cluster densities are a problem in both approaches, since it can be impossible to find a global clustering level that provides the correct classes (for non-trivial systems).

Our recently introduced sequential superparamagnetic clustering (SSC) method [4] does not suffer from the disadvantages of the other methods. It neither requires any a priori information about the number of clusters, nor does it have a cluster shape bias. Moreover, it provides us with an intrinsic criterion for the identification of natural classes. This allows us to find the most natural clustering resolution by using locally optimised levels. A performance comparison of this approach with the standard methods, using the example of fingerprint-coded chemical compounds, has clearly demonstrated the power and superiority of our method [4].

In this contribution, we first provide a short introduction to the SSC algorithm (for a detailed exposition of the method, we have to refer to [4]). By comparing to K-means and Ward's clustering, using a nontrivial two-dimensional toy model, we give an intuitive understanding of the advantages of SSC. Finally, as a real world application, we enter the field of bioinformatics where we extract the natural classes from the mitochondrial genomes of 20 mammals. By using the Li sequence similarity measure [5], we obtain the most stable branches of the corresponding phylogeny tree.

2. Sequential Superparamagnetic Clustering (SSC)

Superparamagnetic Clustering (SC), which was introduced in [6], implements clustering as a self-organised process on an inhomogeneous Potts spin system. The sites of the system are given by the data points to be clustered, and a Potts spin variable s_i with $s_i \in \{1, ..., q = 10\}$ [1] is assigned to each site x_i. The Potts spins interaction is described by the Hamiltonian

$$H(S) = \sum_{(i,j)} J_{ij}(1 - \delta_{s_i s_j}). \tag{1}$$

The coupling strength J_{ij} between two sites x_i and x_j is a a decreasing function of the distance $|x_i - x_j|$, considering only the k nearest neighbor sites. The probability to find the system in a certain configuration S is determined by the canonical probability

$$p(S) = \frac{1}{Z} e^{-H(S)/T}, \tag{2}$$

[1] The choice of the q-value does not effect the clustering results [4].

where the partition function $Z = Z(T)$ serves as a normalisation factor. The temperature T acts as a control parameter expressing the average energy of the system. As T is increased, the systems typically undergoes a number of phase transitions: (**I**) For small T, the system is in a ferromagnetic phase, where spins are likely to be aligned. (**II**) For an intermediary T-range, a superparamagnetic phase occurs. Strongly coupled spins tend to be aligned, whereas weakly coupled spins behave independently. Thus, clusters of aligned spins reflect the regions of similar data points. By further increasing T, these clusters generally cascade into smaller clusters, so that a hierarchy of classes and subclasses is obtained. (**III**) For high T, the system enters the paramagnetic phase, where any order disappears and only singleton clusters remain.

Clustering aims at the superparamagnetic phase, where clusters emerge that are stable over large temperature intervals. Among the data points, clusters can be identified via the pair correlation criterion: Two points x_i and x_j belong to the same cluster, if the pair correlation G_{ij} exceeds a threshold Θ,

$$G_{ij} = \sum_{S} p(S)\delta_{s_i s_j} > \Theta, \tag{3}$$

where Θ can be freely chosen from the interval $[1/q, 1 - 2/q]$. To calculate G_{ij}, we need to scan the configuration space for a series of temperatures $T = \{T_{min}, T_{min} + \Delta T, ..., T_{max}\}$. For that purpose, a Monte Carlo approach is performed for each T. We use the Swendsen-Wang algorithm [6], as this algorithm is able to sweep a representative subset of configurations without getting stuck in local energy minima. Stable results are obtained within 200 Monte Carlo steps.

As we have determined the clusters for different temperatures, we get back to the central question: How can the most natural classes be identified? Natural classes (as visibly discernable in Figure 1) are stable over whole ranges of temperature. Usually, dense clusters are stable over broader ranges of T, but sometimes they occur for high T only, in particular when two clusters are too tight to be separated at smaller temperatures. In the worst cases, such superclusters abruptly break up into singletons, without going through an intermediate state. Less dense clusters generally only occur for small T. It may even happen that they do not emerge at all, especially when the coupling is too short-ranged (e.g., only two nearest neighbors couplings).

Data sets where the mean distances within some clusters matches the distance between some of the clusters, are most problematic. It is not absolutely obvious, which temperatures are best for the identification of clusters and how to choose the coupling between the sites (i.e. the number of nearest neighbours). In this case, as was stated in the introduction, only a local resolution, i.e., a local variation of T, is able to solve the problems. As T is a global parameter, characterising the thermal equilibrium of the system, in order to achieve a locally regulated resolution, we introduce a sequential procedure.

For this procedure, we first cluster the original set using SC. Then we extract the most stable cluster, i.e., the cluster which is stable over the broadest temperature range. After extraction, this cluster, as well as the residual set, are clustered by SC separately. Again, the most stable clusters are extracted and so on. As a result, we obtain sequences of sets of increasing homogeneity. This allows to extract the classes that appear most natural in their local context. Natural classes are stable over a broad T-range, but do not show stable sub-clusters over a broad range. Thus, the procedure provides us with a good criterion for natural classes. The procedure also has the advantage that the results are largely independent of the number of coupled neighbours [4]. For typical data sets, the time complexity of the algorithm is as for the Ward's method ($O(N^2)$).

3. Toy System

Figure 1 poses a two-dimensional nontrivial clustering problem.

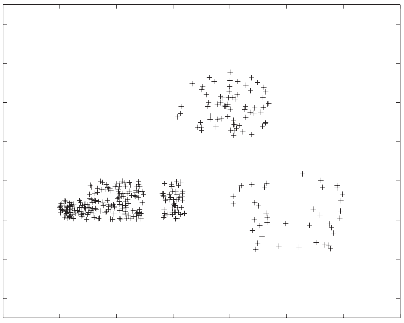


Figure 1: A two-dimensional nontrivial clustering problem.

By construction, four clusters are present. One dense cluster consists of 180 points and has a nontrivial shape (left). Another dense cluster, made of 50 points, is close to the first cluster. A less dense cluster made up of 70 points, extends towards the positive y-direction. Another sparse cluster of 40 points extends towards the lower right corner. Based on a randomly drawn rectangular distribution, it could possibly be subdivided into a left and a right subcluster, making clustering ambiguous to some degree. Additionally, there are two more points that pose severe problems: Firstly, the first cluster has a nontrivial shape, and, secondly, the gap between the two dense clusters is of the order of a typical distance between points of a less dense cluster. This last point indicates that a local dependence of the clustering resolution is necessary for successful clustering. In the following, we compare the clustering results of our algorithm to results of Ward's method and K-means clustering, where the latter are provided by a Matlab code (adapted by the authors).

Ward’s method cannot correctly distinguish among the classes. At no stage of the hierarchy, do the four natural clusters coexist. (Figure 2). In fact, the major problem of the algorithm is due to the shape, and the neighbourhood, of the dense clusters. Ward’s clustering subdivides the big cluster into two units at a very early stage (Figure 3). Even when Ward’s method is capable of clustering correctly, to find a natural level criterion is an unresolved issue.

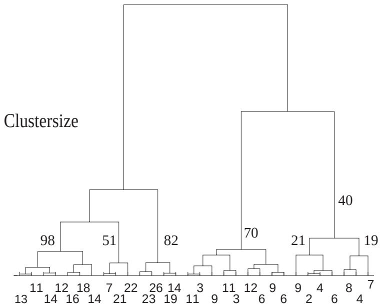
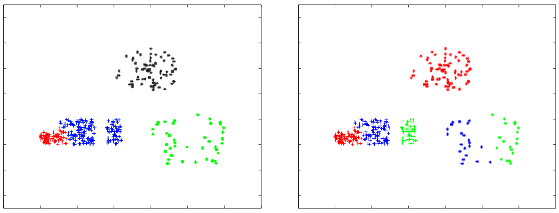


Figure 2: Dendrogram for Ward’s clustering method. The numbers indicate cluster sizes.



(a) Four clusters stage

(b) Six clusters stage

Figure 3: Clusters defining the four and six clusters stage of Ward’s clustering.

For judging K-means clustering, we have to assume that the number of clusters K is known. Still, we are unable to identify the four most natural clusters. K-means clustering also fails when dealing with the shape and the close neighbourhood of the dense clusters (Figure 4). A larger K would rather split up a sparse cluster than separate the two dense clusters. The solution for four clusters is independent of the cluster centre initialisation. Nonetheless, it is optimal in the sense of K-means clustering, demonstrating the principal shortcoming of this method.

SSC, however, is able to exactly extract the four clusters that intrinsically are most natural (Figure 5). As can be seen in Figure 6, the four clusters are already present at $T = 0$, when the entire data set is clustered (no ferromagnetic phase). This demonstrates that superparamagnetic clustering easily deals with clusters of arbitrary shape, by establishing a global (spin) order on the basis of local interaction (via coupling). However, due to the different densities of the clusters, their stability against an increase in T is an individual property, making less dense clusters marginally stable only. In the run of the sequential clustering procedure, these clusters are rendered ever more stable (as they are no longer compared to other denser clusters, but to their local background). Finally, they become easily detectable as natural classes.

Figure 5b shows the typical dendrogram achieved by SSC. The earlier a cluster extracted, the more compact it is. Each extracted cluster is iteratively clustered, to detect potential substructures. For the thin cluster of 40 points, two possible sub-clusters can be detected. Their T-stability provides the information on how ‘natural‘ these clusters are. In the present case, they are rather unstable. As a consequence, it is natural to consider the cluster a whole natural entity. The same holds for the detection of marginally stable subclusters within the other detected classes.

4. Mitochondrial Genomes

In the problem of phylogeny, the goal is to determine the evolutionary relationships among species. On the one hand, the construction of phylogeny trees allows the re-

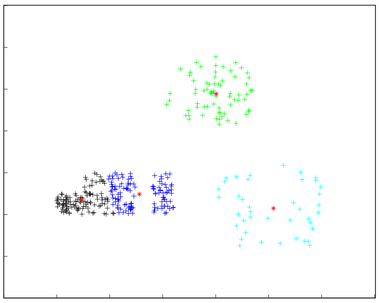
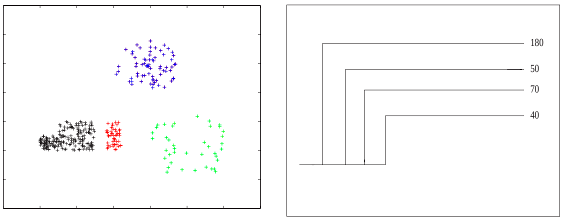


Figure 4: Solution for K-means clustering, assuming 4 clusters. *‘s denote the final position of the cluster centres.



(a) Clusters obtained by SSC

(b) SSC-dendrogram.

Figure 5: Results for SSC.

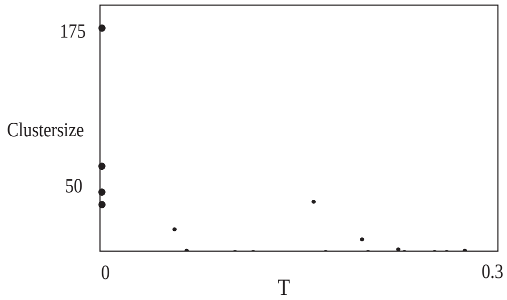


Figure 6: Cluster-size diagram for the first step performed by SSC, where the size of occurring clusters is drawn in dependence of T. Larger dots mark the four initial clusters.

construction of the evolutionary career of species, on the other hand, identifying natural classes helps to identify preserved cousinhoods among species. A challenging question is how to appropriately compare two genomes, i.e., how to define the similarity measure. Li et al. [5] introduced an information-based measure with metric property, that appropriately handles complete genomes. For two sequences x, y they define the distance $d(x, y)$ as

$$d(x, y) = 1 - \frac{K(x) - K(x|y)}{K(xy)}, \tag{4}$$

where $K(x|y)$ is the conditional Kolmogorov complexity, $K(x) = K(x, \{\})$ and xy denotes the concatenation of the two strings. $K(x|y)$ is the length of the shortest program on an universal computer with input y and output x. As such, it is not computable, but an upper bound can be found using the Lempel-Ziv algorithm [7]. Compression programs (such as GENCOMPRESS[3]) implement this approach.
To demonstrate the efficacy of SSC, we took the distance matrix of the mitochondrial genomes of 20 mammals (provided by [5]). Figure 7 summarises the obtained 'natural' classes. Note that the SSC-dendrogram does not directly provide a phylogeny tree, but the most natural classes in order of their inherent compactness. A phylogeny tree between classes could be constructed afterwards. As the data set is rather small, this application is more of a toy system. Real world applications would aim at much bigger data sets, for instance, of bacterial genomes.

5. Conclusion

To summarise, sequential superparamagnetic clustering is an exceptionally reliable tool for clustering analysis. In all nontrivial problems investigated, it yielded far better results than the standard methods. The method intrinsically provides a criterion for an unbiased extraction of the most natural classes, by determining the best clustering resolution on a semi-local scale. In addition, the system temperature T can be seen as an attention, or resolution, parameter, which places clustering within the general context of human, or artificial, cognition. In fact, identifying 'natural' classes is a vital aspect to order perceptional data. Like human beings, the presented method is able to find such classes in an unsupervised manner and gives a first and quick insight into the perceptual scene. However, once identified, the classes could possibly be further subdivided by increased attention or resolution. Thus, this clustering method mimics human perception.

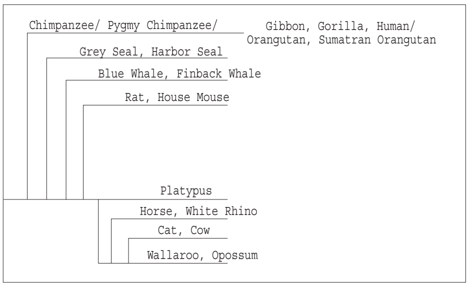


Figure 7: SSC-dendrogram for the genomes set.

References

[1] J.-J. van der Vyver, M. Christen, N. Stoop, T. Ott, W.-H. Steeb, R. Stoop, "Towards genuine machine autonomy," *Robotics and Autonomous Systems*, vol.64(3), pp.151–157, 2004.

[2] G. M. Downs, J. M. Barnard, "Clustering methods and their uses in computational chemistry," *Reviews in Computational Chemistry*, vol.18, pp.1–40, 2002.

[3] Y. Zhao, G. Karypis, "Clustering in life sciences," *Methods Mol Biol.* vol. 224, 183–218, 2003.

[4] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby, R.Stoop, "Sequential superparamagnetic clustering for unbiased classification of high-dimensional chemical data," *J. Chem. Inf. Comput. Sci.*, (in press), 2004.

[5] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol.17(2), pp.149–154, 2001.

[6] M. Blatt, S. Wiseman, E. Domany, "Superparamagnetic clustering of data," *Phys. Rev. Lett.*, vol.76, pp.3251–3254, 1996.

[7] W.-H. Steeb, R. Stoop, "Exact complexity of the logistic map," *Int. J. Theor. Phys.* vol. 36, 943–946, 1997.