Cost functions for pairwise data clustering

Leonardo Angelini, Mario Pellicoro, Sebastiano Stramaglia
Dipartimento Interateneo di Fisica, Università di Bari and I.N.F.N., Bari, Italy

Luigi Nitti
Dipartimento dell’Emergenza e dei Trapianti di Organi, Sezione Fisica Medica, Università di Bari and I.N.F.N., Bari, Italy

Cost functions for non-hierarchical pairwise clustering are introduced, in the probabilistic autoencoder framework, by the request of maximal average similarity between the input and the output of the autoencoder. The partition provided by these cost functions identifies clusters with dense connected regions in data space; differences and similarities with respect to a well known cost function for pairwise clustering are outlined.

Clustering methods aim at partitioning a set of data-points in classes such that points that belong to the same class are more alike than points that belong to different classes. These classes are called clusters and their number may be preassigned or can be a parameter to be determined by the algorithm. There exist applications of clustering in such diverse fields as pattern recognition, astrophysics, communications, biology, business and many others. Two main approaches to clustering can be identified: parametric and non-parametric clustering.

Non-parametric approaches make few assumptions about the data structure and, typically, follow some local criterion for the construction of clusters. Typical examples of the non-parametric approach are the agglomerative and divisive algorithms that produce dendrograms. In the last years non-parametric clustering algorithms have been introduced employing the statistical properties of physical systems. The Super-Paramagnetic approach by Domany and coworkers exploits the analogy to a model granular magnet: the spin-spin correlation of a Potts model, living on the data-points lattice and with pair-couplings decreasing with the distance, is used to partition points in clusters. The synchronization properties of a system of coupled chaotic maps are used in to produce hierarchical clustering.

Parametric methods make some assumptions about the underlying data structure. Generative mixture models treat clustering as a problem of density estimation: data are viewed as coming from a mixture of probability distributions, each representing a different cluster, and the parameters of these distributions are adjusted to achieve a good match with the distribution of the input data. This can be obtained by maximizing the data likelihood (ML) or the posterior (MAP) if additional prior information on the parameters is available.

Many parametric clustering methods are based on a cost function: the best partition of points in clusters is assumed to be the one with minimum cost. Often cost functions incorporate the loss of information incurred by the clustering procedure when trying to reconstruct the original data from the compressed cluster representation: the most popular algorithm to optimize a cost function is K-means. Starting from a statistical ansatz and invoking maximum likelihood leads to a cost function which has been observed to work for clustering financial time series.

It is important to stress the difference between central clustering, where it is assumed that each cluster can be represented by a prototype, and pairwise clustering where data are indirectly characterized by pairwise comparison instead of explicit coordinates; pairwise algorithms require as input only the matrix of dissimilarities. Obviously the choice of the measure of dissimilarity is not unique and it is crucial for the performance of any pairwise clustering method. It is worth remarking that it often happens that the dissimilarity matrix violates the requirements of a distance measure, i.e. the triangular inequality does not necessarily holds.

Folded Markov chains are used in the Probabilistic Autoencoder Framework to derive cost functions for clustering. Some examples of two-stage folded Markov chains, and the corresponding algorithms for clustering and topographic mapping, are thoroughly analyzed in, where it is also shown that the cost function for pairwise clustering, introduced in, may be seen as a consequence of Bayes’ theorem and the requirement of minimal average distortion in a probabilistic autoencoder.

It is the purpose of this work to introduce a new class of cost functions for pairwise clustering which can be obtained, in the autoencoder frame, by requiring maximal similarity instead of minimal distortion. We show that the cost functions here introduced provide a non-hierarchical clustering of points where dense connected regions of points in the data space are recognized as clusters.

Let us now discuss autoencoders described by one-stage folded Markov chains. Let us consider a point $x$, in a data space, sampled with probability distribution $P_0(x)$: a code index $\alpha \in \{1, \ldots, q\}$ is assigned to $x$ according to conditional probabilities $P(\alpha|x)$. A reconstructed version of the input, $x'$, is then obtained by use of the Bayesian decoder:
\[ P(x'|\alpha) = \frac{P(\alpha|x') P_0(x')}{P(\alpha)}. \quad (1) \]

The joint distribution of \( x \), \( x' \) and \( \alpha \), describing this encoding-decoding process, is

\[ P(x,x',\alpha) = P_0(x) P(\alpha|x) P(x'|\alpha); \quad (2) \]

owing to (1), the joint distribution reads:

\[ P(x,x',\alpha) = \frac{P_0(x) P_0(x') P(\alpha|x) P(\alpha|x')}{P(\alpha)}. \quad (3) \]

The conditional probabilities \( \{P(\alpha|x)\} \) are the free parameters that must be adjusted to force the autoencoder to emulate the identity map on the data space.

Let \( d(x,x') \) be a measure of the distortion between input and output of the autoencoder. The average distortion is then given by:

\[ D = \sum_{\alpha=1}^{q} \int dx \int dx' P_0(x) P_0(x') P(\alpha|x) P(\alpha|x') d(x,x'). \quad (4) \]

Moreover, let \( s(x,x') \) be a measure of the similarity between input and output; the average similarity is then given by

\[ S = \sum_{\alpha=1}^{q} \int dx \int dx' P_0(x) P_0(x') P(\alpha|x) P(\alpha|x') s(x,x'). \quad (5) \]

It is natural to postulate a one-to-one mapping between values of distortion and similarity, \( s = F(d) \), with \( F \) a strictly decreasing function. A good autoencoder is obviously characterized by a low value of \( D \) and high value of \( S \). However we remark that the two requirements \( \text{Min}(D) \) and \( \text{Max}(S) \), for reasonable choices of \( F \), are not generally equivalent.

Now we turn back to the clustering problem. Given a data-set \( \{x_i\} \) of cardinality \( N \), partitioning these points in \( q \) classes corresponds, in this frame, to design an autoencoder, with \( q \) code indexes, acting on data space. We choose the encoder to be deterministic:

\[ P(\alpha|x) = \delta_{\alpha \sigma(x)}, \quad (6) \]

\( \sigma(x) \in \{1,\ldots,q\} \) being the code index associated to \( x \). The estimate for the average distortion [1], based on the data set at hand, is given by \( D = NH_d[\sigma] \), where we introduce the hamiltonian \( H_d \) for the Potts variables \( \{\sigma_i\} \):

\[ H_d[\sigma] = \sum_{\alpha=1}^{q} \sum_{i,j=1}^{N} \delta_{\alpha \sigma_i} \delta_{\alpha \sigma_j} d_{ij} / \sum_{k=1}^{N} \delta_{\alpha \sigma_k}, \quad (7) \]

where \( \sigma_i = \sigma(x_i) \), \( d_{ij} = d(x_i,x_j) \). It turns out that \( H_d \) is equivalent to the cost function for pairwise clustering, influential in the clustering literature, introduced in [13].

The estimate for the average similarity is, similarly, given by \( S = -NH_s[\sigma] \), where we introduce the hamiltonian \( H_s \):

\[ H_s[\sigma] = - \sum_{\alpha=1}^{q} \sum_{i,j=1}^{N} \delta_{\alpha \sigma_i} \delta_{\alpha \sigma_j} s_{ij} / \sum_{k=1}^{N} \delta_{\alpha \sigma_k}. \quad (8) \]

If we choose the autoencoder by minimizing the average distortions, then the best partition of the data-set in \( q \) classes corresponds to the ground state of \( H_d \). If we choose it by maximizing the average similarity, then the ground state of \( H_s \) must be sought for, instead. Since both \( \{d_{ij}\} \) and \( \{s_{ij}\} \) may be taken positive, it follows that \( H_d \) is characterized by antiferromagnetic couplings between the Potts variables, while \( H_s \) is made of ferromagnetic couplings. Denominators in both \( H_d \) and \( H_s \) serve to enforce the coherence among the \( q \) clusters. In particular, without the denominator the ground state of \( H_s \) would correspond to a single big cluster.

The form of the function \( F \), determining the relation between \( s \) and \( d \), has to be specified. In what follows we consider two forms of this relation. A scale-free relation

\[ s_{ij} = F_\gamma(d_{ij}) = \left( \frac{d_{ij}}{\langle d \rangle} \right)^\gamma, \quad (9) \]

depending on the exponent \( \gamma \), and a scale-dependent relation

\[ s_{ij} = F_a(d_{ij}) = \exp \left( - \frac{1}{2a^2} \left( \frac{d_{ij}}{\langle d \rangle} \right)^2 \right), \quad (10) \]

dependent on the scale \( a \). In the formulas above, \( \langle d \rangle \) is the average dissimilarity over all the pairs of data-set points. The exponent \( \gamma \) will be restricted to assume small values so as to characterize the corresponding Potts model by long-range ferromagnetic couplings; the scale parameter \( a \) will be bounded in \([0,1]\).

At this point it is worth stressing that minimization of the distortion and maximization of the similarity yield, in the autoencoder frame, different cost functions. The hamiltonian \( H_d \) embodies the requirement that pairs of distant points (large \( d_{ij} \)) should belong to different clusters. On the other hand, the hamiltonian \( H_s \), for reasonable choices of \( F \), concentrates on pairs of close points (small \( d \)) and forces them to belong to the same cluster. In other words, \( H_s \) may be seen to implement the idea that clusters should be searched for as dense connected regions in the data space.

We describe now the application of the variational criterions for clustering, described above, to some artificial and real data-sets. We consider two optimization algorithms to find the configuration of minimum cost: simulated annealing [17] and mean-field annealing [18]. Both approaches associate a Gibbs probability distribution to the functional to
be optimized. Simulated annealing is a Monte-carlo technique which samples the Gibbs distribution as the temperature is reduced to zero, while mean-field annealing attempts to track an approximation, to the mean of the distribution, known as mean field approximation \[19\]. We remark that an efficient mean-field annealing algorithm for cost function \[6\], based on the EM scheme \[20\], is described in \[13\]: the generalization of that algorithm to \[8\] is straightforward.

\[\epsilon = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1,j\neq i}^{N} (\delta_{\eta_i,\eta_j} - \delta_{\eta_i} - \delta_{\eta_j})^2 \]  

which counts the number of pairs of points upon which the two partitions disagree. Using the scale-dependent \(F_d\), we find the ground state of \(H_d\) to differ from those of \(H_d\) by \(\epsilon < 0.01\) varying \(a\) in \([0.05, 1]\). Analogously, using the scale-free \(F_s\), with \(\gamma \in [0.1, 1.5]\), we find \(\epsilon < 0.02\) when we compare the ground state of \(H_s\) with those of \(H_d\). Hence, on this data set, the cost functions introduced above work similarly within wide ranges of \(\gamma \) and \(a\) values.

We find a similar behaviour with respect to the famous IRIS data of Anderson \[22\]. This data set has often been used as a standard for testing clustering algorithms: it consists of three clusters (Virginica, Versicolor and Setosa) and there are 50 objects in \(\mathbb{R}^4\) per cluster. Two clusters (Virginica, Versicolor) are very overlapping. The clustering result, with \(q = 3\) and minimizing \(H_d\), consists of three clusters of 61, 39 and 50 points respectively, with 90% of correct classification percentage. We obtain exactly the same partition by minimizing \(H_s\) using a scale-free \(F\) (with \(\gamma \in [0.15, 1.45]\)), and using a scale-dependent \(F\) (with \(a \in [0.25, 1]\)). For \(a \in [0.1, 0.25]\) we obtain, in the scale-dependent case, a slightly different partition with clusters' sizes 58, 42, 50 and correct classification percentage 93.3%. These results show that also in the IRIS case the pairwise clustering procedures by distorsion minimization and similarity maximization are almost equivalent.

A typical situation resulting in different answers from \(H_d\) and \(H_s\) is depicted in Fig.2a. This two-dimensional data-set is made of an elongated cluster and a Gaussian distributed circular one. It is evident that two dense connected regions are present, and that the farthest pairs of points belong to the same connected region. This is the type of data-set such that minimizing the distorsion is not equivalent to maximizing the similarity. In Fig.2b the partition we obtain minimizing \(H_d\) is depicted: it fails to recognize the structure in the data-set. Let us now consider the ground state of \(H_s\) with the scale-dependent \(F\). For \(a < 0.7\) the ground state, depicted in Fig.2c, recognizes with 99% accuracy the data structure. At \(a \sim 0.7\) a transition phenomenon occurs: the configuration depicted in
Fig. 2c ceases to be the global minimum, the new ground state (Fig. 2d) being very close to the solution by $H_d$.

![Figure 3](image3.png)

**FIG. 3.** (a) The efficiency (percentage of correctly classified points) versus $\gamma$, obtained on the data-set depicted in Fig. 2 by minimizing $H_s$ with scale-dependent $F$. The dashed line is the efficiency obtained by minimization of $H_d$. (b) The $\epsilon$ parameter, (see the text) between partitions corresponding to adjacent values of $\gamma$, is plotted versus $\gamma$. (c) The size of the two output clusters versus $\gamma$.

In Fig. 3a we depict the efficiency of the classification versus the resolution parameter $a$, for the scale dependent $F$, while in Fig. 3b we consider a sequence of $\epsilon$-values and we plot the $\epsilon$ between partitions corresponding to adjacent values of $\gamma$. The peak at $a = 0.7$ is the indicator of the transition between global minima. Finally, in Fig. 3c the size of the two clusters, versus $\gamma$, is depicted. Concerning the scale-free $F$, in Fig. 4 the same plots as in Fig. 3 are depicted, showing that the good minimum is stable for a wide range of $\gamma$. The choice of the optimization algorithm deserves a comment. All the results described above are obtained by simulated annealing; we also apply the mean-field annealing scheme, described in [13], and we always find a configuration very close to the one from simulated annealing, while spending less computational time. This confirms that optimization algorithms rooted on mean-field theory yield quickly a good solution on these problems [13].

In summary, we address non-hierarchical pairwise clustering and, working in the probabilistic autoencoder frame, we introduce a class of cost functions arising from the introduction of an adaptive relation between distortion and similarity, i.e. the function $s = F(d)$ might be depending on the properties of the data-set in a neighbourhood of the pair of points under consideration. It will be also important to develop cluster-validity criterions to provide a means to choose an optimal $q$ value in situations where the number of classes is ambiguous.

![Figure 4](image4.png)

**FIG. 4.** (a) The efficiency (percentage of correctly classified points) versus $\gamma$, obtained on the data-set depicted in Fig. 2 by minimizing $H_s$ with scale-independent $F$. The dashed line is the efficiency obtained by minimization of $H_d$. (b) The $\epsilon$ parameter, (see the text) between partitions corresponding to adjacent values of $\gamma$, is plotted versus $\gamma$. (c) The size of the two output clusters versus $\gamma$.

---

[1] B.D. Rypley, *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge U.K., 1996.

[2] R.O. Duda, P.E. Hart, *Pattern Recognition and Scene Analysis*. Wiley, New York, 1973.

[3] A. Dekel, M.J. West, *Astrophys. J.* 228, p. 411 (1985).

[4] Y. Linde, A. Buzo, R.M. Gray, *IEEE Trans. on Communications* 28, p. 84 (1980).

[5] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra,
[6] L. Kullmann, J. Kertesz, R.N. Mantegna, Physica A 287, p. 412 (2000).

[7] M. Blatt, S. Wiseman, E. Domany, Phys. Rev. Lett. 76, pp. 3251-3255 (1996).

[8] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, S. Stra-maglia, Phys. Rev. Lett. 85, pp. 554-557 (2000).

[9] C.M. Bishop, Neural networks for pattern recognition. Clarendon Press, Oxford, 1995.

[10] A. Utsugi, Network 7, p. 727 (1996).

[11] L. Giada, M. Marsili, 'Data clustering and noise undressing of correlation matrices', preprint cond-mat/0101237.

[12] K. Rose, E. Gurewitz, G. Fox, Phys. Rev. Lett. 65, pp. 945-948 (1990).

[13] T. Hofman, J.M. Buhmann, IEEE Trans. P.A.M.I. 19, pp.1-14 (1997).

[14] S.P. Luttrell, Neural Computation 6, p. 767, 1994.

[15] C.M. Bishop, M. Svensen, C.K.I. Williams, Neural Computation 10, p.215 (1997).

[16] T. Graepel, Statistical Physics of clustering algorithms, Diplomarbeit, Technique Universitat, FB Physik, Institut fur Theoretishe Physik, Berlin, April 1998.

[17] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220, p.671 (1983).

[18] See, e.g., A.L. Yuille, J.J. Kosowsky, Neural Computation 6, pp. 341-356 (1994), and references therein.

[19] G. Parisi, Statistical Field Theory, Addison Wesley, California 1988.

[20] A.P. Dempster, N.M. Laird, D.B. Rubin, Jour. Royal Stat. Soc. 39, p.1, (1977).

[21] In the text we use an operational definition of ground state as the best output over a number (10-50) of simulated annealing runs. The true ground state might be found only by an unpractical exhaustive search.

[22] E. Anderson, Bull. Amer. Iris Soc. 59, p.2 (1935).