A Nonparametric Information Theoretic Clustering Algorithm

Lev Faivishevsky LEVTEMP@GMAIL.COM
Jacob Goldberger GOLDBEJ@ENG.BIU.AC.IL
School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel

Abstract

In this paper we propose a novel clustering algorithm based on maximizing the mutual information between data points and clusters. Unlike previous methods, we neither assume the data are given in terms of distributions nor impose any parametric model on the within-cluster distribution. Instead, we utilize a non-parametric estimation of the average cluster entropies and search for a clustering that maximizes the estimated mutual information between data points and clusters. The improved performance of the proposed algorithm is demonstrated on several standard datasets.

1. Introduction

Effective automatic grouping of objects into clusters is one of the fundamental problems in machine learning and in other fields of study. In many approaches, the first step toward clustering a dataset is extracting a feature vector from each object. This reduces the problem to the aggregation of groups of vectors in a feature space. Then various clustering algorithms are applied on these feature vectors. The specific form of the feature space along with possible additional information about cluster structure determine a class of algorithms that may be used to group the vectors. According to the required form of input, three major kinds of clustering algorithms may be defined.

The first kind of algorithms assumes that the feature vectors are given as points in a finite-dimensional space R^d without additional information on the clusters structure. Distances between vectors may naturally give rise to pairwise data point similarities. The class of methods that cluster vectors in R^d includes the spectral clustering algorithms (Ng et al., 2002), (Zelnik-Manor & Perona, 2005), that have attracted much attention in recent years. The second class of clustering algorithms also admits input in the form of vectors in R^d but in addition implicitly or explicitly assumes certain types of in-cluster distribution (e.g. applying the EM algorithm to learn a Gaussian mixture density). Although these iterative methods can suffer from the drawback of local optima, they provide high quality results when the data clusters are organized according to the anticipated structures, in this case in convex sets. When the data are arranged in non-convex sets (e.g. concentric circles) these algorithms tend to fail. It follows that in a certain sense the second kind of alogrithms is a subset of the first kind. The third kind of algorithms corresponds to the case of distributional clustering. Here each data point is described as a distribution. In other words the feature representation of each data point is a parametric description of the distribution. Both discrete and continuous distributions may be considered. The former case is illustrated by a generic example of document clustering. In a continuous setup we can consider the problem where each object is a Gaussian distribution and we want to cluster similar Gaussians together. In all these cases the cluster distribution is a (possibly weighted) average of the distributions of the objects that are assigned to the cluster. Hence the third kind of algorithms is a subset of the second kind.

The relative entropy or the Kullback-Leibler divergence is a natural measure of the distance between distributions. Therefore this quantity is of particular importance in the field of distributional clustering. Given such a choice for distance, the mutual information becomes an optimal clustering criterion (Banerjee et al., 2004). In practice, the mutual information is computed between cluster labels and feature representations of data points in terms of distributions (Dhillon et al., 2003). Similar ideas gave rise to the Information Bottleneck approach (Tishby et al., 1999). The mutual information has been proven to be a powerful clustering criterion for document clustering (Slonim &

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

Tishby, 2000),(Slonim et al., 2002) and clustering of Gaussians (Davis & Dhillon, 2007). However all the above methods are limited to the domain of distributional clustering since they require an explicit parametric representation of data points. In all these methods there is an explicit assumption regarding the parametric structure of the intra cluster distribution.

In this paper we extend the information theoretic criterion to a general domain of clustering algorithms whose inputs are simply vectors in R^d. In particular, we maximize the mutual information between cluster labels and features of data points without imposing any parametric model on the cluster distribution. Our method computes this target in an intuitive straightforward manner using a novel non-parametric entropy estimation technique (Faivishevsky & Goldberger, 2009). We show that this results in an efficient clustering method with state-of-the-art performance on standard real-world datasets. The reminder of this paper is organized as follows. Section 2 discusses the mutual information criterion of clustering in detail. Section 3 introduces the Nonparametric Information Clustering (NIC) algorithm. Section 4 reviews related work. Section 5 describes numerical experiments on several standard datasets.

2. The Mutual Information Criterion for Clustering

A (hard) clustering of a set of objects $X = \{x_1, ..., x_n\}$ into n_c clusters is a function $C : X \to \{1, ..., n_c\}$. Denote the cluster of x_i by c_i. Denote the number of points assigned to the j cluster by n_j. Given a clustering score function $S(C)$, the task of clustering the set X is finding a clustering $C(X)$ that optimizes $S(C)$. A clustering task is defined by the object description and the score function $S(C)$.

In this paper we consider the data points as independent samples of a distribution that can be either given as part of the problem statement or unknown. The clustering C is a function of the random variable X and therefore $C(X)$ is also a random variable. Hence we can define the mutual information $I(X; C)$ based on the joint distribution. Since $I(X; C) = H(X) - H(X|C)$ and $H(X)$ does not depend on the specific clustering, we can use the conditional entropy $H(X|C)$ as a measure of the clustering quality:

$$S_{MI}(C) = H(X|C) = \sum_{j=1}^{n_c} \frac{n_j}{n} H(X|C = j) \qquad (1)$$

The mutual information score function that measures the intra-cluster entropy resembles the k-means score that measures the intra-cluster variance. Clearly, the MI provides a more robust treatment for various cases of differently distributed data as discussed below. This measure is intuitive; we expect that in a good clustering the objects in the same cluster will be similar, whereas similar objects will not be assigned to different clusters. Expressing this intuition into information theory terminology, we expect that the average entropy of the object distribution in a cluster will be small. This is obtained by maximizing $I(X; C)$.

To compile the MI cost function into a clustering algorithm we have to tackle the technical issue of computing the within-cluster entropy terms $H(X|C = j)$. The simplest case is when the objects all belong to a finite set. In this case the distribution $p(X|C = j)$ is discrete and the entropy can be computed based on the frequency histogram of the objects in the cluster. We demonstrate this on the generic problem of unsupervised document clustering. Utilizing the bag of words paradigm (Salton & McGill, 1983), each document is viewed as a bag containing all the words that appear in it and each cluster can be viewed as a bag containing all the words from all the documents that are mapped into that cluster. More formally, each document i is represented by a vector $\{n_1^i, n_2^i, ..., n_M^i\}$, where n_w^i is the number of instances of word w in the document and M is the size of the word dictionary. Given a document clustering C we can easily compute the word statistics in the cluster. Defining the average frequency of word w occurrence in the cluster j by:

$$p(w|C = j) \propto \sum_{i|c_i=j} n_w^i \qquad (2)$$

we arrive at the within cluster entropy:

$$H(X|C = j) = -\sum_{w=1}^{M} p(w|C = j) \log p(w|C = j) \quad (3)$$

The criterion $I(X; C)$ in this context is also known as the Information-Bottleneck (IB) principle (Tishby et al., 1999). (In the usual definition of the IB we further assume a uniform prior over the document, which means that if we want to make the above framework consistent with IB we need to weight each word to be inversely proportional to the document size).

There is a subtle point here that needs to be clarified. The task we want to perform here is clustering the documents in the corpus such that all the documents in a given group are related to the same topic. However, in the mutual information framework described above, technically the objects to be clustered are the words. The entropy $H(X|C = j)$ we compute in the expression (3) is the word entropy in the cluster. The document structure is used to place a semi-supervised constraint that all the words in a given document will be assigned to the same cluster.

The situation is more complicated if the objects we want to cluster do not belong to a finite set but instead are 'feature' vectors in the Euclidean space R^d. A simple assumption we can impose is that the conditional density $f(x|C = j)$ is Gaussian. The mean and covariance of the Gaussian distribution are taken as the empirical average and variance of the points assigned to the cluster. Since there is a closed-form expression for the entropy of a Gaussian distribution we can compute the cluster score $I(X; C)$ given the within-cluster Gaussian assumption. To simplify the model we further assume that the cluster covariance matrices are all scalar matrices. The differential entropy of a d-dimensional Gaussian distribution with a covariance matrix $\sigma^2 I$ is $\frac{d}{2}\log(2\pi e\sigma^2)$. Hence, the conditional entropy part of the mutual information clustering criterion where the within-cluster distribution is a spherical Gaussian is:

$$H(X|C = j) = \frac{d}{2} \log \frac{1}{n_j} \sum_{i|c_i=j} \|x_i - \mu_j\|^2 \qquad (4)$$

where μ_j is the empirical average of the points assigned to the j-th cluster. Similar to the discrete case, here we can also consider a semi-supervised setup with the additional constraints that points in given subsets (e.g. a list of pair of points) should be assigned to the same cluster (see e.g. (Shental et al., 2004)). We can take the continuous semi-supervised case one step further and consider the clustering problem where each object is a Gaussian distribution and we want to cluster similar Gaussians together (see e.g. (Goldberger & Roweis, 2005)) and represent all the Gaussians in the same clustering with a single ('collapsed version') Gaussian distribution. One rationale for such a clustering is simplifying a mixture of a large number of Gaussians into a mixture of a fewer components. In this case since we assume that the intra-cluster distribution is Gaussian we can still explicitly compute the MI clustering criterion.

In all the examples described above the cluster distribution is predefined and is part of the problem setup and therefore the cluster entropy can be explicitly computed. In the general case of clustering points in R^d we do not have any prior knowledge on the within cluster distribution. However, assuming that the intra cluster distribution is Gaussian is not always a good choice since by utilizing a Gaussian distribution to describe the density we implicitly assume a unimodal blob type shape which is not always the case.

3. Nonparametric Mutual Information Clustering

Assume that a dataset X is represented by a set of features $x_1, ..., x_n \in R^d$ without any additional information on the feature distributions either for individual objects or for objects that are in the same cluster. In the MI clustering criterion $I(X; C)$ the relevant term is not the within-cluster distribution but the within cluster entropy. The key point is that by using a mutual information clustering criterion we do not need to have an explicit representation of the intra-cluster distribution. We only need to compute the cluster entropy. In what follows we propose to use a nonparametric estimation of in-cluster entropy in order to benefit from the MI clustering score function (1).

Classical methods for estimating the mutual information $I(X; C)$ require the estimation of the joint probability density function of $(X, C(X))$. This estimation must be carried out on the given dataset. Histogram- and kernel-based (Parzen windows) pdf estimations are among the most commonly used methods (Torkkola, 2003). Their use is usually restricted to one- or two-dimensional probability density functions (i.e. pdf of one or two variables). However, for high-dimensional variables histogram- and kernel-based estimators suffer dramatically from the curse of dimensionality; in other words, the number of samples needed to estimate the pdf grows exponentially with the number of variables. An additional difficulty in kernel based estimation lies in the choice of kernel width.

Other methods used to estimate the mutual information are based on k-nearest neighbor statistics (see e.g. (Victor, 2002),(Wang et al., 2009)). A nice property of these estimators is that they can be easily utilized for high dimensional random vectors and no parameters need to be predefined or separably tuned for each clustering problem (other than determining the value of k). There are a number of non-parametric techniques for the (differential) entropy estimation of random vectors $x_1, ..., x_n \in R^d$ which are all variants of the following estimator (Kozachenko & Leonenko, 1987):

$$H_k = \frac{d}{n} \sum_{i=1}^{n} \log \epsilon_{ik} + \text{const(k)} \qquad (5)$$

where ϵ_{ik} is the Euclidean distance from x_i to its k-th nearest neighbor. The constant in Eq. (5) is:

$$\psi(n) - \psi(k) + \log(c_d)$$

where $\psi(x)$ is the digamma function (the logarithmic derivative of the gamma function) and c_d is the volume of the d-dimensional unit ball. The H_k entropy estimator is consistent in the sense that both the bias and the variance vanish as the sample sizes increase. The consistency of the 1-NN estimator was proven in (Kozachenko & Leonenko, 1987) and the consistency of the general k-NN version was shown in (Goria et al., 2005).

In the case of iterative clustering algorithms we need to compute $I(X;C)$ of many clusterings and since the neighbors depend on the clustering we have to recompute the neighbors for each clustering in the optimization process. Because they are non-parametric, the kNN estimators (5) rely on order statistics. This makes the analytical calculation of the gradient of $I(X;C)$ cumbersome. It also leads to a certain lack of smoothness of the estimator value as a function of the sample coordinates.

Here we utilize the MeanNN differential entropy estimator (Faivishevsky & Goldberger, 2009) due to its smoothness with respect to the coordinates of data points. The MeanNN estimator exploits the fact that the kNN estimation is valid for every k and therefore averaging estimators (5) for all possible values of k leads itself to a new estimator of the differential entropy:

$$H_{mean} = \frac{1}{n-1}\sum_{k=1}^{n-1} H_k = \frac{d}{n(n-1)}\sum_{i\neq l}\log\|x_i - x_l\| + \text{const} \tag{6}$$

This estimator computes the entropy based on the pair-wise distances between all the given data points and thus eliminates calculation of nearest neighbors. Applying this estimator to in-cluster entropy estimation yields:

$$H(X|C=j) \approx \frac{d}{n_j(n_j-1)}\sum_{i\neq l|c_i=c_l=j}\log\|x_i - x_l\|$$

Plugging this estimation into the score function $S_{MI}(C)$ (1) yields the following form of the clustering quality measure:

$$S_{NIC}(C) = \sum_j \frac{d}{n_j-1}\sum_{i\neq l|c_i=c_l=j}\log\|x_i - x_l\| \tag{7}$$

NIC stands for Nonparametric Information Clustering, which is how we dub our approach.

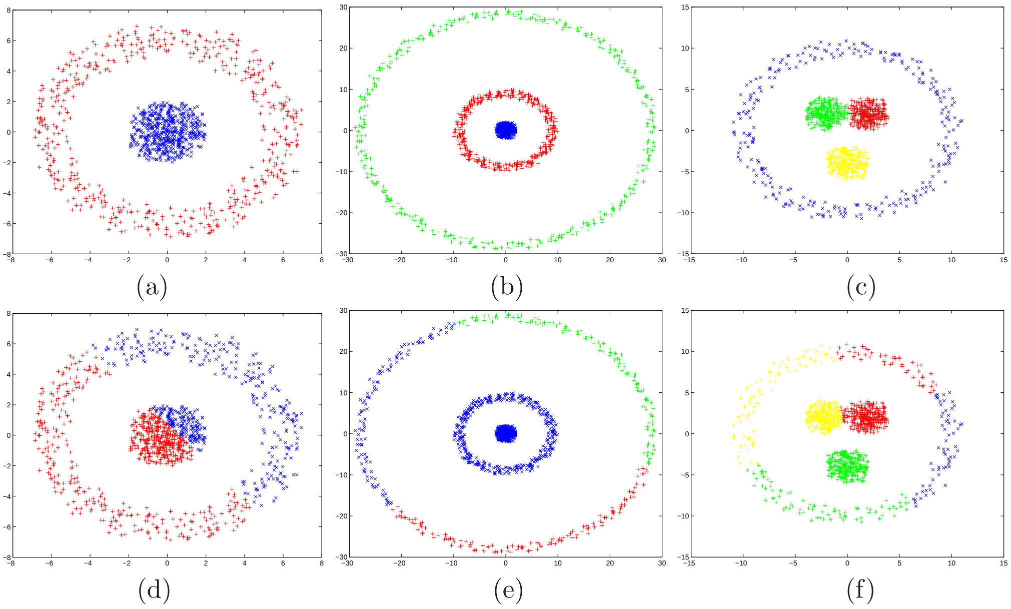
To optimize the score function $S_{NIC}(C)$ we can apply a greedy sequential algorithm that resembles the sequential version of the k-means algorithm (Slonim et al., 2002). The sequential greedy algorithm is known to perform well in terms of both clustering quality and computational complexity. The sequential clustering algorithm starts with a random partition of the data points into clusters. Then, it goes over the n points in a cyclical manner and for each point checks whether moving it from its current cluster to another one improves the score function $S_{NIC}(C)$. This loop may be iterated until either we reach a local optimum (i.e., a stage in which no point transition offers an improvement) or the local improvements of the score function become sufficiently small. As there is no guarantee that such a procedure will find the global optimum, it may be repeated several times with different random partitions as the starting point in order to find the best local optimum among the repeated searches.

```
Input: Data vectors $X = \{x_1, x_2, ..., x_n\} \subset R^d$, number of clusters $n_c$.
Output: Clustering assignment $\{c_1, c_2, ..., c_n\}$, $c_i \in \{1, ..., n_c\}$
Method:
    1. Apply data whitening via multiplying the data by the matrix $Cov(X)^{-\frac{1}{2}}$
    2. Randomly initialize assignment $C(X)$.
    3. Calculate score:
           $S_{NIC}(C) = \sum_i \frac{1}{n_j - 1} \sum_{i\neq l|c_i=c_j=j} \log\|x_i - x_l\|$
       where $n_j$ is the size of the $j$-th cluster.
    4. Do until convergence
        • Go over the points in a circular manner.
        • For data point $x_i$ calculate scores of all possible reassignments of $x_i$ to different clusters.
        • Update current assignment $C$ by choosing label $c_i$ that leads to the minimal score.
```

Figure 1. The Nonparametric Information Clustering (NIC) algorithm.

To find the cluster re-assignment of a data point x_i we need to compute the updated entropy of each cluster after adding x_i to that cluster. To do so we need to calculate the log-distance of x_i to all the other members of the cluster. Hence the complexity of reassigning x_i to a new cluster is $O(n)$ and the overall computational complexity of the algorithm is $O(n^2)$.

The iterative clustering algorithm may be prefaced by whitening of the input data. This numerical procedure imposes a linear transformation that may contribute to the numerical robustness of the computations, see

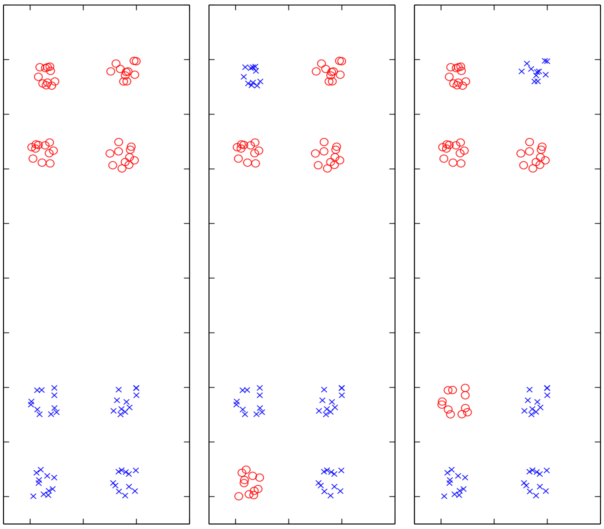


(a) (b) (c)

(d) (e) (f)

Figure 2. Comparison of the proposed clustering method NIC and the k-means clustering algorithm on three synthetic cases. (a)-(c) NIC, (d)-(f) k-means.

e.g. (Wang et al., 2009). Since the pre whitening is accomplished as multiplication of input data by the invertible matrix matrix $A = Cov(X)^{-1/2}$ the mutual information between the datapoints and the labels is not changed. The Nonparametric Information Clustering (NIC) algorithm is summarized in Fig. 1.



(a) (b) (c)

Figure 3. Three possible clusterings (into two clusters) of the same dataset: (a) 'correct' clustering, (b) and (c) erroneous clusterings. Using MeanNN as the MI estimator, the MI clustering score favors the correct solution while using the kNN yields the same score for all the three clusterings.

4. Related work

The commonly used k-means algorithm addresses objects X as vectors in R^d. The k-means score function measures the sum of square-distances between vectors assigned to the same cluster. Observing that:

$$\sum_{i|c_i=j} \|x_i - \mu_j\|^2 = \frac{1}{2n_j} \sum_{i\neq l|c_i=c_l=j} \|x_i - x_l\|^2$$

where μ_j is the average of all data points in cluster j, we can rewrite $S_{kmeans}(C)$ as follows:

$$S_{kmeans}(C) = \sum_{j=1}^{n_c} \frac{1}{n_j} \sum_{i\neq l|c_i=c_l=j} \|x_i - x_l\|^2 \qquad (8)$$

It is instructive to compare the k-means score with the mutual information score based on a Gaussian within-cluster density (4) and the proposed S_{NIC} score (7):

$$\begin{aligned} S_{kmeans}(C) &= \sum_{j=1}^{n_c} \frac{1}{n_j} \sum_{i\neq l|c_i=c_l=j} \|x_i - x_l\|^2 \\ S_{GaussMI}(C) &= \sum_{j=1}^{n_c} \log \frac{1}{n_j} \sum_{i\neq l|c_i=c_l=j} \|x_i - x_l\|^2 \\ S_{NIC}(C) &= \sum_{j=1}^{n_c} \frac{1}{(n_j - 1)} \sum_{i\neq l|c_i=c_l=j} \log \|x_i - x_l\|^2 \end{aligned} \qquad (9)$$

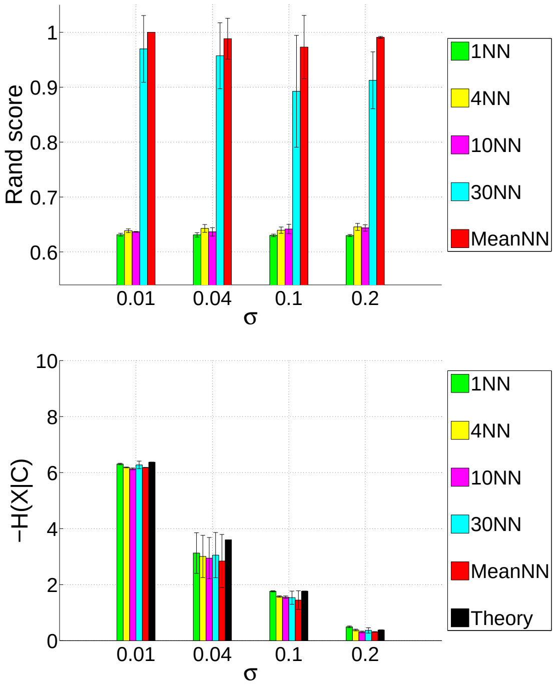


Figure 4. Comparison of MeanNN and kNN on MoG data. (Top) Clustering results for sevreal values of k. (Bottom) Unsupervised estimation of conditional entropy $H(X|C)$.

Note that the log shifts into a more internal position when we consider more general setups, namely in the first special case that optimizes in-cluster variances the log is absent. In the case of a clustering score based on the in-cluster Gaussian density assumption, the log is in the middle of summation. In the most general case with no assumption on in-cluster distributions the log is in the most internal position. If the clusters are well separated, the clustering assignments are the same for both methods. This holds in contrast to the general case that leads to different optimal assignments. Generally log weighting provides robustness to outliers. In addition S_{NIC} is able to cluster non-convex sets correctly whereas k-means fails in these cases. On the basis of these advantages the proposed method manages to correctly cluster cases where k-means is known to fail, see Fig. 2 for several such examples.

Recently there have been a number of attempts to generalize mutual information based criteria to general feature spaces. To address this issue, in (Tishby & Slonim, 2001) a random walk over the data points has been defined, serving to transform input data to a transition probability matrix that could be further analyzed via the IB algorithm. A recent work (Slonim et al., 2005) suggests using the mutual information between different data points as part of a general information-theoretic treatment of the clustering problem. However, both of the approaches have limitations. The former involves various non-trivial steps in the data pre-processing and the latter requires a sufficiently high dimensionality d of input space R^d for reliable estimation of the information relations between data points, as observed in (Seldin et al., 2007). Yet another recent method was presented in (Seldin et al., 2007). Here the feature representations of data points were practically allowed to obtain categorical values. However this still did not permit the direct application of the method to problems with input data in R^d. Inevitably heuristic techniques should be used to transform continuous feature values into a categorical form, e.g. quantization. Such heuristics make it difficult to apply to data that are represented as feature vectors in R^d.

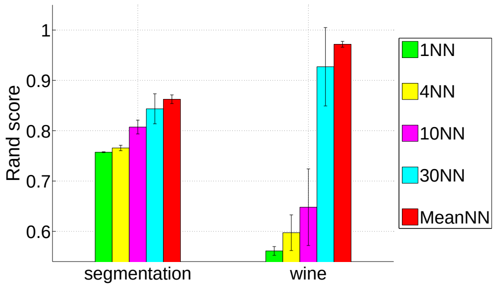


Figure 5. Comparison of the MeanNN and kNN estimators for the NIC method by UCI datasets: segmentation and wine. Statistics are shown for 10 repetitions.

5. Experiments

In this section we describe two sets of experiments. In the first experiment we concentrate on comparing two candidate non-parametric entropy estimators for information-theoretic clustering algorithm. In the second set of experiments we compare the performance of the proposed NIC algorithm with the performance of several standard clustering algorithms on datasets from the UCI repository.

5.1. Comparing between MeanNN and kNN

We compared the NIC clustering algorithm based on the MeanNN estimator to the same iterative algorithm based on kNN estimator. The kNN estimator has a significant drawback in that the contribution of each data point to the final score is defined only by distances from the point to its k nearest neighbor. Therefore the data in a cluster may be subdivided into small subgroups with k or slightly more points in each, and the

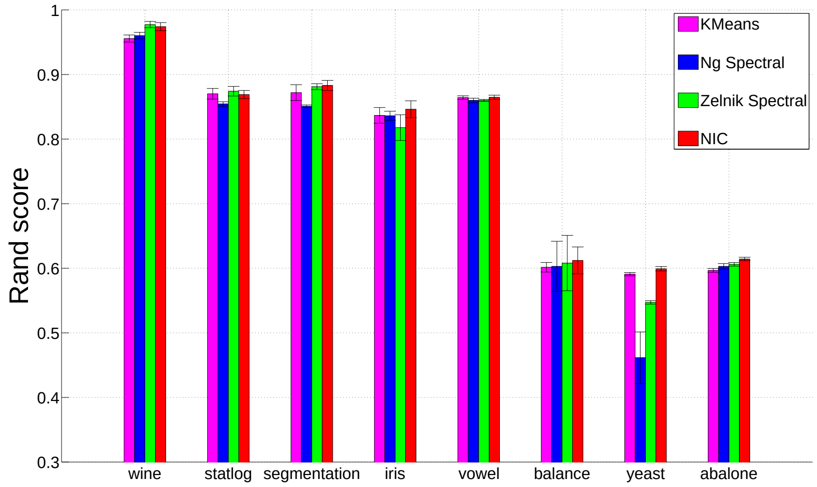


Figure 6. Performance of several clustering methods on UCI datasets: iris, statlog, segmentation, vowel, wine, abalone, balance and yeast. Statistics are shown for 10 repetitions.

joint interaction of such groups will be failed to traced by the kNN estimator. This leads then to counter intuitive clustering results for small and moderate values of k, see Fig. 3 for an illustration of this behavior. The MeanNN leads to a better clustering score for the correct clustering assignment, whereas the kNN erroneously leads to equal scores for all the three clusterings shown in Fig. 3 for every k from 1 to 10.

Next we compared the two alternatives on data set sampled from a mixture of four 2D Gaussians. The Gaussians centers were on the vertices of the unit square, and they shared the same scalar covariance matric $\sigma^2 I$. In this case $H(X|C) = \log(2\pi e\sigma^2)$. We randomly generated 400 samples for values of $\sigma = 0.01, 0.04, 0.1, 0.2$. For each value of σ we made 10 repetitions. In each run we applied MeanNN estimator as well as kNN for k=1,4,10,30. Performance was evaluated using the Rand Index score (Rand, 1971) which is a standard non-parametric measure of clustering quality. MeanNN performed significantly better than kNN estimators, see Fig. 4 (Top). On the other hand both MeanNN and kNN score functions lead to similar results in recovering true conditional entropy, see Fig. 4 (Bottom). Here the entropy estimation was done in an unsupervised manner using labels from result clustering assignments.

Fig. 5 shows clustering results on two datasets from the UCI repository (Asuncion & Newman, 2007). In each run we randomly selected 95% from a dataset and applied MeanNN estimator as well as kNN for k=1,4,10,30. There were 10 repetitions for each dataset. Performance was evaluated using the Rand Index. The MeanNN is shown to be more adequate for the clustering task than kNN. For the kNN estimator, smaller values of k generally led to worse average performance. This can be attributed to the locality of kNN estimators which does not allow them to treat more global structures correctly, as was shown above. On the other hand, using larger k improves the average performance but leads to higher variance of the estimation. This example shows the advantage of the MeanNN estimator that takes into account data interdependencies on all scales. Being asymptotically correct, nearest neighbors techniques may fail in adaptations to the problem specific spatial scale.

5.2. Comparison to other clustering algorithms

Next we compared the proposed method with three well known clustering techniques for clustering data points represented as vectors in R^d. The first is the widely used k-means method (Lloyd, 1982). The second is the standard spectral clustering method (Ng et al., 2002) and the third is rotation spectral clustering with local scaling (Zelnik-Manor & Perona, 2005). We evaluated the performance of the algorithm on eight standard datasets from the UCI repository. In each run we again randomly selected 95% from a dataset and applied all four methods. There were 10 repetitions for each dataset. Performance was evaluated using the Rand Index score. Since spectral clustering methods are sensitive to their intrinsic parameters such as scale (Ng et al., 2002) and number of neighbors (Zelnik-Manor & Perona, 2005) we used a logarithmic grid of scale and neighborhood values in each repetition. Spectral clustering results are reported for the best scale and neighborhood. The clustering results are summarized in Fig. 6. Note that for the NIC method there are no parameters that needed to be tuned. It is clear that NIC achieves state-of-the-art performance in these cases.

6. Conclusion

We proposed a new clustering method dubbed NIC based on the maximization of mutual information between data points and their labels. As opposed to other information theory based clustering methods our approach does not require knowledge of the within cluster distributions. We showed that using the MeanNN to estimate the within cluster entropy yields a clustering algorithm that achieves state-of-the-art accuracy on the standard datasets and has the advantage of relatively low computational complexity. We do not claim, however, that the MeanNN is the optimal entropy estimation for clustering. Future research can concentrate on finding entropy estimators that are more suitable for clustering tasks. In addition to a specific clustering algorithm we defined a general information-theoretic framework that unifies and extends several previously proposed clustering algorithms.

References

Asuncion, A. and Newman, D.J. UCI machine learning repository. 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. Clustering with Bregman divergences. In *Journal of Machine Learning Research*, pp. 1705–1749, 2004.

Davis, J. V. and Dhillon, I. Differential entropic clustering of multivariate Gaussians. *Advances in Neural Information Processing Systems 19*, 2007.

Dhillon, I., Mallela, S., Kumar, R., Mallela, S., Guyon, I., and Elisseeff, A. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:2003, 2003.

Faivishevsky, L. and Goldberger, J. ICA based on a smooth estimation of the differential entropy. *Advances in Neural Information Processing Systems 21*, 2009.

Goldberger, J. and Roweis, S. Hierarchical clustering of a mixture model. *Advances in Neural Information Processing Systems 17*, 2005.

Goria, M., Leonenko, N., Mergel, V., and Inverardi, P. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparam. Statist.*, pp. 277–297, 2005.

Kozachenko, L. and Leonenko, N. On statistical estimation of entropy of random vector. *Problems Infor. Transmiss.*, 23 (2), 1987.

Lloyd, S. P. Least squares quantization in pcm. *Special issue on quantization, IEEE Trans. Inform. Theory*, pp. 129–137, 1982.

Ng, A. Y., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, 2002.

Rand, W. Objective criteria for the evaluation of clustering methods. *J. Amer.Statist. Assoc. 66*, pp. 846–850, 1971.

Salton, G. and McGill, M. J. Introduction to modern information retrieval. *McGraw-Hill Book Company*, 1983.

Seldin, Y., Slonim, N., and Tishby, N. Information bottleneck for non co-occurrence data. *Advances in Neural Information Processing Systems 19*, 2007.

Shental, N., Bar-Hillel, A., Hertz, T., and Weinshall, D. Computing gaussian mixture models with EM using equivalence constraints. *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Slonim, N. and Tishby, N. Document clustering using word clusters via the information bottleneck method. *Proc. of the 23rd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2000.

Slonim, N., Friedman, N., and Tishby, N. Unsupervised document classification using sequential information maximization. *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2002.

Slonim, N., Atwal, G. S., Tracik, G., and Bialek, W. Information-based clustering. *Proc. of the National Academy of Science (PNAS)*, 102:182971830, 2005.

Tishby, N. and Slonim, N. Data clustering by markovian relaxation and the information bottleneck method. *Advances in Neural Information Processing Systems 13*, 2001.

Tishby, N., Pereira, F., and Bialek, W. The information bottleneck method. *Allerton Conf. on Communication, Control and Computing*, 1999.

Torkkola, K. Feature extraction by non-parametric mutual informtaion maximization. *Journal of Machine Learning Research*, 2003.

Victor, J. D. Binless strategies for estimation of information from neural data. *Physical Review*, 66, 2002.

Wang, Q., Kulkarni, S. R., and Verdu, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Information Theory*, pp. 2392–2405, 2009.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems 17*, 2005.