Towards a Statistical Theory of Clustering*

Ulrike von Luxburg[1] and Shai Ben-David[2]

[1] Fraunhofer IPSI, Darmstadt, Germany
[2] School of Computer Science, University of Waterloo, Canada

Abstract. The goal of this paper is to discuss statistical aspects of clustering in a framework where the data to be clustered has been sampled from some unknown probability distribution. Firstly, the clustering of the data set should reveal some structure of the underlying data rather than model artifacts due to the random sampling process. Secondly, the more sample points we have, the more reliable the clustering should be. We discuss which methods can and cannot be used to tackle those problems. In particular we argue that generalization bounds as they are used in statistical learning theory of classification are unsuitable in a general clustering framework. We suggest that the main replacements of generalization bounds should be convergence proofs and stability considerations. This paper should be considered as a road map paper which identifies important questions and potentially fruitful directions for future research about statistical clustering. We do not attempt to present a complete statistical theory of clustering.

1 Introduction

Clustering is one of the most widely used techniques for exploratory data analysis. Across all disciplines, from social sciences over biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. Despite this popularity of clustering, distressingly little is known about theoretical properties of clustering. One of the main reasons is that it is very difficult to evaluate the quality of a partition of some given data set other than with ad-hoc measures. Contrary to data analysis methods such as regression or classification, for clustering there exists no "ground truth". We wish to consider the common situation in which clustering takes place without having any significant prior knowledge about the subject data set. The very reason for performing clustering is that we want to discover a structure in the data which we did not know about before. Consequently, if a clustering algorithm does not achieve "good" results we do not know whether the reason is that the algorithm performs poorly or whether there is just no group structure in our data. In this paper we argue that even though the question "what clustering is" is difficult to answer in such generality, there are important sub-questions which are well defined and can and should be investigated in a general statistical framework.

* Submitted version, May 20, 2005

2 Asking the right questions

There are two completely different issues with clustering: the problem "what a cluster actually is" and the problem "how such a thing can be found in the data". More precisely:

1. How does a desirable clustering look like if we have complete knowledge about our data generating process?
2. How can we approximate such an optimal clustering if have incomplete knowledge about our data, or if we have limited computational resources?

To emphasize the difference between those two questions let us consider a simple example. Assume that we are astronomers and want to cluster a given set of stars in different categories. We assume that we know everything about those stars, that is we know the true values of all parameters describing the stars: how large they are, how bright they are, which materials they consist of, etc. To answer Question 1 we now have to define how we would like to cluster stars given all those parameters. Formally, the answer to this question is a function which for all combinations of parameter values returns a partition of the set of stars. In true life, we do not know the exact parameters of those stars, we only have some measurements which might be noisy and inaccurate. Furthermore, a full analysis of the data we have may require unacceptable computational time. This is the setting which Question 2 refers to. It asks for a method by which we can still discover or approximate the correct clustering even though we do not have the full information about the stars.

More abstractly, both questions can be formulated in a statistical setting where the data is generated according to some probability distribution. The first question asks how a given probability space should be clustered (as a function of the probability distribution). The second question asks how this clustering can be approximated if we only know a finite sample of data points from this probability distribution. Each question has a different nature and refers to different aspects of clustering. The first question is a *conceptual* question about *clusterings* of a given data space. The second question has two aspects to it, *algorithmic* and *statistical*, asking for a resource-efficient *clustering algorithm*. Answering Question 1 "simply" requires to make a *definition*. To answer Question 2, we have to propose an algorithm and then *prove* that this algorithm has the desired properties. The algorithmic Question 2 bears much similarity to classical theoretical machine learning questions. Like classification or regression tasks, it may be divided into two aspects – information and computation. That is, how can one overcome having incomplete data, and what is the complexity of carrying out the required computational tasks. In this paper we focus on the statistical aspects of clustering: Do the partitions constructed by a given clustering algorithm "converge" for growing sample size? If yes, how fast? What is the "limit clustering"? How far is the clustering of a given sample from the limit clustering? Which algorithm is the best one to use on my given finite sample?

On a given data set, does our algorithm fit random noise or does it discover true structure?

3 Complete information case

In this section we briefly discuss Question 1: if we have the full information about our data generating process, how do we cluster the space of objects? We denote the underlying probability space by $(\mathcal{X}, P)$. To answer Question 1 we have to define a function which for all probability distributions P tells us how the data space $\mathcal{X}$ should be clustered. The easiest way to do this is by defining a *quality measure* q of clusterings. In the literature there is an overwhelming number of quality measures for clusters on discrete sets, and most of them can also be used to define a clustering of the whole probability space $(\mathcal{X}, P)$. For example we can define the clustering induced by P as the partition of the space $\mathcal{X}$ into k sets which minimizes the expected distances of the individual points to the cluster centers. Or we maximize the expected ratio between the intra-cluster similarity to the inter-cluster similarity. Such quality measures are plausible heuristics, but often rather ad-hoc.

A definition implicitly present in many clustering algorithms is that the goal of clustering is to identify high density regions which are separated by low density regions (e.g., Cuevas et al. (2001), Stuetzle (2003), and references therein). Of course, for a sound mathematical analysis this definition has to be made precise. The advantage of this definition is that it is geometrically easy to understand. The disadvantage is that it suggests that to perform clustering we first have to estimate the level sets of the underlying distribution. As density estimation is very difficult in general, this does not seem to be a promising approach. As a side remark, note that it is an interesting open question whether clustering in this framework is (provably) easier than density estimation or not.

A third principal way to define clustering is via axiomatic approaches (e.g., Jardine and Sibson (1971), Wright (1973), Hartigan (1975), or Puzicha et al. (2000)). Instead of trying to define clustering explicitly, one states certain axioms which a clustering algorithm should satisfy, such as rotation invariance, invariance with respect to uniform scaling, and so on. The next step is then either to show that a certain clustering algorithm satisfies all axioms, or to the contrary, to prove that it is impossible to construct a clustering respecting all axioms (Kleinberg, 2003).

An approach which we find quite promising is to measure the quality of a clustering by its *"interestingness"*. A clustering is called interesting if it has a large distance to predefined "uninteresting" clusterings such as the trivial clustering. Instead of defining what a clustering should be, this approach defines what a clustering should not be. This concept has not yet been studied from a theoretical point of view, but it has already been implemented in practice, for example in the gap statistic (Tibshirani et al., 2001).

There are many more ways to define clustering, but, for every definition that we are aware of, it is rather obvious that it has major drawbacks. To a large extent, this is due to the fact that clustering is performed for a wide variety of very different purposes. If the goal is to compress the data, an approach maximizing the compression rate should be chosen. If the goal is to separate parts of the data which have been generated by different processes, a mixture model using the high density definition is more appropriate. Furthermore, for many important applications of clustering it is not at all clear what type of clustering is the "right" one (this is the case, for example, with clustering microarray gene expression data to detect functionally related genes, or clustering customers for marketing applications). Since we think that Question 1 cannot be solved in full generality we suggest to start with the Question 2. We will see that many important aspects of this question can be answered even without having an answer to the basic conceptual Question 1. In a sense, some of the answers that we suggest for Question 2 can be viewed as necessary conditions on a satisfactory answer to Question 1.

4 Comparing clusterings

A basic tool for any statistical analysis of clustering is a formal measure of the distance (or similarity) between different clusterings. In the literature there exist many definitions of distances between clusterings *of the same set*. However, to compare clusterings of two independent random samples, one needs to measure the distance between clusterings of *different sets*. We propose to do that by using an extension operator. Given any clusterings C_1 and C_2 of subsets $S_1, S_2 \subset \mathcal{X}$, respectively, we extend each of them to a clustering of the whole space $\mathcal{X}$ and then compare the extended clusterings. For example, an algorithm where such an extension is straight forward to implement is k-means. Given a k center clustering of a sample, each point in the whole data space is attributed to the closest sample cluster center. Such extension operators can also be defined for non-center based algorithms, as an example take the extension of spectral clustering suggested in von Luxburg et al. (2004, 2005).

A completely different method of comparing clusterings can be used if we are working with a quality measure q of a clustering (cf. Section 3). Instead of comparing two clusterings C_1 and C_2 directly, we simply compute the distance $|q(C_1) - q(C_2)|$ between their qualities. One can view such a measure as a projection of the space of clusterings into one aspect - the quality measure at hand. We expect that similar clusterings will have similar quality values, but the reverse statement may often fail - clusterings may have the same quality value while still being very different from each other. One also needs to be careful to make sure that comparing qualities of different clusterings makes. For example, we might need to take into account the sizes of the sets S_1 and S_2.

5 Finite sample case

In this section we assume that the underlying distribution P is unknown, but we are given a finite sample $X_1, ..., X_n$ which has been drawn i.i.d. with respect to P. We denote its empirical distribution by P_n.

5.1 Generalization bounds - classification versus clustering

For classification, generalization bounds are an omni-present tool to answer all kinds of statistical questions about classification. They are used to prove convergence of certain algorithms, they provide convergence rates, they can be used to estimate errors on a given sample, and often they are employed for model selection purposes for a particular sample. Intuitively, classification and clustering are closely related. In both problems we try to partition a given data set into groups, the only difference is that in classification we have additional prior knowledge how to do this. Thus the motivation to derive "generalization bounds for clustering" is obvious. In this section we argue that in some special situations such bounds can also be useful for clustering, but in the general case the classification generalization bounds do not have a corresponding counterpart in clustering.

Let us recapitulate the roles of generalization bounds in the statistical learning theory of classification. The overall goal of classification is to find a classifier f which has a small true risk $R(f) = E_P(\ell(x, y, f(x))$. Here ℓ is a loss function which for each point x in the input space measures the distance between its true label y and the predicted label $f(x)$. The expectation is taken with respect to the underlying probability distribution P. As the latter is unknown, we cannot compute $R(f)$ on a finite sample $(x_i, y_i)_{i=1,...,n}$. Instead we evaluate the empirical risk $R_{\text{emp}}(f) = E_{P_n}(\ell(x, y, f(x)) = 1/n \sum_{i=1}^{n} \ell(x_i, y_i, f(x_i))$. Generalization bounds are now used to bound the distance between the true and the empirical risk:

$$P(|R_{\text{emp}}(f) - R(f)| > \varepsilon) \leq ...$$

Usually, those bounds are worst case bounds which simultaneously hold for all functions f in the function class $\mathcal{F}$ used by the classification algorithm. In practice we then argue as follows. From the set of functions $\mathcal{F}$, the algorithm chooses a function f_n that has small empirical risk. The generalization bound guarantees that the empirical risks $R_{\text{emp}}(f)$ of all functions $f \in \mathcal{F}$ are close to their true risks $R(f)$. In particular this holds for the function f_n. As we already know by the mechanism of the algorithm that f_n has a small empirical risk $R_{\text{emp}}(f_n)$, we can then conclude that it also has a small true risk $R(f_n)$. This is what we wanted to know. Note that statistical learning theory does not tell us anything about how close the risk of the classifier f_n is to the risk of the best possible classifier f_* (the Bayes classifier). We only specify that if we pick a function which has a small empirical risk on the sample, then this function is likely to have a small true risk. This also leads to statements which can compare the risk of a given classifier f_n to the best classifier in the set $\mathcal{F}$, but we never attempt to compare to classifiers which are not in the function class $\mathcal{F}$.

Pursuing the analogy between classification and clustering, one may try to develop a similar notion of a risk for clustering. This risk has to be defined as the expectation of some loss function on individual points. There are a few clustering paradigms where it is natural to define such a loss function. For example in the k-means framework we can define the loss $\ell(x, C)$ of data point x with respect to clustering C as the distance of x to its corresponding cluster center. The quality $q(C)$ of the clustering C is then the expected loss $E_P(\ell(x, C))$, and the empirical quality is then estimated by $q_{\text{emp}}(C) = E_{P_n}(\ell(x, C))$. Then we can work in a framework of "empirical quality maximization" which is completely analogous to empirical risk minimization for classification. The results of Ben-David (2004) are an example of bounds obtained along these lines.

However, the assumption that the quality of a clustering is an expectation over some function of individual samples is hardly ever satisfied in clustering (k-means being a noble exception). In the general case, where the information we care about goes beyond some quality function, this approach breaks down. On a first glance, one possible substitute might be to define the true risk of a clustering C_n as its distance to the (unknown) true clustering C_*, that is $R(C_n) := d(C_n, C_*)$ for some distance function d between clusterings. But in contrast to the classification scenario, here we cannot estimate this risk by some "empirical risk"; such an estimator simply does not exist. In classification we are given some information about the optimal classifier f_* by the training labels. But in clustering we do not have any information on C_*, and hence we cannot build an estimator of $d(C, C_*)$. So the generalization bound toolbox, which allows us to effectively estimate the true risk of a classifier by some empirical risk, does not have any application in the general clustering setting.

In Section 5.3 we will discuss an approach of *stability bounds*, which may be seen as a weak version of generalization bounds.

5.2 Convergence of clustering algorithms

The convergence of a clustering algorithm provides evidence for the intuition that the more data points we get, the more reliable the result of the clustering algorithm should be. An algorithm which does not converge produces rather unpredictable results on any given sample and thus is completely unreliable. Conversely, if an algorithm does converge, it can be investigated whether, at least for some prototypical examples, the limit clustering is a useful clustering of the data space or not.

To define what convergence means, let us fix some clustering algorithm $\mathcal{A}$ and a sequence of sample points $(X_n)_{n \in \mathbb{N}}$ drawn i.i.d according to the unknown probability distribution P. Let C_n be the clustering constructed by algorithm

$\mathcal{A}$ on the first n data points $X_1, ..., X_n$. Let C be some clustering of $\mathcal{X}$, and d some distance function between clusterings. We say that the *sequence of clusterings* $(C_n)_{n \in \mathbb{N}}$ *converges to some limit clustering* C if $d(C_n, C) \to 0$. In the last section we have already seen that it is impossible to derive statistical learning theory type bounds for the distance $d(C_n, C)$ if C is unknown. Thus we have to investigate the question of convergence for each clustering algorithm individually. It is mainly for three classes of non-parametric clustering algorithms that certain convergence properties are known: k-means (Pollard (1981), see Lember (2003) for a recent overview), linkage algorithms (Hartigan, 1981, 1985), and spectral clustering (von Luxburg et al., 2004, 2005). In particular the latter example shows that convergence analysis of clustering algorithms can lead to unexpected insights which are very relevant for practical applications of the algorithm. For spectral clustering it can be seen that one variant of the algorithm always converges, while the other one can fail to converge or can converge to trivial solutions. Hence we advocate for the importance of convergence analysis for clustering algorithms, an issue which has not been taken very seriously in the past.

5.3 Stability bounds for model selection

Convergence results for clustering algorithms lead to a first argument for or against those algorithms in general, but they do not help us further in choosing a suitable algorithm for a particular data set. As we have only very restricted means of measuring the general quality of a clustering (Question 1), we have to resolve to more indirect measures. One such measure is stability. We expect that the results of a "good" clustering algorithm are stable with respect to the sampling process, that is they do not change much if we draw another sample or add or delete some points from our sample. In particular, stability is an indication whether the model proposed by some algorithm fits to the data or not. For example, if our data contains three true clusters, but we use a clustering algorithm which looks for four clusters, the algorithm wrongly needs to split one of the clusters into two clusters. Which of the three true clusters are split might change from sample to sample, and thus the result will not be very stable. Ben-David (2005) demonstrates such behavior on several basic probability distributions. Stability has already been used in practical applications of clustering, see for example Levine and Domany (2001), Ben-Hur et al. (2002), Dudoit and Fridlyand (2002), Lange et al. (2003).

Let S_m and S'_m two independent samples of size m from the underlying distribution P, $C(S_m)$ and $C(S'_m)$ the clusterings constructed by some algorithm $\mathcal{A}$ on the respective samples, and d some a distance measure between clusterings. The *stability* of the clustering algorithm $\mathcal{A}$ for distribution P and m sample points can be defined as

$$\beta(\mathcal{A}, P, m) = E_P(d(C(S_m), C(S'_m))).$$

It measures how much the clusterings differ across different samples of the same size. Note that β actually measures the *in*stability of an algorithm: The smaller β is, the the more stable is the underlying algorithm. Thus maximizing the stability of an algorithm means to minimize the quantity β. As β refers to the unknown distribution P, we need to estimate it by some empirical quantity which can be computed on the given sample. With such an estimator we then can establish an "empirical stability maximization" framework. From some set of given models (that is, clustering algorithms) we chose a model for which the estimated stability is high. Then we need to prove "stability bounds" to make sure that high estimated stability implies that the true stability of the algorithm is high. For example, Ben-David (2005) proves such a bound for a certain measure of distance between clusterings:

Theorem (Ben-David, 2005) *For every probability distribution P and any center-based clustering algorithm $\mathcal{A}$, for any $m \in \mathbb{N}$ and $t > 0$, if S_1, S_2 are two i.i.d. P-random m-size samples, then,*

$$P\left[|d(C(S_1), C(S_2)) - \beta(\mathcal{A}, P, m)| > t\right] < e^{-mt^2}.$$

This theorem tells us that for clusterings C_1, C_2 computed by $\mathcal{A}$ on two independent m-samples, the distance $d(C_1, C_2)$ is a good estimator of the stability $\beta(\mathcal{A}, P, m)$ of the algorithm.

Let us make one remark about the relation between stability and convergence. Of course, if we know that a clustering algorithm converges, we also know that it will be stable in the long run, that is stability for the limit for $n \to \infty$ is guaranteed. Stability as defined above measures changes that occur in a regime of n data points. This is a different concept than convergence in that it is connected to a particular sample size n. An algorithm might be very stable if we only alter one of n data points, but still in the long run might oscillate between different solutions. The other way round, an algorithm might be very unstable if it only knows n data points, but might converge very fast once it gets enough data. Thus the concepts of convergence and stability are complementary aspects. However, as stability for large n is already very close to the convergence of an algorithm we suggest to use the slightly weaker concept of stability bounds in cases where convergence of algorithms is difficult to show. Note that this suggestion implicitly assumes that in the long run, stability is more or less monotonic in the sample size. This is an issue which needs further investigation.

6 Other desirable clustering features and directions for further research

One fundamental problem of the stability approach is how we can avoid choosing trivial solutions. By the definition of stability, an algorithm which always proposes the same or the trivial clustering is very stable. Thus it will not be enough to maximize the stability of a solution, we will also have to ensure that the proposed solution is useful. One idea is that additional to being stable, a good clustering algorithm should also be flexible in that it should be capable of constructing a large variety of different clusterings. We could say that an *algorithm $\mathcal{A}$ is flexible* if there is a big family $\mathcal{F}$ of mutually distant clusterings, and a family of probability distributions $\mathcal{P}$, such that for large enough sample sizes m, for every $C \in \mathcal{F}$ there exists some $P \in \mathcal{P}$ so that if S_m is an m- sample of P then the clustering $C(S_m)$ proposed by $\mathcal{A}$ is close to C, with high probability. The properties of stability and flexibility capture opposing features: algorithms that are insensitive to sample variations are more stable, whereas flexibility requires sensitivity to sample variations. Interestingness as briefly introduced in Section 3 is an independent concept which would very well fit into a framework together with stability and flexibility.

Ultimately, we would like to find a method to achieve stable and interesting results by a flexible clustering algorithm. How such a statement can be cast into a more precise framework is an interesting challenge for future research. We think that some methods of statistical learning theory can be recycled for this purpose, but as we indicated above, some completely new tools might be necessary to achieve this goal. In our opinion, clustering is a great playground for statisticians, and many important results remain to be discovered.

Bibliography

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median clustering. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 415–426. Springer, 2004.

S. Ben-David. A notion of stability for statistical clustering with applications to model selection. Technical Report, University of Waterloo, University of Waterloo, Canada, 2005.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, 2002.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36: 441–459, 2001.

S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, 3(7), 2002.

J. Hartigan. *Clustering algorithms.* Wiley, New York, 1975.

J. Hartigan. Consistency of single linkage for high-density clusters. *JASA*, 76 (374):388–394, 1981.

J. Hartigan. Statistical theory in clustering. *Journal of classification*, 2:63–76, 1985.

N. Jardine and R. Sibson. *Mathematical taxonomy.* Wiley, London, 1971.

J. Kleinberg. An impossibility theorem for clustering. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 446–453. MIT Press, Cambridge, MA, 2003.

T. Lange, M. L. Braun, V. Roth, and J. Buhmann. Stability-based model selection. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 617–624. MIT Press, Cambridge, MA, 2003.

J. Lember. On minimizing sequences for k-centres. *Journal of Approximation Theory*, 120:20–35, 2003.

E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1): 135–140, 1981.

J. Puzicha, T. Hofmann, and J. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 2003.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *J. Royal. Statist. Soc. B*, 63(2):411, 2001.

U. von Luxburg, O. Bousquet, and M. Belkin. On the convergence of spectral clustering on random samples: the normalized case. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 457–471. Springer, 2004.

U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems (NIPS) 17*. MIT Press, Cambridge, MA, 2005.

W. Wright. A formalization of cluster analysis. *Pattern Recognition*, 5:273–282, 1973.