JMLR: Workshop and Conference Proceedings 27:65–79, 2012 Workshop on Unsupervised and Transfer Learning

Clustering: Science or Art?

Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and NICTA, Canberra ACT 0200, Australia

Isabelle Guyon ISABELLE@CLOPINET.COM
ClopiNet, 955 Creston Road, Berkeley, CA 94708, USA

Editor: I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver

Abstract

We examine whether the quality of different clustering algorithms can be compared by a general, scientifically sound procedure which is independent of particular clustering algorithms. We argue that the major obstacle is the difficulty in evaluating a clustering algorithm without taking into account the context: why does the user cluster his data in the first place, and what does he want to do with the clustering afterwards? We argue that clustering should not be treated as an application-independent mathematical problem, but should always be studied in the context of its end-use. Different techniques to evaluate clustering algorithms have to be developed for different uses of clustering. To simplify this procedure we argue that it will be useful to build a "taxonomy of clustering problems" to identify clustering applications which can be treated in a unified way and that such an effort will be more fruitful than attempting the impossible — developing "optimal" domain-independent clustering algorithms or even classifying clustering algorithms in terms of how they work.

1. Introduction

Knuth (1974) said of computer programming that *"It is clearly an art, but many feel that a science is possible and desirable."* Whether clustering is art or science is an old question: *"Is taxonomy art, or science, or both?"* asked Anderson (1974) whilst reviewing the state of systematic biological taxonomy. He justifiably went on to claim

> *Discussions of taxonomic theory or practice that refer to the concepts "science" and "art" without finer delineation will be less clear and less productive than discussions that first attempt definitions of specific concepts in the panoply of science such as precision or objectivity or repeatability or confidence and then apply these explicitly in evaluating alternatives and specific steps within the taxonomic process* (Anderson, 1974, p. 59).

The purpose of this paper is to provide such a finer delineation taking as our starting point the means by which clustering algorithms are evaluated. Our original motivation was to develop better benchmark challenge problems for clustering. Our reflections on this lead

us to grapple with the broader issue of the point and purpose of clustering (and abandon the idea of such benchmarks).

Clustering is "unsupervised classification" or "unsupervised segmentation". The aim is to assign instances to classes that are not defined *a priori* and that are (usually) supposed to somehow reflect the "underlying structure" of the entities that the data represents. In the broader, non machine learning literature it is common to use the word "classification" when talking of clustering (Bowker and Star, 1999; Farris, 1981). "Taxonomy" refers to what machine learners would call hierarchical clustering.

Clustering relates data to knowledge and is a basic human activity. Bowker and Star (1999) have argued how fundamental it is in understanding the world[1]. It affects knowledge representation and discovery (Kwasnik, 1999). It defines infrastructures that have real political significance (Bowker and Star, 1999). It forms the basis for systematic biology, and the need for classification remains ever-present (Ruepp et al., 2004). It is pervasive.

Clustering holds a fascination for many mathematicians and engineers and as a consequence there is a large literature on domain independent clustering techniques. However, this literature rarely finds its way to practitioners and has long been criticized for its lack of relevance: Farris (1981) wrote of a 1976 symposium on "Classification and Clustering" that

> *The technical skill shown by many of the contributors to this symposium might well produce valuable new methods, if it could be directed to problems of systematic importance.*

The lack of appreciation is puzzling. *Supervised* classification techniques *are* widely appreciated (and used) for solving real problems. And clustering seems to simply be "unsupervised classification." However, there is a fundamental difference: supervised clustering can be easily made into a well defined problem with a loss function, which precisely formalizes what one is trying to do (and furthermore can be grounded in a rational way in the real underlying problem). The loss function can be viewed as an *abstraction* of the ultimate end-use problem. The difficulty with unsupervised clustering is that there are a huge number of possibilities regarding what will be done with it and (as yet) no abstraction akin to a loss function which distills the end-user intent. Depending on the use to which a clustering is to be put, the same clustering can either be helpful or useless.

It is often presumed that for any situation where clustering may be used there is a single "right" clustering. ("The goal of data clustering ... is to discover the *natural* grouping(s) of

1. The reason why Borges' famous strange classification quoted below strikes us as so bizarre is precisely because it makes us wonder what on earth would it be like to understand the world in that extraordinary manner — "the impossibility of thinking that" (Foucault, 1970).

> These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia called the Heavenly Emporium of Benevolent Knowledge. In its distant pages it is written that animals are divided into (a) those that belong to the emperor; (b) embalmed ones; (c) those that are trained; (d) suckling pigs; (e) mermaids; (f) fabulous ones; (g) stray dogs; (h) those that are included in this classification; (i) those that tremble as if they were mad; (j) innumerable ones; (k) those drawn with a very fine camel's-hair brush; (l) etcetera; (m) those that have just broken the flower vase; (n) those that at a distance resemble flies. (Borges, 1999)

a set of patterns points, or objects" (Jain, 2010, p.3)[2]) Others take this further and maintain that the right answer is determinable by the data (alone, without reference to intended use): "the data should vote for their preferred model type and model complexity" (Buhmann, 2010). The presumption seems to be based on the notion that categories exist independently of human experience and intent (a sort of Platonic "carving nature at its joints"). Such a doctrine of "natural kinds" has been convincingly discredited (see e.g. Gilmour and Walters (1964)). Philosophical analysis of "natural kinds" reveals substantial difficulties (Bird and Tobin, 2010), see also Lakoff (1987) for a cognitive science perspective. The problems of such "absolute" approaches to clustering are also demonstrated in Kleinberg (2003).

Users of classification methods often reject the notion that clustering is a domain-independent subject:

> *I suspect that one of the reasons for the persistence of the view that classification is subject-independent is that classificatory theorists have been largely insulated from sources that would inform them otherwise* (Farris, 1981, p.213).

The same problem occurs in supervised classification: if one is not prepared to commit to a particular loss function (as a formal codification of the use to which your classifier will be put) one can just estimate the underlying probability distribution. But then one is stuck with the question of *how to judge the quality of such a distribution estimate*. There can be no loss-independent way of doing this that is universally superior to other methods; pace the use of the area under the receiver operating characteristic curve (Hand, 2008).

Many of the arguments about the right way to cluster or how to compare clustering methods are side-effects of the fact that there is a very wide diversity of clustering problems. Even within a particular domain of application (say the classification of biological organisms) there can be very diverse and opposing views as to what constitutes a valuable classification: see the account of Hull (1988) of the battles between *phenetics*, which attempts to classify on the basis of observable characteristics ignoring phylogeny, and *cladistics*, which is avowedly phylogenetic. The arguments between adherents of either of these camps are not resolvable (even in principle) by domain-independent means — their conflict stems from a disagreement regarding what is the real problem that needs solving.

The focus of the present paper is on *problem solving* with clustering and how clustering methods are *used* rather than on the algorithmic details of the *techniques*; there are already many comprehensive reviews of techniques available (e.g., Jain et al., 1999; Xu and Wunsch, 2005; Berkhin, 2006). In this paper we do not really care how a clustering algorithm works, as long as it achieves the goal we have set. From this perspective it is pointless to argue whether clustering is *essentially* density-level set estimation, information-compression[3] or an instance of a problem in graph theory. We argue that theoretical or methodological motivations of clustering algorithms alone are insufficient to qualify clustering as a scientific method. Some practitioners think that in the past research focused too much on this methodological side. Johnson (1968, p. 224) expressed his frustration caustically:

2. Ironically, in the same paper Jain recognizes the point, which contradicts any notion of "natural" that "the representation of the data is closely tied with the purpose of the grouping. The representation must go hand in hand with the end goal of the user" (Jain, 2010, p.11). We do not believe there can be a "true" clustering definable solely in terms of the data — truth is relative to the problem being solved.
3. Shannon's prophetic words are still true: "the use of a few exciting words like *information, entropy, redundancy*, do not solve all our problems" (Shannon, 1956).

> *The theoreticians of numerical taxonomy have enjoyed themselves immensely over the past decade (though not without developing several schools with scant respect for each other!). The mushrooming literature is quite fascinating and new developments tumble after each other. Anyone who is prepared to learn quite a deal of matrix algebra, some classical mathematical statistics, some advanced geometry, a little set theory, perhaps a little information theory and graph theory, and some computer technique, and who has access to a good computer and enjoys mathematics (as he must if he gets this far!) will probably find the development of new taximetric methods much more rewarding, more up-to-date, more 'general', and hence more prestigious than merely classifying plants or animals or working out their phylogenies.*

Arguably, it is the most important part of the scientific process to evaluate whether methods serve the end goals they have been designed for. We believe that there is an urgent need for such evaluation procedures for clustering.

2. Deficiencies in current clustering evaluation

In this section we discuss why most of the methods used in the clustering literature to evaluate clustering algorithms are very problematic and do not serve their purpose. The point we want to make is that clustering algorithms cannot be evaluated in a problem-independent way: whether a clustering of a particular data set is good or bad cannot be evaluated without taking into account what we want to do with the clustering once we have it. This insight is remarkably old. Gilmour and Walters (1964, p.5) quote Mercier (1912):

> *The nature of the classification that we make ... must have direct regard to the purpose for which the classification is required. In as far as it serves the purpose, the classification is a good classification, however 'artificial' it may be. In as far as it does not serve this purpose, it is a bad classification, however 'natural' it may be.*

In this section we argue that this insight is completely ignored by most of the current literature on clustering. Scanning the current literature on clustering algorithms one will find that one or several of the following methods are typically used to argue for the success of a clustering algorithm. We think that all these methods are insufficient and can be completely misleading.

Evaluation on artificial data sets. Clustering algorithms are applied to artificial data sets, for example points drawn from a mixture of Gaussians. Then the clustering results are compared against the "ground truth". Such a procedure can make sense to evaluate the statistical performance of a clustering algorithm under particular assumptions on the data generating process. It cannot be used to evaluate the *usefulness* of the clustering — usefulness cannot be evaluated without a particular purpose in mind.

Evaluation on classification benchmark data sets. Clustering algorithms are applied to classification data sets, that is data sets where samples come with class labels. Then the class labels are treated as the ground truth against which the clustering results of

different algorithms are compared (using one out of various scores as the (adjusted) Rand index, misclassification error, F-measure, normalized mutual information, variation of information, and so on). High agreement with the ground truth is interpreted as good clustering performance.

We believe this approach is dangerous and misleading: It is an *assumption* that class labels coincide with cluster structure and that the "best" clustering of the data set coincides with the labels. This assumption might be true for some data sets but not for others. There might even exist a more "natural" clustering of the data points that is not reflected in the current class labels. Or, as it is often the case in high-dimensional data, different subspaces of features support completely different clusters. Consider a set of images that is labeled according to whether it contains a car, but a clustering algorithm decides to cluster the images according to whether they are greyscale or color. In such a case, the clustering algorithm discovers a very reasonable clustering, but achieves a very bad classification error. The classification error by itself cannot be used as a valid score to compare clusterings.

Evaluation on real world data sets. Sometimes people run their algorithm on a real data set, and then try to convince the reader that the clusters "make sense" in the application; this is claimed by some to be "the best way to evaluate clustering algorithms" (Kogan, 2007, p.156)[4]. For example, proteins are grouped according to some known structure. This is more or less a qualitative version of the approach using benchmark data sets. It can make sense if the clustering algorithm is intended for use in exploratory data analysis in this particular application, but does not carry any further meaning otherwise.

Internal clustering quality scores. There exist many internal scores to measure how "good" a clustering is (sum of square distances to cluster centers, ratio of between-cluster to within-cluster similarities, graph cut measures like the normalized cut or the Cheeger cut, likelihood scores, and so on). We argue that all these scores are unsuitable to evaluate the quality of clustering algorithms in an objective way. Such *scores are useful on the level of algorithms* where they can be used as an objective function in an optimization problem, and it is a valid research question how different scores can be optimized efficiently. However, across different algorithms these scores tell only little about the usefulness of the clustering. For every score preferring one clustering over the other one can invent another score which does the opposite. A unique, global, objective score for all clustering problems does not exist.

In supervised classification we are faced with a similar problem. Depending on whether we compare algorithms based on the zero-one loss or the area under the ROC curve, say, we may get different answers (that is different algorithms will be superior depending on the measure used). However, the advantage we have in supervised learning is that we can abstract from the real problem we have to solve by introducing a loss function which can guide the choice of solution. In clustering, this cannot be achieved in a domain-independent function, which makes the situation much worse.

4. Compare Farris (1981, p. 208): "A clustering method is selected in each application for its ability to manufacture a grouping most in accord with the subjective feelings of a 'professional taxonomist.' (That taxonomist, of course, will then claim vindication of his views; they have been verified by an 'objective' method!) One must wonder what value might be attributed to a method chosen primarily for its failure to contradict preconceptions."

Universal "Benchmark data sets". The UCI approach to supervised classification (whereby there is a small fixed collection of data sets, divorced from their real end use, that are used as a simple one-dimensional means of evaluating machine learning solutions) is widely used to compare supervised learning algorithms. However its value can be questioned even in the supervised case. For example, although lip-service is often paid to the idea that a loss function can be derived from utilities (arising in the particular end-use problem one is solving) these benchmark problems are typically only evaluated on a single loss function. Given the diversity of end uses to which clustering is put, any such approach seems hopeless for clustering. One need only look at how problematic it is to study taxonomic repeatability within a particular domain to be daunted (Moss, 1971).

We believe it makes sense to study particular applications of clustering and find out what procedures are good or bad per application. But this is a process of interaction between algorithm developers and practitioners. We need input from practitioners to help judge whether results are good or bad. We do not believe that such a process can be automated — the data sets would need to have a "true classification", and then we are in the situation described above.

3. Proposed method of evaluation: measure the *usefulness* for the *particular task* under consideration

As we have discussed above, clusterings or clustering algorithms cannot be evaluated without taking into account the use the clustering will be put to. To sketch how evaluation procedures might look like if they do take into account the use the clustering is put to, let us consider the two following, very distinct scenarios.

3.1. Evaluating the usefulness of clustering: Two example scenarios

Clustering for data pre-processing. Often "one does not learn structure for its own sake, but rather to facilitate solving some higher level task" (Seldin and Tishby, 2010). Clustering is often used as an automated pre-processing step in a whole data processing chain. For example, we cluster customers and products to compress the contents of a huge sales data base before building a recommender system. Or we cluster the search results of a search engine query to discover whether the search term was ambiguous, and then use the clustering results to improve the ranking of the answers. In such situations, the whole purpose of clustering is to improve the overall performance of the system. This overall performance can usually be quantified by some *problem-dependent* score.

From a methodological point of view it is relatively straightforward and uncontroversial how clustering can be evaluated in this scenario. One can interpret the clustering as just one element in a whole chain of processing steps. Put more extremely, the clustering (algorithm) is just one more "parameter" which has to be tuned, and this tuning can be achieved similarly as for all other parameters, for example by cross validation over the final outcome of our system as a whole. We do not directly evaluate the "quality" of the clustering, and we are not interested in whether the clustering algorithm discovers "meaningful groups". All we care about is the usefulness of the clustering for achieving our final goal. For example, to

build a music recommender system it might be a useful preprocessing step to cluster songs or users into groups to decrease the size of the underlying data set. In this application we do not care whether the clustering algorithm yields a meaningful clustering of songs or users, as long as the final recommender system works well. If it performs better when a particular clustering algorithm is used, this is all we need to know about the clustering step.

If the final application in mind is indeed supervised classification and clustering is performed as a pre-processing step, then one can evaluate the quality of the clustering in terms of the degradation of classification performance, when quantized observations are used in place of the original observations (Bock, 1992), or the improvement in performance arising from the additional information in the class labels, along with the original observations (Candillier et al., 2006).

Clustering for exploratory data analysis. Here, clustering is used to discover aspects of the data which are either completely new, or which are already suspected to exist, or which are hoped not to exist. For example, one can use clustering to define certain sub-categories of diseases in medicine, or as a means for quality control to detect undesirable groupings that suggest experimental artifacts or confounding factors in the data. Exploratory data analysis should "present the data to the analyst such that he can see patterns in the data and formulate interesting hypotheses about the data" (Good, 1983).

As far as we know, no systematic attempt has been made to assess whether clustering in general (or a particular clustering algorithm) is useful for exploratory data analysis. In addition to the technical question concerning how to perform the clustering, this question has a psychological aspect. Ultimately it is a human user who will explore the data and hopefully detect a pattern. One approach to evaluate the performance of a certain clustering algorithm might be to ask humans to use a particular clustering algorithm to generate hypotheses, and later evaluate the quality of the hypotheses on independent data. Major obstacles to this endeavor are how to evaluate whether a hypothesis is "interesting," and how to perform a "placebo clustering" as a null model to compare with.

Data exploration is often performed visually. If clustering and visualization are treated as two independent components of some data exploration software, we believe it likely that the particular choice of the clustering algorithm is not very relevant compared to the design of the human computer interaction interface — the visualization and data manipulation capabilities of the system will likely be responsible for success or failure of the attempt to discover structure in the data. As opposed to treating clustering and visualization independently from each other, it is a promising approach to consider them jointly. For example, it might make sense to sacrifice a bit of accuracy in the clustering algorithm if this leads to a performance gain in the visualization part (consider the t-SNE algorithm of van der Maaten and Hinton (2008) as an example). The evaluation of such a system has to take into account the human user as well.

3.2. Can we optimize the "usefulness" directly?

The information bottleneck approach (Tishby et al., 1999) attempts to directly optimize the usefulness of a clustering. It tries to find a cluster assignment of all input points that is as "informative as possible" (in terms of mutual information) about a particular

"property of interest". At first glance, this framework seems to be exactly what we are looking for. At second glance, one realizes that it is not so obvious how to implement it in practice. Often it is very hard to quantify a "feature of interest". In the exploratory data analysis setting this seems close to impossible. In the data pre-processing setting, we are interested in a high classification accuracy in the end, which is too abstract a target for the information bottleneck approach (one may as well directly optimize classification performance (Bock, 1992)). Furthermore, the method unjustifiably assumes that Shannon information is "intrinsic" and captures the essence of meaning in the data. There are in fact many different notions of information and even just "gathering information" from the data implicitly presumes a loss function (DeGroot, 1962). Formally the choice of a notion of information is tantamount to the choice of a loss function in a supervised learning problem (Reid and Williamson, 2011). Nevertheless we believe that of all the literature on clustering, the information bottleneck is closest in intent to what we are interested in as it at least tries to take into account "what we are interested in."

3.3. How are meta-criteria like clustering stability related to the usefulness?

In a statistical setting, it is assumed that the given data points are samples from some underlying probability distribution. There are many data sets where such an assumption makes sense (customers are samples from the "set of humans"; a particular set of handwritten digits just contains a few instances out of a much larger set of "all possible hand written digits"). In such a setting, it has often been advocated that it is important to ascertain whether a particular clustering just "fits noise" or uncovers "true structure" of the data. There are several different tools that attempt to distinguish between these two cases. Below are four such notions (sorted by increasing stringency).

Stability. The same clustering algorithm is applied repeatedly to perturbed versions of the original data. Then a stability score is computed that evaluates whether the results of the algorithm are "stable" or "unstable". If the results are unstable, they are considered unreliable and unsuitable for further use. See von Luxburg (2010) for a large list of references.

Convergence of clustering algorithms. The question is whether, for increasing amounts of data, the results of a clustering algorithm converge to a particular solution and whether this solution is reasonable. See Pollard (1981) or von Luxburg et al. (2008) for examples.

Generalization bounds. One computes generalization bounds that tell how much the clustering results obtained on a finite sample are different from the ones one would obtain on the full underlying distribution. These bounds are in the tradition of statistical learning theory and depend on the size of the class of models from which the clustering is chosen. See Buhmann (2010) for an approach where the richness of the hypothesis is balanced against the stability of the clustering results and Seldin and Tishby (2010) for PAC-Bayesian generalization bounds for the expected out-of-sample performance of clustering.

Statistical significance. Here the goal is to assign confidences scores to clustering results. They should tell how confident we are that the clustering results significantly deviate

from some null model of "unclustered" data. In many branches of science it is a strong requirement to report such confidence scores. There exist numerous ways to compute confidence scores in the literature, see Efron et al. (1996) for an example.

None of the criteria listed above directly evaluates the quality of a particular clustering of a particular data set; they always take into account the clustering *algorithm* (respectively, the model class from which the clustering solution was chosen). The purpose of all the criteria mentioned above is to handle the statistical uncertainty in the data.

How are these criteria related the usefulness of a clustering? Their importance also depends on the particular use of the clustering. In the pre-processing setting. statistical considerations do not play any role, as long as the system works. If the clustering algorithm gives different results on different samples, but the system works on either of these results, then we are fine. For example, many people use k-means to cluster a huge set of texts, say, into a set of manageable size. If we want to cluster a data set of 10^6 items into 10^3 clusters, it does not really matter whether the clusters correspond to true underlying structure — they only have to serve for data compression. We can have a completely different clustering each time we get a new data set, but the overall system might still work fine on any of these data sets.

Interestingly, it is in the setting of "discovering structure" that statistical significance is important. In the exploratory setting, a user does not have infinite time to inspect all sorts of meaningless clusterings, and we cannot hope to generate a meaningful hypothesis from nothing. In this sense, statistical significance is a necessary (but not sufficient) criterion for an algorithm to be useful. Similarly, significance is important in taxonomic applications, for example when defining species in biology. Here we want to define categories based on underlying structure and not on noise. This insight is quite striking: it is the "soft" exploratory data analysis and discovering structure setting where we have the "hardest" statistical requirements for our clustering algorithms.

It is conceivable that one could define the problem of "structure discovery" precisely enough to allow one to then analyse the statistical significance of structures so discovered, but we do not have any concrete ideas concerning this. We imagine that at best it would be "discovery" from a predetermined set of structures. Ideally such a test would condition on the data (Reid, 1995).

4. A suggestion for future research

We have seen that clustering is used in a variety of contexts with very different goals and that clustering results cannot be evaluated without taking into account this context. We are left with the question as to what can be done. We take our lead from Tukey (1954):

> *Difficulties in identifying problems have delayed statistics far more than difficulties in solving problems. This seems likely to be the case in the future too. Thus it is appropriate to be as systematic as we can about unsolved problems. . . . Different ends require different means and different logical structures. . . . While techniques are important in experimental statistics, knowing when to use them and why to use them are more important.*

4.1. A systematic catalog of clustering problems

We believe that it would be valuable (and relatively straight-forward) to compile a table of different clustering problems and corresponding evaluation procedures. A more effective approach might be to come up with a way to treat several clustering problems with similar methods, so one does not have to start from scratch for every new application. We believe that the most effective approach would be to systematically build a taxonomy or catalog of clustering problems (Hartigan, 1977). We believe that this should be done in a solution agnostic way: define the problem in a purely declarative manner without trying to say how it should be solved. We conjecture that such a taxonomy will be of considerable help. This proposal is tantamount to the research program "left to the reader as an exercise" in the self-referentially titled *The Botryology of Botryology* (Good, 1977). This proposal is quite different to the several clusterings of clustering *algorithms* that have appeared Jain et al. (2004); Jain (2010); Andreopoulos et al. (2009) and which do not really address an end-users' concerns. We emphasize that our focus is on distinguishing properties of clustering *problems*, not *algorithms*. Hence, for our catalogue of clustering problems it is irrelevant to take into account distinctions like parametric vs. non-parametric, frequentist vs. Bayesian, model-based vs. model free, information-theoretic vs. probabilistic, etc. Of course different algorithms "solve" different problems — but we are suggesting to attempt a classification of problems in a manner independent of how to solve it. Such a declarative approach has proved valuable in computer programming and other branches of engineering.

We do not yet attempt to suggest how this catalog of clustering problems will look in any detail, but do suggest several "dimensions" of clustering applications which may be important in such an endeavor. These dimensions are largely independent of application domain or clustering algorithm.

Exploratory — confirmatory. We expect that exploratory/confirmatory distinction (Tukey, 1977) mentioned earlier will be central. While exploratory data analysis is used to discover patterns in data and to formulate concrete hypotheses about the data, confirmatory data analysis deals with the question of how to validate a given hypothesis based on empirical data. Clustering is employed in both contexts. Examples for exploratory uses of clustering are: Detecting latent structure, defining categories in data for later use (e.g. different subcategories of a disease), verifying that there is structure in the data, verifying that no unexpected clusters show up (quality control), verifying that expected clusters are there. For confirmative clustering consider the following example from medicine. Assume that one hypothesizes a particular categorization among patients (e.g., according to a syndrome). Then data are collected from a different feature space (say, gene expressions). The confirmation comes from looking at the clusters in the new space and comparing them to the hypothesized categorization.

Qualitative — quantitative. A fundamental distinction is whether the clustering results are used in a quantitative or qualitative context, that is whether we can compute a score to evaluate the overall performance of our system. The exploratory context is often "qualitative". We are interested in certain properties of the clusterings, but not in any final scores. Examples for quantitative uses of clustering are data preprocessing, data compression, or clustering for semi-supervised learning. Here, a final score can be used to evaluate the clustering performance. Note that quantitative clustering is not necessarily

the same as confirmatory data analysis. In this sense, the distinction "exploratory — confirmative" is not necessarily the same as "quantitative — qualitative".

Unsupervised — supervised. There are different degrees of supervision in clustering problems. Situations where one does not have a clue what one is looking for are rare, and in many cases additional information can be used. For example, in exploratory data analysis the analyst usually has a good idea what he is looking for and he might bias the clustering with a particular choice of features or a particular choice of similarity measure between patterns. Further, the assessment of the usefulness of clustering as a preprocessing is often an iterative process: if clustering is found useful in a supervised task, it may be more likely to be useful in similar tasks. For instance, clustering has become a mainstream technique to build codebooks of phonemes or subwords in speech recognition. In some transfer learning methods, a set of clusters selected as a good preprocessing for one supervised task might used for another similar supervised task. In some transduction learning approaches, clustering might be carried out on both labeled and unlabeled data and the classification of labeled data performed according to the majority of labels found in a given cluster.

Bias towards particular solutions. In most uses of clustering, one has a "bias" concerning what one is looking for. This bias affects the type of clusters one tries to construct. In some applications the focus might be to join similar data points; for example, when detecting a chemical compound with similar properties to a given one. In other applications, it might be more important to separate different points, for example to identify emerging topics in a stream of news. Other examples are compact clusters vs. chain-like clusters, peak-based vs. gap-based clusters, flat clustering vs. hierarchical clustering.

Modeling the data-generating process. Do we need to find clusters that represent some understanding of how the data were generated? Not necessarily. For instance, in speech recognition it is common to use vector quantization as preprocessing, a method making no assumption about how data were generated. Similarly, in image processing, connected component methods do not attempt to uncover the mechanism by which data were generated. However, in some applications, clustering can be understood as a method for uncovering latent data structure, and prior knowledge may guide users to select the most suitable approach. Below are three examples illustrating how such prior knowledge could be exploited.

1. Phylogenies are usually modeled as a diffusion process. Clustering approaches to such problems usually attempt to uncover the underlying hierarchy, hence are tackled using hierarchical clustering methods.

2. If instead the data were generated by a shallow process, there is no reason to use a hierarchical model. For instance, handwriting used to be taught with a few methods in the US, hence, handwriting styles could be clustered by assuming just a two level random process: drawing the method, then drawing the writer; Gaussian mixtures or k-means type of algorithms are appropriate for such problems.

3. A third type of model is constraint-based and assumes interactions between samples (like in a Markov random field of an Ising model of magnetism). For instance, dress-code can be assumed to emerge from peer-pressure and result in clusters of people

dressing in a similar way. Graph partitioning methods may lend themselves to such cases.

Evidently, in each of these three cases using clustering methods that do not attempt to model the data generative process might "work well", if all we are interested in is representing groups efficiently for compression or prediction. But using a data generative model may make a lot of difference from the point of view of gaining insight. For instance, we may want to *predict the consequences of actions* or *devise policies to attain a desired goal*. Consider an example from epidemiology. Patients with the same symptoms may be clustered assuming one of the three models described above. The first model (hierarchical model) can be appropriate if a genetic mutation is responsible for a disease. Then, based on the hierarchical model one can try to trace a population to a common ancestor and use this knowledge to diagnose and treat patients. If patients may have a disease because of one categorical factor of variability (environment, diet, etc.) like in the case of scurvy and lack of vitamin C, then a mixture model might be more appropriate. If the disease is transmitted by contagion, then modeling the data with an interaction model (third case above) is appropriate, and a corresponding disease prevention policy can be devised.

We stress that a catalog of clustering problems is likely to be quite complex, which is no different to, say, a catalog of structural engineering problems. Many real problems require multiple factors dealt with at once (a wall keeps out the rain, wind and noise, but also holds up the roof and provides somewhere to hang a painting). Nevertheless these aspects can be described declaratively (in terms of water-resistance, wind load, sound attenuation, static and dynamic load bearing and visual aesthetics).

5. Conclusions

We asked whether clustering was art or science, but concluded that it is meaningless to view clustering as a domain-independent method. We deliberately duck our original question by claiming it unhelpful and irrelevant. If one wants an abstract label, a better one is engineering, which embraces different ways of knowing ("art" and "science"), recognizes the intrinsic psychological component of many problems, has standardized language and problem descriptions to avoid undue technique-focus and most certainly is focussed on solving the end-user problem (Vincenti, 1990).

If clustering researchers want real impact in applications, then it is time to step back from a purely mathematical and algorithmic point of view. What is missing is not "better" clustering algorithms but a problem-centric perspective in order to devise meaningful evaluation procedures.

Acknowledgements

We are most grateful for inspiring discussions and comments on earlier drafts by the following people (without implying that they agree with our opinions or conclusions): Shai Ben-David, Kristin Bennett, Léon Bottou, Joachim Buhmann, Lawrence Cayton, Christian Hennig, Stefanie Jegelka, Vincent Lemaire, Mark Reid, Volker Roth, Naftali Tishby, and one anonymous reviewer from *Machine Learning Journal*.

RW is supported by the Australian Research Council and NICTA through Backing Australia's Ability.

References

S. Anderson. Some Suggested Concepts for Improving Taxonomic Dialogue. *Systematic Zoology*, 23(1):58–70, 1974.

B. Andreopoulos, A. An, X. Wang, and M. Schroeder. A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics*, 10:297–314, 2009.

P. Berkhin. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer, Berlin, 2006.

A. Bird and E. Tobin. Natural kinds. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Summer 2010 Edition)*. Stanford University, 2010.

H.H. Bock. A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In M. Schader, editor, *Analyzing and Modeling Data and Knowledge*, pages 19–36. Springer Verlag, 1992.

J. L. Borges. El idioma analytico de John Wilkins. In *La Nación*. Penguin, London, 8 February 1999. Translated and Republished as "John Wilkins' Analytical Language," pages 229–232 in *The Total Library: Non-fiction, 1922–1986*.

G. C. Bowker and S. Star. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, Massachusetts, 1999.

J.M. Buhmann. Information theoretic model validation for clustering. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1398–1402. IEEE, 2010.

L. Candillier, I. Tellier, F. Torre, and O. Bousquet. Cascade evaluation of clustering algorithms. In *Machine Learning: ECML 2006*, pages 574–581. Springer, 2006.

M.H. DeGroot. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):7085 – 7090, 1996.

J. S. Farris. Classification Among the Mathematicians (Review of "Classification and Clustering," by J. Van Ryzin). *Systematic Zoology*, 30(2):208–214, 1981.

M. Foucault. *The Order of Things: An Archaeology of the Human Sciences*. Random House, 1970.

J.S.L. Gilmour and S.M. Walters. Philosophy and classification. In W.B. Turrill, editor, *Vistas in Botany, Volume IV: Recent Researches in Plant Taxonomy*, pages 1–22. Pergamon Press, Oxford, 1964.

I.J. Good. The botryology of botryology. In J. van Ryzin, editor, *Classification and Clustering: Proceedings of an Advanced Seminar conducted by the Mathematics Research Center, The University of Wisconsin-Madison*, pages 73–94. Academic Press, 1977.

I.J. Good. The philosophy of exploratory data analysis. *Philosophy of Science*, 50(2), 1983.

D. J. Hand. Comment on "The skill-plot: A graphical technique for evaluating continuous diagnostic tests". *Biometrics*, 63:259, 2008.

J. A. Hartigan. Distribution problems in clustering. In J. Van Ryzin, editor, *Classification and Clustering: Proceedings of an Advanced Seminar conducted by the Mathematics Research Center, The University of Wisconsin-Madison*. Academic Press, 1977.

D. L. Hull. *Science as a Process*. University of Chicago Press, 1988.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264 – 323, 1999.

A.K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8): 651–666, 2010.

A.K. Jain, A. Topchy, M.H.C. Law, and J.M. Buhmann. Landscape of clustering algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR04)*, volume 1, pages 260–263, 2004.

L.A.S. Johnson. Rainbow's End: The Quest for an Optimal Taxonomy. *Proceedings of the Linnean Society of New South Wales*, 93(1):1–45, 1968. Reprinted in *Systematic Zoology*, 19(3), 203–239 (September 1970).

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.

D. E. Knuth. Computer Programming as an Art. *Communications of the ACM*, 17(12): 667–673, December 1974.

J. Kogan. *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press, 2007.

B.H. Kwasnik. The role of classification in knowledge representation and discovery. *Library Trends*, 48(1):22–47, 1999.

G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. The University of Chicago Press, 1987.

C. Mercier. *A New Logic*. William Heineman, London, 1912. URL http://www.archive.org/details/newlogic00merciala.

W. W. Moss. Taxonomic repeatability: An experimental approach. *Systematic Zoology*, 20 (3):309–330, 1971.

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

M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731 – 817, 2011.

N. Reid. The roles of conditioning in inference. *Statisitical Science*, 10(2):138–157, May 1995.

A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Münsterkötter, and H. Mewes. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32(18):5539, 2004.

Y. Seldin and N. Tishby. PAC-Bayesian Analysis of Co-clustering and Beyond. *Journal of Machine Learning Research*, 11:3595–3646, 2010.

C. E. Shannon. The bandwagon. *IRE Transactions on Information Theory*, 2(3):3, 1956.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

J. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1), 1977.

J. W. Tukey. Unsolved problems of experimental statistics. *Journal of the American Statistical Association*, 49(268):706–731, December 1954.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

W. G. Vincenti. *What Engineers Know and How They Know It: Analytical Studies from Aeronautical History*. The Johns Hopkins University Press, Baltimore, 1990.

U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555 – 586, 2008.

R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, May 2005.