A PAC-Bayesian Approach to Formulation of Clustering Objectives

Yevgeny Seldin[*†]
[*]Max Planck Institute
for Biological Cybernetics
Tübingen, Germany
seldin@tuebingen.mpg.de

Naftali Tishby[†‡]
[†]School of Computer Science and Engineering
[‡]Interdisciplinary Center for Neural Computation
The Hebrew University of Jerusalem, Israel
tishby@cs.huji.ac.il

Abstract

Clustering is a widely used tool for exploratory data analysis. However, the theoretical understanding of clustering is very limited. We still do not have a well-founded answer to the seemingly simple question of "how many clusters are present in the data?", and furthermore a formal comparison of clusterings based on different optimization objectives is far beyond our abilities. The lack of good theoretical support gives rise to multiple heuristics that confuse the practitioners and stall development of the field.

We suggest that the ill-posed nature of clustering problems is caused by the fact that clustering is often taken out of its subsequent application context. We argue that one does not cluster the data just for the sake of clustering it, but rather to facilitate the solution of some higher level task. By evaluation of the clustering's contribution to the solution of the higher level task it is possible to compare different clusterings, even those obtained by different optimization objectives. In the preceding work it was shown that such an approach can be applied to evaluation and design of co-clustering solutions. Here we suggest that this approach can be extended to other settings, where clustering is applied.

1 Introduction

From early childhood we learn to group and partition (cluster) objects based on different parameters: cubes by their color or shape, animals as domestic and wild, or as mammals and birds, and so on. These types of questions are prominent throughout our lives and we have developed fairly good skills and intuition as to how to approach them. But once we want to address them automatically we have to transform our intuition into rigorous mathematical formulations. At the very least we should be able to compare different solutions to the clustering problem and determine which is better. But how can we compare clustering of cubes by color with clustering of cubes by shape? At a first glance it appears to be a comparison of apples and oranges. However, we argue that it is possible to do such comparison and that the answer depends on the goal and subsequent usage of the clustering. Namely, we should ask: what is the purpose of clustering the cubes? If we would like to pack the cubes into a toy-car, it is probably better to cluster the cubes by shape, as this task is indifferent to color. But if we plan to make some other use of the cubes, wherein color is important, clustering by color is probably a better choice.

2 Related Work

The idea of considering clustering in the context of a higher level task was inspired by the Information Bottleneck (IB) principle [19, 17, 18]. The IB principle considers the problem of extracting information from a random variable X that is relevant for the prediction of a random variable Y. The relevance variable Y defines the high level task. The extraction of relevant information from X is done by means of clustering of X into clusters $\tilde{X}$ [19]. In other words, IB is looking for the structure $\tilde{X}$ of X that is relevant for the prediction of Y. The IB principle was further extended to graphical models in [18]. In the example with the cubes we can define $X_1, .., X_d$ to be various parameters of the cubes, where d is the number of the parameters considered for each cube. For example, X_1 can be the shape of the cube, X_2 cube's color, X_3 weight, etc. The relevance variable Y may be an indicator variable as to whether the cube fits into a certain volume or a real-valued variable, for example the coefficient of light energy absorption. Clearly, each relevance variable corresponds to a different partition (clustering) of the parameter space. It is important to note that the requirement to compress X is an important part of the high level task in IB. In [13, 12] it has been proved that if the high level task is solely the prediction of Y and $d = 1$, direct smoothing of the empirical conditional distributions yields better performance than clustering-based solutions.

The idea to consider clustering as a proxy to a solution of a prediction task was further developed in [8, 7]. In these works Krupka and Tishby analyze a scenario wherein each object has multiple properties, but only a fraction of the properties is observed. Consider the following illustration: assume we are presented with multiple fruits and we observe their parameters, such as size, color, and weight. We can cluster the fruits by their observed parameters in order to facilitate prediction of unobserved parameters, such as taste and toxicity. This approach enables to conduct a formal analysis and derive generalization bounds for prediction rules based on clustering.

In recent years extensive attempts have been made to address the question of model order selection in clustering in terms of its stability [9, 21, 15, 3]. This perspective suggests that for two random samples generated by the same source clustering of the samples should be similar (and hence stable). Otherwise the obtained clustering is unreliable. Although it has been proved that in a large sample regime stability can be used for model order selection [15], no upper bounds on the minimal sample size required for stability estimates to hold can be proved. Furthermore, in certain cases stability indices based on arbitrarily large samples can be misleading [3]. Since in any practical application the amount of data available is limited, currently existing stability indices cannot be used for reliable model order selection. Moreover it is not clear whether the stability indices can be used to compare solutions based on different optimization objectives.

Gaussian ring example. We use the following example from [12] to illustrate that generalization and stability criteria for evaluation of clustering are not equivalent. Assume points in $\mathbb{R}^2$ are generated according to the following process. First, we select a center μ of a Gaussian according to a uniform distribution on a unit circle in $\mathbb{R}^2$. Then we generate a point $x \sim \mathcal{N}(\mu, \sigma^2 I)$ according to a Gaussian distribution centered at μ with a covariance matrix $\sigma^2 I$ for a fixed σ (I is a 2 by 2 identity matrix). Given a sample generated according to the above process we can apply a mixture of Gaussians clustering in order to learn the generating distribution. Note that:

1. Due to the symmetry in the generating process, the solution will always be unstable (the centers of Gaussians in the mixture of Gaussians model can move arbitrarily along the unit circle).

2. By increasing the sample size and the number of Gaussians in the mixture of Gaussians model we can approximate the true data generating process arbitrarily well.

Hence, models with good generalization properties are not necessarily stable. This point should be kept in mind when using generalization as an evaluation criterion in clustering.

3 PAC-Bayesian Analysis of Co-clustering

In [13, 14, 12] it was shown that PAC-Bayesian bounds provide a handful tool for analysis of generalization in co-clustering. We will briefly review these results and in the next section suggest that they can likely be extended to other settings, wherein clustering is applied. Co-clustering is a widely used method for analysis of matrix data by simultaneous clustering of rows and columns of the matrix [2]. It has successfully been applied in multiple domains, including clustering of documents and words in text mining [6], genes and experimental conditions in bioinformatics [4], viewers and movies in recommender systems [12], etc. In [14] it is pointed out that there are actually two different types of tasks that are solved with co-clustering. The first one is discriminative prediction, which corresponds to tasks like collaborative filtering, where we want to predict the rating given a

$\langle$viewer, movie$\rangle$ pair. This task can be considered as a supervised classification task solved via clustering, because the rating can be considered as a label of the $\langle$viewer, movie$\rangle$ pair. The second task is estimation of a joint probability distribution of two variables and it corresponds to tasks such as the analysis of word-document co-occurrence matrices. This task is a purely unsupervised learning task.

PAC-Bayesian bounds [11, 10, 1] are a handful framework for generalization analysis in heterogeneous hypothesis spaces. For a hypothesis class $\mathcal{H}$ the heterogeneity is introduced via a prior $\mathcal{P}(h)$ over $h \in \mathcal{H}$, which has to be selected before we observe the sample. The prior $\mathcal{P}(h)$ determines which hypotheses are considered simple and which are more complex. In many cases there is a natural heterogeneity present in the hypothesis class which can be exploited to construct the prior. For example, the class of decision trees can be divided into subclasses corresponding to tree depth and a preference to shallow trees can be given this way. The PAC-Bayesian bounds consider randomized predictors. For a posterior distribution $\mathcal{Q}$ over $\mathcal{H}$ that can be selected after observing the sample, the randomized predictor associated with $\mathcal{Q}$ acts by selecting a hypothesis $h \in \mathcal{H}$ according to $\mathcal{Q}$ and then applying h to make the prediction. Let $L(h)$ to be the expected loss of a hypothesis h, let $L(\mathcal{Q}) = \mathbb{E}_{\mathcal{Q}(h)} L(h)$ be the expected loss of the randomized prediction strategy $\mathcal{Q}$, let $\hat{L}(\mathcal{Q})$ be its empirical counterpart, and assume that the loss is bounded in the $[0, 1]$ interval. The PAC-Bayesian bounds state that with a probability greater than $1 - \delta$ over the sample selection for all randomized prediction strategies $\mathcal{Q}$ simultaneously:

$$D_b(\hat{L}(\mathcal{Q}) \| L(\mathcal{Q})) \leq \frac{D(\mathcal{Q}\|\mathcal{P}) + \frac{1}{2}\ln(4N) - \ln\delta}{N}, \tag{1}$$

where $D(\mathcal{Q}\|\mathcal{P})$ is the KL-divergence [5] between $\mathcal{Q}$ and $\mathcal{P}$ and $D_b(\hat{L}(\mathcal{Q})\|L(\mathcal{Q}))$ is the KL-divergence between two Bernoulli variables with biases $\hat{L}(\mathcal{Q})$ and $L(\mathcal{Q})$. In [14, 12] this result is further extended to a result of a similar form that applies to discrete density estimation.

In [13, 14, 12] PAC-Bayesian bounds were applied to derive generalization bounds for discriminative prediction and density estimation based on co-clustering. The bounds suggest that the expected performance of such models is optimized through the optimization of a trade-off between empirical performance and the mutual information [5] that the cluster variables preserve on the rows and columns of the data matrix. In addition to its theoretical importance, this result also has a practical value. By optimization of the trade-off in [12], state-of-the-art performance was achieved in prediction of missing ratings in the MovieLens collaborative filtering dataset.

4 Discussion and Future Directions

The main message of this abstract is to suggest the external instead of internal evaluation of clustering. Namely, instead of evaluation of different properties of clustering itself we suggest to evaluate the contribution of clustering to the solution of a more general task. The approach to external evaluation of clustering that was illustrated on the co-clustering task in [13, 14, 12] can be extended to other clustering tasks. For example, we can evaluate clustering of points in $\mathbb{R}^d$ with a Gaussian mixture model by its ability to predict new points generated by the same distribution that generated the training set. It should be possible to apply the PAC-Bayesian framework to analyze this formulation of clustering objective. In particular, if the posterior distribution $\mathcal{Q}$ is a mixture distribution $\mathcal{Q}(h) = \sum_i \lambda_i \mathcal{Q}_i(h)$ for $\sum_i \lambda_i = 1$, then by convexity of the KL-divergence $D(\mathcal{Q}\|\mathcal{P}) \leq \sum_i \lambda_i D(\mathcal{Q}_i\|\mathcal{P})$. Substitution of this result into (1) supports the intuition behind the "Gaussian ring" example presented earlier: the generalization bound does not depend on the number of Gaussians involved in the solution, but only on their weighted KL-divergence from a prior distribution. To complete the analysis it is required to extend the PAC-Bayesian bounds to continuous loss functions, which is a subject for future work [12].

We further note that clustering is a particular case of structure learning. Hence, we can extend the approach of external evaluation of clustering to evaluation of structure learning (and in particular graphical models) by its contribution to a solution of a more general task. The suggested framework does not limit the high level task to a prediction task. It is possible to analyze the advantages of structure-based models in other contexts. For example:

1. Cases where we are not limited to a single prediction question, but rather a range of questions coming from a predefined family [16].

2. Tasks of control. For example, it may be easier to control or influence a process (e.g., our own hand or some tool) if we have a simple representation of its structure.

Structure-based models can also be preferable when computation or memory constraints are imposed. Moreover, it is not a-priori clear in questions of prediction whether structure-based approaches can outperform unstructured approaches, such as SVMs. Vapnik's well-known postulate states that "one should not solve a harder problem on the way to solving a simpler problem" [20]. Structure learning is an extra effort that is not justified in the context of a pure prediction task. However, we know that as humans we perceive the world around us in a structured manner. Thus there must be advantages that we gain from the knowledge of structure. The most distinctive advantage of structure-based models is the *understanding* and *simplification* of the underlying processes and phenomenons. But in order to build computational models that are able to understand the data it is essential to quantify the notion of understanding or at least to be able to compare the level of understanding gained by different approaches (similar to the way we can measure which of two students understood a course better). We conjecture that quantification of understanding should be done in the context of its potential applications. For example, one student can understand the course better to pass a written exam, but his schoolmate will outperform him in an oral exam because it examines another type of understanding. Thus, the analysis of structure learning in the context of prediction tasks is just a small step towards the quantification of knowledge and development of algorithms that are able to understand the data.

References

[1] A. Banerjee. On Bayesian bounds. In *ICML*, 2006.

[2] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *JMLR*, 8, 2007.

[3] S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In *COLT*, 2008.

[4] H. Cho and I. S. Dhillon. Co-clustering of human cancer microarrays using minimum sum-squared residue co-clustering. *IEEE/ACM Trans. on Computational Biology and Bioinformatics (TCBB)*, 5:3, 2008.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[6] I. Dhillon, S. Mallela, and D. Modha. Information-theoretic co-clustering. In *ACM SIGKDD*, 2003.

[7] E. Krupka. *Generalization from Observed to Unobserved Features*. PhD thesis, The Hebrew University of Jerusalem, 2008.

[8] E. Krupka and N. Tishby. Generalization from observed to unobserved features by clustering. *JMLR*, 9, 2008.

[9] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability based validation of clustering solutions. *Neural Comp.*, 2004.

[10] A. Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.

[11] D. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1), April 2003.

[12] Y. Seldin. *A PAC-Bayesian Approach to Structure Learning*. PhD thesis, The Hebrew University of Jerusalem, 2009.

[13] Y. Seldin and N. Tishby. Multi-classification by categorical features via clustering. In *ICML*, 2008.

[14] Y. Seldin and N. Tishby. PAC-Bayesian generalization bound for density estimation with application to co-clustering. In *AISTATS*, 2009.

[15] O. Shamir and N. Tishby. On the reliability of clustering stability in the large sample regime. In *NIPS*, 2009.

[16] J. Shawe-Taylor and A. Dolia. A framework for probability density estimation. In *AISTATS*, 2007.

[17] N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2002.

[18] N. Slonim, N. Friedman, and N. Tishby. Multivariate information bottleneck. *Neural Comp.*, 18, 2006.

[19] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control and Computation*. 1999.

[20] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[21] U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.