Clustering Stability: An Overview

Clustering Stability: An Overview

Ulrike von Luxburg

Max Planck Institute for
Biological Cybernetics
Tübingen
Germany
ulrike.luxburg@tuebingen.mpg.de

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is U. von Luxburg, Clustering Stability: An Overview, Foundation and Trends® in Machine Learning, vol 2, no 3, pp 235–274, 2009

ISBN: 978-1-60198-344-2

Foundations and Trends® in Machine Learning

Volume 2 Issue 3, 2009

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2009, Volume 2, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 2, No. 3 (2009) 235–274
© 2010 U. von Luxburg
DOI: 10.1561/2200000008

Clustering Stability: An Overview

Ulrike von Luxburg

Max Planck Institute for Biological Cybernetics, Tübingen, Germany, ulrike.luxburg@tuebingen.mpg.de

Abstract

A popular method for selecting the number of clusters is based on stability arguments: one chooses the number of clusters such that the corresponding clustering results are "most stable". In recent years, a series of papers has analyzed the behavior of this method from a theoretical point of view. However, the results are very technical and difficult to interpret for non-experts. In this monograph we give a high-level overview about the existing literature on clustering stability. In addition to presenting the results in a slightly informal but accessible way, we relate them to each other and discuss their different implications.

Contents

1

Introduction

Model selection is a difficult problem in non-parametric clustering. The obvious reason is that, as opposed to supervised classification, there is no ground truth against which we could "test" our clustering results. One of the most pressing questions in practice is how to determine the number of clusters. Various ad hoc methods have been suggested in the literature, but none of them is entirely convincing. These methods usually suffer from the fact that they implicitly have to define "what a clustering is" before they can assign different scores to different numbers of clusters. In recent years a new method has become increasingly popular: selecting the number of clusters based on clustering stability. Instead of defining "what is a clustering", the basic philosophy is simply that a clustering should be a structure on the data set that is "stable". That is, if applied to several data sets from the same underlying model or of the same data-generating process, a clustering algorithm should obtain similar results. In this philosophy it is not so important how the clusters look (this is taken care of by the clustering algorithm), but that they can be constructed in a stable manner.

The basic intuition of why people believe that this is a good principle can be described by Figure 1.1. Shown is a data distribution with four

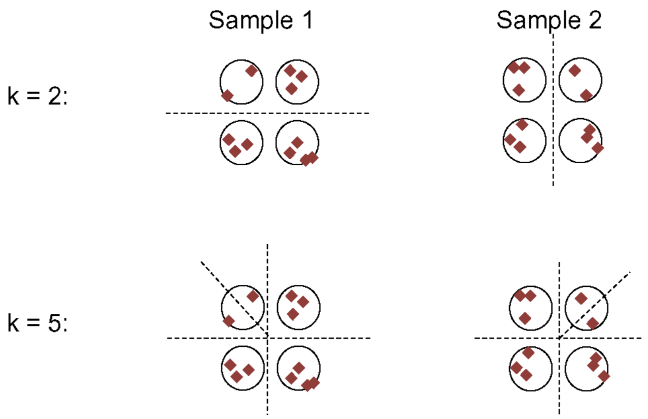

Fig. 1.1 Idea of clustering stability. Instable clustering solutions if the number of clusters is too small (first row) or too large (second row). See text for details.

underlying clusters (depicted by the black circles), and different samples from this distribution (depicted by red diamonds). If we cluster this data set into $K = 2$ clusters, there are two reasonable solutions: a horizontal and a vertical split. If a clustering algorithm is applied repeatedly to different samples from this distribution, it might sometimes construct the horizontal and sometimes the vertical solution. Obviously, these two solutions are very different from each other, hence the clustering results are instable. Similar effects take place if we start with $K = 5$. In this case, we necessarily have to split an existing cluster into two clusters, and depending on the sample this could happen to any of the four clusters. Again the clustering solution is instable. Finally, if we apply the algorithm with the correct number $K = 4$, we observe stable results (not shown in the figure): the clustering algorithm always discovers the correct clusters (maybe up to a few outlier points). In this example, the stability principle detects the correct number of clusters.

At first glance, using stability-based principles for model selection appears to be very attractive. It is elegant as it avoids to define what a good clustering is. It is a meta-principle that can be applied to any basic clustering algorithm and does not require a particular clustering model. Finally, it sounds "very fundamental" from a philosophy of inference point of view.

However, the longer one thinks about this principle, the less obvious it becomes that model selection based on clustering stability "always works". What is clear is that solutions that are completely instable should not be considered at all. However, if there are several stable solutions, is it always the best choice to select the one corresponding to the most stable results? One could conjecture that the most stable parameter always corresponds to the simplest solution, but clearly there exist situations where the most simple solution is not what we are looking for. To find out how model selection based on clustering stability works we need theoretical results.

In this monograph we discuss a series of theoretical results on clustering stability that have been obtained in recent years. In Section 2 we review different protocols for how clustering stability is computed and used for model selection. In Section 3 we concentrate on theoretical results for the K-means algorithm and discuss their various relations. This is the main section of the paper. Results for more general clustering algorithms are presented in Section 4.

References

[1] S. Ben-David, “A framework for statistical clustering with constant time approximation algorithms for K-median and K-means clustering,” *Machine Learning*, vol. 66, pp. 243–257, 2007.

[2] S. Ben-David, D. Pál, and H.-U. Simon, “Stability of k-Means Clustering,” in *Conference on Learning Theory (COLT)*, (N. Bshouty and C. Gentile, eds.), pp. 20–34, Springer, 2007.

[3] S. Ben-David and U. von Luxburg, “Relating clustering stability to properties of cluster boundaries,” in *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, (R. Servedio and T. Zhang, eds.), pp. 379–390, Springer, Berlin, 2008.

[4] S. Ben-David, U. von Luxburg, and D. Pál, “A sober look on clustering stability,” in *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, (G. Lugosi and H. Simon, eds.), pp. 5–19, Springer, Berlin, 2006.

[5] A. Ben-Hur, A. Elisseeff, and I. Guyon, “A stability based method for discovering structure in clustered data,” in *Pacific Symposium on Biocomputing*, pp. 6–17, 2002.

[6] A. Bertoni and G. Valentini, “Model order selection for bio-molecular data clustering,” *BMC Bioinformatics*, vol. 8(Suppl 2):S7, 2007.

[7] A. Bertoni and G. Valentini, “Discovering multi-level structures in biomolecular data through the Bernstein inequality,” *BMC Bioinformatics*, vol. 9(Suppl 2), 2008.

[8] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts,

V. Sondak, M. Hayward, and J. Trent, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, pp. 536–540, 2000.

[9] S. Bubeck, M. Meila, and U. von Luxburg, "How the initialization affects the stability of the k-means algorithm," Draft, http://arxiv.org/abs/0907.5494, 2009.

[10] S. Dasgupta and L. Schulman, "A probabilistic analysis of EM for mixtures of separated, spherical gaussians," *JMLR*, vol. 8, pp. 203–226, 2007.

[11] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[12] J. Fridlyand and S. Dudoit, "Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method," Technical Report 600, Department of Statistics, University of California, Berkeley, 2001.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[14] M. K. Kerr and G. A. Churchill, "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *PNAS*, vol. 98, no. 16, pp. 8961–8965, 2001.

[15] T. Lange, V. Roth, M. Braun, and J. Buhmann, "Stability-based validation of clustering solutions," *Neural Computation*, vol. 16, no. 6, pp. 1299–1323, 2004.

[16] J. Lember, "On minimizing sequences for k-centres," *Journal of Approximation Theory*, vol. 120, pp. 20–35, 2003.

[17] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural Computation*, vol. 13, no. 11, pp. 2573–2593, 2001.

[18] M. Meila, "Comparing clusterings by the variation of information," in *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*, (B. Schölkopf and M. Warmuth, eds.), pp. 173–187, Springer, 2003.

[19] U. Möller and D. Radke, "A cluster validity approach based on nearest-neighbor resampling," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, pp. 892–895, Washington, DC, USA: IEEE Computer Society, 2006.

[20] D. Pollard, "Strong consistency of k-means clustering," *Annals of Statistics*, vol. 9, no. 1, pp. 135–140, 1981.

[21] D. Pollard, "A central limit theorem for k-means clustering," *Annals of Probability*, vol. 10, no. 4, pp. 919–926, 1982.

[22] O. Shamir and N. Tishby, "Cluster stability for finite samples," in *Advances in Neural Information Processing Systems (NIPS) 21*, (J. Platt, D. Koller, Y. Singer, and S. Rowseis, eds.), Cambridge, MA: MIT Press, 2008.

[23] O. Shamir and N. Tishby, "Model Selection and Stability in k-means clustering," in *Proceedings of the 21rst Annual Conference on Learning Theory (COLT)*, (R. Servedio and T. Zhang, eds.), 2008.

[24] O. Shamir and N. Tishby, "On the reliability of clustering stability in the large sample regime," in *Advances in Neural Information Processing Systems 21 (NIPS)*, (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), 2009.

[25] M. Smolkin and D. Ghosh, “Cluster stability scores for microarray data in cancer studies,” *BMC Bioinformatics*, vol. 36, no. 4, 2003.

[26] A. Strehl and J. Ghosh, “Cluster ensembles — A knowledge reuse framework for combining multiple partitions,” *JMLR*, vol. 3, pp. 583–617, 2002.

[27] N. Vinh and J. Epps, “A novel approach for automatic number of clusters detection in microarray data based on consensus clustering,” in *Proceedings of the Ninth IEEE International Conference on Bioinformatics and Bioengineering*, pp. 84–91, IEEE Computer Society, 2009.