Combinatorial Approaches to Clustering and Feature Selection

Michael E. Houle

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
http://www.nii.ac.jp/en/
meh@nii.ac.jp

Extended Abstract

1 Introduction

One of the most serious difficulties in the analysis of high-dimensional data sets involves the treatment of measures of similarity. Although similarity measures often retain some discriminative ability as the dimension increases, the similarity values themselves are often difficult to interpret. Methods for search, clustering and feature selection that perform quantitive tests of similarity values (as opposed to comparative tests) are particularly susceptible to this problem. This presentation will be concerned with combinatorial models of clustering based on shared neighbor information, and their application to feature selection, subspace clustering, and multiple clustering. The models assume a secondary, derived form of similarity measure based on the intersection properties of neighborhoods defined according to the original similarity measure. The use of secondary similarity has been recently shown to offer solutions that are more robust and more scalable with respect to the dimension of the data.

2 Secondary Similarity Measures

For similarity search and their applications, the distance measures commonly used in practice are known to be sensitive to local variations within the data distribution, as well as the number of data features involved (the dimension). These dependencies can severely limit the the efficiency and accuracy of the search, and ultimately the quality of the solution — a phenomenon often referred to as the *curse of dimensionality.* Generally speaking, as the number of data features increases, pairwise distance values tend to concentrate tightly about their mean, reducing the overall discriminative ability of the similarity measure. The effect occurs for a broad range of data distributions and similarity measures, and can be so pronounced as to cast doubt upon whether efficient nearest neighbor search is even achievable in higher dimensions [1]. However, when a data set is composed of many well-formed clusters, the concentration effect will typically be less severe across cluster boundaries, with distances from a cluster member to other cluster members being relatively easy to distinguish from distances to non-members, especially when the clusters are well separated [1–3].

In general, any improvement in the discriminative ability of the similarity measure employed can be expected to yield improvements in the performances of solutions based on it. Some simple enhancement strategies involve the use of *shared neighbor* (SN) information, in which a *secondary* similarity between two points v and w is defined in terms of data objects in the common intersection of neighborhoods based at v and w, where the neighborhoods themselves are determined according to a supplied *primary* similarity measure. The primary measure can be any function that determines a well-defined ranking of the data objects relative to the query. Recent studies have shown that secondary similarity measures based on SN information are generally more robust in higher dimensions than their associated primary distance measures, since the neighborhoods of object pairs drawn from a common cluster tend to have significantly more items in common than to pairs drawn from different clusters [4, 5]. Furthermore, recent advances in approximate similarity search allow for neighborhood information to be generated accurately and efficiently for many practical applications [6, 7].

3 Multi-Source RSC Clustering

Shared-neighbor information has been used to guide clustering algorithms for almost 40 years [8–10]. However, early methods required that the neighborhood size k be fixed in advance by the user. The use of fixed values of k can introduce a very significant bias on the sizes and other characteristics of clusters that can be produced by the methods, in that they tend to favor the discovery of groups with size of roughly the same order as k.

In order to account for the effects of varying k, the Relevant-Set Correlation (RSC) model for clustering was proposed [11]. The RSC model provides a consistent and comprehensive framework for the assessment of cluster quality, based on the statistical significance of a form of correlation between the neighborhood sets of its members. More precisely, the model tests the significance of any grouping against the assumption that the neighborhoods contain zero information (that is, against the assumption that they were generated by means of random selection). The greater the extent to which the assumption is violated, the greater the significance of the grouping.

The RSC model quantifies the quality of cluster candidates of any arbitrary size (allowing the comparison of any two candidates regardless of their size), the degree of association between pairs of cluster candidates, and the degree of association between clusters and individual data items. An efficient greedy selection strategy, GreedyRSC, has been developed based on RSC, and was shown to be very competitive in practice [11]. It requires only two user-supplied parameters, describing the minimum acceptable cluster size, and the size of the maximum acceptable overlap between two clusters. Both of these parameters can be chosen in a natural way with no knowledge of the nature of the data or its distribution. The number of clusters is not specified by the user.

This presentation will be concerned with an extension of the RSC model to account for multiple sources of neighborhood information. Each of these sources is assumed to have its own similarity measure based on its own collection of data features (which may or may not contain features also used by other sources). Like the original RSC model, the extended model relies only on the neighborhood rankings produced according to the sources, and has no knowledge of the nature of the similarity measure or features involved.

The extension of RSC will be seen to have implications for subspace clustering and feature selection, as well as multiclustering. In particular, the discussion will include the following potential applications of the extended model:

- The significance of data sets can be simultaneously assessed with respect to object membership as well as the number of sources of neighborhood information. If each source is associated with its own collection of features, the model in effect assesses the significance of the association of a particular group of objects with a collection of features.
- Under the model, the combination of sources that are most strongly associated with a putative cluster can be identified very efficiently.
- In applications for which multiple clusterings of the data have been generated, the model can be used to decide to which clustering a particular candidate cluster is best aligned. This can potentially serve as a foundation upon which multiple clustering criteria can be designed.

References

1. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is “nearest neighbor” meaningful? In: Proc. ICDT. (1999)
2. Bennett, K.P., Fayyad, U., Geiger, D.: Density-based indexing for approximate nearest-neighbor queries. In: Proc. KDD. (1999)
3. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM TKDD **3**(1) (2009) 1–58
4. Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor-distances defeat the curse of dimensionality? In: Proc. SSDBM. (2010)
5. Bernecker, T., Houle, M.E., Kriegel, H.P., Kröger, P., Renz, M., Schubert, E., Zimek, A.: Quality of similarity rankings in time series. In: Proc. SSTD. (2011)
6. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: Symp. Foundations of Computer Science. (2006) 459–468
7. Houle, M.E., Sakama, J.: Fast approximate similarity search in extremely high-dimensional data sets. In: Proc. ICDE. (2005)
8. Jarvis, R.A., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE TC **C-22**(11) (1973) 1025–1034
9. Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes,. Inform. Sys. **25** (2000) 345–366
10. Ertöz, L., Steinbach, M., Kumar, V.: Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proc. SDM. (2003)
11. Houle, M.E.: The relevant-set correlation model for data clustering. Stat. Anal. Data Min. **1**(3) (2008) 157–176