ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Hybrid Hierarchical Clustering: Cluster Assessment via Cluster Validation Indices

Mark J. Embrechts[1], Jonathan D. Linton[2], and Christopher J. Gatti[1] *

1 - Rensselaer Polytechnic Institute - Industrial and Systems Engineering
Troy, NY 12180 - USA

2 - University of Ottawa - Telfer School of Management
Ottawa, ON K1N 6N5 - Canada

Abstract. This paper introduces a hybrid hierarchical clustering method, which is a novel method for speeding up agglomerative hierarchical clustering by seeding the algorithm with clusters obtained from K-means clustering. This work describes a benchmark study comparing the performance of hybrid hierarchical clustering to that of conventional hierarchical clustering. The two clustering methods are compared for 16 benchmark data sets based on the cluster validation index signature, an aggregation of several cluster indices. In most cases, the cluster signatures indicate similar clusterings for unseeded and seeded hierarchical clustering.

1 Introduction

Data clustering refers to an automated data partitioning process where the data objects are lumped into groups of similar objects and there is a clear dissimilarity between groups. Data clustering is one of the key preliminary data analysis tools for a number of reasons, including the partitioning large data sets into smaller and more manageable sets of data, and the visualization of results which often provides valuable insight to the data. Agglomerative hierarchical clustering is one such clustering technique that results in a tree-like data structure based on the similarity between the data. Hierarchical clusterings are most often displayed in the form of a dendrogram, which shows the relationships of the data both within and between clusters. The interpretation of the results of hierarchical clustering can be domain-specific, such as in biology where data may follow an evolutionary path, and such features are captured in the dendrogram, also known as a phylogenetic tree in this case.

While this method has proven to be quite useful, the computing time for the hierarchical clustering algorithm using the average-link linkage criteria scales as N^3 with the number of data N. For each of the $N-1$ cluster agglomerations, $\frac{N(N-1)}{2}$ cluster dissimilarities are searched to determine the two most similar clusters to merge. A second drawback to this algorithm is that the distance matrix must be stored in memory. Applying hierarchical clustering to large datasets that are now widely available therefore becomes impractical.

The purpose of this work is to introduce a novel hybrid, agglomerative hierarchical clustering strategy for large data sets to speed up the clustering algorithm.

*The authors acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada in conducting this research.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

This method consists of seeding hierarchical clustering with clusters obtained from K-means clustering and performing hierarchical clustering as usual thereafter. We also introduce a cluster quality assessment method using multiple cluster validation indices, as the cluster seeding has a direct impact on the final hierarchical clustering. Additionally, we demonstrate how cluster seeding affects cluster quality using a wide variety of common benchmark data sets.

2 Hierarchical clustering

In this work, we focus on agglomerative hierarchical clustering, and more specifically, the Sequential Agglomerative Hierarchical Non-overlapping clustering algorithm (SAHN) [12]. This algorithm proceeds in an iterative fashion in which the N data points are considered the initial clusters. At each iteration $i = 1, ..., N-1$, two clusters are combined to form a new cluster, and thus this algorithm builds clusters from the bottom up resulting in a final, single cluster with N objects. The selection of clusters to be combine is based on a pair-wise group method in which, at each iteration, the two closest clusters, as defined by some dissimilarity metric, are grouped.

This sequential algorithm results in a set of $[H_1, ..., H_Q]$ partitions of objects, where H_1 is the disjoint partition, H_Q is the conjoint partition, and H_j is a refinement of H_i for all $1 \leq i < j \leq Q$. Thus, the iterative nature of this algorithm generates H_{i+1} from H_i for all $1 \leq i \leq Q$. More formally, SAHN produces a nested partition of the data, $\mathbf{X}_{NM}$, with N data instances and M features, into non-overlapping subgroups H_l such that:

$$
\begin{aligned}
&\mathbf{X}, H = \{H_1, ..., H_Q\} \ \ (Q \leq N), \text{ s.t.} \\
&1)\ H_q \neq \varnothing, \ q = 1, ..., Q \\
&2)\ H_Q = \mathbf{X} \\
&3) \text{ if } C_l \in H_q \text{ and } C_m \in H_r, \ q > r \Rightarrow \\
&\quad C_l \subset C_m \text{ or } C_l \cap C_m = \varnothing \ \forall \ l, m \neq l, \text{ and } m, l = 1, ..., Q
\end{aligned}
$$

As described, the SAHN algorithm is based on a dissimilarity metric which describes the proximity between two clusters, and there are variations on this approach. Common metrics include the Euclidean distance, the Manhattan distance, and the maximum distance, as these are easy to understand and are commonly used in practice. Additionally, the SAHN algorithm is dependent on a linkage criteria; that is, a criteria that is used to define what constitutes the 'closest' clusters. Common linkage criteria include the complete (maximum), single (minimum), or average linkage criteria.

3 Hybrid hierarchical clustering via cluster seeding

A hybrid clustering scheme was first presented by Kwon and Han [6] who used hierarchical clustering to seed a traditional K-means clustering algorithm in

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

order to improve and stabilize cluster performance. In the current work, we take the reverse approach and seed the hierarchical clustering algorithm with clusters obtained from K-means clustering with the goal of improving the time-scaling of average-link hierarchical clustering.

Cluster seeds are first generated by specifying a proportion $p \in (0, 1]$ of the number of data X_{NM} (N instances and M features) to be used as seeds. The number of cluster seeds c are then taken as $\lfloor pN \rfloor$. The standard K-means algorithm is run with c clusters resulting in c clusters seeds, which are defined by cluster centers $Y_i \in \mathbb{R}^M$ for $i = 1, ..., c$. The centers of these seeds Y_i are then used as inputs to average-link hierarchical clustering (using the Euclidean distance metric in the current work), which proceeds as described in Section 2.

The drawback of such an approach of seeding hierarchical clustering is that the final clustering is likely different than the clustering obtained by implementing the standard SAHN algorithm without cluster seeding. Clustering seeding therefore approximates the actual data. As $p \to 0$, the number of cluster seeds becomes small and the seeding essentially generalizes the data set, and thus the computing time is small. As $p \to 1$, the number of seeds tends towards the number of data N and the underlying fidelity of the data is retained, although the decrease in computing time relative to unseeded SAHN becomes small.

4 Cluster quality assessment

One of the most common questions in clustering concerns the estimation of the natural number of clusters present in a data set. The use of cluster validation indices are one attempt to shed light on this issue by assessing the structure of the clusters. Each index is a single aggregate measure that attempts to summarize the clustering, for example, by comparing intra- and inter-cluster distances among all clusters. The true number of clusters is taken to be that which either minimizes or maximizes the index, depending on the index that is used.

While numerous indices have been developed, we focus on four in this work: the Davies-Bouldin (DB) cluster index[1]; the Dunn index [2]; the cluster SilHouette Width index ($SHWI$) [8]; and Hubert and Arabie's adjusted Rand Index (ARI) [4]. The first three indices are internal validation indices that are computed based only the labels assigned from the final clustering, whereas ARI is an external validation index that also takes into account the target data labels. We refer the reader to the literature for a more complete overview of these metrics [5, 13].

When using several validation indices, one is often tempted to try to determine the index that performs best over many data sets. We avoid this task and, in the spirit of Sledge et al. [10], take a holistic view of the cluster validation indices. We interpret the set of validation indices as a fingerprint that is specific to the clustering algorithm and data set, rather than use them to determine the natural number of clusters. As noted, cluster seeding results in a hierarchical clustering that is an approximation, and is therefore different from an unseeded hierarchical clustering. The impact of cluster seeding was evaluated by com-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

puting and comparing the cluster validation index fingerprint for unseeded and seeded hierarchical clustering. Furthermore, we quantify the difference in the validation index fingerprint in the following manner. For each of the four cluster validation indices $i = 1, ...4$, let Z_i^* denote the vector of cluster validation indices (i.e., a cluster validation 'curve') for unseeded hierarchical clustering cut into an arbitrary number of clusters. Similarly, let $Z_{i,p}$ denote the vector of cluster validation indices for seeded hierarchical clustering with seed proportion p. An aggregate measure of the cluster fingerprint similarity, C_p, was computed by averaging the correlation coefficient (over the four validation indices) between the validation index curves of the unseeded and seeded clustering:

$$C_p = \frac{1}{4} \sum_{i=1}^{4} Corr(Z_i^*, Z_{i,p})$$

5 Evaluation of cluster seeding

The influence of using cluster seeds with hierarchical clustering was evaluated on 16 benchmark data sets (Table 1). These benchmark data sets were chosen because they exhibit different characteristics including multi-class data, variation in clustering difficulty, overlapping classes, large numbers of data samples, large numbers of features, etc. Most of these data sets are frequently cited in the literature [3]. Additionally, some of the authors' or collaborators' data sets are introduced here, including: Santos' 2-D clustering data [9], Linton's *Journal of Business Ethics* (JBE) data [7], and Karen Smith's Microglia data [11].

The standard SAHN algorithm was compared to the SAHN algorithm seeded with clusters from K-means clustering for two different seed proportions p of $\frac{N}{2}$ and $\frac{N}{4}$ on the 16 data sets mentioned above. For these assessments, cluster seeds were selected as those that had the smallest cluster dispersion out of 100 different initializations of K-means clustering. The time savings for hierarchical clustering can be easily computed due to the cubic scaling dominance of SAHN: seeding with $p = \frac{N}{2}$ and $p = \frac{N}{4}$ should take $1/8^{th}$ and $1/64^{th}$ of the computing time, respectively, compared to clustering without seeding.

Table 1 lists C_p for $p = \frac{N}{2}$ and $p = \frac{N}{4}$ for each data set. For some data sets, C_p is high and indicates that the cluster validation fingerprint is largely unchanged when using clustering seeding (e.g., clock, anvils, and iris data sets), whereas other data sets have lower values of C_p which indicate that the fingerprint changes (e.g., spiral and microglia data sets). One would expect that the values of $C_{N/2}$ would be greater than $C_{N/4}$, although this is not always the case, thus indicating that a coarse seeding produced a cluster fingerprint that was more similar to that of the unseeded case.

Figure 1 shows two examples of cluster signatures (i.e., cluster validation indices versus number of clusters) for the microglia and clock data using SAHN without seeding and SAHN with seeding with $C_{N/4}$ (443 and 31 seeds, respectively). The clock data exhibit a very different validation index signature compared to the microglia data, indicating that cluster signature is unique to the

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Table 1: Data sets and cluster seeding correlation.

Data set name	Data size[a,b]	# classes[b]	$C_{N/2}$	$C_{N/4}$
Fischer's iris data	150 × 4	3	0.9117	0.8952
Italian olive oil	572 × 8	9	0.7447	0.6039
JMDS Portuguese rock data	134 × 18	6	0.6708	0.5474
Kohonen's animal data	16 × 16	V	0.8788	0.8788
Leukemia	38 × 7129	2	0.6336	0.5191
Linton's JBE data (1 word)	29 × 340	4	0.8500	0.5404
Linton's JBE data (2 word)	29 × 521	4	0.5338	0.4949
Many Gaussians	V × 2	V	0.8620	0.8620
Microglia	1772 × 300	3	0.2285	0.3167
Tobacco	26 × 16	2	0.8161	0.8940
Two Gaussians	V × 2	2	0.7918	0.5464
Wieland's spiral data	194 × 2	2	0.3051	0.4279
Santos 2-D clustering data				
clock	126 × 2	3	0.9415	0.9353
anvils	201 × 2	2	0.9348	0.8997
bermuda	182 × 2	4	0.3053	0.7485
beans	140 × 2	2	0.7373	0.4869

[a] # data × # features; [b] V = variable number of samples or classes.

data set. These plots also show that the cluster signature for each of the 4 validation indices is grossly similar for each data set. However, although they may appear visually similar, the shape of each cluster validation index curves can be quite different based on C_p as in Table 1, such as for the microglia data.

6 Conclusions

This paper introduces a novel hybrid clustering approach using K-means cluster seeding in order to mitigate the poor scaling of the computing time for average-link hierarchical clustering. For the 16 data sets evaluated, it was found that in most cases, cluster seeding had little impact on the final clustering based on the visual similarity and correlation of the cluster signature. The similarity between the cluster signatures with and without seeding suggest that cluster seeding may be used to significantly improve the computing time of hierarchical clustering while still providing a good approximation to the true clustering. In this work the difference between unseeded and seeded clustering was quantified using the average correlation coefficient, although other metrics could be used (e.g., mean absolute percent differences) and these should be explored in future work.

References

[1] Davies, D.L. and Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Analysis Mach. Intell. **1**, pp. 224–227 (1971)

[2] Dunn, J.: Well separated clusters and optimal fuzzy partitions. J. Cybern., **4**, pp. 95–104 (1974)

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

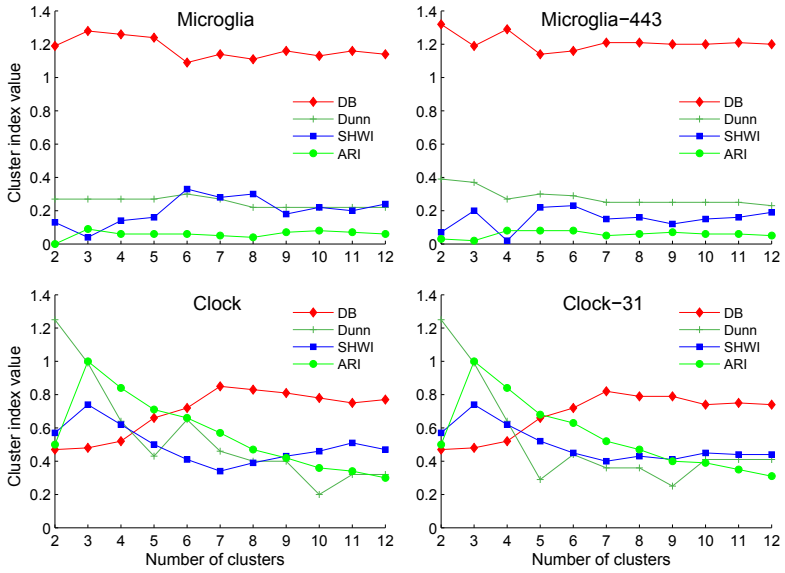

Fig. 1: Illustration of clustering signatures (i.e., cluster evaluation indices versus number of clusters) for the Santos clock data and the microglia data for hierarchical clustering with (right) and without (left) cluster seeding.

[3] Embrechts, M.J., Gatti, C.J., Linton J.D., and Roysam B.: Hierarchical Clustering for Large Data Sets. in *Advances in Signal Processing and Machine Learning: Theory and Applications* Ludmilla Mihaylova, and Petia Georgieva, Eds. (Springer, Berlin, 2012).

[4] Hubert, L. and Arabie, P.: Comparing partitions. J. Classif. **2**, pp. 193–218 (1985)

[5] Kaufman, L. and Rousseeuw, P.: *Finding Groups in Data* (Wiley Interscience, 1990)

[6] Kwon, S. and Han, C.: Hybrid clustering method for DNA microarray data analysis. Gene Inform. **13**, pp. 258–259 (2002)

[7] Linton, J. and Chen M.-N.: Working paper: Analysis of the Evolution of the Field of Business Ethics through Text Mining, University of Ottawa. (2011).

[8] Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J. Comput. Appl. Math. **20**, pp. 53–65 (1987)

[9] Santos, J.: *Data Classification with Neural Networks and Entropic Criteria.* Ph. D. Dissertation, (School of Engineering, University of Porto FEUP, 2007)

[10] Sledge, I.J., Havens, T.C., Bezdek, J.C., and Kelleher, J.M.: Relational cluster validity. In Aranda, J. and Xambó, S. (eds.), *World Congress on Computational Intelligence* (Barcelona, Spain), pp.151–185 (2010)

[11] Smith, K.: Private Communication, June21, 2011, Wadsworth Center, Albany, NY (2011)

[12] Sneath, P.H.A. and Sokal, R.R.: *Numerical Taxonomy.* (W. H. Freeman, 1973)

[13] Xu, R. and Wunsch II, D.: *Clustering. IEEE Press Series on Computational intelligence* (Wiley, 2008)