Linear Storage and Potentially Constant Time Hierarchical Clustering Using the Baire Metric and Random Spanning Paths

Fionn Murtagh and Pedro Contreras

Abstract We study how random projections can be used with large data sets in order (1) to cluster the data using a fast, binning approach which is characterized in terms of direct inducing of a hierarchy through use of the Baire metric; and (2) based on clusters found, selecting subsets of the original data for further analysis. In this work, we focus on random projection that is used for processing high dimensional data. A random projection, outputting a random permutation of the observation set, provides a random spanning path. We show how a spanning path relates to contiguity- or adjacency-constrained clustering. We study performance properties of hierarchical clustering constructed from random spanning paths, and we introduce a novel visualization of the results.

1 Introduction

In our current era of Big Data, and given the central importance of hierarchical clustering for so many application domains, there is a need to improve computationally on standard quadratic time algorithms (i.e. $O(n^2)$ for n observation vectors, see, e.g., Murtagh 1985). In Murtagh (2004), we even discuss constant time hierarchical clustering, which presupposes that our data is, naturally or otherwise, embedded in an ultrametric topological space. In this article, we take further our work in Contreras and Murtagh (2012) and Murtagh et al. (2008). In those works, we demonstrated the effectiveness of linear computational time hierarchical clustering, using a range of examples, including from astronomy and chemistry.

F. Murtagh (✉)
Department of Computing, Goldsmiths University of London, London SE14 6NW, UK

De Montfort University, Leicester, UK

Department of Computing and Mathematics, University of Derby, Derby, UK
e-mail: fmurtagh@acm.org

P. Contreras
Thinking Safe Ltd., Egham, Surrey TW20 0EX, UK
e-mail: pedro.contreras@acm.org; pedro@cs.rhul.ac.uk

A.F.X. Wilhelm, H.A. Kestler (eds.), *Analysis of Large and Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-25226-1_4

In particular we focus on very high dimensional data. We have demonstrated in many clustering case studies that random projection can work very well indeed. Random projection is a first stage of the processing, which allows both computationally efficient and demonstrably effective hierarchical clustering, using the Baire metric (Murtagh et al. 2008; Contreras and Murtagh 2012). The Baire metric is simultaneously an ultrametric. In this article, we further develop the theory and the practice of random projection in very high dimensional spaces. We are seeking a computationally efficient clustering method for massive (large n, number of rows), very high dimensional, very sparse data. Massive high dimensional data are typically sparse (i.e. containing many non-presence or 0 terms).

To help the reader to reproduce our results, some ancillary material, including R code, is available at http://www.multiresolutions.com/HiClBaireRanSpanPaths

2 Data

We present our methodology using a case study. We used the textual content of 34,352 research funding proposals, that were submitted to, and evaluated by, a research funding agency in the years 2012–2013. We refer to these proposals as proposals or documents. Because it permits search, and basic clustering, we used the Apache Solr software (Solr 2013), which is based on the Apache Lucene indexing software. Clustering in Solr is nearest neighbour-based, and is termed MLT, "more like this". Similarity scores between pairs of documents, based on textual content, are produced. Murtagh (2013) provides a short description of the MLT similarity. (Murtagh 2013, is available with this article's ancillary material.) Our documents were indexed by Solr, and MLT similarity coefficients were generated for the top 100 matching proposals. A selection of 10,317 of these proposals constituted the set that was studied. Our major aim in this work was prototyping our approach, based on the results provided by Solr. The R sparse matrix format (Matrix Market 2013) was used for subsequent R processing. The maximum MLT score (i.e. similarity coefficient value) was 3.218811. In matrix terms, we have 10,317 proposals (rows) crossed by 34,352 proposals (columns). Non-zero values accounted for 0.2854 % of the elements of this matrix.

Figure 1 serves to describe the properties of this data: a somewhat skewed Gaussian marginal distribution for the proposals, and a power law for the similar or matching MLT proposals. In Murtagh et al. (2008), we also find such Gaussian and power law behaviour for high dimensional chemical data.

In this article, we will seek to cluster the 10,317 proposals, using their similarities with the fuller set of 34,352 proposals as features. Justification for this feature space perspective, rather than directly using the MLT similarities, is that MLT similarities are asymmetric.

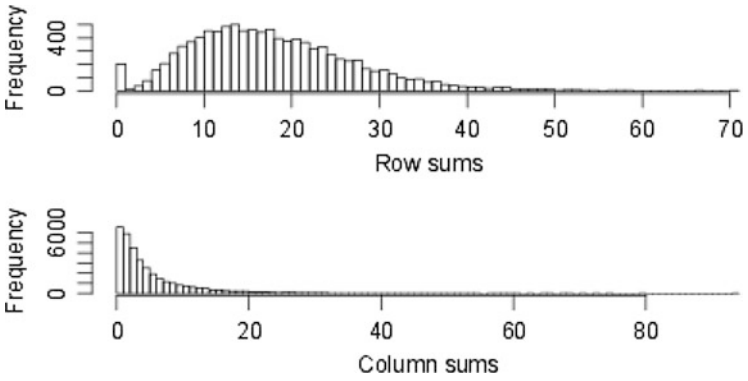

Fig. 1 Marginal distributions by row and by column. Numbers of rows, columns: 10,317, 34,352

3 Random Projection and Use for Clustering

For a discussion of random projections used for clustering, and also description and use of the Baire distance, see Contreras and Murtagh (2012). See also Sect. 4 below. Given our input data, i.e. a cloud of points in k-dimensional space, conventional random projection uses a random valued linear mapping in order to yield a much reduced dimensionality space: $f : \mathbb{R}^k \rightarrow \mathbb{R}^\ell$ where $\ell \ll k$. In our approach we are not seeking to use this ℓ-dimensional subspace, but rather we take a consensus ranked set of positions from the values of points on the ℓ axes. For this, we use a set of ℓ random projections, each onto a one-dimensional subspace. Therefore we consider ℓ random axes. The theory underpinning this, relative to conventional (Kaski 1998) random projection, is provided in Murtagh and Contreras (2015).

To deal with variability of outcomes in random projections used for clustering, Fern and Brodley (2003) project to a random subspace, apply a Gaussian mixture model, using expectation maximization, then an ensemble-based data aggregation matrix collects interrelationship information, which is submitted to an agglomerative hierarchical clustering. In Boutsidis et al. (2010), a random projection subspace is Shown to provide bounds on k-means clustering properties. The objective in Kaski (1998) is to determine the subspace of best metric fit to the original space.

For Urruty et al. (2007): "We begin by clustering the points of each of the selected uni-dimensional projections." And: "in the second phase we refine the clustering by using two processes: bimodulation and cluster expansion." (The former term introduced by those authors is for cluster specification using multiple random projections; and the latter term used by those authors uses hyper-rectangles to find the largest density cluster.) In this article, we develop in particular the early phase of clustering on uni-dimensional projections, and we relate such clustering to the Baire hierarchical clustering. Our objective is to develop a fast multiresolution hashing approach to clustering, rather than the optimal fit of proximity relations in $\mathbb{R}^\ell$, relative to proximity relations in $\mathbb{R}^k$.

4 Baire Clustering of a Random Spanning Path

4.1 Random Spanning Paths

Consider a random projection into a one-dimensional subspace, i.e. onto a random axis, of our set of documents. Such a random projection defines a permutation of the object set. It thereby defines a random spanning path. Spanning paths are useful and beneficial for data analysis. An optimal (i.e. minimum summed weight) spanning path has been used as an alternative to a minimal spanning tree (Murtagh 1985, ch. 4). The spanning path is the solution of the travelling salesman problem (Murtagh 1985, ch. 1). Braunstein et al. (2007) consider bounds for random path lengths relative to the optimal path length in the case of Erdős–Rényi and scale-free networks.

4.2 Inducing a Hierarchy through Endowing the Data with the Baire Metric

Our algorithm is as follows. Determine a random projection of our data. Induce a Baire hierarchy, using a regular 10-way tree. At level 1, the clusters will be labelled by 0, 1, 2, ...9. At level 2, the labelling is 00, 01, ...99. Full details of the Baire metric, and ultrametric, that endows the data with a hierarchy, is described in Contreras and Murtagh (2012). Our data values are univariate. Without loss of generality, take our values as being bounded by 0 and 1. An immediate consequence of the Baire metric is that, at level 1, all values that start with 0.3 will be in the same cluster; as will all values that start with 0.4; and so on for the 10 clusters at level 1. The Baire metric is a longest common prefix metric.

A random projection onto a one-dimensional axis provides a view of the relationships in the data, and hence a view of the clustering properties. See Contreras and Murtagh (2012). The random projected values are found to be quite similar in their interrelationships for different random vectors. We demonstrate this below. We determine the consensus or majority set of neighbourhood relationships from a sufficiently large set of random projections.

Now consider a given random projection. We determine a partition into clusters of the observables, following projection onto the random vector. A set of partitions can be sought, with their clusters ordered by inclusion.

Traditional approaches to clustering use pairwise dissimilarities, between adjacent clusters of points. (In partitioning, k-means takes a set of cluster centres and stepwise refines this set of cluster centres, together with their cluster assignments. Hierarchical clustering determines, stepwise, the smallest set of dissimilarities and agglomerates the associated pair of clusters.)

A direct reading of a partition is the alternative pursued here. Let the distance defined between adjacent clusters be a p-adic or m-adic distance (where typically

p refers to a prime number, and m refers to a non-prime integer). We define a cluster by an m-adic ball: $U_r(a) = \{x : |x-a|_m \leq r\}$. A Baire distance is associated with an ultrametric (a distance defined on a tree, rather than the real number line). Balls are either disjoint or are ordered by inclusion. It follows that for given r a partition is defined. For a set of values of r the set of associated partitions have clusters that are hierarchically structured, i.e. the associated set of clusters is a partially ordered set.

To address variability in results furnished by different random projections, we adopt the following approach: first, determine a stable, mean random projection. Then use it as the basis for a Baire clustering.

Parenthetically, let us address a comment sometimes made in regard to the m-adic distance used by us here. (We use $m = 10$; p-adic distances, where p is a prime, lead to an alternative to the real number system.) Consider two real measurements with values $2.99999\ldots$ and $3.0000\ldots$. These would be mapped onto different clusters in our approach. The following remark is however an appropriate one here: "two points on a complex protein may be close in Euclidean space but distant in terms of chemical reaction propensity" (Manton et al. 2008, pp. 81–82). In other words, if our digits have some form of inherent meaning, then it may well be fully appropriate to consider very similar real values to be quite separate and distinct.

4.3 Stability of Random Spanning Path

In Fig. 2 we assess the convergence, based on the first random projection, and the successive means of 2, 3, 4, ..., 98 random projections. The squared error is between the mean of these random projections, each normalized by its maximum value, and the mean of the 99 random projections, also normalized. We note the fast and stable (although not uniform) convergence. (The R code carrying out this processing is available on the web site containing our ancillary material.)

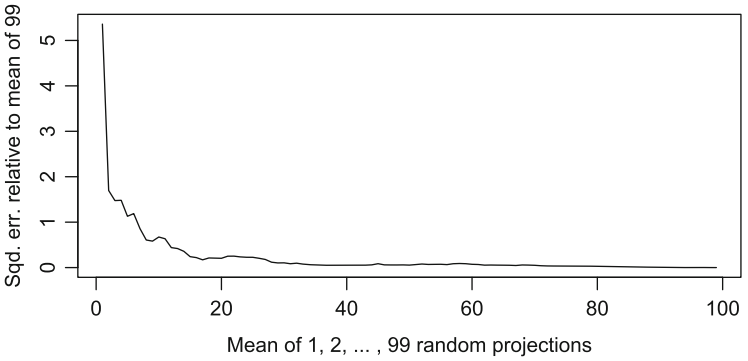

Fig. 2 Squared error of the mean of 1, 2, 3, ..., 98 random projections, relative to the mean of 99 random projections

4.4 Random Spanning Paths are Highly Correlated

In the case of random projection sets (mean of 99 realizations, sorted), for reproducibility we set the initialization seeds. For seeds 1471 and 3189, we had a correlation coefficient of 0.9999919.

Another random projection set (mean of 99 realizations, sorted) was generated with seed 7448. The correlations with the first two random projections were 0.9999937 and 0.9999905.

A further random projection was generated (seed 8914), and correlations with the first three sets were: 0.9999933, 0.9999898, 0.9999933. We conclude that using a given random projection set, in our work here resulting from 99 realizations, this is a fully sufficient basis for further cluster analysis.

In Murtagh and Contreras (2015) we consider theoretical properties of one-dimensional random projection in very high dimensional spaces.

5 Applying the Baire Distance to Obtain the Hierarchical Clustering

5.1 The Baire Metric and Ultrametric

As we have noted, the Baire distance is a longest common prefix distance which is also an ultrametric, or distance defined on a tree. In a given random projection, we can read off clusters using their Baire distance properties. Consider four adjacent, in rank order, projected values: 3.493297, 3.493731, 3.499185, 3.499410. The maximum value found for this particular random vector was 35.21912. We fix, in this instance, the projected values to be 8 digit values (viz., 2 digits in the integer part, and 6 digits in the fractional part, with zero padding if necessary). We define the Baire distance, with base 10, as 10 to the negative power of the last common, shared, digit.

The first two of our projected values above have Baire distance equal to 10^{-4} (because they share these digits: 3.493). The second two of our projected values above have Baire distance equal to 10^{-4}. The Baire distance between the second and third of our projected values above is 10^{-3}. The first and the fourth of our projected values have this same Baire distance, 10^{-3}.

Having defined the Baire distance between projected values, we next consider the Baire distance between clusters of projected values. Consistent with our consideration of adjacency of projected values, a cluster is a segment or succession of adjacent values. A singleton cluster is a single projected value. By considering the agglomeration of adjacent values 3.493297, 3.493731 at Baire distance 10^{-4}, and furthermore the adjacent values 3.499185, 3.499410 also at Baire distance 10^{-4}, we have the agglomeration of these two clusters, or segments, at Baire distance 10^{-3}, since the digits 3.49 are shared.

Computational analysis is as follows. For a random projection, we have the product of a (sparse) $k \times \ell$ matrix and a vector. Before taking sparsity into account, this gives $O(k\ell)$ time. The mean of a fixed number of random projections requires $O(k)$ time. The potentially linear Baire distance clustering comes from reading the mean random projection values, with assignment of each in turn to cluster nodes in the Baire hierarchy. In this way, a linked list of cluster (or node in the Baire hierarchy) members is built up.

5.2 A Theorem Ensuing from the Baire Ultrametric

The agglomeration of clusters takes a cluster or segment of ordered values $(x_{\text{lo}} \ldots x_{\text{hi}})$ to be agglomerated with a cluster of ordered values $(y_{\text{lo}} \ldots y_{\text{hi}})$. Based on adjacency in the clustering of random projections, and our definition of Baire distance between clusters, we have the following, where d_B is the Baire distance: $d_B(x_{\text{lo}}, y_{\text{hi}}) = d_B(x_{\text{hi}}, y_{\text{lo}})$.

If the two adjacent clusters are labelled c_x and c_y, then $\max\{d_B(i,j) \mid i \in c_x, j \in c_y\} = d_B(x_{\text{lo}}, y_{\text{hi}})$. Call this Baire distance $d_{\max}(c_x, c_y)$. Similarly, $\min\{d_B(i,j) \mid i \in c_x, j \in c_y\} = d_B(x_{\text{hi}}, y_{\text{lo}})$. Call this Baire distance $d_{\min}(c_x, c_y)$. What we availed of here was: $x_{\text{lo}} < x_{\text{hi}} < y_{\text{lo}} < y_{\text{hi}}$. From the foregoing description, the following theorem holds.

Theorem for Baire distance, *d*: $d_{\min}(c_x, c_y) = d_{\max}(c_x, c_y)$ for all contiguous clusters, c_x, c_y.

To show this, we start with singleton clusters, and the definition of the Baire distance, d. Following cluster formation, the cardinalities of the clusters will grow. By induction this theorem is extended to clusters c_x, c_y of any cardinality. A simple example ensues from the 4 points, together with their projected values, that were discussed above in Sect. 5.1. Given the terms "single link" and "complete link", as used in traditional hierarchical clustering, this theorem establishes that single and complete link agglomerative criteria are identical. This finding is consistent with having endowed our data not just with a metric, but with an ultrametric.

It has been noted how a random projection on a one-dimensional subspace is a random spanning path. This also establishes a contiguity or adjacency relationship between all points that we are analysing. So our hierarchical clustering can also be considered as a contiguity-constrained hierarchical clustering.

In Murtagh (1985) two contiguity-constrained hierarchical clustering algorithms were discussed. Proofs were provided that both would guarantee that no inversions could arise in the hierarchy, that is, there could be no non-monotonic change in cluster criterion value. One algorithm, also developed by other authors, Ferligoj and Batagelj (1982) and Legendre and Legendre (2012), was contiguity constrained complete link clustering: the pairwise most distant set of (by requirement, contiguous) cluster members determines the inter-cluster dissimilarity: $d_{\max}(c_x, x_y)$. The other contiguity-constrained hierarchical clustering was single link, where

inter-cluster dissimilarity is defined as the pairwise closest set of cluster members ($\min\{d_{ij}|i \in c_1, j \in c_2\}$), subject to the contiguity constraint.

By virtue of the theorem above, for all adjacent and agglomerable clusters c_x, c_y, $d_{\min}(c_x, c_y) = d_{\max}(c_x, c_y)$, we also have that the above described contiguity-constrained complete link and the contiguity-constrained single link hierarchical clustering methods are identical. This holds because of the Baire distance.

These perspectives add to the importance, in practice and in its theoretical foundations, of the theorem for the Baire distance.

5.3 *Visualization of Baire Hierarchy*

Using a regular 10-way tree, Fig. 3 shows a Baire hierarchy with nodes colour-coded (rainbow colour lookup table used), and with the root (a single colour, were it shown), comprising all clusters, to the bottom. The terminals of the 8-level tree are at the top. Ancillary material for this article, as noted in the "Introduction", is available. The R code used for Fig. 3 is listed there, and the code for the subsequent analysis of clusters extracted from the hierarchy.

The first Baire layer of clusters, displayed as the bottom level in Fig. 3, was found to have 10 clusters. (8 are very evident, visually.) The next Baire layer has 87

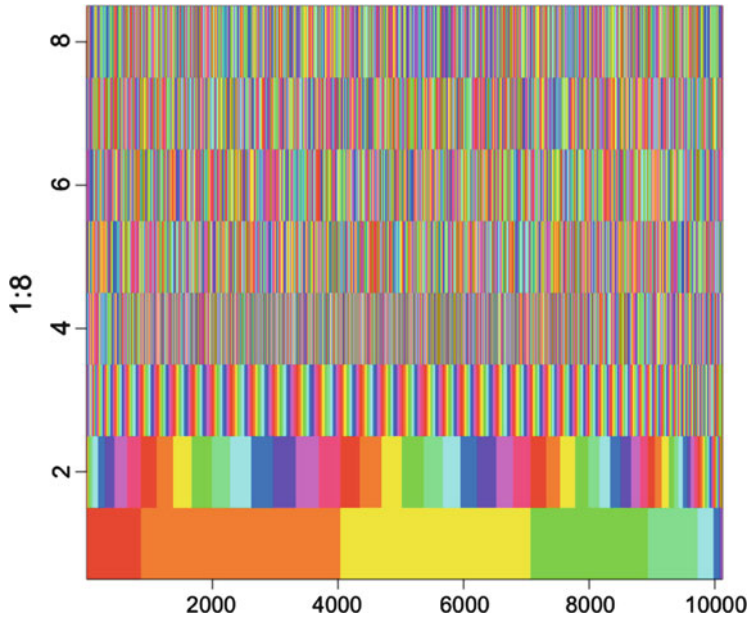

Fig. 3 Means of 99 random projections. Abscissa: the 10,118 (non-empty) documents are sorted (by random projection value). Ordinate: each of 8 digits comprising random projection values

clusters, and the third Baire layer has 671 clusters. See our ancillary material for a study of the clusters at layers 1 and 2.

6 Conclusions

In Contreras and Murtagh (2012), there is reporting on analysis of clusters found using the methodology developed here (in application domains that include astronomy and chemistry) and there is comparison with other, alternative processing approaches.

We can state that our work is oriented towards inter-cluster analysis, rather than intra-cluster analysis. That is to say, we want candidate observation classes, and furthermore we seek to be selective about what we derive from the data, in order to carry on to further use of the selected, derived clusters. Such overall processing is very suitable for big data analytics. The theorem stated in Sect. 5.2 points to the major importance of the Baire viewpoint. Further theoretical results are presented in Murtagh and Contreras (2015).

Acknowledgements We are grateful to Paul Morris for initial discussions related to this work.

References

Boutsidis, C., Zouzias, A., & Drineas, P. (2010). Random projections for k-Means clustering. *Advances in Neural Information Processing Systems, 23*(iii), 298–306.

Braunstein, L. A., Zhenhua W. U., Chen, Y., Buldyrev, S. V., Kalisky, T., Sreenivasan, S., Cohen, R., López, E., Havlin, S., & Stanley, H. E. (2007). Optimal path and minimal spanning trees in random weighted networks. *International Journal of Bifurcation and Chaos, 17* (7), 2215–2255.

Contreras, P., & Murtagh, F. (2012). Fast, linear time hierarchical clustering using the baire metric. *Journal of Classification*, **29**, 118–143.

Ferligoj, A., & Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika, 47*, 413–426.

Fern, X. Z., Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In T. Fawcett & N. Mishra (Eds.), *Proceedings 20th International Conference on Machine Learning* (pp. 186–193).

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *IJCNN'98, IEEE International Joint Conference on Neural Networks* (Vol. 1, pp. 413–418).

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd ed.). Amsterdam: Elsevier.

Manton, K. G., Huang, H. & Xiliang G. U. (2008). Chapter 3 - Molecular basis of CNS aging, frailty, fitness and longevity: A Model based on cellular energetic. In J. P. Tsai (Ed.), *Leading-edge cognitive disorders research*, New York: Nova Science, Hauppauge.

Matrix Market (2013). Matrix market exchange formats, http://math.nist.gov/MatrixMarket/formats.html

Murtagh, F. (1985). *Multidimensional clustering algorithms*. Heidelberg and Vienna: Physica-Verlag.

Murtagh, F. (2004). On ultrametricity, data coding, and computation. *Journal of Classification, 21*, 167–184.

Murtagh, F. (2013). MoreLikeThis and Scoring in Solr, report, 4 pp., 26 May 2013. http://www.multiresolutions.com/HiClBaireRanSpanPaths

Murtagh, F., & Contreras, P. (2015). Constant time search and retrieval in massive data with linear time and space setup, through randomly projected piling and sparse p-adic coding, article in preparation.

Murtagh, F., Downs, G., & Contreras, P. (2008). Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding. *SIAM Journal of Scientific Computing, 30*, 707–730.

Solr (2013). Solr, Apache Lucene based search server, http://lucene.apache.org/solr

Urruty, T., Djeraba, C., & Simovici, D. A. (2007). Clustering by random projections, *Advances in data mining. Theoretical aspects and applications lecture notes in computer science* (Vol. 4597, pp. 107–119).