Fast Hierarchical Clustering from the Baire Distance

Pedro Contreras[1] and Fionn Murtagh[1,2]

[1] Department of Computer Science. Royal Holloway, University of London. 57 Egham Hill. Egham TW20 OEX, England. pedro@cs.rhul.ac.uk
[2] Science Foundation Ireland. Wilton Place, Dublin 2, Ireland. fmurtagh@acm.org

Summary. The Baire or longest common prefix ultrametric allows a hierarchy, a multiway tree, or ultrametric topology embedding, to be constructed very efficiently.

The Baire distance is a 1-bounded ultrametric. For high dimensional data, one approach for the use of the Baire distance is to base the hierarchy construction on random projections.

In this paper we use the Baire distance on the Sloan Digital Sky Survey (SDSS, http://www.sdss.org) archive. We are addressing the regression of (high quality, more costly to collect) spectroscopic and (lower quality, more readily available) photometric redshifts. Nonlinear regression is used for mapping photometric and astrometric redshifts.

Key words: Ultrametric, Hierarchy, Clustering, Chemoinformatics, Astronomy.

1 Introduction

In this work we introduce a novel (ultrametric) distance called Baire and show how it can be used to produce clusters through grouping data in "bins". We seek to find inherent hierarchical structure in data, rather than fitting a hierarchy structure to data (as is traditionally used in multivariate data analysis) in an inexpensive computational way.

This paper is structured as follows: firstly we give a definition of the Baire distance; secondly we apply that distance to a chemoinformatics dataset; thirdly we apply the Baire distance to an astronomy dataset; finally we present our conclusions.

2 Longest Common Prefix or Baire Distance

2.1 Ultrametric Baire space and distance

A Baire space consists of countably infinite sequences with a metric defined in terms of the longest common prefix: the longer the common prefix, the closer

a pair of sequences. What is of interest to us here is this longest common prefix metric as defined in [1]. The longest common prefixes at issue are those of precision of any value. For example, consider two such values, x_{ij} and y_{ij}, which, when the context easily allows it, we will call x and y.

Without loss of generality we take x and y to be bounded by 0 and 1. Each are of some precision, and we take the integer $|K|$ to be the maximum precision. We pad a value with 0s if necessary, so that all values are of the same precision.

Thus we consider ordered sets x_k and y_k for $k \in K$. In line with our notation, we can write x_k and y_k for these numbers, with the set K now ordered. So, $k = 1$ is the first decimal place of precision; $k = 2$ is the second decimal place; . . . ; $k = |K|$ is the $|K|\,th$ decimal place. The cardinality of the set K is the precision with which a number, x_k , is measured.

Consider as examples $x_k = 0.478$; and $Y_k = 0.472$. In these cases, $|K| = 3$. For $k = 1$, we find $x_k = y_k = 4$. For $k = 2$, $x_k = y_k$. But for $k = 3$, $x_k \neq y_k$.

We now introduce the following distance (case of vectors x and y, with 1 attribute):

$$d_B(x_K, y_K) = \begin{cases} 1 & \text{if } x_1 \neq y_1 \\ \inf\ 2^{-n} & \quad x_n = y_n \quad 1 \leq n \leq |K| \end{cases} \tag{1}$$

We call this d_B value Baire distance, which can be shown to be an ultrametric [1, 2, 3, 4, 5].

Note that the base 2 is given for convenience. When dealing with binary data 2 is the chosen base. When working with real numbers the chosen base is 10.

3 Application to Chemoinformatics

In the 1990s, the Ward minimum variance hierarchical clustering method became the method of choice in the chemoinformatics community due to its hierarchical nature and the quality of the clusters produced. This method reached its limits once the pharmaceutical companies tried processing datasets of more than 500,000 compounds mainly due to its processing requirement of $O(n^2)$.

Datasets of half a million compounds are normal in today's world. There are different ways of encoding a compound to a machine readable form. In chemistry binary fingerprints for chemical compounds are common. The compound is encoded in a fixed length binary string. For details of different encoding systems in chemistry see [6].

In [1] we applied the Baire distance to a chemoinformatics dataset with the following characteristics:

– 1.2 million chemicals crossed by 1052 presence/absence attributes (binary matrix)

- the data matrix is highly sparse, occupancy is $\approx 8.6347\%$
- chemicals per attribute follow a power law with exponent ≈ 1.23
- attributes per chemical are approximately Gaussian.

3.1 Dimensionality reduction by random projection

As mentioned above it is a well known fact that traditional clustering methods do not scale well in very high dimensional spaces. A standard and widely used approach when dealing with high dimensionality is to first apply a dimensionality reduction method. For example, Principal Component Analysis (PCA) is a very popular choice to deal with this problem. It uses a linear transformation to form a simplified data set retaining the characteristics of the original data. PCA does this by means of choosing the attributes that best preserve the variance of the data. This is a good solution when the data allows these calculations, but PCA as well as other dimensionality reduction techniques remain expensive, computationally speaking.

In order to apply the Baire distance our first step was to reduce the dimensionality of the original data. We chose to use random projection [7, 8] not only because of performance but also because of some nice properties of this methodology. Random projection is the finding of a low dimensional embedding of a point set.

In fact random projection here works as a class of hashing function. Hashing is much faster than alternative methods because it avoids the pair-wise comparisons required for partition and classification. If two points (p, q) are close, they will have a very small $\|p - q\|$ (Euclidean metric) value; and they will hash to the same value with high probability. If they are distant, they should collide with small probability.

3.2 Chemoinformatics data clustering

In order to cluster the binary data we did the following:

- normalise the binary data matrix A by column sums; let's call the resulting matrix B
- produce a random vector Z
- project B into Z; let's call the resulting matrix R
- sort the matrix R
- cluster R applying the longest common prefix or Baire distance; then values that are identical fall in the same cluster.

Following the above process, we show in [1] (p. 728) that for this dataset we can get clusters that are very close to the clusters obtained by k-means. This can be due to many reasons: one reason is that data sparsity in a key factor (i.e. in a large sparse dataset groups are likely to be far from each other, and therefore groups are easier to identify).

4 Application to Astronomy

The Sloan Digital Sky Survey (SDSS) [9] is systematically mapping the sky, producing a detailed image of it and determining the positions and absolute brightnesses of more than 100 million celestial objects. It is also measuring the distance to a million of the nearest galaxies and to one hundred thousand quasars. The acquired data has been openly given to the scientific community.

In this work we are interested into four parameters from a subset of the SDSS data release 5 [10]: declination (DEC), right ascension (RA), spectrometric (Z_{spec}) and photometric (Z_{phot}). In particular we look into redshift data that, for either redshift, vary between 0 and 0.6.

DEC and RA give the position of an astronomical object in the sky. Spectrometric and photometric parameters are two different values obtained to measure the redshift. On one hand we have the spectrometric technique that uses the spectrum of electromagnetic radiation (including visible light) which radiates from stars and other celestial objects. On the other hand we have the photometric technique that uses a faster and more economical way of measuring the redshift, but is less precise than the spectrometric method.

Notice that when talking on the context of speed the advantage of using the Baire metric lies on that it can calculated in $O(n)$ time, unlike many of the traditional clustering method that need a higher computational complexity.

4.1 Clustering SDSS data based on a Baire distance

Figure 1 a) shows DEC versus RA, i.e. the object's position in the sky. Figure 1 b) shows the Z_{spec} and Z_{phot} currently used to cluster redshifts. This section of the sky represents approximately 0.5 million coordinate points. As can be observed, various sections of the sky are represented in the data.

Figures 1 c), d), e) and f) show graphically how Z_{spec} and Z_{phot} clusters look at different levels of decimal precision. For example, on the one hand we find that values of Z_{spec} and Z_{phot} that have equal precision in the 3rd decimal digit are highly correlated. On the hand when Z_{spec} and Z_{phot} have only the first decimal digit in common correlation is weaker (as shown in Figure 1 e).

Notice that in Figure 1 f) the data point are scatter around the plot area, these are the data points that have the least information in common, i.e. the data points that do not share any decimal places but the first digit.

Table 1 shows the clusters found for all different levels of precision. In other words this table shows the confidence levels for mapping of Z_{spec} and Z_{phot}. For example, we can expect that 82.49% of values for Z_{spec} and Z_{phot} to have at least two common prefix digits. Additionally we observe that a considerable number of observations share at least 3 digits in common.

In the following section we a take this notion of clusters even further and compare it to results obtained with the k-means clustering algorithm.

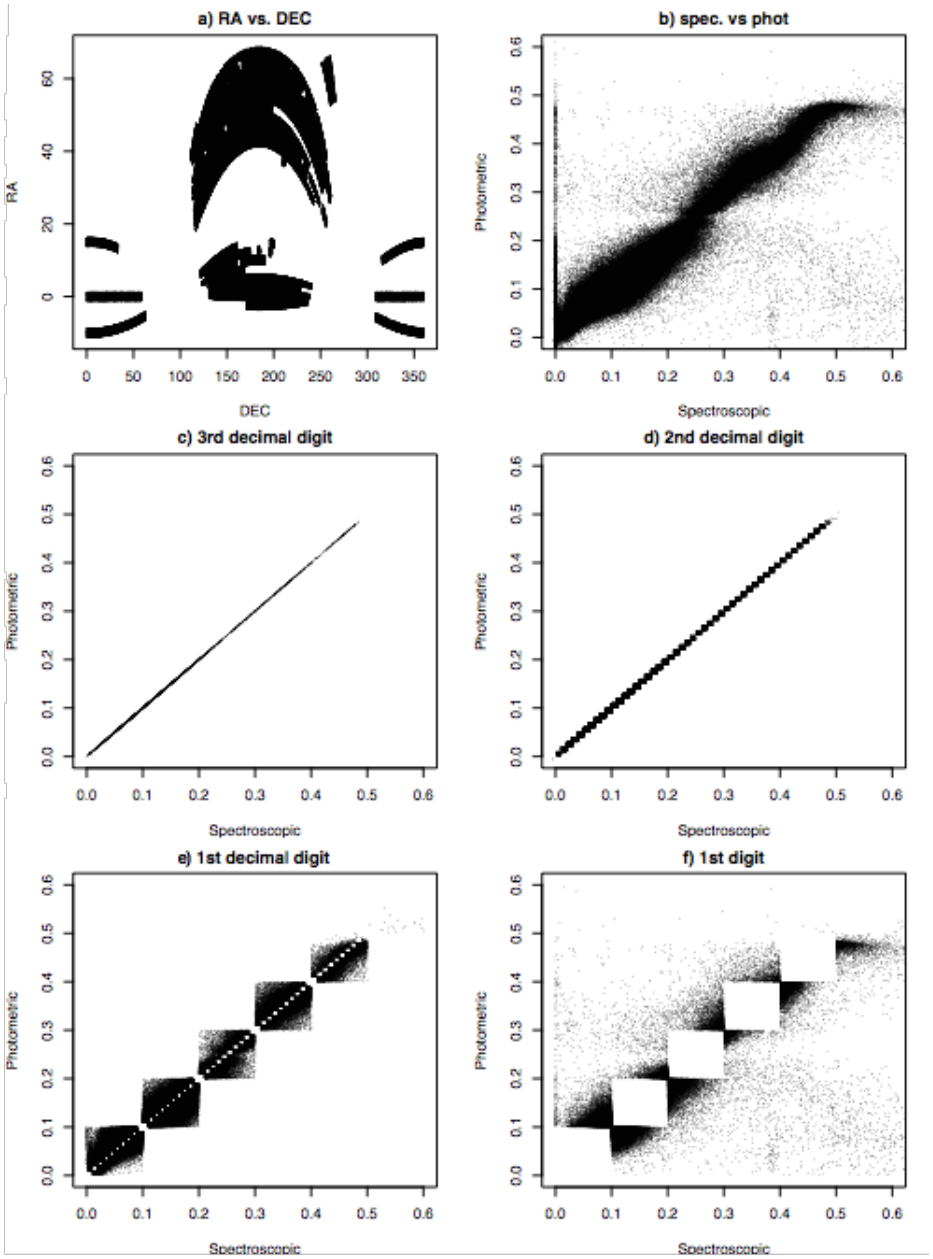

Fig. 1. SDSS data and results for a given precision digit

Table 1. Clusters based on the longest common prefix

Digit	No	%
1	76.187	17.19
2	270.920	61.14
3	85.999	19.40
4	8.982	2.07
5	912	0.20
6	90	0.02
7	4	—
	443.094	100

4.2 Baire and K-means cluster comparison

In order to establish how "good" the Baire clusters are we can compare them with k-means. Let us recall that our data values are in the interval [0, 0.6[(i.e. including zero values but excluding 0.6). Thus when building the Baire based clusters we will have a root node "0" that includes all the observations. For the Baire distance equal to two we have six nodes (or clusters) with indices "00, 01, 02, 03, 04, 05". For the Baire distance of three we have 60 clusters with indices "000, 001, 002, 003, 004,...,059" (i.e. ten children for each node 00,..,05).

We carried out a number of comparisons for the Baire distance of two and three. For example, we know that for $d_B = 2$ there are six clusters, then we took our data set and applied k-means with six centroids based on the Hartigan and Wong [11] algorithm. The results can be seen in Table 2, where the columns are the k-means clusters and the rows are the Baire clusters. From the Baire perspective we see that the node 00 has 97084 data points contained within the first k-means cluster and 64950 observations in the fifth. Looking at node 04, all members belong to the cluster 3 of k-means. We can see that the Baire clusters are closely related to the clusters produced by k-means at a given level of resolution.

Table 2. Cluster comparison based on Baire distance = 2; columns present the k-means clusters (k=6); rows present Baire nodes.

—	1	5	4	6	2	3
00	97084	64950	0	0	0	0
01	0	28382	101433	14878	0	0
02	0	0	0	18184	4459	0
03	0	0	0	0	25309	1132
04	0	0	0	0	0	11116
05	0	0	0	0	0	21

We can push this procedure further and compare the clusters for d_B defined from 3 digits of precision, and k-means with $k = 60$. Looking at the results from the Baire perspective we find that 27 clusters are overlapping, 9 clusters are empty, and 24 Baire clusters are completely within the boundaries of the ones produces by k-means as presented in Table 3.

It is seen that the match is consistent even if there are differences due to the different clustering criteria at issue. We have presented results in such as way as to show both consistency and difference.

Table 3. Cluster comparison based on Baire distance = 3; columns present the k-means clusters (k=60); rows present Baire nodes.

—	21	1	6	38	25	58	32	20	15	13	14	37	17	2	51	4
015	3733	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
004	0	3495	0	0	0	0	0	0	0	0	0	0	0	0	0	0
018	0	0	2161	0	0	0	0	0	0	0	0	0	0	0	0	0
020	0	0	0	1370	0	0	0	0	0	0	0	0	0	0	0	0
001	0	0	0	0	968	0	0	0	0	0	0	0	0	0	0	0
000	0	0	0	0	515	0	0	0	0	0	0	0	0	0	0	0
022	0	0	0	0	0	896	0	0	0	0	0	0	0	0	0	0
034	0	0	0	0	0	0	764	0	0	0	0	0	0	0	0	0
036	0	0	0	0	0	0	0	652	0	0	0	0	0	0	0	0
037	0	0	0	0	0	0	0	508	0	0	0	0	0	0	0	0
026	0	0	0	0	0	0	0	0	555	0	0	0	0	0	0	0
027	0	0	0	0	0	0	0	0	464	0	0	0	0	0	0	0
032	0	0	0	0	0	0	0	0	0	484	0	0	0	0	0	0
030	0	0	0	0	0	0	0	0	0	0	430	0	0	0	0	0
045	0	0	0	0	0	0	0	0	0	0	0	398	0	0	0	0
044	0	0	0	0	0	0	0	0	0	0	0	295	0	0	0	0
039	0	0	0	0	0	0	0	0	0	0	0	0	278	0	0	0
024	0	0	0	0	0	0	0	0	0	0	0	0	0	260	0	0
041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	231	0
042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	225	0
047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	350
048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57
049	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

5 Conclusions

In this work a novel distance called the Baire distance is presented. We show how this distance can be used to generate clusters in a way that is computationally inexpensive when compared with more traditional techniques. This

approach therefore makes a good candidate for exploratory data analysis when data sets are very big or in cases where the dimensionality is large. In addition to the advantage of speed, this distance is an ultrametric which can easily be seen as a hierarchy. We applied the Baire definition of distance to two cases:

- In the chemoinformatics case "good" clusters were obtained in the sense that these are close to those produced by k-means.
- In the astronomy case clusters generated with the Baire distance can be useful when calibrating redshifts. In generally, applying the Baire method to cases where digit precision is important can be of relevance, specifically to highlight data "bins" and some of their properties.

Future direction of work includes applying the Baire metric to other data sets. Our particular interest lies in high dimensional and massive data sets like the ones presented in this paper.

References

1. F. Murtagh, G. Downs and P. Contreras. Hierarchical Clustering of Massive, High Dimensional Data Sets by Exploiting Ultrametric Embedding. Society for Industrial and Applied Mathematics, *SIAM J. Scientific Computing.* Vol. 30, No. 2. pages 707–730. 2008.
2. F. Murtagh. On Ultrametricity, Data Coding, and Computation. *Journal of Classification.* Vol. 21. pages 167–184. 2004.
3. F. Murtagh. Thinking Ultrametrically. Ed. D. Banks and L. House and F.R. McMorris and P. Arabie and W. Gaul. Springer. *Classification, Clustering, and Data Mining Applications.* Springer. pages 3–14. 2004.
4. F. Murtagh. Quantifying Ultrametricity. Ed. J. Antoch. *Proceedings in Computational Statistics, Compstat.* Springer. pages 1561–1568. 2004
5. F. Murtagh. Identifying the Ultrametricity of Time Series. *European Physical Journal B.* Vol. 43. pages 573–579. 2005.
6. N. Brown. Chemoinformatics – An Introduction for Computer Scientists. *ACM Computing Surveys.* Vol. 41, No. 2, Article 8. 2009.
7. E. Bingham and H. Mannila. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. *KDD '01: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. ACM.* San Francisco, California. 2001.
8. S. Vempala. *The Random Projection Method.* DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, Rutgers University, Vol. 65. American Mathematical Society, 2004.
9. SDSS. Sloan Digital Sky Survey. `http://www.sdss.org`. 2008.
10. D. Raffaele, S. Antonino, L. Giuseppe, B. Massimo, P. Maurizio, D. Elisabetta and T. Roberto. Mining the SDSS archive. I. Photometric Redshifts in the Nearby Universe. ArXiv, arXiv:astro-ph/0703108v2. 2007.
11. J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. Applied Statistics 28, 100108. 1979.