Families of dendrograms

Patrick Erik Bradley[1]

Institut für Industrielle Bauproduktion, Englerstr. 7, D-76128 Karlsruhe
`bradley@ifib.uni-karlsruhe.de`

Abstract. A conceptual framework for cluster analysis from the viewpoint of p-adic geometry is introduced by describing the space of all dendrograms for n datapoints and relating it to the moduli space of p-adic Riemannian spheres with punctures using a method recently applied by Murtagh (2004b). This method embeds a dendrogram as a subtree into the Bruhat-Tits tree associated to the p-adic numbers, and goes back to Cornelissen et al. (2001) in p-adic geometry. After explaining the definitions, the concept of classifiers is discussed in the context of moduli spaces, and upper bounds for the number of hidden vertices in dendrograms are given.

1 Introduction

Dendrograms are ultrametric spaces, and ultrametricity is a pervasive property of observational data, and by Murtagh (2004a) this offers computational advantages and a well understood basis for developing data processing tools originating in p-adic arithmetic. The aim of this article is to show that the foundations can be laid much deeper by taking into account a natural object in p-adic geometry, namely the *Bruhat-Tits tree*. This locally finite, regular tree naturally contains the dendrograms as subtrees which are uniquely determined by assigning p-adic numbers to data. Hence, the classification task is conceptionally reduced to finding a suitable p-adic data encoding. Dragovich and Dragovich (2006) find a 5-adic encoding of DNA-sequences, and Bradley (2007) shows that strings have natural p-adic encodings.

The geometric approach makes it possible to treat time-dependent data on an equal footing as data that relate only to one instant of time by providing the concept of *family of dendrograms*. Probability distributions on families are then seen as a convenient way of describing classifiers.

Our illustrative toy data set for this article is given as follows:

Example 1.1 *Consider the data set $D = \{0, 1, 3, 4, 12, 20, 32, 64\}$ given by $n = 8$ natural numbers. We want to hierarchically classify it with respect to the 2-adic norm $|\cdot|_2$ as our distance function, as defined in Section 2.*

2 A brief introduction to p-adic geometry

Euclidean geometry is modelled on the field $\mathbb{R}$ of real numbers which are often represented as decimals, i.e. expanded in powers of the number 10^{-1}:

$$x = \sum_{\nu=m}^{\infty} a_\nu 10^{-\nu}, \quad a_\nu \in \{0,\ldots,9\}, \quad m \in \mathbb{Z}.$$

In this way, $\mathbb{R}$ completes the field $\mathbb{Q}$ of rational numbers with respect to the absolute norm $|x| = \begin{cases} x, & x \geq 0 \\ -x, & x < 0 \end{cases}$. On the other hand, the p-adic norm on $\mathbb{Q}$ with

$$|x|_p = \begin{cases} p^{-\nu_p(x)}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

is defined for $x = \frac{a_1}{a_2}$ by the difference $\nu_p(x) = \nu_p(a_1) - \nu_p(a_2) \in \mathbb{Z}$ in the multiplicities with which numerator and denominator of x are divisible by the prime number p: $a_i = p^{\nu_p(a_i)} u_i$, and u_i not divisible by p, $i = 1, 2$.

The p-adic norm satisfies the *ultrametric triangle inequality*

$$|x + y|_p \leq \max\{|x|_p, |y|_p\}.$$

Completing $\mathbb{Q}$ with respect to the p-adic norm yields the field $\mathbb{Q}_p$ of *p-adic numbers* which is well known to consist of the power series

$$x = \sum_{\nu=m}^{\infty} a_\nu p^\nu, \quad a_\nu \in \{0,\ldots,p-1\}, \quad m \in \mathbb{Z}. \tag{1}$$

Note, that the p-adic expansion is in increasing powers of p, whereas in the decimal expansion, it is the powers of 10^{-1} which increase arbitrarily. An introduction to p-adic numbers is e.g. Gouvêa (2003).

Example 2.1 *For our toy data set D, we have $|0|_2 = 0$, $|1|_2 = |3|_2 = 1$, $|4|_2 = |12|_2 = |20|_2 = 2^{-2}$, $|32|_2 = 2^{-5}$, $|64|_2 = 2^{-6}$, i.e. $|\cdot|_2$ is maximally 1 on D. Other examples: $|3/2|_3 = |6/4|_3 = 3^{-1}$, $|20|_5 = 5^{-1}$, $|p^{-1}|_p = |p|_p^{-1} = p$.*

Consider the unit disk $\mathbb{D} = \{x \in \mathbb{Q}_p \mid |x|_p \leq 1\} = B_1(0)$. It consists of the so-called *p-adic integers*, and is often denoted as $\mathbb{Z}_p$ when emphasizing its ring structure, i.e. closedness under addition, subtraction and multiplication. A p-adic number x lies in an arbitrary closed disk $B_{p^{-r}}(a) = \{x \in \mathbb{Q}_p \mid |x - a|_p \leq p^{-r}\}$, where $r \in \mathbb{Z}$, if and only if $x - a$ is divisible by p^r. This condition is equivalent to x and a having the first r terms in common in their p-adic expansions (1). The possible radii are all integer powers of p, so the disjoint disks $B_{p^{-1}}(0), B_{p^{-1}}(1), \ldots, B_{p^{-1}}(p-1)$ are the maximal proper subdisks of $\mathbb{D}$, as they correspond to truncating the power series (1) after the constant term. There is a unique minimal disk in which $\mathbb{D}$ is contained properly, namely $B_p(0) = \{x \in \mathbb{Q}_p \mid |x|_p \leq p\}$. These observations hold

true for arbitrary p-adic disks, i.e. any disk $B_{p^{-r}}(x)$, $x \in \mathbb{Q}_p$, is partitioned into precisely p maximal subdisks and lies properly in a unique minimal disk. Therefore, if we define a graph $\mathscr{T}_{\mathbb{Q}_p}$ whose vertices are the p-adic disks, and edges are given by minimal inclusion, then every vertex of $\mathscr{T}_{\mathbb{Q}_p}$ has precisely $p+1$ outgoing edges. In other words, $\mathscr{T}_{\mathbb{Q}_p}$ is a $p+1$-*regular* tree, and p is the size of the residue field $\mathbb{F}_p = \mathbb{Z}_p/p\mathbb{Z}_p$.

Definition 2.2 The tree $\mathscr{T}_{\mathbb{Q}_p}$ is called the *Bruhat-Tits tree* for $\mathbb{Q}_p$.

Remark 2.3 *Definition 2.2 is not the usual way to define $\mathscr{T}_{\mathbb{Q}_p}$. The problem with this ad-hoc definition is that it does not allow for any action of the projective linear group $\mathrm{PGL}_2(\mathbb{Q}_p)$. A definition invariant under projective linear transformations can be found e.g. in Herrlich (1980) or Bradley (2006).*

An important observation is that any infinite descending chain

$$B_1 \supseteq B_2 \supseteq \dots \tag{2}$$

of strictly decreasing p-adic disks converges to a unique p-adic number $\{x\} = \bigcap\limits_n B_n$. A chain (2) defines a halfline in the Bruhat-Tits tree $\mathscr{T}_{\mathbb{Q}_p}$. Halflines differing only by finitely many vertices are said to be *equivalent*, and the equivalence classes under this equivalence relation are called *ends*. Hence the observation means that the p-adic numbers correspond to ends of $\mathscr{T}_{\mathbb{Q}_p}$. There is a unique end $B_1 \subseteq B_2 \subseteq \dots$ coming from any strictly increasing sequence of disks. This end corresponds to the point at infinity in the p-adic projective line $\mathbb{P}^1(\mathbb{Q}_p) = \mathbb{Q}_p \cup \{\infty\}$, whence the well known fact:

Lemma 2.4 *The ends of $\mathscr{T}_{\mathbb{Q}_p}$ are in one-to-one correspondance with the $\mathbb{Q}_p$-rational points of the p-adic projective line $\mathbb{P}^1$, i.e. with the elements of $\mathbb{P}^1(\mathbb{Q}_p)$.*

From the viewpoint of geometry, it is important to distinguish between the p-adic projective line $\mathbb{P}^1$ as a p-adic manifold and its set $\mathbb{P}^1(\mathbb{Q}_p)$ of $\mathbb{Q}_p$-rational points, in the same way as one distinguishes between the affine real line $\mathbb{A}^1$ as a real manifold and its rational points $\mathbb{A}^1(\mathbb{Q}) = \mathbb{Q}$, for example. One reason for distinguishing between a space and its points is:

Lemma 2.5 *Endowed with the metric topology from $|\cdot|_p$, the topological space $\mathbb{Q}_p$ is totally disconnected.*

The usual approaches towards defining more useful topologies on p-adic spaces are by introducing more points. Such an approach is the *Berkovich topology*, which we will very briefly describe. More details can be found in Berkovich (1990).

The idea is to allow disks whose radii are arbitrary positive real numbers, not merely powers of p as before. Any strictly descending chain of such disks gives a point in the sense of Berkovich. For the p-adic line $\mathbb{P}^1$ this amounts to:

Theorem 2.6 (Berkovich) $\mathbb{P}^1$ *is non-empty, compact, hausdorff and arc-wise connected. Every point of* $\mathbb{P}^1 \setminus \{\infty\}$ *corresponds to a descending sequence* $B_1 \supseteq B_2 \supseteq \ldots$ *of p-adic disks such that* $B = \bigcap B_n$ *is one of the following:*

1. *a point* x *in* $\mathbb{Q}_p$,
2. *a closed p-adic disk with radius* $r \in |\mathbb{Q}_p|_p$,
3. *a closed p-adic disk with radius* $r \notin |\mathbb{Q}_p|_p$,
4. *empty.*

Points of types 2. to 4. are called *generic*, points of type 1. *classical*. We remark that Berkovich's definition of points is technically somewhat different and allows to define more general *p*-adic spaces. Finally, the Bruhat-Tits tree $\mathscr{T}_{\mathbb{Q}_p}$ is recovered inside $\mathbb{P}^1$:

Theorem 2.7 (Berkovich) $\mathscr{T}_{\mathbb{Q}_p}$ *is a retract of* $\mathbb{P}^1 \setminus \mathbb{P}^1(\mathbb{Q}_p)$, *i.e. there is a map* $\mathbb{P}^1 \setminus \mathbb{P}^1(\mathbb{Q}_p) \to \mathscr{T}_{\mathbb{Q}_p}$ *whose restriction to* $\mathscr{T}_{\mathbb{Q}_p}$ *is the identity map on* $\mathscr{T}_{\mathbb{Q}_p}$.

3 *p*-adic dendrograms

ν_2	0	1	3	4	12	20	32	64
0	∞	0	0	2	2	2	5	6
1	0	∞	1	0	0	0	0	0
3	0	1	∞	0	0	0	0	0
4	2	0	0	∞	3	4	2	2
12	2	0	0	3	∞	3	2	2
20	2	0	0	4	3	∞	2	1
32	5	0	0	2	2	2	∞	5
64	6	0	0	2	2	1	5	∞
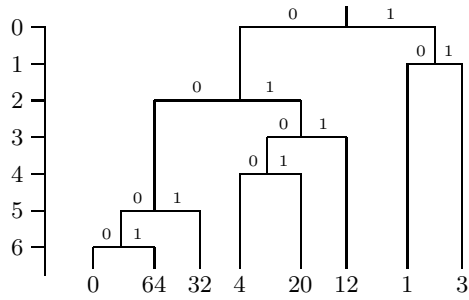
Fig. 1. 2-adic valuations for D.

Fig. 2. 2-adic dendrogram for $D \cup \{\infty\}$.

Example 3.1 *The 2-adic distances within D are encoded in Figure 1, where* $\mathrm{dist}(i,j) = 2^{-\nu_2(i,j)}$, *if* $\nu_2(i,j)$ *is the corresponding entry in Figure 1, using* $2^{-\infty} = 0$. *Figure 2 is the dendrogram for D using* $|\cdot|_2$: *the distance between disjoint clusters equals the distances between any of their representatives.*

Let $X \subseteq \mathbb{P}^1(\mathbb{Q}_p)$ be a finite set. By Lemma 2.4, a point of X can be considered as an end in $\mathscr{T}_{\mathbb{Q}_p}$.

Definition 3.2 The smallest subtree $\mathscr{D}(X)$ of $\mathscr{T}_{\mathbb{Q}_p}$ whose ends are given by X is called the *p-adic dendrogram* for X.

Cornelissen et al. (2001) use p-adic dendrograms for studying p-adic symmetries, cf. also Cornelissen and Kato (2005). We will ignore vertices in $\mathscr{D}(X)$ from which precisely two edges emanate. Hence, for example, $\mathscr{D}(\{0, 1, \infty\})$ consists of a unique vertex $v(0, 1, \infty)$ and three ends. The dendrogram for a set $X \subseteq \mathbb{N} \cup \{\infty\}$ containing $\{0, 1, \infty\}$ is a rooted tree with root $v(0, 1, \infty)$.

Example 3.3 *The 2-adic dendrogram in Figure 2 is nothing but $\mathscr{D}(X)$ for $X = D \cup \{\infty\}$ and is in fact inspired by the first dendrogram of Murtagh (2004b). The path from the top cluster to x_i yields its binary representation $[\cdot]_2$ which easily translates into the 2-adic expansion: $0 = [0000000]_2$, $64 = [1000000]_2 = 2^6$, $32 = [0100000]_2 = 2^5$, $4 = [0000100]_2 = 2^2$, $20 = [0010100]_2 = 2^2 + 2^4$, $12 = [0001100]_2 = 2^2 + 2^3$, $1 = [0000001]_2$, $3 = [0000011]_2 = 1 + 2^1$.*

Any encoding of some data set M which assigns to each $x \in M$ a p-adic representation of an integer including 0 and 1, yields a p-adic dendrogram $\mathscr{D}(M \cup \{\infty\})$ whose root is $v(0, 1, \infty)$, and any dendrogram for real data can be embedded in a non-unique way into $\mathscr{T}_{\mathbb{Q}_p}$ as a p-adic dendrogram in such a way that $v(0, 1, \infty)$ represents the top cluster, if p is large enough. In particular, any binary dendrogram is a 2-adic dendrogram. However, a little algebra helps to find sufficiently large 2-adic Bruhat-Tits trees $\mathscr{T}_K$ which allow embeddings of arbitrary dendrograms into $\mathscr{T}_K$. In fact, by K we mean a finite extension field of $\mathbb{Q}_p$. The p-adic norm $|\cdot|_p$ extends uniquely to a norm $|\cdot|_K$ on K, for which it is a complete field, called a *p-adic number field*. The *integers of K* are again the unit disk $\mathcal{O}_K = \{x \in K \mid |x|_K \leq 1\}$, and the role of the prime p is played by a so-called *uniformiser* $\pi \in \mathcal{O}_K$. It has the property that $\mathcal{O}_K/\pi\mathcal{O}_K$ is a finite field with $q = p^f$ elements and contains $\mathbb{F}_p$. Hence, if some dendrogram has a vertex with maximally $n \geq 2$ children, then we need K large enough such that $2^f \geq n$. This is possible by the results of number theory. Restricting to the prime characteristic 2 has not only the advantage of avoiding the need to switch the prime number p in the case of more than p children vertices, but also the arithmetic in 2-adic number fields is known to be computationally simpler, especially as in our case the so-called *unramified* extensions, i.e. where $\dim_{\mathbb{Q}_2} K = f$, are sufficient.

Example 3.4 *According to Bradley (2007), strings over a finite alphabet can be encoded in an unramified extension of $\mathbb{Q}_p$, and hence be classified p-adically.*

4 The space of dendrograms

From now on, we will formulate everything for the case $K = \mathbb{Q}_p$, bearing in mind that all results hold true for general p-adic number fields K. Let $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{P}^1(\mathbb{Q}_p)$ consist of n distinct classical points of $\mathbb{P}^1$ such that $x_1 = 0$, $x_2 = 1$, $x_3 = \infty$. Similarly as in Theorem 2.7, the p-adic dendrogram $\mathscr{D}(S)$ is a retract of the marked projective line $X = \mathbb{P}^1 \setminus S$. We call $\mathscr{D}(S)$

the *skeleton* of X. The space of all projective lines with n such markings is denoted by $\mathfrak{M}_n$, and the space of corresponding p-adic dendrograms by $\mathfrak{D}_{n-1}$. $\mathfrak{M}_n$ is a p-adic space of dimension $n-3$, its skeleton $\mathfrak{D}_{n-1}$ is a cw-complex of real polyhedra whose cells of maximal dimension $n-3$ consist of the binary dendrograms. Neighbouring cells are passed through by contracting bounded edges as the $n-3$ "free" markings "move" about $\mathbb{P}^1$ without colliding. For example, $\mathfrak{M}_3$ is just a point corresponding to $\mathbb{P}^1 \setminus \{0,1,\infty\}$. $\mathfrak{M}_4$ has one free marking λ which can be any $\mathbb{Q}_p$-rational point from $\mathbb{P}^1 \setminus \{0,1,\infty\}$. Hence,
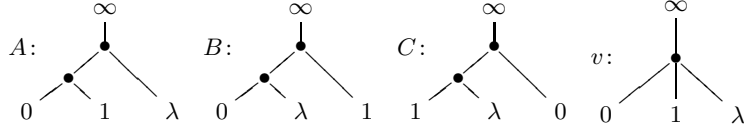
Fig. 3. Dendrograms representing the different regions of $\mathfrak{D}_3$.

the skeleton $\mathfrak{D}_3$ is itself a binary dendrogram with precisely one vertex v and three unbounded edges A, B, C (cf. Figure 3). For $n \geq 3$ there are maps

$$f_{n+1} \colon \mathfrak{M}_{n+1} \to \mathfrak{M}_n, \quad \phi_{n+1} \colon \mathfrak{D}_n \to \mathfrak{D}_{n-1},$$

which forget the $(n+1)$-st marking. Consider a $\mathbb{Q}_p$-rational point $x \in \mathfrak{M}_n$, corresponding to $\mathbb{P}^1 \setminus S$ with skeleton d. Its fibre $f_{n+1}^{-1}(x)$ corresponds to $\mathbb{P}^1 \setminus S'$ for all possible S' whose first n entries constitute S. Hence, the extra marking $\lambda \in S' \setminus S$ can be taken arbitrarily from $\mathbb{P}(\mathbb{Q}_p) \setminus S$. In this way, the space $f_{n+1}^{-1}(x)$ can be considered as $\mathbb{P}^1 \setminus S$, and $\phi_{n+1}^{-1}(d)$ as the p-adic dendrogram for S. What we have seen is that taking fibres recovers the dendrograms corresponding to points in the space $\mathfrak{D}_n$. Instead of fibres of points, one can take fibres of arbitrary subspaces:

Definition 4.1 A *family of dendrograms with n data points over a space Y* is a map $Y \to \mathfrak{D}_n$ from some p-adic space Y to $\mathfrak{D}_n$.

For example, take $Y = \{y_1, \ldots, y_T\}$. Then a family $Y \to \mathfrak{D}_n$ is a time series of n collision-free particles, if $t \in \{1, \ldots, T\}$ is interpreted as time variable. It is also possible to take into account colliding particles by using compactifications of $\mathfrak{M}_n$ as described in Bradley (2006).

5 Distributions on dendrograms

Given a dendrogram $\mathscr{D}$ for some data $S = \{x_1, \ldots, x_n\}$, the idea of a classifier is to incorporate a further datum $x \notin S$ into the classification scheme represented by $\mathscr{D}$. Often this is done by assigning probabilities to the vertices

of $\mathscr{D}$, depending on x. The result is then a family of possible dendrograms for $S \cup \{x\}$ with a certain probability distribution. It is clear that, in the case of p-adic dendrograms, this family is nothing but $\phi_{n+1}^{-1}(d) \to \mathfrak{D}_n$, if $d \in \mathfrak{D}_{n-1}$ is the point representing $\mathscr{D}$. This motivates the following definition:

Definition 5.1 A *universal p-adic classifier $\mathcal{C}$ for n given points* is a probability distribution on $\mathfrak{M}_{n+1}$.

Here, we take on $\mathfrak{M}_{n+1}$ the Borel σ-algebra associated to the open sets of the Berkovich topology. If $x \in \mathfrak{M}_n$ corresponds to $\mathbb{P}^1 \setminus S$, then $\mathcal{C}$ induces a distribution on $f_{n+1}^{-1}(x)$, hence (after renormalisation) a probability distribution on $\phi_{n+1}^{-1}(d)$, where $d \in \mathfrak{D}_{n-1}$ is the point corresponding to the dendrogram $\mathscr{D}(S)$. The similar holds true for general families of dendrograms, e.g. time series of particles.

6 Hidden vertices

A vertex v in a p-adic dendrogram $\mathscr{D}$ is called *hidden*, if the class corresponding to v is not the top class and does not directly contain data points but is composed of non-trivial subclasses. The subforest of $\mathscr{D}$ spanned by its hidden vertices will be denoted by $\mathscr{D}^h$, and is called the *hidden part* of $\mathscr{D}$. The number b_0^h of connected components of $\mathscr{D}^h$ measures how the clusters corresponding to non-hidden vertices are spread within the dendrogram $\mathscr{D}$. We give bounds for b_0^h and the number v^h of hidden vertices, and refer to Bradley (2006) for the combinatorial proofs (Theorems 8.3 and 8.5).

Theorem 6.1 *Let $\mathscr{D} \in \mathfrak{D}_n$. Then*

$$v^h \leq \frac{n + 2 - b_0^h}{2} \quad and \quad b_0^h \leq \frac{n - 4}{3},$$

where the latter bound is sharp.

7 Conclusions

Since ultrametricity is the natural property which allows classification and is pervasive in observational data, the techniques of ultrametric analysis and p-adic geometry are at ones disposal for identifying and exploiting ultrametricity. A p-adic encoding of data provides a way to investigate arithmetic properties of the p-adic numbers representing the data.

It is our aim to lay the geometric foundation towards p-adic data encoding. From the geometric point of view it is natural to perform the encoding by embedding its underlying dendrogram into the Bruhat-Tits tree. In fact, the dendrogram and its embedding are uniquely determined by the p-adic numbers

representing the data. For this end, we give an account of p-adic geometry in order to define p-adic dendrograms as subtrees of the Bruhat-Tits tree.

In the next step we introduce the space of all dendrograms for a given number of data points which, by p-adic geometry, is contained in the space $\mathfrak{M}_n$ of all marked projective lines, an object appearing in the context of the classification of Riemann surfaces. The advantages of considering the space of dendrograms rely on the fact that a conceptual formulation of moving particles as families of dendrograms is made possible, and its simple geometry as a polyhedral complex. Also, assigning distributions on $\mathfrak{M}_n$ allows for probabilistic incorporation of further data to a given dendrogram. At the end, we give bounds for the numbers of hidden vertices and hidden components of dendrograms.

What remains to do is to computationally exploit the foundations laid in this article by developping a code along these lines and apply it to Fionn Murtagh's task of finding ultrametricity in data.

Acknowledgements

References

BERKOVICH, V.G. (1990): *Spectral Theory and Analytic Geometry over Non-Archimedean Fields.* Mathematical Surveys and Monographs, 33, AMS.

BRADLEY, P.E. (2006): Degenerating families of dendrograms. *Preprint.*

BRADLEY, P.E. (2007): Mumford dendrograms. *Preprint.*

CORNELISSEN, G. and KATO, F. (2005): The p-adic icosahedron. *Notices of the AMS, 52, 720–727.*

CORNELISSEN, G., KATO, F. and KONTOGEORGIS, A. (2001): Discontinuous groups in positive characteristic and automorphisms of Mumford curves. *Mathematische Annalen, 320, 55–85.*

DRAGOVICH, B. AND DRAGOVICH, A. (2006): A p-Adic Model of DNA-Sequence and Genetic Code. *Preprint* `arXiv:q-bio.GN/0607018`.

GOUVÊA, F.Q. (2003): *p-adic numbers: an introduction.* Universitext, Springer.

HERRLICH, F (1980): Endlich erzeugbare p-adische diskontinuierliche Gruppen. *Archiv der Mathematik, 35, 505–515.*

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

MURTAGH, F. (2004): Thinking ultrametrically. In: D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.): *Classification, Clustering and Data Mining*, Springer, 3–14.