Ultrametricity of Dissimilarity Spaces and Its Significance for Data Mining

Dan Simovici*, Rosanne Vetro**, Kaixun Hua***

University of Massachusetts Boston
* dsim@cs.umb.edu
** rvetro@cs.umb.edu
*** kingsley@cs.umb.edu

Abstract. We introduce a measure of ultrametricity for dissimilarity spaces and examine transformations of dissimilarities that impact this measure. Then, we study the influence of ultrametricity on the behavior of two classes of data mining algorithms (kNN classification and PAM clustering) applied on dissimilarity spaces. We show that there is an inverse variation between ultrametricity and performance of classifiers. For clustering, increased ultrametricity generate clusterings with better separation. Lowering ultrametricity produce more compact clusters.

1 Introduction

Ultrametrics occur in the study of agglomerative hierarchical clustering algorithms, phylogenetic trees, p-adic numbers, certain physical systems, etc.

Our goal is to evaluate the degree of ultrametricity of dissimilarity spaces and to study the impact of the degree of ultrametricity on performance of classification and clustering algorithms.

Measuring ultrametricity of metric spaces has preoccupied a number of researchers (for example, in (Rammal et al., 1985)); however, the proposed measures are usable for the special case of metrics and are linked to the subdominant ultrametric attached to a metric which requires computing a single-link clustering or a minimal spanning tree. We propose an alternative measure referred to as the weak ultrametricity that can be applied to the more general case of dissimilarity spaces.

A *dissimilarity space* is a pair (S, d), where S is a set and $d : S \times S \longrightarrow \mathbb{R}$ is a function such that $d(x, y) \geqslant 0$, $d(x, x) = 0$, and $d(x, y) = d(y, x)$ for $x, y \in S$. We assume that all dissimilarity spaces considered are finite.

A *triangle* in (S, d) is a triple $(x, y, z) \in S^3$. To simplify the notation, we denote $t = (x, y, z)$ by xyz.

The mapping d is a *quasi-metric* if it is a dissimilarity and it satisfies the triangular inequality $d(x, y) \leqslant d(x, z) + d(z, y)$ for $x, y, z \in S$. In addition, if $d(x, y) = 0$ implies $x = y$, then d is a *metric*.