

Structure géométrique des distances hiérarchiques

Casanova-del-Angel, F.

Section d'Études et de la Recherche de l'École Supérieure d'Ingénierie et d'Architecture,
Unité Professionnelle « Adolfo López Mateos » de l'Institut Polytechnique National, Mexico, Mexique.
fcasanova@ipn.mx et fcasanova49@prodigy.net.mx

Résumé

La structure théorique de la classification hiérarchique nécessaire à la construction de l'arbre ou du dendrogramme est la base pour montrer la relation théorique des distances hiérarchiques géométriques pour une séquence de hiérarchies partielles où, lorsque lors de la suivante élection de classes à ajouter deux hiérarchies partielles égales existent, alors la hiérarchie partielle à ajouter dépendra des distances géométriques présentant les hiérarchies partielles se rapportant à celles de la troisième classe. Le développement théorique est illustré par le biais d'applications avec des données sur l'effet de la corrosion atmosphérique sur l'acier structurel dont est constituée l'infrastructure civile dans la Ville de Mexico, ainsi que sur l'évaluation du rendement des enseignants de troisième cycle au Mexique.

Mots clés : dendrogramme, distances hiérarchiques, relations triangulaires : équilatérales, isocèles et scalènes.

Abstract

Based on the theoretical structure of hierarchical classification to build the tree or dendrogram, is shown the theoretical relationship of geometrical hierarchical distances for a sequence of partial hierarchies where two partial and equal hierarchies exist in the election of classes to be added, then the partial hierarchy to be added depends on geometric distances shown by partial hierarchies regarding the third class. Theoretical development is exemplified through applications with data from the effect of atmospheric corrosion of structural steel in civil infrastructure in Mexico City and the assessment of teaching performance for postgraduate studies in Mexico.

Key words: dendrogram, hierarchical distances, triangular relationships: equilateral, isosceles and scalene.

Introduction

Ci-dessus sont présentées, comme point de départ de la description théorique de la structure géométrique des distances hiérarchiques, quelques commentaires sur les techniques ayant pour but la recherche de caractéristiques similaires au sein d'un ensemble de données afin d'identifier des groupes, des clusters ou des rassemblements d'un ensemble de valeurs caractérisés comme individuelles ou variables (voir [8]). Ces techniques exploratoires impliquent deux types d'algorithmes: les hiérarchiques et les non hiérarchiques ou de partition. Chacun de ceux-ci a différents niveaux de différenciation, c'est à dire, il est possible de distinguer deux grands types de méthodes de classification:

- i. les méthodes non hiérarchiques qui produisent directement une partition en un nombre fixe de classes. Un dendrogramme non hiérarchique se présente sous la forme d'un ensemble de points, où certains éléments sont reliés par des arêtes qui donnent à l'ensemble des individus des propriétés particulières dues aux données. L'un des dendrogrammes les plus intéressants est celui de longueur minimale, lequel est équivalent à une méthode de construction d'une hiérarchie de classes, et
- ii. les méthodes hiérarchiques qui produisent une succession de partitions en classes chaque fois similaires à l'image des célèbres classifications des espèces, genres, familles, ordres, etc.

Il est important de se souvenir qu'avant l'application de n'importe quel algorithme de classification, il est d'abord nécessaire d'effectuer une sélection appropriée des variables, l'élection d'une mesure indiquant une similarité entre les objets, c'est à dire, la distance et la méthode de classification (hiérarchique ou pas). La seconde étape est l'application de l'algorithme de classification et l'analyse des résultats à partir de leur description, où les nœuds ou les agrégats qui contiennent les aînés et les benjamins du nœud, son poids, la valeur de l'indice du niveau hiérarchique et le