

Resumo

A dissertação apresentada visa o estudo de métodos de Classificação Não Hierárquica que se inserem na metodologia VL.

A metodologia VL, Validade da Ligação, acenta no uso de medidas de probabilidade para comparar elementos e/ou classes. Os métodos classificatórios não hierárquicos abordados usam como critério de afectação de um elemento a uma classe, a função de distribuição de uma estatística amostral sobre os valores de comparação entre o elemento a classificar e os elementos da classe. Os métodos VL estudados recorrem a estatísticas amostrais de tendência central, nomeadamente, à média aritmética, à média geométrica, à mediana e ao centro, resultando os métodos NHMEAN, NHVGM, NHVMED e NHVC. Simultaneamente, realiza-se um estudo sobre os correspondentes métodos directos, cujo critério de afectação assenta na aplicação directa da estatística sobre os valores de comparação entre elementos.

Em geral, um procedimento classificatório não hierárquico inicia-se com um conjunto de centros que assumem o papel de pólos de agregação. Com base no critério wise-choice e considerando a estatística definida em cada um dos métodos, propõem-se novos critérios de inicialização.

O desempenho dos métodos foi avaliado face a conjuntos de dados simulados, com estruturas inicialmente impostas, usando como medida de comparação, entre a partição inicial e a estrutura fornecida, o Índice de Rand Modificado. Os métodos foram também aplicados a conjuntos de dados reais já estudados na literatura.

Abstract

The objective of the present dissertation is to study methods of Non-Hierarchical Clustering that come within the VL methodology.

The VL – Validity of the Link – methodology is based in the use of probability measures to compare elements and/or classes. The considered non-hierarchical clustering methods use as criterion to assign an element to a class the distribution function of a sample statistics of the comparison values between the element to assign and the classe's elements. The studied VL methods use sample central statistics, namely, the arithmetic mean, the geometrical mean, the median and the centre, leading respectively to the methods NHMEAN, NHVGM, NHVMED and NHVC. Simultaneously, a study of the corresponding direct methods, whose assignment criteria is based on the direct application of the statistics of the values of comparison between elements, is performed.

Generally a non-hierarchical clustering process is initialized with a set of centres, that act as aggregation poles. New initialization criteria are proposed, based on the wise-choice criterion and considering the specific statistics defined in each of the methods.

The performance of the methods is evaluated by considering simulated data with pre-defined structures, and using the Modified Rand Index to compare the initial and the obtained partitions. The methods are also applied to real sets, previously studied in the literature.