

Resumo

A dissertação aqui apresentada começa por situar os problemas da Classificação e da Classificação Hierárquica no contexto da Estatística Moderna, referindo as fontes matemáticas que lhe dão suporte. Assim a primeira preocupação foi fazer uma síntese das bases matemáticas da Análise Classificatória, e em particular da Análise Classificatória Hierárquica, e um levantamento dos principais problemas de investigação, pura e aplicada, desta área da actualidade. Desta atitude relevou-se o problema da validação e do controlo das estruturas classificatórias.

O resultado de um método de Classificação Hierárquica Ascendente depende de dois aspectos: a eventual existência de uma estrutura nos dados e do(s) método(s) utilizado(s), sendo de admitir que cada método tem um modelo de estrutura subjacente, que lhe é próprio e que ele é capaz de otimizar em cada situação. A aplicação de um critério a um conjunto de dados com estrutura mal definida, ou mesmo ausente, tem como resultado uma estrutura que está associada ao modelo subjacente ao método.

A necessidade de produzir dados com tipos de estruturas, sobre as quais se pudesse actuar, motivou a dedução de processos para a geração aleatória de árvores de classificação, tendo daí surgido a proposta de duas metodologias: a Geração com Parâmetro de Forma, que permite gerar árvores predominantemente de um tipo prefixado, e a Geração Uniforme.

Na sequência da análise dos vários critérios de Classificação Hierárquica Ascendente, foi realizado o estudo de um conjunto de métodos directos e os correspondentes de metodologia VL - Validade da Ligação, que recorrem a estatísticas de tendência central. Foram estudados e programados os métodos baseados na média geométrica, já propostos anteriormente, e os baseados no centro e na mediana.

Numa Classificação Hierárquica Ascendente ocorrem diferentes tipos de estruturas classificatórias, cuja pertinência de comparação é indiscutível. Foi realizado um estudo, de propriedades e de distribuição, dos diferentes coeficientes estatísticos ordinais disponíveis para tal fim e foi feita uma averiguação da possibilidade de obter coeficientes únicos para as diferentes comparações.

Com o conjunto de ferramentas mencionadas foram definidos protocolos capazes de informar sobre os problemas da comparação de índices, de hierarquias e de índices e hierarquias e sobre os efeitos que perturbações aleatórias têm sobre as mesmas.

Abstract

This thesis begins to situate Classification and Hierarchical Classification problems in the Modern Statistics context, relating mathematics sources that support them. Elaborate a summary of the Cluster Analysis and in particular of the Hierarchical Cluster Analysis mathematics bases, has been

the first preoccupation, and one lifting of the main pure and applied actual investigation problems of this area. From this attitude the validation problem and the clustering structures control has been unfolded important.

The Ascendant Hierarchical Cluster method result depends of two aspects: the possible existence of data structure and of the employed method(s). We know that each method has a proper subjacent structure model and that he has capability to optimise in each situation. When a data has no structure or a bad defined structure, the result is a structure that is associated with the subjacent model method.

The data production requirement with structures types about with we could control, motivated the deduction of classification trees random generation processes, two methodologies had there been proposed: a Parameter Form Generation, to generate trees with a preponderate prefixed type, and a Uniform Generation.

From the analysis of Hierarchical Ascendant Cluster several criterions, has been done the study of a direct methods set and the correspondents with the VL - Validity of the Link methodology, that are based in central character statistics. The methods based in geometric mean, proposed previously, in the middle and in the median were studied and programmed.

Different types of clustering structures take place in Ascendant Hierarchical Cluster, and is indisputable to compare them. A distribution and proprieties study, of the available different ordinal statistics coefficients, has been done and is presented, and the possibility to obtain unique coefficients for the different comparisons was searched.

With the mentioned tools set, protocols able to inform about the indexes, about hierarchies, and about indexes and hierarchies comparison problems and about effects that random disturbances have over them were defined.