RESUMO

O recente aumento da dimensão das bases de dados geradas pelas actividades industriais, governamentais, científicas entre outras, ultrapassaram em muito a nossa capacidade de análise e de interpretação de dados, necessárias à tomada de decisões. É por esse motivo que as técnicas de redução de dados investigadas na área de Extracção de conhecimento de Bases de Dados (ECBD) têm ganho certa relevância.

A ECBD é essencialmente um processo que pretende procurar identificar padrões válidos, anteriormente desconhecidos, úteis e inteligíveis, e é constituído por várias fases. A selecção e preparação de dados, redução de dimensões, procura de padrões, avaliação e interpretação dos resultados, constituem as várias fases do processo ECBD. A procura de padrões, realizada pelos algoritmos de aprendizagem e estatísticos, é efectuada tendo como base o tipo de problema (ex. classificação, regressão, agrupamento entre outros) e resulta num determinado modelo. Na fase de redução de dimensões pretende-se diminuir (segundo o número de casos, atributos e valores) o tamanho do conjunto de dados de forma a que este seja eficientemente processado pelos algoritmos de aprendizagem

Nesta Tese de Mestrado abordamos dois métodos de redução de dimensões. O primeiro envolve utilizar apenas um subconjunto de casos, e o outro utilizar apenas um subconjunto de atributos.

Apresentamos um método de redução de casos, que designámos de algoritmo de processamento por partições progressivas (PPP), que visa construir um modelo utilizando apenas um subconjunto do conjunto de dados inicial, de forma a poupar tempo de processamento, não sacrificando os resultados finais do modelo.

Quanto à selecção de atributos desenvolvemos um método que designámos de SAPPP que utilizando os resultados parciais do algoritmo PPP, combina método de selecção de atributos com o método de redução do número de casos.

Avaliámos empiricamente os dois métodos apresentados utilizando vários conjuntos dados e verificámos ser útil aplicá-los em conjuntos de dados com um elevado número de casos, confirmando a nossa expectativa inicial. Este resultado promissor justifica que se dedique mais esforço para comprovar os resultados iniciais e melhorá-los.

SUMMARY

Recent increase in the size of databases generated as a side effect of industrial, governmental, scientific activities often surpass our capacity to analyze and interpret this data and advance thus with decision making. This is why the techniques of data reduction studied within the area of Knowledge Discovery in Databases (KDD), have gained relevance.

The objective of Knowledge Discovery in Databases is to search for, and identify, valid and previously unknown patterns, that are intelligible to the user and also useful. This process consists of various phases, such as selection and preparation of data, dimensionality reduction, search for patterns and evaluation and interpretation of the results. The search for patterns, carried out often with the help of Machine Learning and Statistical algorithms. This process takes into account the problem type (e.g. classification, regression clustering among others) and produces a particular specialized model (e.g. a classifier). The objective of dimensionality reduction is to reduce the size of one (or more) dimensions of the data, such as number of cases, attributes or number of values.

In this M.Sc. Thesis we investigate two methods of dimensionality reduction. The first one involves using only a subset of the given cases (case reduction), and the other a subset of attributes (features). We present a method of case reduction, designated as an *algorithm for processing progressive partitions* (PPP), whose objective is to construct a model using only a subset of given cases, and in that way save time, without sacrificing much the overall performance. As for attribute reduction method, we have developed a method referred to as SAPPP, which combines *attribute selection* with the method of processing progressive partitions (PPP) referred to earlier.

We have carried evaluation of both dimensionality reduction methods on several datasets. We have verified that they are in general useful in datasets with large number

of cases, confirming thus our initial expectation. This promising results justifies that more effort is dedicated to this issue. The aim should be not only verify that the results hold in other situations, but also, to improve the methods further.