HORMONES 2005, 4(4):189-191

Editorial

Data Mining: A New Technique In Medical Research

E. Papageorgiou,[1] I. Kotsioni,[2] A. Linos[3]

[1]MSc, Statistician – Biostatistician, Institute of Preventive Medicine, Environmental and Occupational Health (Prolepsis), Athens, Greece, [2]Project Manager – Economist, Msc, Institute of Preventive Medicine, Environmental and Occupational Health (Prolepsis), Athens, Greece, [3]MD, MPH, FACE, Associate Professor, Department of Hygiene and Epidemiology, Medical School, University of Athens, Greece.

This issue hosts an article that discusses the application of Data mining techniques to the treatment of infertile men with azoospermia, and compares this procedure of categorisation into groups with the corresponding clinical categorisation. Although the dataset used was not very large, the authors have provided an interesting overview of this novel approach involving the use of Data Mining in azoospermia.

Data Mining may be defined as that composite of techniques employed to detect patterns in large datasets so as to extract "hidden" pieces of information. It is a fairly new technique, typically used to predict possible future trends or to discover concealed patterns in the "behaviour" of the data. As the name Data Mining implies, the main objective of these techniques is to search in large databases (databanks) for valuable, and often hidden, items of information. Thanks to the sophistication of modern-day technology, the researcher is enabled to process these vast databases in just a few minutes, a procedure that has been in the past both time-consuming and arduous.

Before the introduction of Data Mining tools, the actual process of pattern detection was extremely protracted and painstaking. While statisticians have for some time been performing Data Mining "manually", recent advances in statistical software, computer power, and storage capabilities as well as the creation of very large databanks have made possible data analysis at very low costs and have thus increased the accuracy of results and the chances of discovering hidden patterns.

Data Mining requires the use of advanced statistical algorithms and modelling. In other words it is the combination of Statistics and Artificial Intelligence (or Machine Learning) which allows us to explore for patterns in large datasets. Moreover, techniques linked to Data Mining may easily be applied on existing software and hardware, upgrading their capabilities and efficiency. Alkharouf et al, for example, demonstrate the use of OLAP, an online analytical processing Data Mining tool in gene-databases[1]. Techniques such as Artificial Neural Networks, Decision Trees, Classification (Nearest Neighbour Classification), Clustering, Data Mining Algorithms, Rule Induction etc, are widely acknowledged by experts as the most frequently utilized and favoured Data Mining techniques.

The benefits of using Data Mining are numerous and the ever-increasing, newly developed applications of technologically enhanced information systems guarantee the establishment of Data Mining techniques as a very powerful and valuable tool for a wide variety of users in different fields. Today, though the primary application of Data Mining is in the financial and marketing sectors, Data Mining usage has recently been expanded to other fields

Address all correspondence to:
XXXX....., Greece
Received 21-09-05, Revised 20-10-05, Accepted 27-10-05

such as Medicine, Biology, Genetics and Biomedical Sciences in general.

In Biomedical Sciences, Data Mining tools are mostly umployed in Genetics and Protein Biology where databanks are very large and complex. Tessier's paper[2] provides a good example of the use of Data Mining techniques in research concerning proteins, extracting fail-safe rules for the prediction of the disulfide-bonding state of cysteines, while Saito et al[3] succeeded in clearly demarcating the distinctions between mutant cell groups by developing a specialized Data Mining tool. Georgii et al[4] were motivated by gene-expression Data Mining techniques to analyze microarray data and came up with an innovative tool for the analysis of microarray gene expression data, while Perez et al[5] implemented in a public web server a method that enabled the ranking of genes in a region of the human genome according to their possible relation to a disease. Korn's research team[6] discovered a novel Data Mining procedure that can identify genes associated with pre-defined phenotypes and/or molecular pathways by using the Cancer Genome Anatomy Project expression Data, while as Bansal[7] eloquently explains, "Hundreds of microbial genomes and many eukaryotic genomes including a cleaner draft of human genome have been sequenced, raising the expectation of better control of microorganisms".

The use of Data Mining algorithms in Medicine might well be one of the most interesting aspects of "computer" application in the field. Pattern search algorithms can search through vast databanks of patient information, providing new insights into conundrums that routinely trouble experts of the biomedical profession. A possible successful application of Data Mining may be in Analytic Epidemiology. Epidemiologists often use massive "global" questionnaires, which, however, are ultimately self-defeating since it is effectively impossible to analyse all the data, while meantime many major breakthroughs are the result of random observations. Data Mining could therefore greatly contribute to the discovery of new etiologic disease associations and provide scientists of the field with valuable (or invaluable?) and previously unavailable knowledge.

Finlay's research group used Data Mining techniques to more accurately detect cardiac abnormalities through body surface potential mapping[8]. Sebban et al[9] demonstrate how Data Mining methods have been applied to the evolutionary genetics and molecular epidemiology of tuberculosis through specification of a technique that reduced some of the experimental constraints and improved the expert's knowledge of unknown patterns. Another example is that illustrated in Wren's and Garner's article[10] which approaches type 2 diabetes though Data Mining and concludes that epigenetic changes within the body are some of the causal factors in the pathogenesis of type 2 diabetes.

To conclude, Data Mining possesses eminent capacity to improve our knowledge of symptoms, diseases, and "virus behaviour patterns". Moreover, Data Mining techniques can also serve to determine patient behaviours. Data Mining has therefore numerous potential applications in most areas of Biomedical Sciences and is considered by experts to be a tool of such technological impact as to rival artificial intelligence. The phrase by Edmund X. DeJesus[11] states most eloquently the obvious: "We may well see the day when the Nobel Prize for a great discovery is awarded to a search algorithm".

REFERENCES

1. Alkharouf N, Jamison C, Matthews BF, 2005 Online analytical processing (OLAP): a fast and effective data mining tool for gene expression databases. J Biomed Biotechnol : Issue 2: 181-188.
2. Tessier D, Bardiaux B, Larre C, Popineau Y, 2004 Data mining techniques to study the disulfide-bonding state in proteins: signal peptide is a strong descriptor. Bioinformatics 20: 2509-2512.
3. Saito T, Sese J, Nakatani Y, et al, 2005, ata Mining tools for the Saccharomyces cerevisiae morphological database. Nucleic Acids Res 33: (Web Server Issue): W753-W757.
4. Georgii E, Richter L, Ruckert U, Kramer S, 2005 Analyzing microarray data using quantitative association rules. Bioinformatics 21: Suppl 2: 123-129.
5. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA, 2005 G2D: a tool for mining genes associated with disease. BMC Genet 6: 45.
6. Korn R, Rohrig S, Schulze-Kremer S, Brinkmann U, 2005 Common denominator procedure: a novel approach to gene-expression data mining for identification of phenotype-specific genes. Bioinformatics 21:

2766-2772.
7. Bansal AK 2005 Bioinformatics in microbial biotechnology – a mini review. Microbial Cell Fact 4: 19.
8. Finlay D, Nugent CD, McCullagh PJ, Black ND, 2005 Mining for diagnostic information in body surface potential maps: a comparison of feature selection techniques. BioMed Eng Online: 4: 51.
9. Sebban M, Mokrousov I, Rastogi N, Sola C, 2001 A data mining approach to spacer oligonucleotid typing of Mycobacterium tuberculosi. Bioinformatics 18: 235-242.
10. Wren J, Garner H, 2004 Data-mining analysis suggests an epigenetic pPathogenesis for type 2 diabetes. J Biomed Biotechnol : Issue 2: 104-112.
11. DeJesus E, October 1995 Editorial, Data Mining, State of Art, http://www.byte.com/art/9510/sec8/art1.htm , assessed on 12.10.2005.