Data Mining and Official Statistics

Gilbert Saporta
Chaire de Statistique Appliquée, Conservatoire National des Arts et Métiers
292 rue Saint Martin, F-75141 Paris cedex03,
saporta@cnam.fr

Abstract: Data mining is a new field at the frontiers of statistics and information technologies (database management, artificial intelligence, machine learning, etc.) which aims at discovering structures and patterns in large data sets. We examine here its definitions, tools, and how data mining could be used in official statistics.

1.What is Data Mining ?

Data Mining is often presented as a revolution in information processing. Here are two definitions taken from the literature:
U.M.Fayyad : *« Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data »*
D.J.Hand : *« Data Mining (consists in) the discovery of interesting, unexpected, or valuable structures in large data sets»*.

The development of Data Mining is related with the availability of very large databases and the need of exploiting these bases in a new way. As J.Friedman (1997) pointed out, the Data Base Management community has become recently interested in using DBMS for decision support. They realized that data which have been collected for management organisation such as recording transactions, might contain useful information for e.g. improving the knowledge of, and the service to customers.

One important feature is thus that Data Mining is concerned with data which were collected for another purpose: it is a *secondary analysis* of data bases. See for instance J.W.P.F. Kardaun and T.Alanko (1998) who speak of VLDB-NPFA, which stands for : *« Very Large Data Bases with data that are collected Not Primarily For Analysis, but for the management of individual cases »*.

The metaphor of Data Mining means that there are treasures (or nuggets) hidden under mountains of data, which may be discovered by specific tools. Of course this is not a new idea and there has been many developments in statistical methodology which were oriented towards the discovery of patterns or models through techniques like EDA (Exploratory Data Analysis) and Multivariate Exploratory Analysis (dimension reduction methods : component, correspondence and cluster analysis among many others). J.Kettenring (former ASA president) defines statistics as *« the science of learning from data »*.

What is new in Data Mining is :

- the availability of very large data bases, which makes obsolete classical definitions of large samples : billions of records, terabytes of data are not unusual. More and more data are collected automatically.

- the massive use of new techniques coming from the computer science community like neural networks, decision trees, induction rules.
- commercial interests in valorising existing information in order to propose individual solutions to targeted customers.
- new software packages, more user-friendly, with attractive interfaces, directed as much towards decision makers as professionals analysts, but much more expensive !

It is interesting to remark that statisticians (those who are not reluctant to this new trend…) prefer the expression Data Mining, while computer scientists use Knowledge Discovery, but it is more a matter of expressing a belonging to a community than a real difference.

2.Goals and tools of Data Mining

The purpose of Data Mining is to find structures in data. Following D.Hand (2000), there are two kinds of structures which are sought in data mining activities : models and patterns.

Building models is a major activity of many statisticians and econometricians, especially in NSI's and it will not be necessary to elaborate too long on this. Let's say that a model is a global summary of relationships between variables, which both helps to understand phenomenons and allows predictions. Linear models, simultaneous equations are widely used. But a model is generally chosen on an *a priori* basis, based upon a simplifying theory. Exploration of alternative models, possibly non-linear or in a non closed mathematical form is made feasible by DM algorithms. DM appears then as a collection of tools presented usually in one package, in such a way that several techniques may be compared on the same data-set.
DM does not reduce to decision trees, or neural networks, or genetic algorithms, as some software vendors claim.
DM algorithms offers extensive possibilities of finding models relating variables together : we have mentioned neural networks, decision trees, which finds non-linear models, graphical models (or Bayesian belief networks) etc. give a valuable representation of relations between variables.

In contrast to the global description given by a model, a pattern is often defined as a characteristic structure exhibited by a few number of points : for instance a small subgroup of customers with a high commercial value, or conversely highly risked. Tools involved here are often cluster analysis, which is a well known statistical technique, but also Kohonen self-organising maps, based on neural networks.
Association rule discovery, or market basket analysis, is one of the favourite and perhaps new tools of Data Mining : its origin is in analysing purchases in supermarket : one is interested in the percent of customers whose purchase simultaneously two goods, more precisely in identifying couples of goods A,B with high conditional probabilities of purchasing A, given that B is purchased. Of course the result is interesting only if P(A/notB) is low.

The problem in pattern discovery is to be sure that patterns are real and useful : the probability of finding any given pattern increases with the size of the database. It is thus necessary to develop validation rules, and in this respect processing the whole database is not the best thing to do : sampling (or splitting the base into several sets) will be better than exhaustive processing. One should test if a model or a pattern remains valid in an other par of the base than the one which has been explored. Usefulness should also be tested : association are only

correlations, and not causation and it is not sure that promoting B will mechanically increase the purchase of A…

3.New mines : texts, web, symbolic data…

Besides classical data-base where data are usually presented in form of a rectangular array, new kinds of data are now present :
Symbolic data, such as fuzzy data or intervals : the data is not known precisely but belongs to an interval, with or without a probability distribution. Eurostat has financed the SODAS project (symbolic data analysis for official statistics), a consortium of 17 European research teams, and a software is available which may be used for Data Mining purposes, see Hebrail (1997) and Bock and al. (2000).

Text mining : most of the information which circulates is now digitalized as wordprocessor documents, and powerful techniques are currently available for a wide variety of applications :for example, one can categorise information from news-agencies, analyse patent portfolios, customer complaints, classify incoming e-mail according to predefined topics, group related documents based on their content, without requiring predefined classes, or assign documents to one or more user-defined categories.

Web mining is a new concept for statistical analysis of website information (text mining) or frequentation, as well as of the behaviour of the websurfers.

4.Applications in official statistics

Of course in National Statistical Institutes there has always be some use of Exploratory Data Analysis, or of model choice algorithms, but as far as I know it seems that there are few, if not none, known applications of data mining techniques in the meaning of trying to discover new models or patterns in their databases by using the new tools described before.
It is not surprising because the main task of NSI's is data production, and analysis is often done by different institutes. Furthermore it seems that the idea of exploring a database with the objective of finding unexpected patterns or models is not familiar to official statisticians who have to answer precise questions and make forecasts. Statistical analysis are done generally if they can be repeated in a production framework. But as far as NSI's manage large data bases on population, trade, agriculture, companies, there is certainly great potentialities in exploiting their mines of data.

Let us point out a few fields where data mining tools could be useful :
- business statistics with special mentions for innovation policy, financial health
- household equipments and savings
- health statistics, more precisely mortality and morbidity, in order to detect unexpected risk factors
- analysis of metadata information by means of text mining

The question remains of who will analyse these databases, the NSI's or other researchers ? Data Mining techniques use individual records ; since confidentiality issues are crucial, it seems that NSI's are the best place, but it implies new directions of research, and likely new recruitments.

Let us mention that Eurostat is funding research projects on Data Mining and Knowledge Extraction, specially designed for the NSI's such as SODAS, KESO (Knowledge Extraction for Statistical Offices), SPIN! (Spatial Mining for Data of Public Interest). All these projects aim to provide specific software and it is very likely that within three years, specific Data Mining tools will be made available to official statisticians.

Conclusion

Data Mining is a growing discipline which originated outside statistics in the data base management community, mainly for commercial concerns. The gap is now filled and Data Mining could be considered as the branch of exploratory statistics where one tries to find unexpected and useful models and patterns, through an extensive use of classic and new algorithms. The expression « unexpected » should not be misleading : one has greater chances to discover something interesting when one is previously acquainted with the data.
Caution is of course necessary to avoid drawing wrong conclusions and statisticians, because they are familiar with uncertainty and risk, are the right people to derive validation tests.
A completely automatic process of knowledge extraction is also a misleading idea : even with very efficient software, human expertise and interventions are necessary.
Official statistics should be a field for Data Mining, giving a new life and profitability to its huge databases, but it may implies a redefinition of the missions of NSI's.

References :

J.H.Friedman (1997) Data mining and statistics : what's the connection ? *http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps*

H.H.Bock, E.Diday (eds) (2000) *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*, Springer-Verlag

D.J.Hand (1998a) Data mining: statistics and more ?, *The American Statistician*, **52**, 112-118

D.J.Hand (1998b) Data mining-reaching beyond statistics, *Research in Official Statistics*, **2**, 5-17

D.J.Hand (2000) Methodological issues in data mining, *Compstat 2000 : Proceedings in Computational Statistics*, ed. J.G.Bethlehem and P.G.M. van der Heijden, Physica-Verlag, 77-85

G.Hebrail (1997) Statistical Data Analysis, Machine Learning, Data Mining, or Symbolic Data Analysis? , *ASMDA 1997* :*VIIIth International Symposium on Applied Stochastic Models and Data Analysis*, Invited&Specialised Sessions Papers, 145-154.

J.W.P.F. Kardaun, T.Alanko (1998) Exploratory Data Analysis and Data Mining in the setting of National Statistical Institutes, *proceedings of NTTS98 conference*, *http://europa.eu.int/en/comm/eurostat/research/conferences/ntts-98*

G.Saporta (1998) The Unexploited Mines of Academic and Official Statistics, in *Academic and Official Statistics Co-operation*, Eurostat, 11-15