International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 5 , September 2017,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

QUANTITATIVE DATA MINING: STATIESTICS TECHNIQUES & OPERATIONS RESEARCH APPROACHES

Bhanudas Suresh Panchabhai, R.C.Patel ACS College, Shirpur

ABSTRACT: This research paper reviews the main applications of statistics and operations research techniques to the quantitative and aspects of Knowledge Discovery (KD) and Data Mining (DM), fulfilling a demanding need. Data Mining, one of the most important phases of the Knowledge Discovery in Databases activity, is becoming ubiquitous with the current information explosion. As a result, there is an increasing need for training professionals to work as analysts or to interface with these. On the other hand, such professionals already exist. Statisticians and operations researchers combine three skills widely used in Data Mining: computer applications, systems optimization and data analysis techniques. This review alerts them about the challenging opportunities that, with little extra training, await them in Data Mining.

Key words: Data Mining, data, data quality, applied statistics, data analysis

INTRODUCATION

At the beginning there was data – or at least there was an effort to collect it. But data collection was a very expensive activity in time and resources. The advent of computers and the Internet made this activity much cheaper and easier to undertake. Organization, always aware of the practical value of databases and of extracting information from them, was finally able to start collecting and using data. Data has become so plentiful that corporations have created data warehouses to store them and have hired statisticians to analyze their information content.

However, the traditional and manual procedures to find, extract and analyze information are no longer sufficient. Fortunately incoming data is now available in computerized format which provided a unique opportunity to mass-process data sets of hundreds of variables with millions of cases, in a way it was not possible before. In addition, analyses approaches are also different.

For, now the problem's research hypotheses are no longer clear and sometimes not even known. Establishing the problem's research hypotheses is now an intrinsic part of the data analysis itself.

This situation has encouraged the development of new tools and paradigms. The result is what we now know as Data Mining (DM) and Knowledge Discovery in Databases (KDD). However, there are many discussions about what DM and KDD activities really are, and what they are not.

On one hand, Bradley et al.statedthat "KDD refers to the overall process of discovering useful knowledge form data, while data mining refers t0 a particular step in this process. Data Mining is the application of specific algorithms for extracting structure from data. The additional steps in the KDD process include data preparation, selection, cleaning, and incorporationof appropriate prior knowledge". On the other, Balasubramanian et al. state: " Data Mining is the process ofdiscovering meaningful new correlation patterns and trends by sifting through vast amounts of data stored in repositories(….) using pattern recognition statistical and mathematical techniques.

DataMining is an interdisciplinary field with its roots in statistics, machine learning, pattern recognition, databases and visualization." Finally, some in the IT community state that Data Mining goes beyond merely quantitative analysis, including other qualitative and complex relations in data base structures such as identifying and extracting information from different data sources, including the Internet.

We will use the first of the above three definitions and limit our discussions to the quantitative aspects of Data Mining. Hence, in this paper DM will concentrate in the quantitative, statistical and algorithmic data analysis part of the more complex KDD activity.

The large divergence in opinions about what Data is or is not, has also brought up other discussion topics. Balasubramanian proposes the following questions: (i) Query against a large data warehouse or against a number of database? (ii) In a massively parallel environment? (iii)Advanced information retrieval through intelligent agents? (iv) Online analytical processing (OLAP)? (v) Multidimensional Database Analysis (MDA)? (vi) Exploratory Data Analysis or Advanced graphical visualization? (vi) Statistical processing against a data warehouse? The above considerations only show how Data mining is a multi-phased activity, characterized by the handling of huge masses of data. The quantitative data analysis is undertaken via statistics, mathematical and other algorithmic methods, without previously establishing research hypotheses. In fact, one defining data mining characteristic is that research hypotheses and relationships between data variables are obtained as a result of (instead of as a condition for) the analyses activities. Hereon, we will refer to this entire multiphase activity as DM/KDD. The information contained (or of interest) in a Database may not necessarily by quantitative. For, we may be interested in finding, counting, grouping or establishing say a relationship between entries of a given type (e.g. titles, phrases, names) as well as in listing their corresponding sources. The latter (qualitative) analysis is another very valid from of DM/KDD and requires a somewhat different treatment, but this is not the main objective of the present paper. From all the above, we conclude that overall, DM/KDD is a fast growing activity in dire need of good people and that particularly well prepared to undertake quantitative DM/KDD work.

The main objective of this paper is to provide a targeted review for professionals in statistics and operations research. Such document will help them to better understand its goals, applications and implication, facilitating a swifter and easier transition to quantitative DM/KDD. For, statisticians and operations researchers combine three skills widely used in Data Mining: computer applications, systems optimization and data analysis techniques. This paper alerts them about the challenging opportunities that, with little extra training await them in Data Mining. In addition, it provides other Data Mining professionals from different backgrounds, a clearer view of the capabilities that statisticians and operations researchers bring to the DM/KDD arena. We will first examine the quantitative DM/KDD process as a sequence of five phases. In the first two phases i.e. data preparation and data mining, we discuss some problems of data definitions and of the applications of several statistical, mathematical, artificial intelligence and genetic algorithm approaches to data analyses. Finally, we overview some computer and other considerations and provide a short list of references.

2. Phases in a DM/KDD study

Accordingly, there are five phases in a quantitative DM/KDD study, which are not very different from those of any comprehensive software engineering or operations research project. They are: (i) Determination of objectives, (ii) Preparation of the data, (iii) Mining the data, (iv) Analysis of results and (v) Assimilation of the knowledge extracted,

(I) Determination of Objectives

Having a clear problem statement strengthens any research study. Establishing such statement constitutes the “determination of objectives” Phase. We thoroughly review the basic information with our client, re-stating goals and objectives in a technical context, to avoid ambiguity and confusion. We select, gather and review ‘he necessary background literature and information, including contextual and subject matter expert opinion on data, problem, component definitions, etc. With all this information we prepare a detailed project plan with deadlines, milestones, reviews and deliverables, including project staffing, costing, management plan, etc. Finally, and most important we obtain a formal agreement from our client about all these particulars.

(II) Preparation of the Data

Many practitioners agree that data preparation is the most time-consuming of these five phases. A figure of up to 60% of total project time has been suggested. Balasubramanian et al. divide the data preparation phase into three subtasks that we will discuss here, too.

Selection of the Data is a complex subtask in itself. It first includes defining the variables that provide the information and identifying the right data sources. Then, we need to understand and define each component data element such as data types, possible values, formats, etc. Finally, we need to retrieve the data warehouse or the Wed. Internet searches, frequent in qualitative DM/DKK applications, may produce a large number of matches, many of which are irrelevant to the query. In such context, information storage and retrieval issues need to be considered very carefully. Another related information management issue is the role of data model in the management of knowledge (KM) which could be defined as aggregating data with context, for a specific purpose. Hence, the importance of analyzing database design and usage issues, as part of the Preparation of the Data phase.

Web forecasting has two main components: the Internet and the user. Establishing indicators (variables) that accurately characterize and relate these two entities is not simple. There are many variables that measure Internet Web page usage which include: (i) Hits, page requests, page views, download; (ii) Dial ups, permanent connections, unique visitors; (iii) Internet subscribers, domain names, permanent connections; (iv) Web site (internal) site (internal) movements (e.g. Pages visited) and (v) Traffic capacity, speed, rate, bandwidth. Such information can be captured by special programs, from four types of web

Logs: (i) access logs (which include dates, times and IP addresses); (ii) agent logs (which include browser information); (iii) error logs (which include abort downloads) and (iv) referrer log. These include information about where the users come from, what previous Web site have they visited and where will they go to, next. Most of these measures present serious definition problems. For example a hit, recorded in the site's Log file, is loosely defined as "the action of a site's Web server passing information to an end user" a page. So, when is then a "hit", a valid "visit" to a Web site? And, if not all hits are valid visits how can we distinguish between different types of hits and count them differently?

Page requests, page views, downloads, etc. pose analogues definition problems as the ones outlined above. The real objective here is counting the number of "visitors" behind these hits, or downloads, etc. For, their count provides the basic units for a model that forecasts Web usage.

On the other hand, we also need to gather information about the user and about their use of the Internet sites. For characterizing and counting the Internet user base we need demographic data, frequently gathered via user surveys and on-line data collection. These are very different data sources: automatically collected Internet data, user survey data, Census data, etc. We must validate, coordinate and put coherently together their respective information.

The data pre-processing task includes ensuring the quality of the selected data. In addition to statistical and visualizing quality control techniques, we need to0 perform extensive background checks regarding data sources, their collection procedures, the measurements used, verification methods, etc. An i9n-depth discussion about data, its quality and other related statistical issues (specifically on materials data, but valid to data collection in general) can be found in (5). Data quality can also be assessed through pie charts, plots, histograms, frequency distributions and other graphical methods. In addition, we can use statistics to compare data values with known population parameters. For example, correlations can be established between well-studied data variables (e.g. height and weight) and used to validate the quality of the data collected.

A data transformation subtask may also be necessary if different data come in units incompatible with each other (e.g. Meters and inches). Data may be given in unusable format (e.g. mm/yy, male/female, etc.) that must be first converted to values handled by statistical software. Data may be missing or blurred and need to be estimated or recovered. Or, simply for statistical modeling modeling reasons the data needs to be transformed.

The above discussion stresses the importance of implementing an exhaustive and carefully planned data quality control and assurance phase, and assurance phase, prior to the data analysis one. They also underline the need for improving the data collection procedures. All these time-consuming activities have a crucial importance in the DM/KDD process to avoid the GIGO model problems.

(III) Mining the Data

There are many approaches and algorithms for mining data, according to whether our analysis objective is to study, classify, classify or predict. We study the data to uncover patterns and associations that may help us better understand it. We can then use those patterns to classify the data into similar groups or to forecast future system behavior. Three recent references in this area are Balasubramanian et al. , the Special Issue on Data Mining of the INFORMS Journal on Compute ring and the new on-line technical journal Data Mining and Knowledge Discovery. Following Bradley et al. We divide the data mining techniques into five classes, according to the objectives that the analysis pursues. These classes are:

1. Predictive modeling, if the goal is to Fra recast one variable/attribute based on others. For example, one may predict sales of an article based on; customer sex, income and age brackets, geographical region, season of the year, etc. Regression analysis is widely used for these problems, but is neither the only nor necessarily always the best approach.
2. Clustering or segmentation, if the goal is to group similar data items into subsets. For example, one may want to group elements by variables such as sex, race, age, income, etc. Or one may want to summarize several such demographic variables into two or three conceptual ones that provide the same information. Cluster and factor analyses are two statistical methods frequently employed in these types of problems. On the other hand, groups or clusters may be pre-specified. In such case, specific group characteristics (variables) may be sought instead and discriminant analysis becomes an appropriate method.
3. Dependency modeling, if the goal is to model the joint probability density function of two or more variables that have a causal or deterministic relationship. Multivariate density function estimation methods are used to solve problems found through this approach.
4. Data summarization, if it finding interesting summaries of parts of the data. Here the goal of the analysis is to find patterns that describe data subsets (summaries). Statistical exploratory data analysis (EDA) methods prove useful in such cases.
5. Change and deviation detection, if we want to find behavior differing from the norm. The analysis objective may be to detect patterns (or to detect individuals) that differ from those of the majority in the data/population. Such differences may exist in the ordering or sequencing of events, as well as in certain specific aspects (e.g. component variables).

For simplicity, we regroup and overview data mining approaches in three categories: Mathematically based, statically based and “Mixed” algorithms, because in most of the five categories above, we may use more than one method, either from this vast tool set. And in this paper, we seek to emphasize the main analytical methods and their applications.

Hence, we will call mathematically based algorithms those that rely heavily in a deterministic approach. Among them we include mathematical programming like linear, non-linear, integer and network methods and memory-based reasoning approaches. We will call statistically based algorithms those that rely heavily in a stochastic approach. Among them we include regression, discrimination, time series and factor analyses. Finally, we will call “mixed” approaches, those that borrow heavily from both, the algorithmic and the stochastic components. Among them we include decision trees, neural networks, clustering and genetic algorithms.

Mathematical By Based Algorithms

Mathematical (linear, nonlinear. Integer) programming: are three, powerful and well known. Operations research optimization tools. Recently, they have been reintroduced for DM/KDD analysis, as can be read in the research paper by Bradley et al. the basic tenant of such approaches consists in defining an objective function to optimize (maximize/minimize) subject to a set of constraints.

Network analyses; includes the two approaches known as ling and affinity analyses. In link analysis, We reconstruct patterns of behavior (sequences) of the system entities and then search for similarities among them. The data is represented as a network, where the nodes are the pieces of information and the links provide the order (sequence)in which they appear. Graphical results facilitate the analyses by subject matter experts.in affinity analysis (subject of the former) sequences of patterns of actions are associated in some special form. The "market basket" problem is classical example. Customers usually buy associated products. The analysis approach seeks to identify and exploit such linked patterns by analyzing the vendor's transaction database.

Memory-based reasoning: also known as "nearest neighbor classifiers" obtains some type of distance between elements in the database then; the algorithm uses this information to form subset of "elements (those separated by a small distance). Initially. Sets or subgroups (training set) are used to establish the corresponding distances. But there are several problems, which are as follows.

First, we need to define the specific metric we will use, to measure the distances between similar data fields (components). Then, we need to define how to combine the different metrics used for each field, into a single distance between two data elements (vectors). Finally, we need to define the number of data elements with which we will compare each new database element. This is the way in which we classify every new element (vectors). Finally. We need to define the number of data elements with which we will compare each new database element. This is the way in which we classify every new element into one of the existing subgroups. Analysis results are thus, highly dependent on the three above –mentioned subjective considerations, as well as on the specific data sample selected for the initial training set. There last two considerations are not unique to this last algorithm, but rather applicable to all of them.

Mixed Algorithms

Neural networks: is a modeling approach that mimics the way the brain is configured and function. Neural nets are composed of simple processing units (cells) structured into hierarchical networks (trees) linked together by paths. Each unit performs a simple task, consisting of three steps: receiving N inputs (submitted by units from the previous level), evaluating them and providing one output (which is sent up to the next level). The base units, instead of receiving from other units, receive their input directly from an outside medium (model input). The top unit provides the model output. Networks differ widely. There are many neural network algorithms, dealing with the evaluation of input/output processes. In addition, network topologies, the number and size of layers and types of connections also differ.

Finally, training/learning strategies define how networks. This occurs by exercising them through sets of values, who are either similar or whose output is known and by adjusting the network node outputs to fit the desired results. Such training activities produce the node weights for the neural net model to work. After it is tuned up, a neural net can be successfully used to predict or classify. They served the same model objectives regression and discrimination do, without having to rely on their stringent statistical assumptions. The down side is the need for reliable data to train the neural net, as well as the model's dependency on the many subjective decisions dealing with the network topology.

Genetic algorithms: is another modeling approach that mimics human behavior by simulating the way biological genetics operates in human evolution (survival of the fittest). An initial set (vector) of inputs is successively recombined via several genetic operations (combination, mutation, crossover, etc.). Genetic operations follow strict probabilistic rules, provided by the model

builder, yielding the offspring of the following generation. The model assesses each generation result via a fitness function, also defined by the model builder. The process terminates after several generations and a solution is obtained via a convergence process.
A pre-defined stopping rule tells the genetic algorithm when to terminate. Since initial conditions vary, so may the final genetic algorithm solution. Hence, convergence to a unique or optimum value is not guaranteed. These algorithms can be combined with neural nets.

Decision tree: in a way, is a sort of mirror image of neural nets. For here, we start with a single input (instead of several, as we do in nets) and work our way down a hierarchical network (tree) to produce one of several possible model outputs. We select each path by following a set of probabilistic rules that direct our choices at every layer of the network topology. As in neural nets, a training set of data is partitioned into similar subgroups and used to establish the probabilistic rules. Decision probabilities at each step are established by comparing the resulting values with the ones expected from the training set. We adjust the model selection probabilities, until the expected agree with those of the training set. We use decision tree models to forecast or classify an observation into one of several final possible pre-established states.

Clustering methods: are algorithms for dividing the population of similar entities. A sample is drawn from the database and used to define subgroups. The determination of the number and components of subgroups is part of the problem. The definition of the metric, through which distances to assess whether two elements are similar or not, is also part of the problem statement. In an elementary context, it is feasible to extract a data sample, recognize the subgroups and separate the data into these subgroups. In large databases, with millions of record, clustering procedures are more complex. Hence, more sophisticated algorithms that use probabilistic methods (such as hierarchical clustering schemes) may have to be used. Through them, we select the number of clusters in the populations, as well as their component elements.
Since the key issue in DM/KDD is the absence of pre-established hypotheses we cannot assume, a-priori, the characteristics of clusters nor of their members. These cluster results should come out of the DM/KDD exercise. The clustering algorithms are closely related to factor analysis techniques.

Statistically Based Algorithms

Regression: is a powerful and widely used statistical procedure that allows us to model a continuous attribute (or dependent variable) as a function of several correlated continuous and discrete attributes (or independent variables). When they are implemented in conjunction with a variable selection procedure, regression models can reduce an initially large of group of explanatory or independent variables down to a smaller subset. This, in turn, is composed of those most significant variables that better explain the relationship. Regression is one of the most effective methods in data analysis because it helps explain or forecast one attribute (say, purchases) as a function of others (say, title, author, price, publisher, etc.).
Discrimination: is applicable when the database can be partitioned a priori into a pre-specified number of subgroups. In this case, a regression-like modeling approach is applied to a set of pre-established attributes (variables). The procedure yields (if it exists) the subset of these attributes that better help separate/classify database elements in their corresponding subgroups. Discrimination produces three important results: establishes the existence of different (pre-specified) subgroups, identifies which variables, among those analyzed, better differentiate items among subgroups and provided an equation for the classification of future observations.
Time series: is applicable when the data comes in a time sequence and helps detect whether (and how) here is a modifying time effect. This is a black-box type of approach, since the real modifying causes are never identified. But, when the causes are similarly affected by time, then the approach is valid. For example, as time increases more people acquire computer systems, become aware of the Internet and start surfing it. Hence, number of web page hits also increases.

Factor analysis: is applicable which each record or transaction vector is composed of multiple attributers (variables) that are associated among them and we are interested in detecting such variable associations. Factor analysis will find an equivalent set of abstract factors, combinations of the original (real) attributes that describe the same problem variability. Such equivalent set will be composed of uncorrelated variables, which can be sorted in decreasing importance of their variance. Removal of the less variable among them simplifies the model for, the less components implies smaller dimension. One advantage of a smaller model is its visualization in a two or three –dimensional space. However, there is a cost: the reduced model explains less problem variability. Factor analysis helps reduce problem complexity and groups records into clusters of similar behavior. Analyzing these, we may extract interesting patterns and other useful information.

(IV) Analysis of Results

In a way, quantitative DM/KDD can be thought of as enhanced form of EDA or exploratory data analysis. However, and due to their huge dimensions. DM/KDD problems are interactive processes. Once each iteration is over, its results are used in one of two ways: as the input for a new integration (if necessary) or as a stage- final (i.e. Final for the current stage). That is, until new data comes in and a re-analysis is required.

DM/KDD is team effort. Hence analysts, domain experts and all other interested team members work together and jointly interpret the analysis results. The team approach provides one of the main reasons for the iterative nature of DM/KDD and one of its strongest features. For example, some team members may not be satisfied with the analysis results and would like to enhance them, or to experiment with other variants of the problem, based on such results. This situation re-initiates (and enriches) the whole DM/KDD analysis process.

(V) Assimilation of Knowledge Extracted

The main objective of a quantitative DM/KDD exercise is problem solving. For, there is a specific situation at hand and also a belief that, analyzing system data, we can obtain information to help resolve it. The analyses of results stage provides such information, which we now need to convert into adequate courses of action.

Computer/System Considerations: The size of the problem (database) and the nature of the analysis (algorithms) compel the DM/KDD process to develop computer-based approaches. All the theory on methods and algorithms overviewed above will not suffice to attack and resolve the DM/KDD problems. We also need software programs to run them in high-performance (parallel) computers. Quantitative DM/KDD software tools, however, are much more than mere statistical and mathematical analyses programs. They do include these algorithmic procedures as part of their tasks. But they also have other very important functions that, combined with the algorithms, do the complete work. These functions include data management and display capabilities as well as semi-automated statistical and mathematical diagnosis and special selection functions. With these added expert functions we can analyze huge masses of data, detect unspecified patterns or associations, uncover research hypotheses and identify the models we then need to apply.

CONCLUSIONS

This paper seeks to alert the statistics and operations research professionals about how close their background is to DM/KDD and how they can partake in this challenging work. Toward this end, we overviewed the main statistics and O.R. techniques for quantitative DM/KDD activities and discussed the DM/KDD paradigm, from problem statement and data considerations to derivation of the final conclusions and their implementation. We gave examples of their uses, problems and limitations, and provided references for the reader up on these topics. Finally, we have also provided information about the work that statisticians and operation s researchers can do in DM/KDD, for those other team members who come from different professional fields.

REFERENCES

[1]. Bradley, P.S., et al.: Mathematical Programming for Data Mining. INFORMS Journal on computing. 11(3), 1999, 217-238.
[2] Balasubramanian, U.et al.: Data Mining, Critical Review and Technology Assessment Report. IATAC/DTIC. Spring 2000.
[3] Anderson, T.W. An Introduction to Multivariate Statistical Analysis. Wiley, NY.1984.
[4] Zenas, A: Text mining and its application, MIT Press, 2003.
[5] Taha, H. A. Operations Research; an introduction. Prentice Hall. New Jersey. 1997.
[6] Special Issue on Data Mining. INFORMS Journal on Computing. Vol. 11, Number 3. 1999.