Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 132 (2019) 53360-53362

ISSN: 2229-712X

Data Mining – Business Statistics

M. SubramaniaKumar[1] and P. Radha[2]

[1]School of Computer Studies, Rathnavel Subramaniam College Arts and Science.

[2]Department of Computer Science, Shri Krishna Arts and Science College.

ARTICLE INFO

Article history:

Received: 12 March 2019;

Received in revised form:

01 July 2019;

Accepted: 11 July 2019;

Keywords

Databases,

Knowledge Discovery,

Telecommunications.

ABSTRACT

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. The amount of data being generated and stored is growing exponentially, due in large part to the continuing advances in computer technology. From the financial sector to telecommunications operations, companies increasingly rely on analysis of huge amounts of data to compete. A new generation of techniques and tools is emerging to intelligently assist humans in analysing mountains of data. New problems arise, partly as a consequence of the sheer size of the data sets involved, and partly because of issues of pattern matching. However, since statistics provides the intellectual glue underlying the effort, it is important for statisticians to become involved. Our goal here is to provide a brief overview of the key issues in knowledge discovery in an industrial context and outline representative applications.

Introduction

Data mining is a process that takes data as input and outputs knowledge. Data mining, on the other hand, is entirely concerned with secondary data analysis. It is important to point out that there is some ambiguity about the term “data mining”, which is in large part purposeful. This term originally referred to the algorithmic step in the data mining process, which initially was known as the Knowledge Discovery in Databases (KDD) process. Note that predictive models can be descriptive (to the degree they are understandable), and descriptive models can be used for prediction. Examples of key predictive methods include regression and classification (learning a function that maps a new example into one of a set of discrete classes). Key description methods include clustering, summarization, visualization, and change and deviation detection. Methods like dependency modeling (e.g., market basket analysis) can be either.

Data Mining Process

The data mining process is an iterative process, although this is not explicitly.After the initial run of the process is complete, the user will evaluate the results and decide whether further work is necessary or if the results are adequate. For example, additional data records may be acquired, additional fields (i.e.,variables) may be generated from existing information or obtained (via purchase or measurement), manual cleaning of the data may be performed, or new data mining algorithms may be selected.

Data Mining developed as a new discipline for several reasons. First, the amount of data available for mining grew at a tremendous pace as computing technology became widely deployed. Specifically, high speed networks allowed enormous amount of data to be transferred and rapidly decreasing disk costs permitted this data to be stored cost-effectively.

A key thing to note about a realistic knowledge discovery process is that it is not simple and linear, but thoroughly iterative and interactive. The results of analysis are fed back into the modelling and hypothesis derivation process to produce improved results on subsequent iterations. Statistics as a discipline has a poor record for timely recognition of important ideas. A common pattern is that a new idea will be launched by researchers in some other discipline, will attract considerable interest (with its promise often being subjected to excessive media hype.which can sometimes result in a backlash), and only then will statisticians become involved. Clean data is a necessary prerequisite for most statistical analyses. Entire books, not to mention careers, have been created around the issues of outlier detection and missing data. An ideal solution, when questionable data items arise, is to go back and check the source. In the data mining context, however, when the analysis is necessarily secondary, this is impossible.

Knowledge Discovery Applications

Knowledge discovery applications and prototypes have been developed for a variety of domains, including marketing, finance, banking, manufacturing, and telecommunications. A majority of the applications use a predictive modelling approach, although a few notable applications use other methods. In market basket analysis one studies conditional probabilities of purchasing certain goods, given that others are purchased. One can potentially interesting patterns as those which have high conditional probabilities as well as reasonably large marginal probabilities for the conditioning variables. A computer program can identify all such patterns with values over given thresholds and present them for consideration by the client. A key thing to note about a realistic knowledge discovery process is that it is not simple and linear, but thoroughly iterative and interactive. The results of analysis are fed back into the modelling and hypothesis derivation process to produce improved results on subsequent iterations. This activity takes time, and if applied to data generated on a regular basis (e.g., quarterly or yearly results), it can have a long lifespan. Systems like IDEA are beginning to support more of the infrastructure aspects of the process (e.g., it supports sequences of operations), allowing reuse of complex analyses on variant datasets.

Data Mining Resources & Tools

A good electronic resource for staying current in the field is KDnuggets (http://kdnuggets.com/), a website that provides information on data mining in the form of news articles, job postings, publications, courses, and conferences, and a free bimonthly email newsletter.

There are a number of general textbooks on data mining. Those who have some background in computer science and are interested in the technical aspects of data mining, including how the data mining algorithms operate, should consider the texts by Han and Kamber (2006) , Tan, Steinbach and Kumar (2006) and Liu (2007). Those with a business background, or whose primary interest is in how data mining can address business problems, may want to consider the texts by Berry and Linoff (2004) and Pyle (2003).

Application Development

While much data mining technology is well developed, its practical application in industry is affected by a number of issues:

• Insufficient training: Graduates of business schools are familiar with verification-driven analysis techniques, occasionally with predictive modelling but rarely with other discovery techniques.

• Inadequate tool support: Most available data mining tools support only one of the core discovery techniques, typically prediction. Other methods, such as clustering, deviation detection, visualization, and summarization, are also needed, as are methods for dealing with exceptions (rare cases) that may be significant in some applications. The tools must also support the complete knowledge discovery process and provide a user interface suitable for business users rather than for other technologists.

• Data unavailability: For a given business problem, the required data is often distributed across the organization in a variety of formats, and the data is often poorly organized or maintained. For this reason, data acquisition and preprocessing usually play a significant part in any knowledge discovery project.

Data warehousing is becoming widespread and can potentially alleviate such problems. Spurious Relationships and Automated Data Analysis To statisticians, one thing will be immediately apparent from the previous examples. Because the pattern searches will throw up a large numbers of candidate patterns, there will be a high probability that spurious (chance) data configurations will be identified as patterns. How might this be dealt with? There are conventional multiple comparisons approaches in statistics, in which, for example, the overall experiment wise error is controlled, but these were not designed for the sheer numbers of candidate patterns generated by data mining. This is an area from some careful thought. It is possible that a solution will only be found by stepping outside the conventional probabilistic statistical framework. Possibly using scoring rules instead of probabilistic interpretations. The problem is similar to that of overstating of statistical models, an issue which has attracted renewed interest with the development of extremely flexible models such as neural networks.

Several distinct but related strategies have been developed for easing the problem, and it may be possible to develop analogous strategies for data mining.

Graphical Methods

Graphical methods**,** e**s**pecially dynamic and interactive graphical methods, also have a key role to play here. Such tools allow one to take advantage of the particular power of the human eye and mind at digesting very complex information. The dynamic graphical display known as the World Tour projecting multivariate data down into a two dimensional projection and letting the direction of projection vary.is an example of this. At present such methods have their limitations. Sitting watching such a display for any length of time can be a mind-numbing experience.

However, here again the computer can come to our aid. We can dense measures of interestingness for a scatterplot and let the machine apply these measures as it produces the projections. We are back at projection pursuit. This, of course, requires us to articulate and dense beforehand what we consider .interesting.. But we can go further. We can present the machine with a series of projections, telling it which ones we _nd interesting and which we do not, and (provided we have given it a basic alphabet of structures) we can let it learn appropriate internal representations of .interesting for itself.

The Potential for KDD Applications

Domains suitable for knowledge discovery are information- rich, have a changing environment, do not already have existing models, require knowledge based decisions, and provide high payoff for the right decisions. Given a suitable domain, the costs and benefits of a potential application are affected by the following factors:

• Alternatives: There should be no simpler alternative solutions.

• Relevance: The key relevant factors need to be present in the data.

• Volume: There should be a sufficient number of cases (several thousand at least). On the other hand, extremely large databases may be a problem when the results are needed quickly.

• Complexity: The more variables (fields) there are, the more complex the application. Complexity is also increased for time-series data.

• Quality: Error rates should be relatively low.

• Accessibility: Data should be easily accessible; accessing data or merging data from different sources increases the cost of an application.

• Change: Although dealing with change is more difficult than not dealing with change, it can be more rewarding, since the application can be automatically and regularly retrained on up-to-date data.

• Expertise: The more expertise available, the easier the project. It should be emphasized that expertise on the form and meaning of the data is as important as knowledge of problem-solving in the domain.

Web Usage Mining

The content and structure of the Web provide significant opportunity for web mining, as described above. Usage of the Web also provides tremendous information as to the quality, interestingness, and effectiveness of web content, and insights into the interests of users and their habits.

By mining clickstream data and other data generated by users as they interact with resources on one or more web sites, behavioral patterns can be discovered and analyzed.

Discovered patterns include collections of frequent queries or pages visited by users with common interests.

Conclusion

Some authors (e.g., Fayyad 1997) see data mining as a .single step in a larger process that we call the KDD The American Statistician, May 1998 Vol. 52, No. 2 117 process.. .KDD. here stands for Knowledge Discovery in Databases. Other steps in this process include data warehousing; target data selection, cleaning, preprocessing; transformation and reduction, data mining, model selection (or combination); evaluation and interpretation, consolidation and use of the extracted knowledge.

Given the commercial interest in data mining, it is hardly surprising that a number of software tools have appeared on the market. Some are general tools, similar to powerful statistical data exploration systems, while others essentially seek to put the capacity for extracting knowledge from data in the hands of the domain expert rather than a professional data analyst.

All “knowledge workers” in our information society, particularly those who need to make informed decisions based on data, should have at least a basic familiarity with data mining. This chapter provides this familiarity by describing what data mining is, its capabilities, and the types of problems that it can address.

References

1.November 1996/Vol. 39, No. 11 COMMUNICATIONS OF THE ACM. 42-48

2.Gary M. Weiss, Ph.D., Brian D. Davison, Ph.D., Data Mining, To appear in the Handbook of Technology Management, H. Bidgoli (Ed.), John Wiley and Sons, 2010. 1-17

3.David J. HAND, Data Mining: Statistics and More? The American Statistician, May 1998 Vol. 52, No. 2. 112-118

4.Babcock, C. (1994), “Parallel Processing Mines Retail Data, Computer World, 6.

5.Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., and Ramanujam, K. (1997), “Advanced Scout: Data Mining and Knowledge Discovery in NBA Data,” Data Mining and Knowledge Discovery, 1, 121-125.

6.Box, G., and Hunter, W. (1965), “The Experimental Study of Physical Mechanisms,” Technometrics, 7, 57–71.

7.Copas, J.B., and Li, H.G. (1997), Inference for Nonrandom Samples(with discussion), Journal of the Royal Statistical Society, Series B, 59,55,–95.

8.Cortes, C., and Pregibon, D. (1997), Mega-monitoring, unpublished paper presented at the University of Washington/Microsoft Summer ResearchInstitute on Data Mining, July 6–11, 1997.

9.Cox, D.R. (1990), “Role of Models in Statistical Analysis,” Statistical Science, 5, 169-174.

10.Fayyad, U. (1997), “Editorial,” Data Mining and Knowledge Discovery, 1,5-10.

11.Matheus, C., Piatetsky-Shapiro, G., and McNeill, D. Selecting and reporting what is interesting: The KEFIR application tohealthcare data. In Advances in Knowledge Discovery and DataMining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy, Eds. AAAI Press/The MIT Press, Cambridge,Mass., 1996, pp. 495–516.

12.Selfridge, P.G., Srivastava, D., and Wilson, L.O. IDEA: Interactive Data Exploration and Analysis. In Proceedings of SIGMOD-96 (Montréal, June 1996). ACM Press, New York, 1996, pp.24–34.

13.Senator, T.E., Goldberg, H.G., Wooten, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P., andWong, R.W.H. The Financial Crimes Enforcement Network AI System (FAIS): Identifying potential money laundering fromreports of large cash transactions. AI Mag. 16, 4 (Winter 1995),21–39.