RESEARCH PAPER
Computer Science
Volume : 5 | Issue : 9 | September 2015 | ISSN - 2249-555X

Data Mining, Extensive Use in Statistics

KEYWORDS

Dr.Kishor H Atkotiya
Head, Dept. of Computer Science, J.H. Bhalodia Women's College, Rajkot

ABSTRACT *The emerging field of "data mining" is (according to Pregibon (1996)) a blend of statistics, artificial intelligence, and database research. It is projected to be a multibillion dollar industry by the year 2000. I propose the name "statistical methods mining" for the process of developing (and marketing) various maps of the world of statistical knowledge in order to apply to data mining the enormous virtual encyclopedia of statistical knowledge that exists in the literature. I propose that planning modern core, applied, and computational statistical activities should be based on statistical history whose mottos are \the purpose of statistical computing is insight not numbers" and \the purpose of statistical history is insight (about future influences) not details (of past influences)". The goal of statistical methods mining is to promote the optimum practice of analysis of diverse (possibly massive) data sets by enabling researchers in all fields to be made aware that statisticians have important roles to play to provide expertise about classical and modern statistical methods, theory, practice, applications, and computational techniques for data mining and statistical inference (extraction of information and identification of models from data). Contents of this paper are: 1. Future of statistics is bright, future of statisticians needs planning; 2. W1hat went wrong in the recent history of statistical science; 3. Balancing core research and interdisciplinary research; 4. Statistical education and statistical modeling are analogous iterative cycles; 5. Statistical history was never neglected in past thinking about the future of statistics.*

1. Future of statistics is bright, future of statisticians needs planning

Why should the future of statistics, statisticians, statistical methods, and statistical science deserve to be discussed at an Interface Symposium whose themes are data mining and the analysis of massive data sets? I believe that data mining requires a diversity of scientific, computing, statistical, and numerical analytical skills to provide visual (graphical) diagnostic presentations that can be used to make decisions and discover relations between variables. In order to accurately distill information, we need to apply and integrate the extensive range of classical and modern statistical methods. We propose that researchers in data mining and massive data set analysis can benefit from *statistical methods mining* which I define as providing a vision of the diversity of statistical methods which enables one to defer learning their details until they are used in applications. A brilliant example of such a paper is Elder and Pregibon (1996).

2. What went wrong in the recent history of Statistical Science

In the 21st century every individual, every organization, every discipline is expected to prepare formal plans for the future which justify their economic (not just their scholarly) relevance in a world of downsizing and outsourcing. To plan for the future of statistics and statisticians we propose the use of "statistical methods mining" based on "statistical historical thinking", defined as the history of statistical applications, methods, theory, computations, and science, studied as guides to planning future influences rather than study of past influences.

We can identify the outstanding fact about the history of American statistics (according to ads for the American Statistical Association 30 minute documentary video "Statistical Science: 150 Years of Progress"); it is the explosion of statistical science that occurred around World War II. We propose planning that the *next outstanding historical event will be the explosion of statistical methods mining* (around the year 2000).

3. Balancing core research and interdisciplinary research

I define core statistical research to be about mathematically synthesizing ideas drawn from many analogous applications. The goal is to create general statistical methods that provide technology transfer between statistical innovations arising in different disciplines that apply statistical methods. Education in core statistics is needed to teach methods that are elegant and applicable, and help statisticians promote the case for their expertise to be part of teams in the broad range of applied statistical practice.

Statistical historical thinking teaches us that every core statistical method began with real practical problems which stimulated theoretical problems (discipline generated problems) for theoretical study. Can statisticians plan their futures if the history of the stimulus of applied research to core research is not an explicit part of our research, education, service, public relations? We need to continuously mentor "whole statisticians" who achieve in their careers a balance between applications (applied research), theory (abstract core research), and computing. We need to award prizes to continuously increase recognition by scientific and academic publics (and academies of science) of the roles and existence of statisticians as "experts about the discipline and practice of statistics".

In my approach to the study of the history of statistics, which views it as a guide to statistical methods mining and planning modern core, applied, and computational statistical research activities, I propose that two kinds of skill need to be balanced:

1. Technical competence (analogous to internal skills), hard work to accomplish goals and implement decisions;
2. Vision to imagine alternative courses of action (analogous to external skills), inspiration to infer information about where to guide one's technical power.
4. Statistical education and statistical modeling are analo-

gous iterative cycles

Practicing statistical methods mining requires understanding about strategies for solving statistical problems and learning statistical methods. We propose that both require a cycle of steps which one usually repeats (iterates) several times before reaching a satisfactory conclusion.

The cycle of statistical model building (whose motto is \no model is true, only useful") consists of four stages:

Stage 1 (S): Specify very general class of models.

Stage 2 (I): Identify tentative parametric model.

Stage 3 (E): Estimate parameters of tentative models.

Stage 4 (T): Test goodness of _t, diagnose improved models.

The PDCA cycle of statistical problem solving (called in quality circles Shewhart's or Deming's wheel) consists of four stages:

Stage 1 (P): Plan; pose the question, form expectations.

Stage 2 (D): Do; collect the data, make observations.

Stage 3 (C): Check; analyze the data, compare observations and expectation.

Stage 4 (A): Act; interpret the results, find the best theory or decision that find the data.

We think of both cycles as EOCI (Expect, Observe, Compare, and Interpret).

Reformers of mathematics education recommend that teachers should communicate the four aspects of learning which cognitive sciences recommend for success:

1. Simple recall,
2. Algorithmic learning,
3. Conceptual learning, and
4. Problem solving strategies.

In statistical teaching we can make these cognitive concepts more concrete by teaching that statistical concepts (such as the sample mean or sample variance) have three aspects:

1. How to define it (mean of sample distribution);
2. How to compute it (average the sample quintile function (values arranged in increasing order));
3. How to interpret it (estimate location parameter of sample).

The fourth aspect of statistical learning consists of ideas about combining concepts to conduct an iterative statistical investigation whose output is data models, which can be applied to simulate more data with the same distribution as the originally observed data.

5. Statistical history was never neglected in past thinking about the future of statistics

This paper is about the future of statistics in the 21st century and the potential of statistical methods mining to stimulate a new explosion of statistical science. I would like to emphasize that we should read proceedings of past conferences on the future of statistics. We give some examples to show that historical thinking was never neglected.

From the Proceedings of the Conference on Directions for Mathematical Statistics, organized by Ghurye (1975) at the University of Alberta in 1974, we quote Mark Kac (p. 6) and Herbert Robbins (p. 116, a provocative essay "Wither mathematical statistics?"). Kac: Johannsen took a large number of beans, weighed them and constructed a histogram; the smooth curve fitted to this histogram was what my teacher introduced to us as the Quetelet curve. That was my first encounter with the normal distribution and the name Quetelet.

REFERENCE Billard, Lynne. (1996) "A Voyage of Discovery," Journal of the American Statistical | Association, Vol. 92, No. 437, Presidential Address. pp. 1{12. | Box, G. E. P. (1980) Comment (Preparing statisticians for careers in industry: report | of ASA Section on Statistical Education committee on training of statisticians for | industry). The American Statistician, 34, 65{80. | Elder, John and Daryl Pregibon. (1996) "A statistical perspective on knowledge | discovery in databases" in Usama M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, | R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, MIT Press, | 83{113. | Fienberg, S. E. (1992) A brief history of statistics in three and one-half chapters: a | review essay. Statistical Science 7, 208{225. | Ghurye, S. G. (1976) Proceedings of the conference on directions for mathematical | statistics. Advances in Applied Probability Supplement, 7, September 1975. | Goodman, Arnold (1975) "American Statistical Association," Encyclopedia of Information | Science and Technology, Volume 1 (Edited by Jack Belzer, Albert Holzman | and Allen Kent), Marcel Dekker. | Goodman, Arnold, (1994) "Interface Insights: From Birth into the Next Century," | Proceedings of Interface '93: 25th Symposium on the Interface of Computing Science | and Statistics, Interface Foundation of North America. | | |