Significant Role of Statistics in Computational Sciences

Rakesh Kumar Singh
Scientist-D
G.B. Pant Institute of
Himalayan Environment &
Development
Kosi-Katarmal, Almora
Uttarakhand, India

Neeraj Tiwari
Professor & Head
Department of Statistics
Kumaun University
SSJ Campus, Almora
Uttarakahnd, India

R.C. Prasad
Scientist-F
G.B. Pant Institute of
Himalayan Environment &
Development
Kosi-Katarmal, Almora
Uttarakhand, India

Abstract: This paper is focused on the issues related to optimizing statistical approaches in the emerging fields of Computer Science and Information Technology. More emphasis has been given on the role of statistical techniques in modern data mining. Statistics is the science of learning from data and of measuring, controlling, and communicating uncertainty. Statistical approaches can play a vital role for providing significance contribution in the field of software engineering, neural network, data mining, bioinformatics and other allied fields. Statistical techniques not only helps make scientific models but it quantifies the reliability, reproducibility and general uncertainty associated with these models. In the current scenario, large amount of data is automatically recorded with computers and managed with the data base management systems (DBMS) for storage and fast retrieval purpose. The practice of examining large pre-existing databases in order to generate new information is known as data mining. Presently, data mining has attracted substantial attention in the research and commercial arena which involves applications of a variety of statistical techniques. Twenty years ago mostly data was collected manually and the data set was in simple form but in present time, there have been considerable changes in the nature of data. Statistical techniques and computer applications can be utilized to obtain maximum information with the fewest possible measurements to reduce the cost of data collection.

Keywords: Statistics, Data Mining, Software Engineering, DBMS, Neural Networks, etc

1. INTRODUCTION

Statistics is a scientific discipline having sophisticated methods for statistical inference, prediction, quantification of uncertainty and experimental design. From ancient to modern times statistics has been fundamental to advances in computer science. The statistics encompasses a wide range of research areas. The future of the World Wide Web (www) will depend on the development of many new statistical ideas and algorithms. The most productive approach is involve with statistics are: computational and mathematical. Modern statistics encompasses the collection, presentation and characterization of information to assist in both data analysis and the decision-making process. Statistical advances made in collaboration with other sciences can address various challenges in the field of science and technology. Computer science uses statistics in many ways to guarantee products available on the market are accurate, reliable, and helpful[1][2].

- ***Statistical Computing:*** The term "statistical computing" to refer to the computational methods that enable statistical methods. Statistical computing includes numerical analysis, database methodology, computer graphics, software engineering and the computer-human interface[1].
- ***Computational Statistics:*** The term "computational statistics" somewhat more broadly to include not only the methods of statistical computing but also modern statistical methods that are computationally intensive. Thus, to some extent, "computational statistics" refers to a large class of modern statistical methods. Computational statistics is grounded in mathematical statistics, statistical computing and applied statistics. Computational statistics is related to the advance of statistical theory and methods through the use of computational methods. Computation in statistics is based on algorithms which originate in numerical mathematics or in computer science. The group of algorithms highly relevant for computational statistics from computer science is machine learning, artificial intelligence (AI), and knowledge discovery in data bases or data mining. These developments have given rise to a new research area on the borderline between statistics and computer science[1].
- ***Computer Science vs. Statistics:*** Statistics and Computer Science are both about data. Massive amounts of data is present around today's World. Statistics lets us summarize and understand it with the use of Computer Science. Statistics also lets data do our work for us[2].

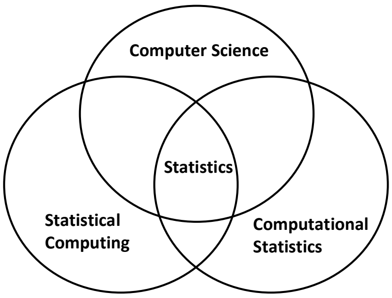

Fig.1. Relation between statistics, computer science, statistical computing and computational statics.

2. STATISTICAL APPROACHES IN COMPUTATIONAL SCIENCES

Statistics is essential to the field of computer science in ensuring effectiveness, efficiency, reliability, and high-quality products for the public. Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility and general uncertainty associated with these discoveries. The following terms are a brief listing of areas in computer science that use statistics to varying degrees at various times[6][7][8]:

- ***Data Mining:*** Data mining is the analysis of information in a database, using tools that look for trends or irregularities in large data sets. In other words "finding useful information from the available data sets using statistical techniques".
- ***Data Compression:*** Data compression is the coding of data using compact formulas, called algorithms, and utilities to save storage space or transmission time.
- ***Speech Recognition:*** Speech recognition is the identification of spoken words by a machine. The spoken words are turned into a sequence of numbers and matched against coded dictionaries.
- ***Vision and Image Analyses:*** Vision and image analyses use statistics to solve contemporary and practical problems in computer vision, image processing, and artificial intelligence.
- ***Human/Computer Interaction:*** Human/Computer interaction uses statistics to design, implement, and evaluate new technologies that are useable, useful, and appealing to a broad cross-section of people.
- ***Network/Traffic Modeling:*** Network/Traffic modeling uses statistics to avoid network congestion while fully exploiting the available bandwidth.
- ***Stochastic Optimization:*** Stochastic optimization uses chance and probability models to develop the most efficient code for finding the solution to a problem.
- ***Stochastic Algorithms:*** Stochastic algorithms follow a detailed sequence of actions to perform or accomplish a task in the face of uncertainty.
- ***Artificial Intelligence:*** Artificial intelligence is concerned with modelling aspects of human thought on computers.
- ***Machine Learning:*** Machine learning is the ability of a machine or system to improve its performance based on previous results.
- ***Capacity Planning:*** Capacity planning determines what equipment and software will be sufficient while providing the most power for the least cost.
- ***Storage and Retrieval:*** Storage and retrieval techniques rely on statistics to ensure computerized data is kept and recovered efficiently and reliably.
- ***Quality Management:*** Quality management uses statistics to analyze the condition of manufactured parts (hardware, software, etc.) using tools and sampling to ensure a minimum level of defects.
- ***Software Engineering:*** Software engineering is a systematic approach to the analysis, design, implementation, and maintenance of computer programs.
- ***Performance Evaluation:*** Performance evaluation is the process of examining a system or system component to determine the extent to which specified properties are present.
- ***Hardware Manufacturing:*** Hardware manufacturing is the creation of the physical material parts of a system, such as the monitor or disk drive.

3. STATISTICS IN SOFTWARE ENGINEERING

Software engineering aims to develop methodologies and procedures to control the whole software development process. Nowadays researchers attempt to bridge the islands of knowledge and experience between statistics and software engineering by enunciating a new interdisciplinary field: *statistical software engineering*. Design of Experiments (DOE) uses statistical techniques to test and construct models of engineering components and systems. Quality control and process control use statistics as a tool to manage conformance to specifications of manufacturing processes and their products. Time and methods engineering uses statistics to study repetitive operations in manufacturing in order to set standards and find optimum (in some sense) manufacturing procedures. Reliability engineering uses statistics to measures the ability of a system to perform for its intended function (and time) and has tools for improving performance. Probabilistic design uses statistics in the use of probability in product and system design. Essential to statistical software engineering, is the role of data: *wherever data are used or can be generated in the software life cycle, statistical methods can be brought to bear for description, estimation, and prediction.* The department of software engineering and statistics trains multiskilled engineers in the processing of information, both in its statistical and computational forms, for use in various business professions.

4. STATISTICS IN HARDWARE MANUFACTURING

The hardware manufacturing companies are applying statistical approaches to create a plan of action that will work more efficiently for forecasting the future productivity of the hardware enterprise[8]. Adopted statistical approaches for:

- Forecasting production, when there is a stable demand and uncertain demand.
- Pinpoint when and which inputs of a specific model will be the cause of uncertainty
- Calculate summary statistics in order to set sample data.
- To make market analysis and process optimizations.
- Statistical tracking and predicting for quality improvement

5. STATISTICS IN DATABASE MANAGEMENT

Databases_are packages designed to create, edit, manipulate and analyze data. To be suitable for a database, the data must consist of records which provide information on individual cases, people, places, features, etc. Optimizer statistics are a collection of data that describe more details about the database and the objects in the database. The optimizer statistics are stored in the data dictionary. They can be viewed using data dictionary views. Because the objects in a database can be constantly changing; statistics must be regularly updated so that they accurately describe these database objects. These statistics are used by the query optimizer to choose the best execution plan for each SQL statement[5]. Optimizer statistics include the following:

- Table Statistics
  - Number of rows
  - Number of blocks
  - Average row length
- Column Statistics

  - Number of distinct values (NDV) in column
  - Number of nulls in column
  - Data distribution (histogram)
- Index Statistics
  - Number of leaf blocks
  - Levels
  - Clustering factor
- System Statistics
  - I/O performance and utilization
  - CPU performance and utilization

Statistical packages for databases are SAS, SPSS, R, etc. and these are available over a wide range of operating systems. Numerous other packages have been developed specifically for the PC DOS environment. S is a commonly available statistical package for UNIX

6. STATISTICS IN ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is the intelligence exhibited by machines or software. Popular AI approaches include statistical methods, computational intelligence, machine learning and traditional symbolic AI. The goals of AI include reasoning, knowledge, planning, learning, natural language processing, perception and the ability to move and manipulate objects. There are a large number of tools used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics, and many others[4]. The simplest AI applications can be divided into two types:

- ***Classifiers:*** Classifiers are functions that use pattern matching to determine a closest match. A classifier can be trained in various ways; there are many statistical and machine learning approaches. The most widely used classifiers is the neural network.
- ***Controllers:*** Controllers do however also classify conditions before inferring actions, and therefore classification forms a central part of many AI systems.

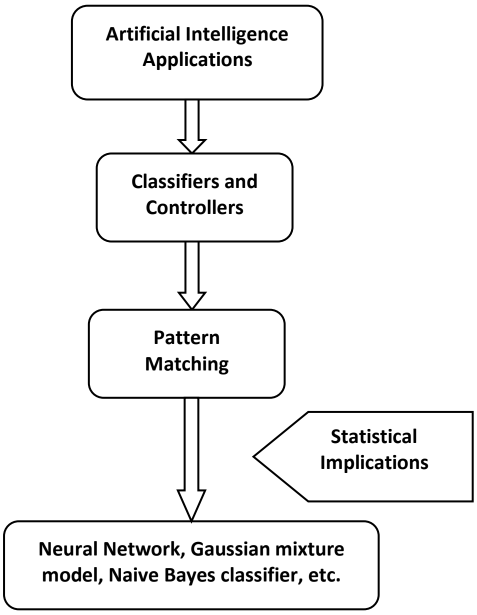
Artificial Intelligence Applications
Classifiers and Controllers
Pattern Matching
Statistical Implications
Neural Network, Gaussian mixture model, Naive Bayes classifier, etc.

Fig.2. Graphical approach of Artificial Intelligence.

7. STATISTICS IN NEURAL NETWORK

Neural network had been used to refer to a network of biological neurons and artificial neural networks used to refer to a network of artificial neurons or nodes. Biological neural networks are made up of real biological neurons that are connected or functionally related in the peripheral nervous system or the central nervous system. Artificial neural networks are made up of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). Artificial neural networks may either be used to gain an understanding of biological neural networks or for solving artificial intelligence problems without necessarily creating a model of a real biological system. Because the inner product is a linear operator in the input space, the Perception can only perfectly classify a set of data for which different classes are linearly separable in the input space, while it often fails completely for non-separable data. While the development of the algorithm initially generated some enthusiasm, partly because of its apparent relation to biological mechanisms, the later discovery of this inadequacy caused such models to be abandoned until the introduction of non-linear models into the field[4].

8. STATISTICS IN BIOINFORMATICS

Bioinformatics is the application of "computational biology" to the management and analysis of biological data. Concepts from computer science, discrete mathematics and statics are being used increasingly to study and describe biological systems. Bioinformatics would not be possible without advances in computer hardware and software: analysis of algorithms, data structures and software engineering. To elaborate algorithms on computers increased the awareness of more recent statistical methods. Statistical analysis for differently expressed genes are best carried out via hypothesis test. More complex data may require analysis via ANOVA or general linear models[8].

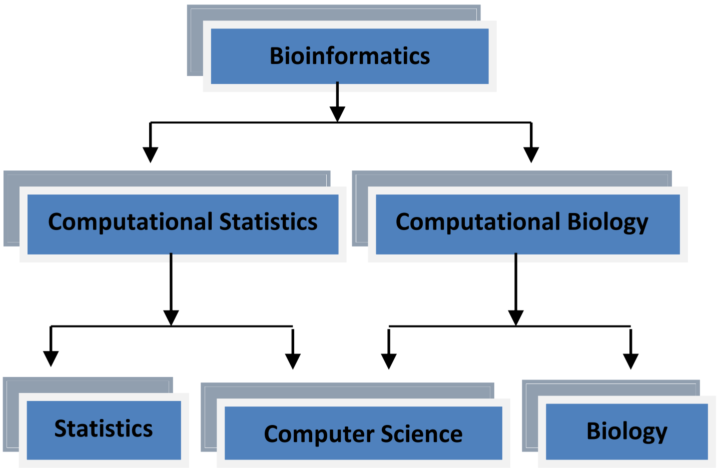

Fig.3. Taxonomy of Bioinformatics.

9. STATISTICS IN DATA MINING

Data Mining is a process of discovering previously unknown and potentially useful hidden pattern in the data. Advances in information technology have resulted in a much more data-based society. Data touch almost every aspect of our lives like commerce on the web, measuring our fitness and safety, doctors treat our illnesses, economic decisions that affect entire nations, etc. Alone, data are not useful for knowledge discovery. Data mining are transitioning from data-poor to data-rich by using the methods like data exploration, statistical inference and understanding of variability and uncertainty[5].

Statistical Elements Present in Data Mining

- Contrived serendipity, creating the conditions for fortuitous discovery**.**
- Exploratory data analysis with large data sets, in which the data are as far as possible allowed to speak for themselves, independently of subject area assumptions and of models which might explain their pattern**.** There is a particular focus on the search for unusual or interesting features**.**
- Specialised problems**:** fraud detection**.**
- The search for specific known patterns**.**
- Standard statistical analysis problems with large data sets**.**

Data Mining from Statistical Perspective

- Data sets which are relatively large and homogeneous might be reasonable to us mainstream statistical techniques on the whole or a very large subset of the data.
- All analyses done by mainstream statistics have intended outcome like set of data to a small amount of readily assimilated information.
- The outcome may include graphs, or summary statistics, or equations that can be used for prediction or a decision tree.
- Large volume of data without loss of information be reduced to a much smaller summary form, this can enormously aid the subsequent analysis task.
- It becomes much easier to make graphical and other checks that give the analyst assurance that predictive models or other analysis outcomes are meaningful and valid

Statistics vs. Data Mining

Feature	Statistics	Data Mining
Type of Problem	Well structured	Unstructured / Semi-structured
Inference Role	Explicit inference plays great role in any analysis	No explicit inference
Objective of the Analysis and Data Collection	First – objective formulation, and then - data collection	Data rarely collected for objective of the analysis/modeling
Size of data set	Data set is small and hopefully homogeneous	Data set is large and data set is heterogeneous
Paradigm/Approach	Theory-based (deductive)	Synergy of theory-based and heuristic-based approaches (inductive)
Type of Analysis	Confirmative	Explorative
Number of variables	Small	Large
Methods/Techniques	- Dependence Methods: Discriminant analysis, Logistic regression - Interdependence Methods: Correlation analysis, Correspondence analysis, Cluster analysis	- Predictive Data Mining: Classification, Regression - Discovery Data Mining: Association Analysis, Sequence Analysis, Clustering

10. PROPERTIES OF STATISTICAL PACKAGES

Statistical packages offer a range of types of statistical analysis[3]. Statistical packages includes:

- Database functions, such as editing, printing reports.
- Capabilities for graphic output, particularly graphs but many also produce maps.
- Common packages are SAS, SPSS, R, etc.
- Available over a wide range of operating systems.
- Some have been "ported" to (rewritten for) the IBM PC.
- Numerous other packages have been developed specifically for the PC DOS environment.
- S is a commonly available statistical package for UNIX

11. CONCLUSION

In this paper, many areas of computer science have been described in which statistics plays a very vital role for data and information management. Statistical thinking fuels the cross-fertilization of ideas between scientific fields (biological, physical, and social sciences), industry, and government. The statistical and algorithmic issues are both important in the context of data mining. Statistics is an essential and valuable component for any data mining exercise. The future success of data mining will depend critically on our ability to integrate techniques for modeling and inference from statistics into the mainstream of data mining practice.

12. REFERENCES

[1] Lauro, C. (1996). Computational Statistics or Statistical Computing, is that the question? *Computational Statistics and Data Analysis*, Vol. 23, pp.191–193.

[2] Billard, L. and Gentle, J.E. (1993). The middle years of the Interface. *Computing Science and Statistics*, Vol. 25, pp.19–26.

[3] Yates, F (1966). Computers: the second revolution in statistics. *Biometrics*, Vol. 22.

[4] Cheng, B. and Titterington, D. M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science*, Vol. 9, pp.2-54.

[5] Elder, J. F. and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. *Advances in Knowledge Discovery and Data Mining*, MIT Press, pp.83-115.

[6] Gentle, J.E. (2004). Courses in statistical computing and computational statistics. *The American Statistician*, Vol. 58, pp.2–5.

[7] Grier, D.A. (1991). Statistics and the introduction of digital computers. *Chance*, Vol. 4(3), pp.30–36.

[8] Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, Vol. 9, pp.123-143.