DOI: 10.18720/SPBPU/2/id20-60

ENGLISH PREFACE

***Xenia Alexandrovna Naidenova**, Cand. of Techn. Sci., Seniour Researcher, Military Medical Academy, Russia, St.Petersburg, Akademika Lebedeva str., house 6, BOX 194044, ksennaidd@gmail.com.*

***Konstantin Vladimirovich Shvetsov**, Cand. of Econ. Sci., professor, Peter the Great St. Petersburg Polytechnic University, Russia, St.Petersburg, Politekhnicheskaya str., house 29, BOX 195251, Konstantin.Shvetsov@spbstu.ru*

***Alexander Viktorovich Yakovlev**, Cand. of Techn. Sci., Seniour Researcher, Military Medical Academy, Russia, St.Petersburg, Akademika Lebedeva str., house 6, BOX 194044; Associate Professor, Saint-Petersburg State University of Aerospace Instrumentation, Russia, St.Petersburg, Bolshaya Morskaia str., house 67, BOX 190000. sven-7@mail.ru.*

***Vladimir Andreevich Parkhomenko**, software engineer, Peter the Great St.Petersburg Polytechnic University, Russia, St.Petersburg, Politekhnicheskaya str., house 29, BOX 195251, Vladimir.Parkhomenko@spbstu.ru.*

The book is devoted to machine learning. We consider *machine learning* as a branch of the smart data processing techniques, characterized by learning how to use examples of multiple solutions, rather than a priori given algorithms to solve a problem. Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Machine learning methods may be referred to predictive analytics or predictive modelling.

For developing machine learning algorithms the following methods are used: mathematical statistics, numerical analysis, optimization methods, probability theory, graph theory, extracting from the data logical rules to recognize given classes of objects, building ontology, constructing knowledge bases and different types of knowledge structures and other approaches.

There are two types of learning: training by precedents and identifying empirical patterns in data. In the second case, there are many objects (situations) and many possible answers (responses, reactions). There is some dependency between answers

and objects, but it is unknown. Only the finite set of «object, answer» pairs, called a training sample, is known. Based on this data, you need to build an algorithm capable of giving an accurate classification response for any possible input object.

This dependency is not necessarily expressed analytically, for example, neural networks implement approximation of almost any complex function that cannot be expressed analytically. An important feature of learning systems is the ability to generalize, that is, to adequately respond to data that goes beyond the training sample.

The task of learning is a generalization of the classic tasks of function approximation. In the classical tasks of approximation, objects are real numbers, vectors, temporal sequences. Approximation of object classifications is used in the approach to machine learning based on inferring good classification (diagnostic) tests from data (see chapter 9 in the book). In real-world applications, input data can be incomplete, inaccurate, non-numerical, heterogeneous. These features lead to a wide variety of machine learning methods.

The machine learning methods are quite well developed. However, modern technical means of measurement, experimental research, and observations have led to the accumulation of huge amounts of data in science, manufacturing, business, transport, healthcare. Therefore, far many difficulties arise in connection with forecasting, management, and decision-making tasks because data requires to be pre-processed, structured, decomposed, selected or classified previously and so on. Thus, the computer science key problems have shifted towards the processing and analysis of big data, in which machine learning is often only the «tip of a huge iceberg». Today, the term «big data» has gained great popularity and is actively used in various fields. But this concept is not uniquely determined.

However, all the definitions agree that «big data» is a data analysis technology designed to extract new useful knowledge from such volumes of data that people cannot handle. Analysis of literature shows that «big data» possesses two fundamental features, apart from its large volume:

– combining data from a variety of sources; it can be a set of different databases, observations or measurements, digital archives of medical images, results of the population screening on the basis of the telemedicine systems, the Internet, etc.;

– the need to use fundamentally new and complex methods of analyzing data based on machine learning algorithms in greater degree than earlier.

The editors' mission was to find such works, which focused on the following problems for solving the prediction and diagnostic tasks in medical-biological and socio-economic studies:

– practical application of machine learning algorithms;
– approaches and practical methods of the formations of scales for evaluating features in solving machine learning problems;
– large-volume of multi-modal data collection and their pre-processing.

Publication policy. Chapters were submitted via the Easychair system. All the chapters included in the monograph have undergone two iterations of double-blind peer reviewing. All the chapters have also been reviewed by one of the book editors (X. A. Naidenova, K. V. Shvetsov and A. V. Iakovlev).

Separate chapters have been reviewed in advance by the authors of other chapters of the book. Unfortunately, some of the submitted papers were not included in the book. These papers were related to valuable but not priority topics of the book. The editors are very grateful to all who submitted their papers for this edition.

The *manuscript's content* is freely available via Internet using DOI. The DOI of the book is given in its bibliographic description. Each chapter has its own DOI, which works as a url-link, in the epigraph before the chapter's title. This link leads directly to the appropriate pdf.

The *table of contents* is presented both in English and Russian languages (as well as index, preface and some other book elements). Special sections of each chapter called «introduction», «conclusion», «acknowledgments» and «references» are omitted in the table of contents.

The following is a brief outline of the chapters, which are conditionally divided into two parts. The authors succeeded in providing a large amount of new and valuable materials related to data processing and machine learning.

0.1. Machine learning in biomedical data analysis

The first part of the book begins with Chapter 1 «Application of machine learning methods in medicine» by Oleg V. Senko and Anna V. Kuznetsova. They give a

survey of machine learning technologies to solve diagnostic and predictive problems in medicine. These technologies allow you to generate algorithms that calculate diagnostic or predictive solutions automatically based on accumulated amounts of clinical data. The chapter provides examples of the use of different machine learning techniques to solve specific medical problems. In addition to widespread technologies, some original methods based on making collective decisions based on regularities extracted from data (regular patterns) are also considered. The issues of correct evaluation of the effectiveness of diagnostic and forecast algorithms are discussed.

Chapter 2 «Spiking recurrent neural networks for the classification of electrocardiograms by the type of arrhythmia» by Kirill V. Nikitin uncovers one of the most actively developing areas in the neural computing theory. Most of the existing models of neural networks (in what follows – NN) can be conventionally divided into two classes depending on the model of neuron – the artificial NN and the biologically oriented NN. In the first class of NN, mathematical models of neurons describing some features of their natural biological prototypes are used – as a rule, neuron inputs are multiplied by weights, summarized and a transmission function, most often sigmoidal, is applied to them. In the second class of NN, neurons are modeled as realistic as possible considering the dynamics, physical and chemical properties and processes occurring in real neurons. One of the main such properties is the generation of short bursts or spikes of activity by neurons. Therefore, such models are often called spiking models.

When modeling spiking NN, a system of differential equations is compiled for each neuron and then this system is approximately solved in the time area. Many papers show that spiking NN have more potential than artificial ones, especially when dealing with time-changing signals. Thus, the spiking NN is now receiving much attention in a wide variety of areas. In this chapter the author's program environment for the simulation of spiking NN is described, the results of its research are given, and the analysis of different models of spiking neurons and chemical synapses is presented. The results obtained are used to develop techniques for building the appropriate NN as part of an unstable state machine to solve dynamic image recognition problems. An example of using the techniques and models of NN to classify electrocardiograms with respect to the type of arrhythmia is considered.

In Chapter 3 «Ensemble learning method based on Oracle Data Mining for cardiovascular diseases prediction», the authors Alexandra D. Soboleva and Oleg Yu. Sabinin consider the existing methods of machine learning for solving the problem of prediction and analyze of their disadvantages. They have proposed and justified a method for solving the problem of prediction by the help of aggregating results of two machine learning algorithms (Generalized Linear Model and Support Vector Machine) opposite by nature. The problem of early diagnosing cardiovascular diseases is considered. The method proposed was examined using three different sets of patient analyses and their histories to assess the risks of cardiac diseases. This examination conforms the effectiveness of the developed approach to solve the problem of early diagnosis of cardiovascular diseases.

Chapter 4 «Applicability of computational fluid dynamics in research of intracranial aneurysms hemodynamics» is an example of a comprehensive approach to complex medical problems requiring the integration of knowledge in the fields of analytical medicine, diagnostics, mathematics, physics, data processing and programming. The authors Andrey V. Gavrilov, Daria D. Dolotova, Evgeniya R. Blagosklonova, Ivan V. Arhipov, Elena V. Grigorieva, Natalia A. Polunina, and Vladimir V. Krylov use modern methods of processing 2D/3D images, as well as methods of mathematical modeling based on computational hydrodynamics to assess the hemodynamic features of intracranial aneurysms.

Choosing a treatment regime for patients with this pathology is a complex task, which must take into account the ratio of risks of aneurysm rupture in its conservative treatment with intraoperative risks in different methods of surgical treatment. Until now, assessing risk of aneurysm rupture is an unresolved problem. The authors propose a system of risk analysis and aneurysms diagnosis based on patient-specific three-dimensional reconstructions of vascular bed section with aneurysms. In this case, the data of CT-angiography are used with the subsequent performing hydrodynamic calculations by taking into account the speed characteristics of a particular patient blood flow. Risk calculations take into account both morphometric and hemodynamic indicators, which cannot be evaluated by existing diagnostic methods in vivo. The chapter presents the richest material of collecting, analyzing and processing a large amount of data in a specific medical

task. The use of machine learning methods is a promising development of the proposed analytical diagnostic system.

Chapter 5 «Automated system for obtaining and mining data for machine learning in evaluation of the specialists' functional state» by Alexander V. Yakovlev, Viacheslav O. Matytsin, Xenia A. Naidenova, and Vladimir A. Parkhomenko is dedicated to the problem of synchronizing measurement of a large number of primary indicators of different natures, their collecting in the database for further processing in order to extract secondary informative indicators to assess the functional state of specialists. Estimating the functional state of specialist's organism means measuring the totality of characteristics of his physiological functions and psychophysiological reactions determined in the different environment conditions of his professional and behavioral activity. Diagnosis and prediction of the functional condition and professional reliability of specialists is an urgent task in many areas of human activity: management of complex human-machine systems, driving, piloting planes, working in space and more.

The authors propose the instrumental measuring stand for multimodal registration of data from which the evaluation of the human functional state derived. The proposed solution uses technical and software tools to register big data sets of three types: audio, video and physiological data obtained from polygraph sensors. The algorithms were selected and the performing of primary data transformations is tested to get a set of structured multimodal indicators, expressed in numerical form. A pilot experiment was conducted to assess the characteristics of the database developed and the capabilities of the stand. The stand can be utilized not only for diagnostic tasks but also for fundamental investigation in physiology.

Chapter 6 «The formation of an external criterion for machine learning based on medical-biological data» by Nella A. Shchukina is concerned with one of the important machine learning problems, namely, the formation of external criterion as an external standard classification of objects to create training samples. In biomedical research, a serious obstacle to using machine learning methods is the ambiguity of external criteria as the basis of a «well-understood» image in pattern recognition tasks. The chapter proposes a method of creating a meaningful scale that establishes a correspondence between the quantitative feature (characteristic) and its qualitative assessments in the absence of a well-defined external criterion.

The method is illustrated by an example of selecting the respondents for a specialty by their level of physical fitness. An external criterion based on a qualitative assessment of their anaerobic ability, which is physiologically related to the degree of their physical fitness, is modeled. The stages of empirical modeling the external criterion, based on an indirect scale of respondents' physiological properties are described. This work can also be widely used not only in life sciences, but also in sociological research.

0.2. Machine learning in the analysis of socio-economic data

Chapter 7 opens the second part of the book. This chapter, called «Machine learning for subgroup discovery under treatment effect» by Aleksey V. Buzmakov, deals with the problem of estimating the effect of impact on the individual level. This problem is relevant in many areas of knowledge from marketing to medicine. The term «treatment effect» originates in a medical literature concerned with the causal effects of binary, yes-or-no «treatments», such as an experimental drug or a new surgical procedure. But this term is now used much more generally. In marketing, sending offers only to people who are interested in a product, perhaps reduce the cost of advertising campaign. The author examines existing mathematical methods for assessing the effects of impact at the individual level, gives a critical analysis of these methods and shows the necessity for the development of new effective methods in this area of knowledge.

Chapter 8 «The function of desirability and the Harrington preference scale in psychological studies» by Nella A. Shchukina proposes an original solution to the problem of integral assessment of psychological profiles of respondents in field of professional selection. The difficulty of this problem lies in forming a comprehensive indicator based on test indicators of different semantic and psychophysical nature.

In this paper, it is proposed to introduce a unified artificial metric based on Harrington's desirability function and to assign this metric to a standard scale, namely, Harrington's preference scale. A way to form the desired integral indicator using machine learning for encoding the test scores of respondents has been developed. The chapter outlines the principles of creating an integral indicator and the rules for its application. The proposed method can be successfully applied for

a wide range of problems in psychophysiological, sociological and socio-economic studies.

Chapter 9 «Application of a logical-combinatorial network to symbolic machine learning tasks» by Xenia A. Naidenova, Vladimir A. Parkhomenko, Konstantin V. Shvetsov acquaints the reader with one of branches of symbolic machine learning related to inferring good classification (diagnostic) tests from data. The main tasks of symbolic machine learning are formulated for inferring logical rules and dependencies from data including functional and implicative dependencies, association rules, key patterns satisfying some special requirements and many others. The Apriori algorithm solving the tasks listed above is described. A neural-like logical-combinatorial network is proposed to effectively realize Apriori-like universal inductive method for extracting logical dependencies from data.

Functioning the neural-like logical-combinatorial network is described for the task of inferring good maximally redundant and irredundant classification tests from data. New approaches are proposed for reducing the computational complexity of algorithms. They are easily implemented on neural-like logical-combinatorial networks.

Acknowledgments

The editors are grateful to authors for the papers submission, to reviewers for all affords to improve the book content, to the administration and staff of Peter the Great St. Petersburg University (SPbPU) for the support of the publishing process.

Editors pay deep respect and thank corresponding member of Russian Academy of Science, SPbPU vice-rector for research V. V. Sergeev for the given opportunity to try new russian-english format of the manuscript for its wider distribution within english readers.

Editors make separate acknowledgment to SPbPU associate professor A. V. Schukin for the valuable comments. Without his support this book would hardly have been born.

Editors are very grateful to the following SPbPU workers for the help with the development of the manuscript publishing requirements: A. V. Vankovich,

V. M. Yakubson, T. V. Baboshina, N. V. Sokolova, V. A. Lysenko, S. V. Shutova, T. P. Naumova.

The book publishing process was partially supported by the Academic Excellence Project 5-100 proposed by Peter the Great St. Petersburg Polytechnic University.

Sincerely yours,

Editors