Chao, Florenly and Liena. Computational Methods for Data Analysis in Clinical Medicine: Issues and Rules. International Journal of Business Society. Special 2021, 10-13

Journal Homepage: www.ijo-bs.com

International Journal of Business Society

Contents lists available at: https://www.ijo-bs.com/issue.html

COMPUTATIONAL METHODS FOR DATA ANALYSIS IN CLINICAL MEDICINE: ISSUES AND RULES

Ding Chao[1]; Florenly[2*], Liena[3]

[1]*Master student, Faculty of Management, Universitas Prima Indonesia*
[2]*Faculty of Management, Universitas Prima Indonesia, florenly@unprimdn.ac.id*
[3]*Faculty of Management, Universitas Prima Indonesia,*
[*]*Corresponding Author*

Information of Article

Article history:
Received: 4 Nov 2021
Revised: 5 Nov 2021
Accepted: 30 Nov 2021
Available online: 1 Dec 2021

Keywords:
Data mining
Analytical models
Clinical medicine

ABSTRACT

The wide availability of new computation tools to perform data analysis and analytical modeling makes medical informatics selection methodical. The professionals must consider this to ensure the most appropriate approach to address difficulties and issues of clinical prediction. Above all, the so-called "data mining" methods could offer methodological resolutions to consider the analysis of medical data and the construction of predictive models. A wide variety of these methods require up-front rules to support physicians and doctors in the most suitable selection of data mining (DM) tools, to construct and validate the predictive models, and the distribution of predictive models in clinical settings and surroundings. The scope and role of the field of research in analytical DM have been discussed in this paper, which is the main aim of this paper. Analytical DM has become a vital tool for researchers, clinicians, and physicians. With the integration of partial and clinic-related data, recently, genomic medicine has been efficiently paid attention to and grown to gain drive. Besides, this field has also gained new groups of complicated problems to be solved.

1. Introduction

Recently, "data mining" was not used more and more in much of the literature in medicine. All in all, this term has never been linked to a specific interpretation, but almost to the types of cross-thinking regarding its meaning: the use of new methods and tools to analyze large amounts of data. Data mining (DM) has been successfully applied in various fields of human endeavor, including marketing, customer management, engineering, and various scientific fields. However, despite high hopes, its application to medical data analysis was relatively flawed until recently. This is primarily true for real-world clinical medicine applications, which can benefit from DM tactics capable of running predictive models, exploiting knowledge available in the clinical field, and explaining proposed decisions simply. Some models support clinical decisions. The aim of analytical DM in clinically performed medicine is that models can be designed to use patient-specific information to predict the desirable outcomes and thus care about the clinical decision making. Analytical data mining approaches can be useful for decision models composition for procedures, for example, forecast, diagnosis, and the planning for the treatment, which can be integrated with clinical information systems after procedures are completed. An assessment and validation have been carried out. This paper reviews data mining procedures that focus on analyzing data and highlights several questions most appropriate to clinical medicine applications.

DM is a process by which a large amount of data can be selected, explored, and designed to realize unfamiliar designs or associations to deliver a pure and valuable outcome for the data specialist [1]. The DM is identical to the knowledge finding in database where it emphasizes analyzing data rather than using factual investigation and analysis-related approaches and procedures [2]. DM-related issues could be somehow resolved by utilizing various methods and tactics derived via the sciences of computational methods using computers and statistics. DM is used in various fields such as machine learning, green computation, data visualization, validation, grouping, and classification. This paper aims to list several methods and technologies applied to data processing in the science of DM by reviewing several research studies. The remaining sections of this paper are organized as follows: Section 2 reviews a literature review; Section 3 focuses on analytical DM process: tasks and guidelines; Section 4 provides a comprehensive discussion on the findings, and Section 5 represents conclusion which includes important concluded remarks for research directions.

2. Literature Review

2.1. Analytical DM Technologies Applied to Classification Purposes

The decision trees use iterative data hashing and push translucent classifiers. The performance could be affected by the hashing of data - documents in the decision trees might have limited cases to help for the unfailing estimates. The computational density of training processes is short as a result of the strong exploratories. The current DM groups mostly

https://dx.doi.org/10.30566/ijo-bs/2021.special.03

consist of alternatives from the Decision Tree Induction Algorithms (DTIA) [3]. The logistic regression is a robust and fixed statistical method [4]. It can be a postponement of normal regression where results can be modelled with two values that generally represent the existence or non-occurrence for such an event. The basic probability model is a multiple of [5] with the naive Bayesian classifier. Still, it has more usages related to the complex method utilising the maximum probability assessment to regulate its probability equation coefficients. Dealing with missing attribute values is not easy. The model could be well characterised by graph [6]. The decision rules can have the procedure of "if based on condition - attribute values then result from value" can be created or derivative straight using the data as with CN2 algorithms [7]. Even if the algorithms can share the greatest portion of the performance(s) characteristics with the decision tree, they can be further affluent in terms of the computation. Artificial neural networks (ANN) have been lately the most common AI-based algorithm for the data modelling used in medicine clinically. That has been undoubtedly a result of the upright analytical act. Even though there might be more than a few shortcomings, especially the very highly rated sensitivity of the method's parameters, the network structure determination is inclusive of the cost of the computations [8]. Train and extrapolate models that are difficult to interpret at best by experts in the field. Neural networks can model complex nonlinear relationships, which has a benefit over humbler the methods used for the modelling, such as the Naïve Bayesian classifier or logistic regression [9, 10]. The Supportive Vector Machines (SVM) is possibly one of the greatest influential algorithms used and applied for classification purposes today, where its analytical accuracy is of importance and consideration. It can be based on compact scientific and math-measured basics and the theory of statistical learning. The goal of SVM has been to discover the hyper-plane that could split the instances of the mixed results. SVMs are primarily designed to solve problems of two classes, finding a superfast plane with the extreme distance to the nearest point of the two situations. This super level can be so-called the optimum level. The set of states neighbouring the highly optimum level can also be called the support vector. In order to find the optimum hyper-plane, there is a must to provide at least a linear classifier [11].

In addition to these linear seeds, SVMs are similarly regularly useful to other nonlinear seeds, and that essentially alter the novel feature range into an innovative advanced dimensional range where the linear classifier can be most probably incidental. Common kernel purposes are, for example, polynomial, circular, and sigmoid functions. The process of selecting the fitting kernel can be, in principle, based on the characteristics of the data and problem field. Naive Bayesian is simple, and its performance can be comparable to other competitive methods used for classification purposes with additional and extra complicated ones. The rapid extrapolation of the classifications would result in the usage of a basic procedure in qualified studies. Once they are replaced in analytical performance(s) by other competitive research works, a further complicated set of rules regularly indicate the attendance of nonlinear relations between several features [12, 13]. The Bayesian networks can be probability graphs and can adequately prompt a common distribution of the possibility across a series of parameters thru a group of conditional distributions of the possibility. A Bayesian network can be a fixed non repeated chart. A piece of point can represent random parameters where the curves can characterise a probability dependence, including the node and the related mothers. Every parameter says x(i) cannot rely on non-issues to provide a group of mothers, pa (xi). In this circumstance, with the Markov hypothesis, the common distribution of the possibility considering the whole set of variables (xs) could be utilised with the help of chain rule [14-16].

2.2. Analytical DM and Genomic Medicine

Newly, prospecting for data prediction has received a solid enhancement from molecular environmental science research. The methods related to the DM, including hierarchical grouping or the SVMs, are commonly used to analyse the microarray data or plentiful quantity mass spectrometry [17]. Interestingly, in recent years, many articles have emphasised several likely analytical DM to originate clinics- relevant models from molecular data. Thus, decision support for the genomic medicine’s areas could be provided. Today, clinicians can possess three types of molecular data: (1) data of genotype, frequently signified by a set of single nucleotide polymorphisms (SNPs), the sequence of the DNA variations that can happen if only happen one nucleotide in a sequence has been changed to the genome. In the meantime, all specific elements include numerous SNPs. The arrangement constitutes of an exclusive shape of DNA for the relevant people; (2) data related to the DNA segment expression, that are probably tested by using microarrays to get a photo of the movement of the whole group of the DNA segment inside a material during a provided period of times also by using polymerase chain reaction (PCR) based procedures, where the PCR in real times, a very rare DNA segment needs to be expressed to a more accurate measurement; (3) Data related to the protein expression, that might contain the full range of profiles of the protein obtained using mass spectrometry techniques, or some proteins tags that might probably be tested and calculated using custom essays.The number of traits can vary commonly from many dozens of classical problems in clinical medicine to thousands of genomics.

3. Analytical DM Process: Issues and Rules

DM is considered an application that performs several procedures from unlike corrections aiming to discover exciting shapes that could represent the data. Given the wide diversity of available technologies and many branches and fields, it is not amazing that the DM could habitually be seen as a difficult trade to acquire and difficult to control. It was stated above that various procedural reproductions were proposed to present engineering basics, procedure organization, and specific DM tasks. A DM procedure that gives the idea to be widespread acquisition acceptance has been introduced in

https://dx.doi.org/10.30566/ijo-bs/2021.special.03

line with these settings. Since DM lists many techniques that can be useful for DM jobs performance(s), it has not planned to provide detailed guidance to benefit from the evaluation schemes and statistics techniques. All of these should be specific to the problem area, the DM tasks, and the type of data under study. Analytical DM related to medicine in a clinical environment is an instance of an accurate and job-specific guide that can address various aspects of clinical data analysis to go together with the DM model and make it additional beneficial in the field. The following description of the analytical DM process can observe the DM scheme and many questions, conclusions, and rules. Those will be included for a potential quota of medical analytical DM applications.

3.1. Defining the Problem and Setting Goals

Analytical DM analyses the data sets that be made of numerous examples of the data (for example, states or several remarks), while every case has been considered by a series of features (it could be called forecasters, characteristics, issues, or variables illumination). An additional distinct feature is called a result parameter, a classified item, non-independent variable, or response parameter. The job of extracting analytical data is to discover the finest-fit model that can tell the traits to the consequence. Unlike the standard DM sets, the sets of the date related ot the medicine area can be lesser. The number of cases typically ranges from numerous tens to numerous thousand.

4. Discussion

Compared to DM in business, marketing, medical DM applications have many distinctive characteristics. It is important to note that medicine is a critical security context in which decision-making activities must always be supported by interpretations. This means that the value of each data may be higher than in other settings: trials can be expensive due to the involvement of staff, the use of expensive devices and the potential discomfort for patients concerning. In clinical exploration, data sets may be small and signal non-reproducible cases. Data can be further affected by various sources of uncertainty, such as measurement errors or missing data, or errors in encoding information hidden in text reports. Doctors and researchers face these difficulties by exploiting their knowledge of the field. Likewise, DM can solve these problems by carefully applying the selection of variables and models, correctly evaluating the resulting models, explicitly coding this knowledge, and using it in data analysis.

Today, DM is a very diverse field with several techniques that can serve the same purpose and also work. It may not be practical to explore all the alternative methods when exploring a given dataset, while the choice of techniques used is often driven by the instinct of data miners. While it is improbable that with the present diversity of approaches the community will be able to create recipe-books, general job descriptions will be provided and a simple set of guidelines that can be used will be applied to the construction of analytical clinical models using analytical techniques and DM. In general, the ideas are presented and summarized in the following list:

- Success criteria in advance are defined. The acceptable ranges are determined for the aim of evaluation statistics before modelling.
- If possible and for reference, the performance results with those obtained from classical statistical models are compared.
- Probability model, not belonging to the net category. Methods that report confidence intervals are chosen.
- To avoid overfitting, models with the data that was used to build them are not ever tested. For small data sets, cross-validation techniques used to obtain evaluation statistics.
- If possible, the resulting model to be tested on an isolated and independent data set.
- Performance scores with confidence intervals are reported.
- Modelling techniques that reveal relationships and can present them in a legible way are preferred. If the goal of DM is to find relationships, black box models is avoided.
- If performance is still acceptable, simple modelling techniques are preferred, perhaps those derived from models that can be verified and evaluated by specialists.
- Feature classification, feature selection, constructive extrapolation, etc., as well as estimation of any parameters, are part of the modelling and should be tested by cross-validation. Its use in pre-processing that occurs before cross-validation increases throughput.
- The project does not end when a good model is found. Consider how to include your model in a decision support or clinical information system. If possible, do a cost / benefit study.
- Explicit evaluation of the application of the model and the possibility of generalization. Consider here in particular the type of data collection (retrospective, prospective, derived from clinical trials or routine), the amount of data available, and the performance of the model.

These guidelines relate to newly emerging issues in personalized and genomic medicine. Today, the construction of reliable analytical models may require the integration of data drawn from heterogeneous sources that include clinical, laboratory, genetic, genomic, and proteomic data. The full availability of data repositories and warehouses able to concurrently provide such information about a single patient, and the methods to integrate it within a decision support system are issues which remain to be resolved.

https://dx.doi.org/10.30566/ijo-bs/2021.special.03

5. Conclusion

Today, numerous mature analytical data mining (DM) approaches have been positively useful to various practical problems in clinical medicine. Data mining is particularly effective with plenty of data. In this setting, the analysis of clinical information from repositories, epidemiological studies, and emerging studies in genomics and proteomics are very important to be included. Data mining approaches have enhanced related methodological procedures such as prior knowledge usage, interesting, uncovered procedures, analytical issues that are interpretable, relationships that produce significance, and models that depend on the rule conception are built or designed. Symbolic tools and models that experts can examine and analyze are crucial for these types of data. Besides, data mining could significantly help discover models that can provide interpretation for prediction purposes. Lastly, data mining contributes to decision-makers producing relational models and it could support such an application of analytical models for daily clinical observes and needs. Data mining and computation methods for knowledge-incentive have become a must. They require very advanced levels of knowledge regarding the research and real applications that are up-to-the-minute. This criterion is set because genomic medicine has offered promises with the upcoming needs that could be integrated with clinical data. This has produced such critical needs in the so near future. To finish, data mining in clinical departments with the arrangements of "bedside" matters, models could be predicted and designed for patient outcomes. Thus, the issues and prediction cost and expenses are very important to be considered together with the procedure of decision making that depends on the prediction model to come out with a good analysis of the results.

References

[1] P. Giudici, *Applied data mining: statistical methods for business and industry*. John Wiley & Sons, 2005.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, no. 3, pp. 37-37, 1996.

[3] J. Quinlan, "C4. 5: ProgramsforMachineLearning, MorganKaufmannPublishers," ed: Inc, 1993.

[4] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[5] J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[6] J. Lubsen, J. Pool, and E. Van der Does, "A practical device for the application of a diagnostic or prognostic function," *Methods of information in medicine,* vol. 17, no. 02, pp. 127-129, 1978.

[7] R. S. Michalski and K. A. Kaufman, "Learning Patterns in Noisy Data: The AQ Approach," in *Machine Learning and Its Applications: Advanced Lectures*, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 22-38.

[8] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in medicine,* vol. 19, no. 4, pp. 541-561, 2000.

[9] D. P. Clark, F. R. Schwartz, D. Marin, J. C. Ramirez-Giraldo, and C. T. Badea, "Deep learning based spectral extrapolation for dual-source, dual-energy x-ray computed tomography," *Medical Physics,* https://doi.org/10.1002/mp.14324 vol. 47, no. 9, pp. 4150-4163, 2020/09/01 2020, doi: https://doi.org/10.1002/mp.14324.

[10] C. Wang *et al.*, "Improving Generalizability in Limited-Angle CT Reconstruction with Sinogram Extrapolation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Cham, M. de Bruijne *et al.*, Eds., 2021// 2021: Springer International Publishing, pp. 86-96.

[11] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[12] S. Long, J. Chen, A. Hu, H. Liu, Z. Chen, and D. Zheng, "Microaneurysms Detection in Color Fundus Images based on Naive Bayesian Classification," ed: Research Square, 2020.

[13] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing,* vol. 77, no. 5, pp. 5198-5219, 2021/05/01 2021, doi: 10.1007/s11227-020-03481-x.

[14] B. G. Marcot and T. D. Penman, "Advances in Bayesian network modelling: Integration of modelling technologies," *Environmental Modelling & Software,* vol. 111, pp. 386-393, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.envsoft.2018.09.016.

[15] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Computational Statistics,* vol. 36, no. 3, pp. 2009-2031, 2021/09/01 2021, doi: 10.1007/s00180-020-00999-9.

[16] X. Zhang and S. Mahadevan, "Bayesian network modeling of accident investigation reports for aviation safety assessment," *Reliability Engineering & System Safety,* vol. 209, p. 107371, 2021/05/01/ 2021, doi: https://doi.org/10.1016/j.ress.2020.107371.

[17] J. Yu *et al.*, "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics,* vol. 21, no. 10, pp. 2200-2209, 2005.

https://dx.doi.org/10.30566/ijo-bs/2021.special.03