KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express xxx (xxxx) xxx

www.elsevier.com/locate/icte

Practical considerations regarding classification learning for clinical diagnosis and therapy advice in oncology

Flavio S. Correa da Silva[a,b,*], Frederico P. Costa[c,b], Antonio F. Iemma[d,b]

[a] *Department of Computer Science, University of Sao Paulo, 05508090, Brazil*
[b] *Autem Medical, Bedford, NH 03110, USA*
[c] *Oncology Center, Hospital Sirio Libanes 01308050, Brazil*
[d] *Department of Exact Sciences, University of Sao Paulo 13418900, Brazil*

Received 31 December 2019; received in revised form 15 March 2020; accepted 16 March 2020
Available online xxxx

Abstract

In the present article the relationship between machine learning and medicine is reviewed, in order to assess the potential for the practical use of machine learning for diagnosis and therapy advice. The considerations built herein are particularly suitable for oncology, in which early diagnostics is particularly important for the success of treatments, and therapy is often based on chemotherapy and radiotherapy, which have harmful side effects.

Keywords: Machine learning in medicine; Supervised learning; Artificial intelligence

1. Introduction

Artificial intelligence (AI) and medicine have a longstanding relationship, possibly started with the development of MYCIN for therapy advice [1]. Medicine has provided AI with challenging problems in clinical diagnosis (*given a set of signs, select the best diagnosis*) and therapy advice (*given a diagnosis, select the best therapy plan*). AI, in turn, has offered promising technologies for problem solving in medicine [2].

Oncology is particularly fit for AI [3,4], as cancer can be treated most effectively if identified early, but symptoms in oncology are difficult to identify at early stages, hence technologies for early diagnosis are welcome; and cancer therapy is often based on chemotherapy and/or radiotherapy, which have severe side effects, hence technologies that can refine therapy plans to minimise side effects and provide unequivocal evidence of the efficacy of innovative therapies are welcome.

The practical use of AI in medicine has occurred more often in management of supporting information, rather than in direct support of activities of healthcare professionals: medical doctors have augmented their capabilities with systems for knowledge representation and processing and automated assistants to process large data sets, but automated diagnosis and therapy advice have only recently started to move outwith academic research [5]. Possible explanations for this observation are because medicine is strongly regulated by organisations and norms such as the Food and Drug Administration in the US and the European Medicines Agency and CE Mark regulation in the European Economic Area. Quality assurance and transparency levels required by these organisations are costly and time consuming; and empirical validation of novel methods and techniques for automated diagnosis and therapy advice requires clinical trials which are costly, labor and time consuming, frequently outwith the scope of academic initiatives.

In recent years, AI in medicine has steered towards machine learning (ML) [6], based on claims that ML can provide smaller subjectivity in comparison with other techniques, because domain knowledge and expertise are replaced by statistically grounded data analysis. Domain modeling, however, is still at the core of ML, particularly when explainability is a strong requirement, as is the case in medicine given the transparency requirements posed by regulatory organisations and norms. Moreover, the demanding requirements of clinical

* Corresponding author at: Department of Computer Science, University of Sao Paulo, 05508090, Brazil.
E-mail address: fcs@usp.br (F.S. Correa da Silva).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2020.03.004

Please cite this article as: F.S. Correa da Silva, F.P. Costa and A.F. Iemma, Practical considerations regarding classification learning for clinical diagnosis and therapy advice in oncology, ICT Express (2020), https://doi.org/10.1016/j.icte.2020.03.004.

trials bring to fore the necessity to control sample complexity (in other words, extracting as much information as possible from samples which are kept as small as possible to train learning algorithms), which leads to requirements of knowledge elicitation for appropriate domain modeling as foundational steps to build systems that can be relevant in practice.

The two main goals of this article are (1) to dispel unrealistic expectations that ML could outgrow expert knowledge in importance, and (2) to provide evidence that AI for direct support of activities of healthcare professionals is feasible.

Section 2 sketches a method to build systems for diagnosis and therapy advice, with particular focus on oncology, structured in such way that the relevance of domain knowledge is clarified. Section 3 characterises lower bounds for sample complexity given expected precision and reliability requirements *wrt* ground truth provided by medical experts. The obtained bounds determine conservative requirements for sample sizes to be employed in clinical trials. Finally, Section 4 presents a discussion and conclusions.

2. Learning, diagnosis and therapy advice

Clinical diagnosis can be seen as a set of steps based on a slice of the patient journey:
(1) Patient p comes to doctor. Doctor selects signs S based on (a) expert knowledge and (b) tacit selection of a reference population P such that $p \in P$.
(2) Given signs S, doctor builds preliminary hypotheses D about diagnostics for p. Expert knowledge and reference population determine unknown yet defined upper bounds on precision and reliability of diagnostics, corresponding to the best available diagnosis (ground truth).
(3) D is ranked according to risk, based on (a) strength of evidence and (b) severity of corresponding diseases. Following order induced by rank, for each $d \in D$ *(a) another set of signs* S'_d *is selected; (b) doctor tags d as either possible or discarded on the face of* S'_d.
(4) Doctor performs fusion for final diagnostics.

The automation of steps 1 to 3 using supervised classification learning can be characterised as follows (Fig. 1):
(1) Given patient $p \in P$, *signs S are selected.*
(2) Oracle $\mathcal{O}_{\hat{P}}$, $\hat{P} \subseteq P$ *is retrieved from a database of oracles. An oracle is a collection of pairs* $\langle S_i, d_i \rangle$ *in which* S_i *are signs observed in* $\hat{p}_i \in \hat{P}$ *and* d_i *are diagnostics. The cardinality of* $\hat{P}$ *must be sufficiently large to ensure appropriate precision and reliability of diagnostics for p, which correspond to a sufficient similarity between empirical classifiers and the best available diagnosis.*
(3) Correlation between signs and hypotheses is characterised wrt functions that best capture how decision procedures can be optimised. In machine learning and statistics jargon, such functions are kernels [7]. *The choice of the appropriate kernel is based on inspection of correlations and expert knowledge about the methods and techniques used to build learning models. The choice of a kernel determines an upper bound on the precision and reliability of empirical classifiers.*
(4) Hypotheses D are built and ranked based on $\mathcal{O}_{\hat{P}}$, *the selected kernel and S. For each hypothesis* $d \in D$, *a second set of signs* S'_d *is selected, based on expert knowledge and the reference population.*
(5) For each S'_d, *a kernel is selected given the correlation between* S'_d *and the corresponding hypothesis.*
(6) Samples of appropriate cardinality are selected from P, considering the required precision and reliability of diagnostics that can ensure sufficient similarity between empirical classifiers and the best available diagnosis.
(7) Automated decision procedures are built for the diagnosis of p employing a sample $\hat{P}$ *of appropriate cardinality and the sets of signs S and* S'_d, $d \in D$.
(8) Decision procedures support medical diagnostics.

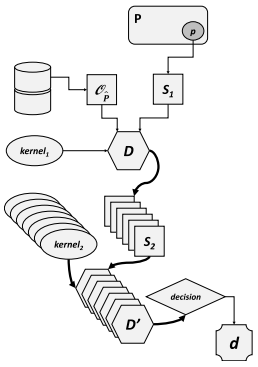

Fig. 1. Procedure — diagnosis.

Therapy advice can also be seen as steps:
(1) Patient p has a previously identified most likely diagnostics. Doctor selects a set of tests T to choose a therapy plan, based on expert knowledge and tacit selection of a reference population P such that $p \in P$.
(2) Given the outcomes of T, doctor builds therapy plan for p. This plan can contain additional decision points in the form of IF-THEN rules.
(3) Treatment is assessed based on observation of attributes, experience and reference population.

General steps to automate this procedure can be characterised as follows (Fig. 2):
(1) Given reference population P, patient $p \in P$ *and corresponding most likely diagnostics d, alternative therapy plans* $t_i \in T$, *in which T is a set of plans, are ranked according to previous knowledge. Ranking is based on correlation analyses of different plans and their corresponding effectiveness, which are built using samples* $\hat{P} \subseteq P$ *of appropriate cardinality, such that precision and reliability can be ensured wrt the best available information about therapy plans. In oncology, given the high mortality related to certain types of tumor, these empirical results can be based on relatively small numbers of cases which are, in turn, described in great detail.*
(2) The most highly ranked therapy is applied on p.

This characterisation of clinical diagnosis and therapy advice in steps helps in the identification of limitations in precision and reliability of diagnosis and therapy advice. The selection of the sets of signs S and S'_d, therapy plans T and reference population P depends upon *expert medical knowledge*. The choices of kernels to characterise correlations

Please cite this article as: F.S. Correa da Silva, F.P. Costa and A.F. Iemma, Practical considerations regarding classification learning for clinical diagnosis and therapy advice in oncology, ICT Express (2020), https://doi.org/10.1016/j.icte.2020.03.004.

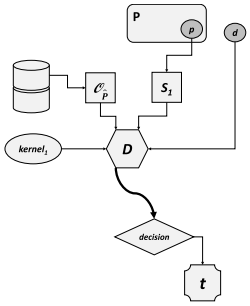

Fig. 2. Procedure — therapy advice.

between signs and hypotheses and between diagnostics and therapy plans depend upon *expert statistical knowledge*. These limitations are imposed upon the best possible diagnosis and therapy decision. Additionally, the cardinality of the samples used to build empirical estimates for the best possible diagnosis and plans is determined given lower bounds provided by statistical analysis.

Limits for precision and reliability of diagnostics and therapy plans as stated in previous paragraphs are at place in standard medical practice, and machine learning techniques, when properly used, can at best ensure the provision of results which are provably similar to the best possible diagnoses and plans. As detailed in the following section, limitations in precision and reliability of decisions due to sample cardinality can be safely bounded provided that we have access to sufficiently large samples.

3. Domain characterisation

Given a population P with correlations between signs $\in S$ and corresponding diagnoses $\in D$, we wish to determine a minimal cardinality N of samples $\hat{P}$ employed to build oracles $\mathcal{O}_{\hat{P}}$ such that $p_1, \ldots, p_N \in P$, each $\vec{v}_i$ is a tuple of values of signs $\in S$ corresponding to observations about p_i, and $d_i \in \{\top, \bot\}$ is a confirmed diagnostics for p_i *wrt* a disease $d \in D$, in which $\top, \bot$ indicate resp. *confirmed* and *refuted disease*.

High correlation is assumed between values of signs $\vec{v}_i$ and diagnostics d_i. The set $\{\langle\vec{v}_1, d_1\rangle, \ldots \langle\vec{v}_{|P|}, d_{|P|}\rangle\}$ of pairs $\langle$signs, diagnostics$\rangle$ for all $p \in P$ can be *partially inconsistent*, amounting for (unknown yet determined) upper bounds on precision of diagnoses and on optimality of therapy plans. These upper bounds characterise the *best available classifiers*. Empirical classifiers based on oracles $\mathcal{O}_{\hat{P}}$ can be built, which are provably sufficiently similar to the best available classifiers.

Probably Approximately Correct Learning [8] extended to cope with partial inconsistencies [9,10], can provide lower bounds for the cardinality of $\vec{P}$ as a function of (1) $|\vec{\mathcal{V}}|$: the cardinality of the valued signs space. Assuming that each sign $s \in S$ can have a finite set of values $\{v_1, \ldots, v_{n_s}\}$ with cardinality n_s, we have that $|\vec{\mathcal{V}}| = \prod_{s\in S} n_s$; (2) ϵ: precision, i.e. an upper bound for the disagreement between an empirical classifier and the best available classifier, e.g. if $\epsilon = 0.1$, then the probability that, given a tuple of values of signs $\vec{v}$, the empirical classifier and the best available classifier provide the same diagnostics $d \in \{\top, \bot\}$ is at least 90%; and (3) δ: reliability, i.e. an upper bound for the risk to build a classifier whose precision is $< \epsilon$. For example, if $\delta = 0.2$ and $\epsilon = 0.1$, then there is a probability below 20% to select a random classifier built using any oracle $\hat{P}$ with a disagreement below 90% *wrt* the best available classifier.

Table 1
Lower bounds for $|\vec{P}|$ given ϵ, δ assuming $|\vec{\mathcal{V}}| \approx 150$.

$\lvert\hat{P}\rvert$		ϵ		
		0.1	0.2	0.3
δ	0.1	400	366	346
	0.2	100	92	87
	0.3	45	41	39

Following [10], a lower bound can be defined for the cardinality of $\hat{P}$ as $|\vec{P}| \geq \frac{(ln|\vec{\mathcal{V}}|+ln\frac{2}{\delta})}{2\epsilon^2}$. This lower bound can be used to determine constraints in sample size if, for example, *Support Vector Classification* is used for diagnosis and *Support Vector Correlation* is used for therapy advice [10]. As an example, assuming $|\vec{\mathcal{V}}| \approx 150$, we have $ln|\vec{\mathcal{V}}| \approx 5$. Employing this value, estimates can be obtained for $|\vec{P}|$ given values for ϵ and δ as presented in Table 1.

In order to obtain a probability below 20% that a classifier will feature disagreement above 10% with the best available classifier, access to a sample with cardinality $\geq$ 100 is required.

A preliminary experiment was developed to ground these conceptual results: a cohort of 46 patients, including advanced cancer patients and healthy control subjects, participated in data collection sessions, in which hemodynamic measurements were collected for approximately one hour, and self-assessment of quality of life (Q) was obtained using the *EORTC QLQ-C30* questionnaire (https://qol.eortc.org/). Some patients participated in more than one session, with intervals between sessions within two and four weeks. Considering repetitions between patients, a sample of 206 observations was obtained.

Based on medical knowledge, heart rate variability (HRV) was considered the primary source of hemodynamic information, and other parameters were employed only to validate HRV measurements. Also based on medical knowledge, the time complexity of measurements was employed as basis for classification of patients based on likelihood of survival beyond the threshold of 360 days after data collection. Time complexity was assessed using the *Higuchi Fractal Dimension (H)* [11] and *Sample Entropy (E)* [12].

Cross-validation assessment indicated a precision of 74% in prediction of survival above the indicated threshold, given the selected parameters $\langle Q, H, E\rangle$. Given the sample size of 206 and the precision bounds obtained in the previous paragraphs, we can rely on a probability $\geq$ 80% that accuracy of predictions is in the interval $[0.9 \times 0.74 \approx 67\%, 1.1 \times 0.74 \approx 81\%]$.

Please cite this article as: F.S. Correa da Silva, F.P. Costa and A.F. Iemma, Practical considerations regarding classification learning for clinical diagnosis and therapy advice in oncology, ICT Express (2020), https://doi.org/10.1016/j.icte.2020.03.004.

This preliminary result characterises in practice the roles of machine learning – and corresponding precision and reliability bounds – and medical expert knowledge to build a system for diagnostics.

4. Conclusion

This article contains a discussion of how systems for classification learning can be inserted into the activities workflow of a medical doctor to support diagnosis and therapy advice. Given that an important barrier to the application of machine learning techniques in medicine can be the requirements of large volumes of data, which can point to the necessity of building and running prohibitively costly clinical trials, an analysis of sample complexity estimates to build oracles to train systems based on supervised learning has been presented, together with a suggested pathway to build oracles based on clinical trials of viable dimensions.

The article is devoted to the clarification of methodological issues related to the development of intelligent systems for medicine – more specifically, diagnosis and therapy advice in oncology – in order to avoid misconceptions which can be misleading regarding expectations about the autonomy of statistical learning with respect to medical expert knowledge. In order to illustrate these views, preliminary empirical results are included featuring classification of cancer patients with respect to expected survival beyond a fixed, arbitrary threshold.

Future articles shall be devoted to rigorous empirical validation of the propositions laid out here, particularly with respect to therapy planning in oncology. Given that, as discussed here, empirical results can take significant time to be obtained, the authors have considered that the presentation of the conceptual framework developed here could be useful even prior to further empirical corroboration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: the authors are partners of Autem Medical Research Lab (Brazil).

CRediT authorship contribution statement

Flavio S. Correa da Silva: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review and editing. **Frederico P. Costa:** Conceptualization, Funding acquisition. **Antonio F. Iemma:** Conceptualization.

References

[1] Edward H. Shortliffe, Stanton G. Axline, Bruce G. Buchanan, Thomas C. Merigan, Stanley N. Cohen, An artificial intelligence program to advise physicians regarding antimicrobial therapy, Comput. Biomed. Res. 6 (6) (1973) 544–560.

[2] Niels Peek, Carlo Combi, Roque Marin, Riccardo Bellazzi, Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes, Artif. Intell. Med. 65 (1) (2015) 61–73.

[3] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, Dimitrios I Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[4] Henry Kaplan, Anna Berry, Kristine Rinn, Erin Ellis, George Birchfield, Tanya Wahl, Xiaoyu Liu, Mariko Tameishi, JD. Beatty, Patricia Dawson, Vivek Mehta, Anna Holman, Mary Atwood, Shlece Alexander, Candy Bonham, Lauren Summers, Iya Khalil, Boris Hayete, Diane Wuest, Wei Zheng, Yuhang Liu, Xulong Wang, Thomas David Brown, Abstract 5299: Machine learning approach to personalized medicine in breast cancer patients, Cancer Res. 78 (13 Supplement) (2018) 5299.

[5] Jonathan H. Chen, Steven M. Asch, Machine learning and prediction in medicine: beyond the peak of inflated expectations, New Engl. J. Med. 376 (26) (2017) 2507.

[6] Igor Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artif. Intell. Med. 23 (1) (2001) 89–109.

[7] Bernhard Scholkopf, Alexander J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT press, 2001.

[8] Leslie G. Valiant, A theory of the learnable, Commun. ACM 27 (11) (1984) 1134–1142.

[9] David Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. and Comput. 100 (1) (1992) 78–150.

[10] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, Foundations of Machine Learning, MIT press, 2012.

[11] T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, Physica D 31 (2) (1988) 277–283.

[12] Joshua S. Richman, Douglas E. Lake, J. Randall Moorman, Sample entropy, in: Numerical Computer Methods, Part E, in: Methods in Enzymology, vol. 384, Academic Press, 2004, pp. 172–184.

Please cite this article as: F.S. Correa da Silva, F.P. Costa and A.F. Iemma, Practical considerations regarding classification learning for clinical diagnosis and therapy advice in oncology, ICT Express (2020), https://doi.org/10.1016/j.icte.2020.03.004.