Computers in Biology and Medicine 95 (2018) 147–166

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Studies in the extensively automatic construction of large odds-based inference networks from structured data. Examples from medical, bioinformatics, and health insurance claims data

B. Robson [a,b,*], S. Boray [a,b]

[a] Ingine Inc., VA, USA
[b] The Dirac Foundation, Oxfordshire, UK

ARTICLE INFO

Keywords:
Inference net
Bayes Net
Hyperbolic Dirac Net
Machine learning
Big Data
Data mining
Clinical decision support
Bioinformatics
Anomaly detection
Fraud detection

ABSTRACT

Theoretical and methodological principles are presented for the construction of very large inference nets for odds calculations, composed of hundreds or many thousands or more of elements, in this paper generated by structured data mining. It is argued that the usual small inference nets can sometimes represent rather simple, arbitrary estimates. Examples of applications in clinical and public health data analysis, medical claims data and detection of irregular entries, and bioinformatics data, are presented. Construction of large nets benefits from application of a theory of expected information for sparse data and the Dirac notation and algebra. The extent to which these are important here is briefly discussed. Purposes of the study include (a) exploration of the properties of large inference nets and a perturbation and tacit conditionality models, (b) using these to propose simpler models including one that a physician could use routinely, analogous to a "risk score", (c) examination of the merit of describing optimal performance in a single measure that combines accuracy, specificity, and sensitivity in place of a ROC curve, and (d) relationship to methods for detecting anomalous and potentially fraudulent data.

1. Introduction and brief review

1.1. Background

Computer applications for medicine such as clinical decision support often need a large amount of organized knowledge [1,2]. Medical knowledge captured on the Internet began to escalate to many petabytes in 2009 [1], but when the discipline of *Artificial Intelligence* (AI) first arose under that name in the 1950s, a broad spectrum of knowledge usable by computers was much harder to gather, and more weight was given to logic solving and game playing [3]. They are considered weak methods, because they do not scale up to large or difficult instances [4]. A recent review of the development of AI [5] also argued that (a) humans learn as children essentially by the combination of *top-down* and *bottom-up* methods that AI has imitated [5,6], and (b) benefit from, and even need, large amount of input information. *By* "top down" is meant the kind of inference based on given or taught knowledge that is prepared in advance, essentially in a form still readily recognizable by humans, typically probabilistic and sometimes called "Bayesian", and typically associated with *inference nets* [4–6]. An inference net is reasonably defined as the *linkages* between a set of conditions and conclusions selected and executed by an *inference engine* in a system that explicitly represents knowledge in the form of words and symbols [5]. It should be contrasted with the "*bottom up*" and *Deep Learning* approaches that use artificial *neural networks*, usually start empty of information, encode gathered "knowledge" or information somewhat obscurely as weights associated with the simulated neurons, and train to focus on addressing one or few particular issues [5,6]. The present report explores the perceived advantages of new methods for large inference networks introduced here, though focus is only on the top down approaches, not least because they use explicit canonical elements of knowledge (Section 1.3) that are assigned equally explicit, essentially statistical, estimates of probabilities as their degrees of truth or scope (Sections 1.3 and 1.4). They therefore extend well to the more familiar and pressing use cases of probability based clinical decision support, Evidence Based Medicine, public health, and epidemiological analysis [1,2]. Some examples are given of applications in the life science and healthcare insurance sectors because these are pressing areas, but also to illustrate how the methodologies required differ somewhat, particularly in the claims case.

* Corresponding author. Ingine Inc., VA, USA.
E-mail address: robsonb@aol.com (B. Robson).

https://doi.org/10.1016/j.compbiomed.2018.02.013
Received 1 January 2018; Received in revised form 19 February 2018; Accepted 19 February 2018

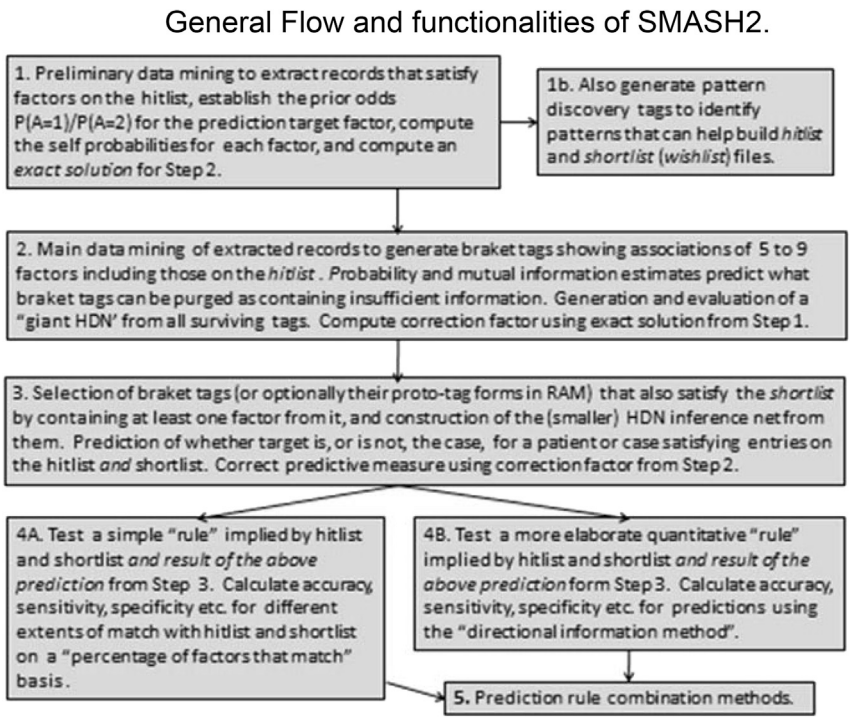

Fig. 1. General Flow and functionalities of SMASH2.

Table 1
Summary of requirements that a prediction for a record be consider a true positive, a true negative, a false positive or a false negative.

PREDICTION.	OBSERVATION.	WORKING MODELS.	AGREEMENT.
Was the brute force prediction using both hitlist and wishlist YES for the target and its value (e.g. 'renal failure'' = 'Y') on first line of hitlist?	Does the record contain target and its value (e.g. 'renal failure'' = 'Y') on first line of hitlist?	Is there sufficient match between *the rest of the record* and hitlist and shortlist, *and* an analogous "degree of match" defined from a prediction using directional information, to make the PREDICTION applicable?	Increment count of true positive TP, or false positive FP,or true negative TN, or false negative FN.
yes	yes	yes	TP
yes	no	yes	FP
yes	no	no (switches PREDICTION)	TN
yes	yes	no (switches PREDICTION)	FN
no	yes	yes	FN
no	no	yes	TN
no	no	no (switches PREDICTION)	FP
no	yes	no (switches PREDICTION)	TP

1.2. Canonical medical knowledge on computers, and reasoning from it

Elements of knowledge accumulated for top down approaches constitute a *knowledge base* [5,7,8], or *knowledge representation store* (KRS) to emphasize when the elements are well rendered into *canonical representations* of knowledge readable both by humans and computers. In the present paper, *KRS elements* are units of knowledge in the KRS that directly represent the building blocks of inference nets. MYCIN [7] and

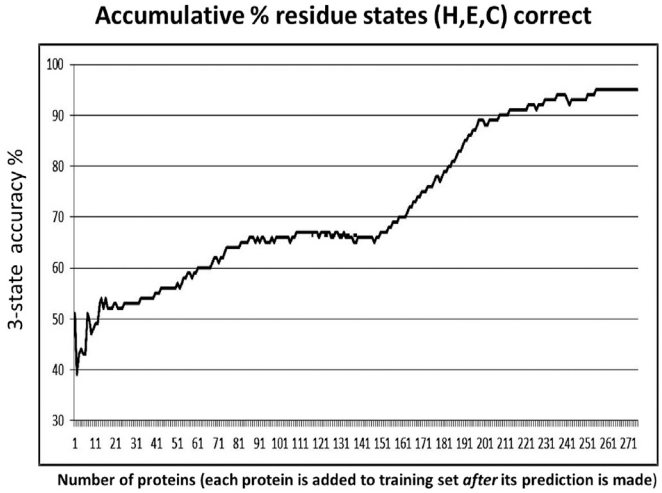

Fig. 2. Example machine learning using the same mining and inference net algorithms applied to the crisper problem of protein secondary structure predictions problem using bioinformatics data.

INTERNIST [8] were pioneering medical *Expert Systems* that exemplify the rapid growth of in size of KRSs in the 1970s. Respectively, they were of some 600 elements of knowledge or "rules" concerning medical microbiology, and some 100,000 concerning internal medicine. Before the Internet significantly impacted on medicine in the mid to late 1980s [8], acquisition of knowledge by debriefing human experts locally was very slow, a problem known as "the Feigenbaum Bottleneck" [9]. Today, the World Wide Web (WWW) links web pages and people [10], and the emerging Semantic Web (SW) is an effort to link all data and knowledge [11]. The SW has gathered more than a trillion KRS elements of knowledge as *semantic triples*, i.e. statements of knowledge in subject-relationship-object form, or interpretable as such [12]. A major feature of the SW for common formats and exchange is the Resource Description Framework (RDF) [11]. An example of a proposal for using the SW and its RDF technology for healthcare information management is