International Journal of Medical Informatics 135 (2020) 104073

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

A patient-similarity-based model for diagnostic prediction

Zheng Jia[a], Xian Zeng[a], Huilong Duan[a], Xudong Lu[a], Haomin Li[b,*]

[a] College of Biomedical Engineering and Instrument Science, Zhejiang University, China
[b] The Children's Hospital of Zhejiang University School of Medicine and the National Clinical Research Center for Child Health, Hangzhou, China

ARTICLE INFO

Keywords:
Patient similarity
Diagnostic prediction
Analogy reasoning
Machine learning

ABSTRACT

Objective: To simulate the clinical reasoning of doctors, retrieve analogous patients of an index patient automatically and predict diagnoses by the similar/dissimilar patients.
Methods: We proposed a novel patient-similarity-based framework for diagnostic prediction, which is inspired by the structure-mapping theory about analogy reasoning in psychology. Patient similarity is defined as the similarity between two patients' diagnoses sets rather than a dichotomous (absence/presence of just one disease). The multilabel classification problem is converted to a single-value regression problem by integrating the pairwise patients' clinical features into a vector and taking the vector as the input and the patient similarity as the output. In contrast to the common k-NN method which only considering the nearest neighbors, we not only utilize similar patients (positive analogy) to generate diagnostic hypotheses, but also utilize dissimilar patients (negative analogy) are used to reject diagnostic hypotheses.
Results: The patient-similarity-based models perform better than the one-vs-all baseline and traditional k-NN methods. The f-1 score of positive-analogy-based prediction is 0.698, significantly higher than the scores of baselines ranging from 0.368 to 0.661. It increases to 0.703 when the negative analogy method is applied to modify the prediction results of positive analogy. The performance of this method is highly promising for larger datasets.
Conclusion: The patient-similarity-based model provides diagnostic decision support that is more accurate, generalizable, and interpretable than those of previous methods and is based on heterogeneous and incomplete data. The model also serves as a new application for the use of clinical big data through artificial intelligence technology.

1. Introduction

Clinical reasoning is the cognitive and noncognitive process by which a healthcare professional interacts with a patient and the environment to collect and interpret patient data, weigh the benefits and risks of actions, and understand patient preferences to determine a working diagnostic and therapeutic management plan with finesse [1,2].

Analogy is a manifestation of clinical reasoning in diagnostic decision making [3]. Analogical ability is a core mechanism in human cognition and a key contributor to higher-order cognition to recognize and reason about common relational structure across different context [4]. Clinical analogy is a process that starting with a corpus of knowledge based on medical theory and past clinical experiences, doctors establish the resemblance of the present patient's signs to previous cases, guess non-observed signs that can be found later, infer links between these signs and, therefore, establish a diagnosis and predict the future evolution of the patient [5]. Analogy plays a central role in two stages of the process of diagnostic decision making: (1) Pattern recognition during data gathering that doctors must be able to discern signs and subsequently to assign them correct weight of relevance. (2) pattern recognition based on both the comparison with theoretical models and with prior experiences, and especially the latter ones for experts than novice doctors [5]. The relevance of a sign within its possible causes is reinforced or weakened by actual cases, where doctors make their own mental schemes from their past experiences, learning the importance of each symptom and creating clusters of symptoms linked to the diseases. For diagnostic decision making, analogy reasoning contains two components: the target, which is the patient that the reasoner is currently attempting to diagnose, and the source, which is the prior patients that is used to compare with [6].Clinical analogy has an inferential nature that allows us not only to evaluates the similarity between the target(the present patient) and source(prior patients), but also to infer additional similarities in non-

* Corresponding author at: Zhouyiqing Building 512#, Yuquan Campus of Zhejiang University, 310027, China.
E-mail address: hmli@zju.edu.cn (H. Li).

https://doi.org/10.1016/j.ijmedinf.2019.104073
Received 4 October 2019; Received in revised form 26 November 2019; Accepted 30 December 2019

Nomenclature

p	Patient
δ	Symptom
I	The total count of distinct symptoms identified from the clinical documents from nephrology
$\boldsymbol{x}_{symptom}$	A feature vector built by symptom similarity
υ	Original numerical lab test value
μ	Mean value
α	Variance value
ζ	Z-score of the numerical lab test value
J	The total count of the lab tests examined in the nephrology department
$\boldsymbol{x}_{labtest}$	A feature vector built by lab test similarity
D	A set of taxonomic diagnoses ICD-10 codes of a patient
d	ICD-10 code of a diagnosis
IC	Information content
CS	Code level similarity
SS	Set level similarity
$\boldsymbol{x}_{preliminary}$	Preliminary diagnosis similarity (a vector that contains only one element)
$\boldsymbol{x}$	A feature vector of a patient pair
y_{diag_sim}	Discharge diagnosis similarity
$\hat{f}_{XGBoost}$	A function supervised machine learning model
$y_{\theta\_min\_sim}$	A criteria of minimum diagnosis similarity
$y_{\theta\_max\_dissim}$	A criteria of maximum diagnoses dissimilarity
k_{p_sim}	Top nearest neighbors for hypotheses generation (positive analogy)
k_{p_dissim}	Top furthest neighbors for hypotheses rejection (negative analogy)
k_{p_kNN}	Top nearest neighbors for kNN method
f_{θ_sim}	A criteria of minimum diagnoses frequency in selected similar neighbors
f_{θ_dissim}	A criteria of minimum diagnoses frequency in selected dissimilar neighbors
k_{d_sim}	Top most frequent diagnoses in selected similar neighbors
k_{d_dissim}	Top most frequent diagnoses in selected dissimilar neighbors
k_{d_kNN}	Top most frequent diagnoses for kNN method
p_{micro}	Micro precision
r_{micro}	Micro recall
$f-1_{micro}$	Micro f-1
Y	The set of gold standard results
$\hat{Y}$	The set of predicted results

observed features(e.g., diagnosis) [5,7].

R[4]. Given a topic, a person may be reminded of a prior analogous case in long-term memory. For example, the percentage of correct solutions nearly tripled when participants were given a hint to “think about what you heard before” [8]. Another experiment showed that surface-similar matches produce more reminding, even though relational matches were rated higher in both soundness and similarity [9]. With massive amounts of clinical data accumulating, the mental bandwidth required to handle such vast amounts of data has clearly exceeded the limitation of human brains, necessitating an increased reliance on machines [10]. In this research, we explored how to retrieve analogous patients of an index patient automatically and predict diagnoses according to his neighborhood. Machine learning algorithms can be harnessed for recommendation of analogous patients in the electronic health records (EHR) datasets and computer-aided diagnoses of chronic diseases and acute events [11].

Predicting diagnoses by examining the numerous clinical data sources to find relevant and useful analogies remains challenging. Related studies on analogy reasoning theory in computer science, such as those on the structure-mapping engine [12], ACME [13], the articulation model [14], the NLP-based method [15] and AN-GAN [16], are very impressive but are still unable to be directly used to measure the similarity between patients by EHR data, which includes unstructured data (e.g., anamnesis), semistructured data (e.g., ICD-10) and structured data (e.g., temperature). The similarity mentioned in these researches is quantitative but not qualitative. However, we are inspired by these theories in psychology and develop a framework for analogical mapping in this research.

From a computer science perspective, the task of diagnostic prediction by using the EHR [17], is a complicated multilabel classification problem that has been studied for more than half of a century [18–21] with methods such as HEME [22], MEDAS [23], and MYCIN [24]. Recently, there has been an increasing number of research studies on diagnostic decision support conducted by developing a patient similarity network paradigm that is naturally interpretable [25,26]. Researchers explored to learn a representation of patient profile by develop a supervised model and solve a constrained optimization problem. Patient similarity was then measured by the geometrical distances between representations, including Euclidean distance [27–29], Hamming distance [30], Manhattan/City Block distance [31], Minkowski distance [32], Cosine distance [33], Correlation distance [34], Mahalanobis distance [28,30,31,35–37], Yule distance [38], Jaccard similarity [39], Dice similarity, Matching similarity [40], distance based on rules [41–43], etc [44]. One of the fundamental difference in this study is that the gold standard in previous researches which applies a supervised machine learning method to determine whether two patients are similar is the incidence of a certain disease, while in our research patient similarity is not dichotomous – it is a continuum. Binary target ignores the fact that patients usually suffer from several diseases and the relationship between multi-morbidities and comorbidities [45]. For example, some diseases occur together more frequently, such as type 1 diabetes and celiac, because of a shared gene that predisposes for these diseases. The prediction models of type 1 diabetes and celiac are not completely unrelated. This is what underlies the difference between our work and the others.

In this paper, inspired by the analogy studies in psychology, we proposed a novel framework to predict diagnoses by using patient similarity. In order to better capture and fuse the semantic information in diagnoses terms and relationship between diseases, we applied a three-step method to compute discharge diagnosis similarity of patient pair and take this decimal value as the outcome of the supervised prediction model. The input of this supervised prediction model is the feature vectors of each patient pair in the cohort, which are built by integrating three parts: symptom similarity, lab test similarity, and preliminary diagnosis similarity. The construction of integrated feature vectors is a process that deriving the attribute similarity. The supervised model is to assign weight to attributes automatically. The prediction of the target is a process that deriving relational similarity. Moreover, we proposed a method to generate diagnoses hypotheses according to Reidhav’s [52] theory that similar patients ought to be diagnosed alike and another method to reject hypotheses according to Reidhav’s theory that dissimilar patients ought not to be diagnosed alike. As our proposed method is actually a general framework, there are several optional targets can be used as the target of the supervised model, such as medication therapy similarity. Extensive experimental results on real-world networks demonstrate that the patient similarity-based model achieves better performances in the diagnostic prediction task.

In conclusion, we make several noteworthy contributions as follows:

- We put forward the idea of a general patient-similarity-based

framework for diagnostic prediction, which is inspired by the structure-mapping theory about analogy reasoning in psychology. The framework is to model human analogy thinking: the attribute similarity is computed at different feature levels; the more complicated relationship similarity is learned by a supervised model.

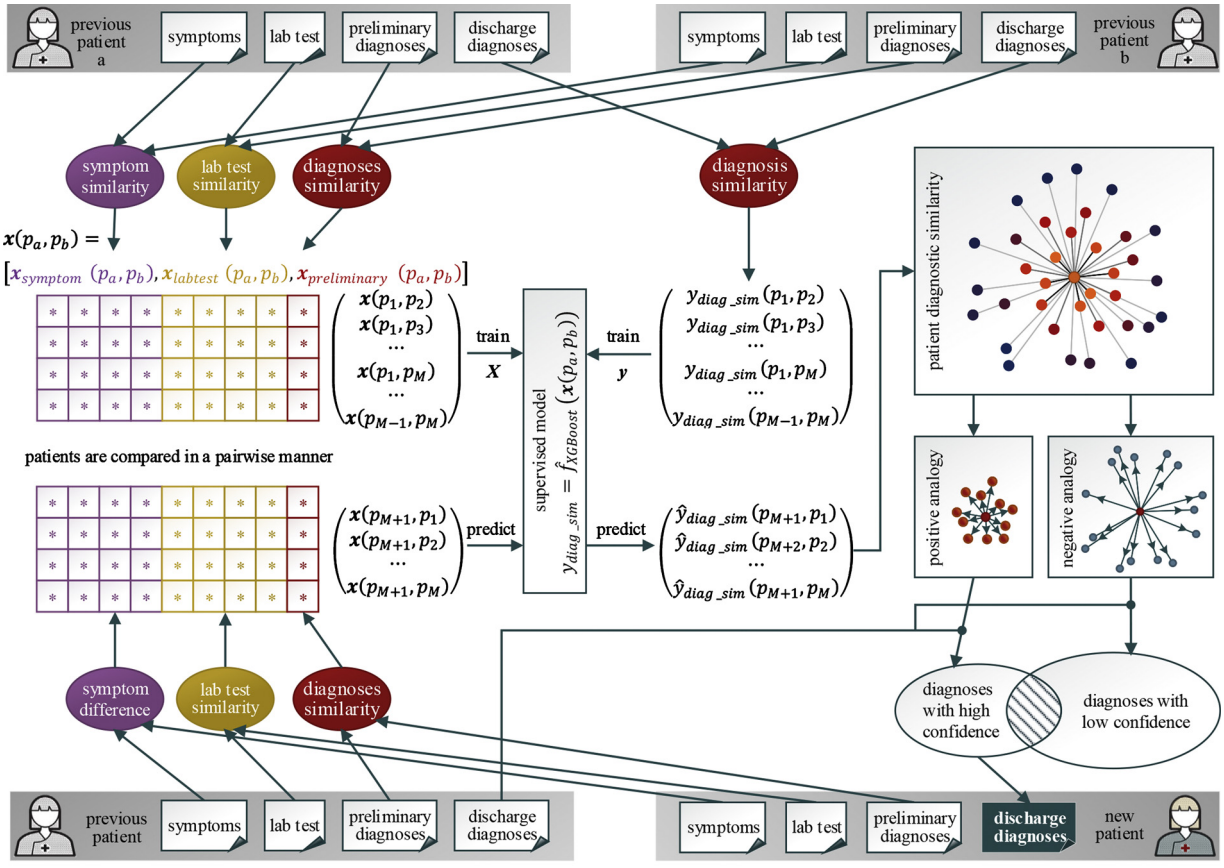
- The measurement of patient similarity by using diagnoses have made it possible to utilize the correlations between diseases to help machine learning models learn in ways that are not possible with existing binary targets alone. In other words, the multilabel classification task is converted into a single-value regression task.
- Rather than attempting to solve the problem of predicting diagnoses just by k-nearest neighbors, we instead explore a two-step method of retrieving positive analogies to generate hypotheses and negative analogies to reject hypotheses.
- Extensive experimental analysis has been conducted on a real-world clinical dataset. The results demonstrate that the proposed model advances the performances on diagnostic prediction task.

The rest of the paper is structured as follows. Section 2 presents the technical details for diagnostic prediction using patient similarity. In Section 3, we empirically evaluate the proposed method on a real-world dataset. We discuss the extensibility, generality, interpretability, limitation and future plan of our work in Section 4. We conclude our work in Section 5.

2. Methods

Gentner, one of the most time-honored prestigious psychologists, claimed in his structure-mapping theory that similarity is like analogy [12,46]. He stated that analogy occurs when comparisons exhibit high degree of relational similarity with little attribute similarity [46]. We are inspired by this theory and convert the analogy identification problem into a similarity measurement problem.

The framework of the method is shown in Fig. 1. Patients are compared in a pairwise manner. The symptoms, lab tests and preliminary diagnoses are integrated into vectors representing the patient pairs. This process is inspired by the attribute similarity in the structure-mapping theory. Clinical features of patient pairs are compared in a pairwise manner, for example, weight vs weight, PLA2R vs PLA2R, vomit P/N vs vomit P/N. The target of the prediction model is the patient similarity defined by the discharge diagnoses. This process is inspired by the relational similarity in the structure-mapping theory. The attribute similarity which limited in specific semantic space are always computable. While the relational similarity which is above attribute similarity and more complicated always require both knowledge and experience to establish. A deep learning model which mimic the human brain was used to learn these complicated relational similarities in this framework. If we only increase the amount of attribute similarity, the comparison will shift towards mere-apperance similarity. After the prediction, similar patients (positive analogy) are used to generate hypotheses of diagnoses, whereas dissimilar patients (negative



$x(p_a, p_b)$ is the feature vector of a patient pair p_a and p_b in the cohort. It is built by integrating three parts: symptom similarity (purple) $x_{symptom}(p_a, p_b)$, lab test similarity (yellow) $x_{labtest}(p_a, p_b)$, and preliminary diagnosis similarity (red) $x_{preliminary}(p_a, p_b)$.

M is the count of patients in the train set. p_{M+1} is a new patient.

The XGBoost model is applied to predict the patient similarity from the input feature vectors: $y_{diag\_sim}(p_a, p_b) = \hat{f}_{XGBoost}(x(p_a, p_b))$

Fig. 1. The framework of the patient-similarity-based approach to predict diagnoses.

analogy) are used to reject hypotheses.

2.1. Attribute similarity

2.1.1. Symptom similarity

A previously reported method [47] by our team that applied a semiautomatic learning-based lexicon construction method to understand clinical documents from nephrology and extract symptoms from Chinese clinical text is used in this study.

First, we construct a dictionary of symptom terms for the nephrology department from clinical documents by using CRF [48]. Then, we use Jieba [49] with the customized symptom dictionary to divide the text into segments. Symptoms are identified, and their negation is detected by applying a finite-state automaton-based negation detection algorithm [50]. The synonyms, i.e., symptoms with similar meanings, are recognized according to the word embedding results of Word2Vec [51]. Manual verification and complementation are also employed to obtain a more precise result.

For the symptom list, $\delta^{(1)}$, $\delta^{(2)}$, ...,$\delta^{(I_1)}$, where I is the total count of distinct symptoms identified from the clinical documents from nephrology, we set $\delta^{(i)} = 1$ if a patient has the i^{th} symptom and 0 otherwise. The symptom similarity of the patients p_a and p_b is defined as

$$\boldsymbol{x}_{symptom}(p_a, p_b) = [\Delta\delta^{(1)}, \Delta\delta^{(2)}, ..., \Delta\delta^{(I)}]$$
$$\Delta\delta^{(i)} = \begin{cases} 1, & \delta_a^{(i)} \neq \delta_b^{(i)} \\ 0, & \delta_a^{(i)} = \delta_b^{(i)} \end{cases}, \ i = 1,2,...,I \tag{1}$$

2.1.2. Lab test similarity

Lab tests and vital signs with numerical results are structured data, e.g., serum creatinine and 24 -h urine protein. The clinical meanings of different values of a certain lab test may not be different. For example, blood urea nitrogen of approximately 3.2–7.1 mmol/L is considered normal.

The Z-score [52] method is implemented to normalize the numerical clinical data. Generally, lab tests with numerical results satisfy a Gaussian distribution $\upsilon \sim N(\mu, \alpha)$. The Z-score is defined as

$$\zeta = \begin{cases} \dfrac{\upsilon - \alpha}{\sigma} - \dfrac{\upsilon_{max} - \alpha}{\sigma}, & if \upsilon > \upsilon_{max} \\ 0, & if \upsilon_{min} \leq \upsilon \leq \upsilon_{max} \\ \dfrac{\upsilon_{min} - \alpha}{\sigma} - \dfrac{\upsilon - \alpha}{\sigma}, & if \upsilon < \upsilon_{min} \end{cases} \tag{2}$$

where α is the mean, σ is the standard deviation, and υ_{max} and υ_{min} are the upper and lower thresholds of the reference interval, respectively. Note that a lab test may have multiple reference intervals for a condition that it relates to sex, age, time, etc. If ζ falls beyond the range of ± 3 standard deviations, it is marked as an outlier and replaced by the critical value.

Other lab tests with categorical results, e.g., abnormal erythrocyte morphology (spherocyte, elliptocyte, stomatocyte, etc.), are converted to dummy vectors by *one hot encoding*.

The lab test list is $\zeta^{(1)}$, $\zeta^{(2)}$, ...,ζ^{J}, where J is the total count of the lab tests examined in the nephrology department. In practice, the list is sparse since patients usually only undergo some of the examinations and lab tests and skip the others. Lab test similarity is defined as the subtraction of the corresponding numerical elements in the vectors of pairwise patients. We can write

$$\boldsymbol{x}_{labtest}(p_a, p_b) = [\Delta\zeta^{(1)}, \Delta\zeta^{(2)}, ..., \Delta\zeta^{(J)}] \tag{3}$$

$$\Delta\zeta^{(j)} = \begin{cases} |\zeta_a^{(j)} - \zeta_b^{(j)}|, & if \zeta_a^{(j)} and \zeta_b^{(j)} \neq null \\ null, & if \zeta_a^{(j)} or \zeta_b^{(j)} = null \end{cases}, j = 1,2, ...,J$$

2.1.3. Diagnosis similarity

Preliminary diagnoses are established after hospital admission and recorded in the admission notes. They are tentative hypotheses and may be not satisfactory. Discharge diagnoses are achieved after treatment, confirmed by the doctors and extracted from the discharge summary notes. They are accurate and definite. Preliminary and discharge diagnoses are categorical and unintelligible for a computer. Fortunately, many categorical clinical concepts have been classified into hierarchical taxonomies, such as ICD-10. Preliminary diagnoses recorded in the admission notes are not encoded with ICD-10. A hybrid approach [53] is used to assign ICD-10 to preliminary diagnoses automatically. Since the quantity of taxonomic codes is excessively large, we aggregate the codes into a manageable quantity according to its hierarchy [54].The preliminary diagnosis similarity and discharge diagnosis similarity are computed according to hierarchical taxonomic ICD-10 codes. D_a and D_b are the sets of taxonomic concepts of patients p_a and p_b, respectively, where $d_a^{(i)} \in D_a$ and $d_b^{(j)} \in D_b$. The similarity of ICD-10 codes between two patients is computed as follows [43,55,56].

$$InformationContent: IC(d) = -\log\frac{\frac{|leaves(d)|}{|subsumers(d)|} + 1}{|leaves(r)| + 1} \tag{4}$$

$$CodeLevelSimilarity: CS(d_a, d_b) = 1 - \frac{2IC(d_c)}{IC(d_a) + IC(d_b)} \tag{5}$$

$$SetLevelSimilarity: SS(D_a, D_b) = \frac{\sum_{d_a \in D_a} \min_{d_b \in D_b} CS(d_a^{(i)}, d_b^{(j)}) + \sum_{d_b \in D_b} \min_{d_a \in D_a} CS(d_a^{(i)}, d_b^{(j)})}{\|D_a\| + \|D_b\|} \tag{6}$$

where r represents the root in the ICD-10 taxonomy and d_c is the least common ancestor of ICD-10 code d_a and d_b.

The preliminary diagnosis similarity equals the set level similarity of two sets of ICD-10 codes of preliminary diagnoses

$$\boldsymbol{x}_{preliminary}(p_a, p_b) = SS(D_{a,preliminary}, D_{b,preliminary}) \tag{7}$$

The discharge diagnosis similarity is

$$y_{diag\_sim}(p_a, p_b) = SS(D_{a,discharge}, D_{b,discharge}) \tag{8}$$

2.2. Prediction model for relational similarity

The feature vectors of each patient pair in the cohort are built by integrating three parts: symptom similarity, lab test similarity, and preliminary diagnosis similarity. A simplified virtual patients example of this process is detailed in the supplementary file.

$$x(p_a, p_b) = [\boldsymbol{x}_{symptom}(p_a, p_b), \boldsymbol{x}_{labtest}(p_a, p_b), \boldsymbol{x}_{preliminary}(p_a, p_b)] \tag{9}$$

The output is $y_{diag\_sim}(p_a, p_b)$, a continuous value ranging from 0 to 1. The original multiclassification problem is converted to a single-value regression problem.

After comparing features of pairwise patients, the supervised machine learning algorithm is applied to learn how to automatically predict the overall similarity between the patients. The model will learn how to weigh each feature and how much it effects on the predicted outcome. The XGBoost model is selected to address this problem.

$$y_{diag_{sim}}(p_a, p_b) = \hat{f}_{XGBoost}(\boldsymbol{x}(p_a, p_b)) \tag{10}$$

First, the matrix made up of input vectors of pairwise patients is sparse and incomplete. Most symptoms not mentioned in the clinical notes are marked as 0, so 0 entries are frequent in the statistics. One hot encoding creates a binary column for each category and returns a sparse matrix. Not all the lab test items are conducted, which directly contributes to missing values. XGBoost is a sparsity-aware algorithm used for handling sparse data. In each tree node, a default direction is learned and added. Second, the clinical features in the input vector are

permuted randomly. Patient similarity is independent of the permutation. The theoretically justified weighted quantile sketch of XGBoost for approximate learning to find the best split is irrelevant to the permutation of clinical features. Third, the size of the data set is expanded from $O(M)$ to $O(M^2)$, which increases the consumption of time and space. XGBoost designs a column block structure that supports parallel learning to optimize the time complexity. The hyperparameters are tuned by using Hyperopt [57].

2.3. Diagnostic prediction

A simplified model of clinical reasoning has three stages [5]: anamnesis and physical examination, provisional hypotheses generation and hypotheses evaluation and verification/rejection. Every provisional hypothesis must be tested; since a link between the signs and diseases does not necessarily exist, the hypotheses with the highest level of confidence are considered the final diagnoses.

Hesse distinguished two types of analogy relations: positive, negative [58]. Reidhav [59] proposed that the normative arguments from analogy can be subdivided into "normative arguments from positive analogy" and "normative arguments from negative analogy". Positive analogy [60] is a list of accepted propositions about the source domain and represents accepted similarities with the corresponding propositions accepted as holding for the target domain. The presence of negative instances justifies a modification of such an analogically created rule [58,61–63]. For clinical analogy reasoning, we follow Reidhav's research and provide a universal generalization of the clinical positive/negative analogy in the following form (Table 1).

The final prediction result is the weighted votes of the hypotheses mentioned above, and confidence = weight * positive + (1-weight) * negative. Empirically, weight is a linear function of the frequency rank in the diagnosis set in the similarome.

2.4. Baseline methods

One method follows the one-vs-all [64] concept and builds one binary data set for each disease. Patients labeled with a discharge diagnosis are considered positive cases whereas patients without the index disease are considered negative cases. XGBoost is applied for binary classification.

Another two methods are based on k-nearest neighbors using Euclidean similarity and Cosine similarity. The top k_{p_kNN} nearest neighbors are selected to compose a similarome. The top k_{d_kNN} most frequent diagnoses are selected as the prediction results. Missing values are imputed by the multivariate imputation by chained equations (MICE) method [65].

3. Result

3.1. Data preparation

We obtained the EHR dataset from a third-level grade A hospital, Shanxi Dayi Hospital. The data set comprises the data from 2714 inpatients from the nephrology department between July 2014 and February 2017. We built a dataset that included the demographics, admission notes, preliminary diagnoses, lab tests, and discharge diagnoses (with ICD-10 codes). The total number of discharge diagnoses in the cohort is 17634, which includes 1115 diagnoses with distinct 6-char codes. On average, a patient suffers from 6.5 diseases. This result indicates that patients included in this study usually suffer from one or more multimorbidities or comorbidities.

The total number of clinical features changes after data preprocessing. A total of 1335 symptoms decreased to 383 by merging synonyms. A total of 157 lab tests with categorical results increased to 349 by one-hot encoding. The 1115 discharge diagnoses with 6-char codes, which are an extended version of ICD-10, decreased to 894 4-char codes, which are defined as subcategories in ICD-10 [66], by replacing the original code with its parent node in the ICD tree, i.e., the 4-char codes in the upper level. The main difference in processing the diagnoses is that for the baseline method, each patient has a 675-item vector to represent his or her preliminary diagnoses, whereas for the patient-similarity-based method, the preliminary diagnosis similarity score is only 1 number computed according to two sets of ICD-10 codes.

3.2. Performance evaluation criteria

In a multilabel classification task, p_{micro}, r_{micro}, and $f-1_{micro}$ are selected to access the performance of the baseline method and patient-similarity-based model. $\hat{Y}$ is the set of predicted results, and Y is the set of gold standard results.

$$p_{micro}(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|\hat{Y}|} \tag{11}$$

$$r_{micro}(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y|} \tag{12}$$

$$f_{micro}(Y, \hat{Y}) = \frac{2 \times p_{micro} \times r_{micro}}{p_{micro} + r_{micro}} \tag{13}$$

The patient-similarity-based model may fail to predict an individual's diseases if there is no eligible similarome or simotyping because of the rigorous screening criteria, $y_{\theta\_min\_sim}$, k_{p_sim}, f_{θ_sim} and k_{d_sim}. The percentage of patients who have diagnoses that are successfully predicted, named the *success percentage*, is also considered.

Table 1
Positive/negative analogy VS hypotheses generation/rejection.

Positive Analogy	Hypotheses generation
i If two patients are relevantly similar, they ought to be diagnosed alike. ii P1 [source patient] ought to be treated as Q. iii P2 [target patient] is relevantly similar to P1 →*P2 ought to be diagnosed as Q.*	Sort the distance between the present patient p_0 and all the historical patients and exclude the patients who do not exhibit a sufficient level of similarity ($\hat{y}_{diag\_sim}(p, p_0) \geq y_{\theta\_min\_sim}$). The top k_{p_sim} nearest neighbors are selected to compose a similarome [56]. Sort the diagnoses of the patients in the similarome by frequency and exclude the diagnoses that do not have a sufficient frequency (*frequency*$\leq f_{\theta\_sim}$). The top k_{d_sim} most frequent diagnoses are reserved as hypothetical diagnoses.
Negative Analogy i If two patients are relevantly dissimilar, they ought not to be diagnosed alike. ii P1 [source patient] ought to be treated as Q. iii P2 [target patient] is not relevantly similar to P1 →*P2 ought not to be diagnosed as Q.*	**Hypotheses rejection** Sort the distance between the present patient p_0 and all the historical patients and exclude the patients who do not have a sufficient level of dissimilarity ($\hat{y}_{diag\_sim}(p, p_0) \leq y_{\theta\_max\_dissim}$). The top k_{p_dissim} nearest neighbors are selected to compose a dissimilarome (i.e., the group of most dissimilar patients) [56]. Sort the diagnoses of the patients in the dissimilarome by frequency and exclude the diagnoses that do not have a sufficient frequency (*frequency* $\leq f_{\theta\_dissim}$). The top k_{d_dissim} most frequent diagnoses are reserved as rejected diagnoses.

3.3. Prediction result

We report the averaged prediction performance of the baseline methods and our methods over 5 independent runs (Table 2). According to the f_{micro} score, our methods outperformed all the baseline methods where the highest f_micro of baselines was 0.661 while our methods achieved 0.698 (positive analogy only) and 0.703 (positive & negative analogy). The one-vs-all baseline method had the highest $p_{micro=0.885}$. Using "negative analogy" to modify the result of "positive analogy" led to better results considering the f_{micro}.

The negative analogues reject some of the hypotheses without adding new hypotheses. This process may change Ture-Positive case to False-Negative case or change False-Positive to True-Negative. According to the calculation formulas, the precision is improved and the recall reduced from 0.6158 to 0.6155 (approximately 0.616).

The success percentage of our method is low for the following reasons: 1. the hypotheses are strictly selected (the parameters for hypotheses generation and rejection), and 2. the dataset is small (a total of 2714 patients). We tried another stricter group of parameters and observed an overall improvement in results despite a lower success percentage. These results suggest that in a larger dataset, better analogies can be found using stricter criteria.

We contrasted the results of the proposed method and the one-vs-all baseline method and identified five diseases with a marked increase in prediction precision: background retinopathy and retinal vascular changes, hyperparathyroidism, metabolic disorder, cataract in other endocrine, nutritional and metabolic diseases, and disorders of phosphorus metabolism and phosphatases. The ratio of positive samples to negative samples of these diseases ranged from 0.55 % to 0.92 %. This finding suggests that the patient-similarity-based model may be more suitable for unbalanced datasets.

4. Discussion

The input vector proposed in this research is extensible. The patient similarity method naturally handles heterogeneous data when the similarity measurement is well defined. Other similarity measures can be appended to the input vectors, such as the similarity between longitudinal sequences of observable measures [67], images [68], drugs [69], genotypes [70].

The patient-similarity-based model is general. Other targets can be predicted by this framework. In a previous study, we proposed a quantitative measurement of clinical drug–drug similarity and medication therapy similarity [39]. If we set the therapy similarity as the target of the supervised model, the therapy of the new patient can be predicted from his neighbors.

Traditional classification approaches often have limited interpretability. Patient similarity is conceptually analogous to a clinical diagnosis, which often involves a physician relating a new patient to previous patients empirically [26]. In addition, boosting trees are considered interpretable because they provide an explicit decision tree of successive choices used in classification.

This promising work allows an accurate prediction performance while guaranteeing a high success percentage when the data of millions of patients are accessible. Eligible similarome is promising even when the screening criteria are strict. Additional experiments based on a large-scale cohort are expected.

The model proposed in this research is a kind of exhaustive comparison with a high computation cost. To effectively evaluate the similarities among pairwise patients and efficiently retrieve similar patients, large-scale patient indexing [71] is essential to handle massive amounts of data.

Our work is a first step toward applying the decimal patient similarity to the diagnostic prediction domain, and we expect this technique to be also suitable for tasks of prediction that are based on other taxonomical clinical concepts such as drug (ATC), operation (ICD-9-CM-3). More generally, patient similarity network could contribute to the design of more accurate and efficient classifiers, such as whether to perform the renal biopsy, through the construction of a network embedding model which aims to learn low-dimensional vector representations for network nodes that fusing structural information about admission diagnoses and content information about anamnesis.

5. Conclusion

In this paper, we explored the retrieve of analogous patients and prediction of diagnoses by using patient similarity. A method for computationally identifying similar/dissimilar patients and predicting diagnoses from a heterogeneous data set that extends beyond numerical features is proposed. Inspired by the structure-mapping theory, we compare the patients in a pairwise manner, integrate clinical features into a vector, compute the diagnostic similarity, develop a supervised machine learning model, and predict diagnoses by positive/negative analogies. We demonstrate that this method allows us to identify analogous patients and predict diagnoses with better performance than the baselines. The f-1 scores of positive-analogy-based prediction and positive-negative-analogy-based prediction are 0.698, 0.703 respectively, while the f-1 scores of the baselines range from 0.368 to 0.661.

Contributors

ZJ, HD and HL contributed to the conceptualization and design. ZJ, XZ, XL and HL contributed to data acquisition and analysis. ZJ and HL drafted the manuscript. All authors provided feedback and final approval of the manuscript.

Data statement

We obtained the EHR dataset from a third-level grade A hospital, Shanxi Dayi Hospital. The data set comprises the data from 3130 inpatients from the nephrology department between July 2014 and February 2017. We built a dataset that included the demographics, admission notes, preliminary diagnoses, lab tests, and discharge diagnoses (with ICD-10 codes). To assess the pros and cons of each method and enhance credibility of the results, patients with overly incomplete data were removed. After filtering, 2714 patients remained. It is composed of data from 1308 females and 1406 males, and the average age of the patients is 53 years for the overall population. The total number of discharge diagnoses in the cohort is 17634, which includes 1115 diagnoses with distinct 6-char codes. On average, a patient suffers from 6.5 diseases.

The total number of clinical features changes after data preprocessing. A total of 1335 symptoms decreased to 383 by merging synonyms. A total of 157 lab tests with categorical results increased to 349 by one-

Table 2
Performance of the baseline methods and our methods.

Method	p_{micro}	r_{micro}	f_{micro}	*success %*	$y_{\theta\_min\_sim}$
one-vs-all	**0.885**	0.527	0.661	–	–
unsupervised k-NN (Euclidean)	0.723	0.247	0.368	–	–
unsupervised k-NN (Cosine)	0.716	0.280	0.402	–	–
patient similarity (positive analogy only)	0.807	**0.616**	0.698	19%	0.2
patient similarity (positive & negative analogy)	0.819	0.616	**0.703**	19%	0.2
patient similarity (positive analogy only)	**0.912**	**0.632**	**0.747**	7%	0.15

Note: Empirically, $k_{p\_sim} = 14$, $f_{\theta\_sim} = 6$, $k_{d\_sim} = 10$, $y_{\theta\_max\_dissim} = 0.7$, $k_{p\_dissim} = 1000$, $f_{\theta\_dissim} = 5$, weight= −0.2 *rank+ 1.6, $k_{p\_kNN} = 14$, $k_{d\_kNN} = 6$. Bold values of the top 5 rows show the best performing under the $y_{\theta\_min\_sim} = 0.2$. The last row show the better perormance achieved when using a more stritic threshold but lower success rate.

hot encoding. The 1115 discharge diagnoses with 6-char codes, which are an extended version of ICD-10, decreased to 894 4-char codes, which are defined as subcategories in ICD-10 [61], by replacing the original code with its parent node in the ICD tree, i.e., the 4-char codes in the upper level. The main difference in processing the diagnoses is that for the baseline method, each patient has a 675-item vector to represent his or her preliminary diagnoses, whereas for the patient-similarity-based method, the preliminary diagnosis similarity score is only 1 number computed according to two sets of ICD-10 codes.

Author statement

There is no experimentation with human subjects.

The privacy rights of human subjects have been observed.

We acknowledge the support of the Shanxi Dayi Hospital (Shanxi, China) for supplying the anonymized clinical data.

Funding

This work was supported by the National Natural Science Foundation of China (81871456) and the National Key R&D Program of China (2016YFC0901905; 2016YFC0901703).

Summary table

What was already known on the topic:

1 Machine learning algorithms can be harnessed for computer-aided diagnoses.
2 Diagnostic prediction is a multilabel classification problem.
3 Analogy is a manifestation of clinical reasoning which allows the doctors to compare patients, identify similarity and infer diagnoses.

What this study added to our knowledge:

1 Patient similarity can be measured by the ICD-10 codes of diagnoses. It's not dichotomous – it is a continuum.
2 The multilabel classification problem of diagnostic prediction can be converted to a single-value regression problem.
3 The retrieval of analogous patients can be implemented by using a supervised model.
4 Diagnoses can be predicted by retrieving similar patients (positive analogy) to generate diagnostic hypotheses and dissimilar patients (negative analogy) to reject diagnostic hypotheses.

Declaration of Competing Interest

The authors declared no conflict of interest.

Acknowledgments

We acknowledge the support of the Shanxi Dayi Hospital (Shanxi, China) for supplying the anonymized clinical data.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2019.104073.

References

[1] J. Higgs, M.A. Jones, Clinical decision making and multiple problem spaces, Clin. Reason. Health Prof. 3rd ed., Elsevier, Amsterdam, 2008, pp. 3–4.
[2] R.L. Trowbridge, J.J. Rencic, S.J. Durning, American College of Physicians, Introduction/Preface, Teach. Clin. Reason., American College of Physicians, Philadelphia, 2015, pp. xvii–xxii.
[3] D. Dumas, P.A. Alexander, L.M. Baker, S. Jablansky, K.N. Dunbar, Relational reasoning in medical education: patterns in discourse and diagnosis, J. Educ. Psychol. 106 (2014) 1021–1035, https://doi.org/10.1037/a0036777.
[4] D. Gentner, F. Maravilla, Analogical reasoning, Routledge Int. Handb. Think. Reason. Routledge, Abingdon, 2017, pp. 186–203, https://doi.org/10.4324/9781315725697.
[5] N. Guallart, Analogical Reasoning in Clinical Practice, (2014), pp. 257–273, https://doi.org/10.1007/978-3-319-06334-8_15.
[6] K. Dunbar, I. Blanchette, The in vivo/in vitro approach to cognition: the case of analogy, Trends Cogn. Sci. 5 (2001) 334–339, https://doi.org/10.1016/S1364-6613(00)01698-3.
[7] G.P. Pena, J. de Souza Andrade-Filho, Analogies in medicine: valuable for learning, reasoning, remembering and naming, Adv. Health Sci. Educ. Theory Pract. 15 (2010) 609–619, https://doi.org/10.1007/s10459-008-9126-2.
[8] M.L. Gick, K.J. Holyoak, Analogical problem solving, Cognit. Psychol. 12 (1980) 306–355, https://doi.org/10.1016/0010-0285(80)90013-4.
[9] D. Gentner, M.J. Rattermann, K.D. Forbus, The roles of similarity in transfer: separating retrievability from inferential soundness, Cognit. Psychol. 25 (1993) 524–575, https://doi.org/10.1006/cogp.1993.1013.
[10] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44.
[11] C.J. Lynch, C. Liston, New machine-learning technologies for computer-aided diagnosis, Nat. Med. 24 (2018) 1304–1305, https://doi.org/10.1038/s41591-018-0178-4.
[12] D. Gentner, Structure-mapping: a theoretical framework for analogy, Cogn. Sci. 7 (1983) 155–170.
[13] K.J. Holyoak, P. Thagard, The analogical mind, Am. Psychol. 52 (1997) 35.
[14] P. Bartha, By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments, Oxford University Press, New York, 2010.
[15] T. Hope, J. Chan, A. Kittur, D. Shahaf, Accelerating innovation through analogy mining, Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, New York, NY, USA, 2017, pp. 235–243, , https://doi.org/10.1145/3097983.3098038.
[16] Y. Hoshen, L. Wolf, Identifying Analogies Across Domains, (2018) https://openreview.net/forum?id=BkN_r2lR-.
[17] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1419–1428, https://doi.org/10.1093/jamia/ocy068.
[18] R.A. Miller, Medical diagnostic decision support systems–past, present, and future: a threaded bibliography and brief commentary, J. Am. Med. Inform. Assoc. 1 (1994) 8–27.
[19] M.E. Matheny, L. Ohno-Machado, Chapter 11 - generation of knowledge for clinical decision support: statistical and machine learning techniques, in: R.A. Greenes (Ed.), Clin. Decis. Support, second ed., Academic Press, Oxford, 2014, pp. 309–337.
[20] R.A. Jenders, Advances in clinical decision support: highlights of practice and the literature 2015–2016, Yearb. Med. Inform. 26 (2017) 125–132.
[21] E.H. Shortliffe, B.G. Buchanan, E.A. Feigenbaum, Knowledge engineering for medical decision making: a review of computer-based clinical decision aids, Proc. IEEE 67 (1979) 1207–1224.
[22] M. Lipkin, J.D. Hardy, Differential diagnosis of hematologic diseases aided by mechanical correlation of data, Science 125 (1957) 551–552.
[23] M. Ben-Bassat, R.W. Carlson, V.K. Puri, M.D. Davenport, J.A. Schriver, M. Latif, R. Smith, L.D. Portigal, E.H. Lipnick, M.H. Weil, Pattern-based interactive diagnosis of multiple disorders: the MEDAS system, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1980) 148–160.
[24] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system, Comput. Biomed. Res. Int. J. 8 (1975) 303–320.
[25] E. Parimbelli, S. Marini, L. Sacchi, R. Bellazzi, Patient similarity for precision medicine: a systematic review, J. Biomed. Inform. 83 (2018) 87–96, https://doi.org/10.1016/j.jbi.2018.06.001.
[26] S. Pai, G.D. Bader, Patient similarity networks for precision medicine, J. Mol. Biol. 430 (2018) 2924–2938.
[27] F. Wang, J. Sun, PSF: a unified patient similarity evaluation framework through metric learning with weak supervision, IEEE J. Biomed. Health Inform. 19 (2015) 1053–1060, https://doi.org/10.1109/JBHI.2015.2425365.
[28] B. Yang, S. Liu, S. Pang, C. Pang, X. Shang, Deep subspace similarity fusion for the prediction of cancer subtypes, 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM (2018) 566–571, https://doi.org/10.1109/BIBM.2018.8621247.
[29] S. Pai, S. Hui, R. Isserlin, M.A. Shah, H. Kaka, G.D. Bader, netDx: interpretable patient classification using integrated patient similarity networks, Mol. Syst. Biol. 15 (2019) e8497, https://doi.org/10.15252/msb.20188497.
[30] J. Lee, J. Sun, F. Wang, S. Wang, C.-H. Jun, X. Jiang, Privacy-preserving patient similarity learning in a federated environment: development and analysis, JMIR Med. Inform. 6 (2018) e20, https://doi.org/10.2196/medinform.7744.
[31] J. Chen, L. Sun, C. Guo, W. Wei, Y. Xie, A data-driven framework of typical treatment process extraction and evaluation, J. Biomed. Inform. 83 (2018) 178–195, https://doi.org/10.1016/j.jbi.2018.06.004.
[32] A. Labellapansa, A. Efendi, A. Yulianti, E.A. Kadir, Lambda value analysis on Weighted Minkowski distance model in CBR of Schizophrenia type diagnosis, 2016 4th Int. Conf. Inf. Commun. Technol. ICoICT (2016) 1–4, https://doi.org/10.1109/ICoICT.2016.7571898.
[33] A. Tashkandi, I. Wiese, L. Wiese, Efficient in-database patient similarity analysis for

personalized medical decision support systems, Big Data Res. 13 (2018) 52–64, https://doi.org/10.1016/j.bdr.2018.05.001.
[34] A. Singh, B. Pandey, Diagnosis of liver disease using correlation distance metric based k-nearest neighbor approach, Intell. Syst. Technol. Appl. 2016, Springer, Cham, 2016, pp. 845–856, https://doi.org/10.1007/978-3-319-47952-1_67.
[35] J. Sun, F. Wang, J. Hu, S. Edabollahi, Supervised patient similarity measure of heterogeneous patient records, SIGKDD Explor. 14 (2012) 16–24.
[36] F. Wang, J. Sun, Survey on distance metric learning and dimensionality reduction in data mining, Data Min. Knowl. Discov. 29 (2014) 534–564, https://doi.org/10.1007/s10618-014-0356-z.
[37] F. Vitali, S. Marini, D. Pala, A. Demartini, S. Montoli, A. Zambelli, R. Bellazzi, Patient similarity by joint matrix trifactorization to identify subgroups in acute myeloid leukemia, JAMIA Open 1 (2018) 75–86, https://doi.org/10.1093/jamiaopen/ooy008.
[38] X. Xu, R. Mazloom, A. Goligerdian, J. Staley, M. Amini, G.J. Wyckoff, J. Riviere, M. Jaberi-Douraki, Making sense of Pharmacovigilance and drug adverse event reporting: comparative similarity association analysis using AI machine learning algorithms in dogs and cats, Top. Companion Anim. Med. (2019) 100366, https://doi.org/10.1016/j.tcam.2019.100366.
[39] X. Zeng, Z. Jia, Z. He, W. Chen, X. Lu, H. Duan, H. Li, Measure clinical drug–drug similarity using Electronic Medical Records, Int. J. Media Inf. Lit. 124 (2019) 97–103, https://doi.org/10.1016/j.ijmedinf.2019.02.003.
[40] Y. Rivault, N.L. Meur, O. Dameron, A Similarity Measure Based on Care Trajectories as Sequences of Sets, Artif. Intell. Med., Springer, Cham, 2017, pp. 278–282, https://doi.org/10.1007/978-3-319-59758-4_32.
[41] A.L. Pineda, A. Pourshafeie, A. Ioannidis, C.M. Leibold, A. Chan, J. Frankovich, C.D. Bustamante, G.L. Wojcik, Medal: a patient similarity metric using medication prescribing patterns, MedRxiv (2019) 19004440, , https://doi.org/10.1101/19004440.
[42] D. Girardi, S. Wartner, G. Halmerbauer, M. Ehrenmüller, H. Kosorus, S. Dreiseitl, Using concept hierarchies to improve calculation of patient similarity, J. Biomed. Inform. 63 (2016) 66–73, https://doi.org/10.1016/j.jbi.2016.07.021.
[43] Z. Jia, X. Lu, H. Duan, H. Li, Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity, BMC Med. Inform. Decis. Mak. 19 (91) (2019), https://doi.org/10.1186/s12911-019-0807-y.
[44] Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert, A survey of binary similarity and distance measures, (n.d.).
[45] D. Riaño, W. Ortega, Computer technologies to integrate medical treatments to manage multimorbidity, J. Biomed. Inform. 75 (2017) 1–13.
[46] D. Gentner, A.B. Markman, Structure mapping in analogy and similarity, Am. Psychol. 52 (1997) 45–56, https://doi.org/10.1037/0003-066X.52.1.45.
[47] Y. Lu, Z. Jia, X. Zeng, C. Feng, X. Lu, H. Duan, H. Li, Renal biopsy recommendation based on text understanding, Stud. Health Technol. Inform. 264 (2019) 689–693, https://doi.org/10.3233/SHTI190311.
[48] M. Ju, H. Duan, H. Li, A CRF-based method for automatic construction of Chinese symptom lexicon, 2015 7th Int. Conf. Inf. Technol. Med. Educ. ITME (2015) 5–8, https://doi.org/10.1109/ITME.2015.90.
[49] S. Junyi, Jieba. Contribute to fxsjy/jieba development by creating an account on GitHub, (2018) https://github.com/fxsjy/jieba.
[50] Z. Jia, H. Li, M. Ju, Y. Zhang, Z. Huang, C. Ge, H. Duan, A finite-state automata based negation detection algorithm for Chinese clinical documents, 2014 IEEE Int. Conf. Prog. Inform. Comput. (2014) 128–132.
[51] Y. Goldberg, O. Levy, Word2Vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method, ArXiv E-Prints 1402 (2014) arXiv:1402.3722.
[52] A. Ghasemi, S. Zahediasl, Normality tests for statistical analysis: a guide for non-statisticians, Int. J. Endocrinol. Metab. 10 (2012) 486–489.
[53] Z. Jia, W. Qin, H. Duan, X. Lv, H. Li, A hybrid method for ICD-10 auto-coding of Chinese diagnoses, Stud. Health Technol. Inform. 245 (2017) 427–431.
[54] L. Li, W.-Y. Cheng, B.S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E.P. Bottinger, J.T. Dudley, Identification of type 2 diabetes subgroups through topological analysis of patient similarity, Sci. Transl. Med. 7 (2015) 311ra174.
[55] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, Knowledge Based Syst. 24 (2011) 297–303.
[56] R. Dogaru, F. Micota, D. Zaharie, Taxonomy-based dissimilarity measures for profile identification in medical data, 2015 IEEE 13th Int. Symp. Intell. Syst. Inform. SISY (2015) 149–154.
[57] GitHub - hyperopt/hyperopt: Distributed Asynchronous Hyperparameter Optimization in Python, (n.d.). https://github.com/hyperopt/hyperopt (Accessed 13 May 2019).
[58] M.B. Hesse, Models and Analogies in Science, University of Notre Dame Press, Notre Dame, 1966.
[59] D. Reidhav, Reasoning by Analogy – A Study on Analogy-based Arguments in Law, thesis/docmono, Lund University, 2007, http://lup.lub.lu.se/record/27171.
[60] J.M. Keynes, A Treatise on Probability, 1921, 8 Collect. Writ, (1973).
[61] Å. Frä ndberg, On analogical use of legal rules, in: Å. Frä ndberg (Ed.), Leg. Order Stud. Found. Juridical Think, Springer International Publishing, Cham, 2018, pp. 153–188, , https://doi.org/10.1007/978-3-319-78858-6_9.
[62] M. Hesse, Analogy and confirmation theory, Philos. Sci. 31 (1964) 319–327, https://doi.org/10.1086/288017.
[63] L.J.A. Juthe, Argumentation by Analogy: a Systematic Analytical Study of an Argument Scheme, (2016) https://dare.uva.nl/search?metis.record.id=534988.
[64] A. Rocha, S.K. Goldenstein, Multiclass from binary: expanding one-versus-All, one-versus-One and ECOC-Based approaches, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 289–302.
[65] M.J. Azur, E.A. Stuart, C. Frangakis, P.J. Leaf, Multiple Imputation by Chained Equations: what is it and how does it work? Int. J. Methods Psychiatr. Res. 20 (2011) 40–49, https://doi.org/10.1002/mpr.329.
[66] World Health Organization, International Statistical Classification of Diseases and Related Health Problems, 10th edition, World Health Organization, 2004.
[67] H. Yin, X. Zhu, S. Ma, S. Yang, L. Chen, A novel similarity comparison approach for dynamic ECG series, Biomed. Mater. Eng. 26 (Suppl 1) (2015) S1095–1105.
[68] Z. Wang, J. Xin, Y. Huang, C. Li, L. Xu, Y. Li, H. Zhang, H. Gu, W. Qian, A similarity measure method combining location feature for mammogram retrieval, J. Xray Sci. Technol. 26 (2018) 553–571.
[69] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Towards personalized medicine: leveraging patient similarity and drug similarity analytics, AMIA Summits Transl. Sci. Proc. (2014) 132–136.
[70] Q. Zou, J. Li, L. Song, X. Zeng, G. Wang, Similarity computation strategies in the microRNA-disease network: a survey, Brief. Funct. Genomics 15 (2016) 55–64.
[71] F. Wang, Adaptive semi-supervised recursive tree partitioning: the ART towards large scale patient indexing in personalized healthcare, J. Biomed. Inform. 55 (2015) 41–54.