Artificial Intelligence In Medicine 107 (2020) 101915

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

CCAE: Cross-field categorical attributes embedding for cancer clinical endpoint prediction

Youru Li[a,b], Zhenfeng Zhu[a,b,*], Haiyan Wu[c], Silu Ding[d], Yao Zhao[a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
[c] Department of Otorhinolaryngology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing, China
[d] Department of Radiation Oncology, The First Hospital of China Medical University, Shenyang, China

ARTICLE INFO

Keywords:
Clinical endpoint prediction
Electronic health records
Categorical variables embedding

ABSTRACT

Patients with advanced cancer are burdened physically and psychologically, so there is an urgent need to pay more attention to their health-related quality of life (HRQOL). With an expected clinical endpoint prediction, over-treatment can be effectively eliminated by the means of palliative care at the right time. This paper develops a deep learning based approach for cancer clinical endpoint prediction based on patient's electronic health records (EHR). Due to the pervasive existence of categorical information in EHR, it brings unavoidably obstacles to the effective numerical learning algorithms. To address this issue, we propose a novel cross-field categorical attributes embedding (CCAE) model to learn a vectorized representation for cancer patients in attribute-level by orders, in which the strong semantic coupling among categorical variables are well exploited. By transforming the order-dependency modeling into a sequence learning task in an ingenious way, recurrent neural network is adopted to capture the semantic relevance among multi-order representations. Experimental results from the SEER-Medicare EHR dataset have illustrated that the proposed model can achieve competitive prediction performance compared with other baselines.

1. Introduction

In advanced cancer treatments, the patient's HRQOL is as important as symptomatic treatment [1]. As a patient-led treatment, palliative care focuses on providing relief from the symptoms, pain, physical stress, and mental stress at any stage of illness to improve the quality of life for cancer patients [2]. Cancer clinical endpoint prediction can estimate the expected treatment effect for patients with advanced cancer, which will be great beneficial to determining the time for providing hospice care and reduce over-treatment effectively [3–6]. In addition, doctors can get more scientific decision-making basis and provide personalized treatment for patient optimally [7].

In general, cancer clinical endpoint prediction is based on patient's basic demographic attributes information, tumor's lesion condition and treatment options (graphical illustration in Fig. 1) [8]. Unlike traditional numerical structured data, cancer patient's EHR usually contains a large amount of categorical information which cannot be directly manipulated per algebraic operations [9], which will unavoidably bring obstacles to the application of effective numerical learning algorithms such as deep learning methods [10,11]. In addition, there is strong semantic coupling among multiple categorical variables [12]. Taking a cancer patient's EHR as an example, it may be obvious that the value 'female' of feature gender in demographic attributes field is highly coupled with the values 'breast' of feature in tumor's lesion condition field. In fact, such case is much popular in cancer patient's EHR. Consequently, how to learn effectively the vectorized representations for categorical variables by utilizing the characteristics of semantic coupling among categorical variables is still a key issue to be solved.

Traditionally, learning representation for categorical data is usually to map discrete attribute value to a numerical vector [13,14]. As one of the most commonly used encoding-based methods, the one-hot encoding [15] can be easily implemented but with weaknesses of high dimensionality and sparseness when handing large number of categorical values. Label Encoding is another way to encode numerical discrete or unstructured texts to numerical label. However, the partial order noise will be introduced into the disordered variables. The IDF based encoding method [16] is popularly used for information retrieval which can learn representation by frequency-inverse document frequency measures, but fails to mine the semantic coupling among categorical variables. Despite that the entity embedding method [17] can

* Corresponding author at: Institute of Information Science, Beijing Jiaotong University, Beijing, China.
E-mail address: zhfzhu@bjtu.edu.cn (Z. Zhu).

https://doi.org/10.1016/j.artmed.2020.101915
Received 2 January 2020; Received in revised form 19 June 2020; Accepted 23 June 2020
Available online 26 June 2020
0933-3657/ © 2020 Elsevier B.V. All rights reserved.

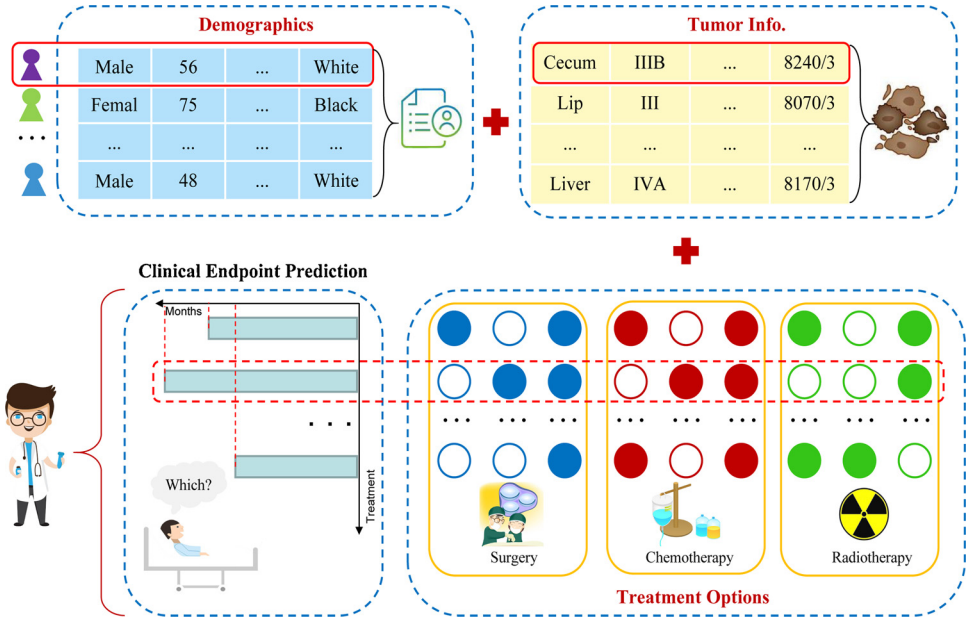

Fig. 1. A demonstration of the significance of cancer clinical endpoint prediction. With the patient's demographics and tumor information, the cancer clinical endpoint corresponding to different treatment options can be predicted respectively. Furthermore, these prediction results can be used by doctors to provide personalized treatment options which can make less over-treatment but better HRQOL for patients.

map similar values close to each other in the embedding space by a neural network during the standard supervised learning, such improvement is limited in capturing with cross-field data. In addition, with the implementation of ordered boosting, CatBoost [18], a widely used method for categorical features processing can achieve better performance compared to some tree-based methods [19,20].

To overcome the challenge to learn vectorized representations for categorical data with strong semantic coupling, a novel cross-field categorical attributes embedding (CCAE) model is proposed. Specifically, to construct a multi-order corpus for each individual categorical variable in EHR, the frequency of the co-occurrence combination is exploited to form a ranking of them by the orders of cross-fields interaction. Based on the obtained corpus, we can learn the embedding-based representations for each categorical variable. Furthermore, RNN is adopted to capture the semantic relevance among multi-order representations.

In general, our main contributions in this paper are:

- A generalized multi-order corpus with co-occurrence combination is constructed by exploiting the semantic coupling among categorical variables.
- A novel model named CCAE is proposed to learn the embedding-based representation from categorical variables by the multiple orders of cross-fields interaction.
- By transforming the order-dependency modeling into a sequence learning task in an ingenious way, recurrent neural network is adopted to capture the semantic relevance among multi-order representations.

2. Preliminaries

Let's first give some notations and formulations used in this paper. For all categorical variables given in EHR, we divide these into three fields (A, B and C) according to the demographics, tumor and treatments. For each patient $m = 1, 2, \ldots, M$, it can be represented as $X_m = (x_m^A, x_m^B, x_m^C) \in \mathbb{R}^{N_A+N_B+N_C}$, where M denotes the size of samples, categorical variables are denoted by $x_m^A(i) \in \{A_j^i | j = 1,\ldots,N_{A_i};\ i = 1,\ldots,N_A\}$, $x_m^B(i) \in \{B_j^i | j = 1,\ldots,N_{B_i};\ i = 1,\ldots,N_B\}$, $x_m^C(i) \in \{C_j^i | j = 1,\ldots,N_{C_i};\ i = 1,\ldots,N_C\}$, N_A, N_B and N_C denote the number of attributes in each field respectively, and N_A, N_B and N_C are the number of values in attributes i for each field. Furthermore, as an representation learning problem for categorical variables with strong semantic coupling, we define the concept of 'order' as the number of cross-fields to better illustrate how we deal with the correlation among variables.

Typically, for a patient m, the patient-level embedding-based representation in order r can be defined as $e^r(X_m) = e^r(x_m^A, x_m^B, x_m^C) \in \mathbb{R}^{(N_A+N_B+N_C)\times d}$, where d is the dimension of the learned representation for each categorical variable. Generally, we learn a nonlinear mapping function to make a prediction $\tilde{y}_m$ with $e^r(X_m)$, $r = 1, 2, 3$ by the following formulation:

$$\tilde{y}_m = f(e^1(X_m), e^2(X_m), e^3(X_m)) \quad (1)$$

where $f(\cdot)$ is the nonlinear mapping function we take for making prediction.

3. Methodology

In this section, we will present the proposed model in detail. The overview of the framework for CCAE is shown in Fig. 2, as we can see, it mainly consists of learning vectorized representation for categorical variables and modeling order-dependency among multi-order representations.

3.1. Cross-field categorical attributes embedding

Usually, cancer patient's EHR are represented by categorical variables with strong semantic coupling. In this part, we will introduce how to utilize the multi-order co-occurrence relationship of attributes to encode categorical variables into special language pattern, and then use the embedded-based method to learn the numerical representation in attribute-level for each patient with the encoded pattern.

3.1.1. Field-awareness co-occurrence encoding

To get more explicit information from the limited implicit information is an important issues to address. In order to better explore the interaction of categorical features in EHR, we divided attributes into three fields: demographics, tumor and treatments. In fact, not all attributes in the original EHR are categorical. Therefore, in order to learn the interaction between attributes more effectively, we discretize the continuous variables into equal-sized buckets based on rank or sample quantiles. Specifically, age, tumor size and extension in EHR are the continuous variables. We transformed the age of patients into discrete groups every 5 years, such as 20–25, 25–30. For tumor size and

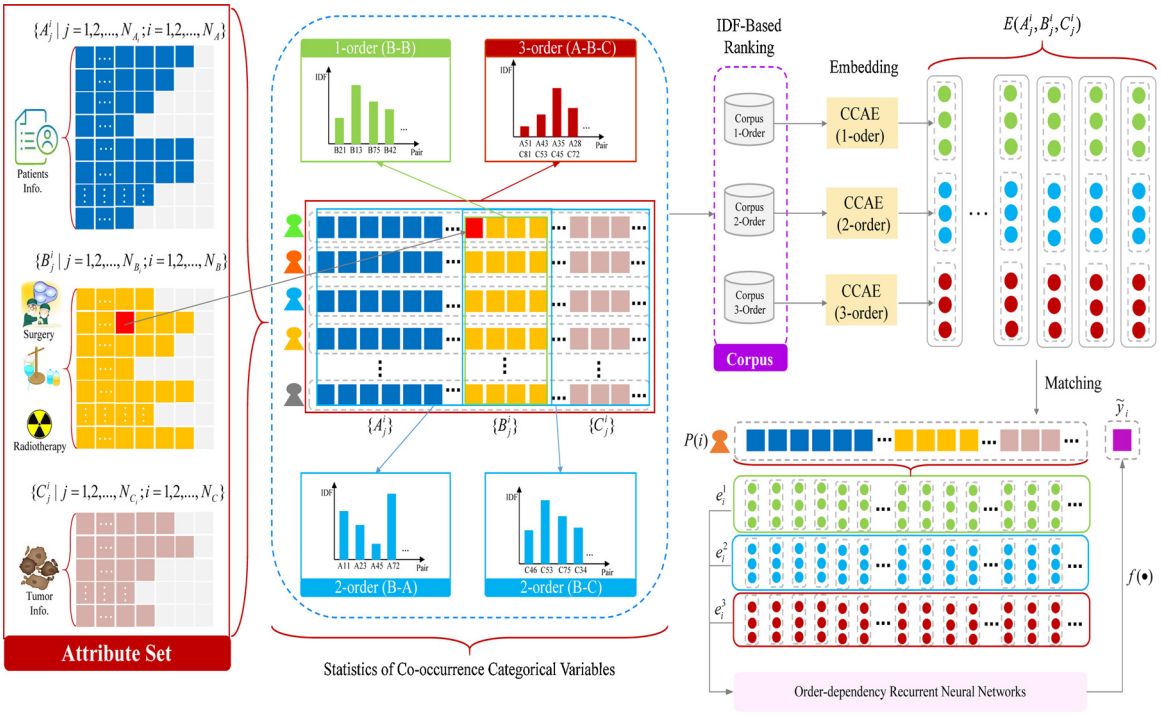

Fig. 2. Graphical illustration of learning representation from data with strong semantic coupling between cross-field categorical variables recorded in the EHR. For all categorical variables in EHR, we divide it into three fields according to the patient's demographics information, tumor information, and treatments information. According to different orders of attributes in EHR, the IDF-based ranked corpus can be constructed for categorical embedding by co-occurrence statistics. Thus, a patient-level representation is obtained by attribute matching of the learned multi-ordered representation to the corresponding attribute for each patient. Furthermore, an order-dependency recurrent neural network is introduced to model the order-dependency among multi-order representations, as well as make the prediction of cancer clinical endpoint.

extension, we use the principle of consistent frequency to group the samples. For example, as categorical attribute $X_m^B(i)$ in the tumor field, its first-order, second-order and third-order co-occurrence combination can be respectively defined as $\{X_m^B(i) \times X_m^B(j)\}$, $\{X_m^A(i) \times X_m^B(j), X_m^B(i) \times X_m^C(j)\}$, $\{X_m^A(i) \times X_m^B(j) \times X_m^C(k)\}$, where i, j, $k = 1, 2, \ldots, N_A + N_B + N_C$ and $i \neq j \neq k$. For an instance, in Fig. 2, the co-occurrence combination $B21$ can be produced by $\{X_m^B(2) \times X_m^B(1)\}$. Further, we also defined a list of associated attributes corresponding to an attribute as a set of co-occurrence combinations of all different orders. However, since the co-occurrence combination of each attribute value may contain a large number of elements and introduce noise redundancy (some of them are strongly correlated, others are not), we will face great difficulties when we attempt to describe the context of attribute values by using the information reflected in the associated attribute list. So, the effect of noise can be reduced effectively, if the lists are sorted in a certain way and done to simplify. Referring to the skills of processing word frequency, we have proposed a clever way to solve the above issue. Specifically, the associated attribute list can be treated as an independent document d_j, and the corresponding inverse document frequency for each attributes idf_i value can also be calculated separately as follow:

$$\mathrm{idf}_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \tag{2}$$

where $|D|$ is the total number of associated attributes list and : means the number of list containing current attribute t_i.

Since the inverse document frequency can filter out more important attribute words, we can select the Top K association attribute and sort them according to importance to get a sequence of ordered attribute descriptions in different order for each attribute. It should be noted that the sequence of the ordered association attribute corresponding to each attribute is completed in field order respectively.

3.1.2. Attribute-level embedding

The core of the learning embedded-based representation of categorical attributes is the construction of generalized corpus. The corresponding attribute mentioned above with sequence of the ordered associated attribute can be seen as a generalized corpus used for representation learning by CCAE, outlined in Algorithm 1. In addition, the sequence of associated attributes sorted by importance can be used to represent the contextual information of the current attribute semantics which is a good indicator of the semantic coupling among the categorical variables. Furthermore, we can learn the vectorized representation of the ordered associated context series used to describe the current attribute by introducing methods of sentences embedding [21]. Specifically, given a general ordered associated context series in order r: $L_i^r = (\ldots, \underbrace{C_{k-l}, \ldots, C_k, \ldots, C_{k+l}}_{2l+1}, \ldots)$, the objective of CCAE is to maximize the average log probability

$$\frac{1}{K} \sum_{t=l}^{K-l} \log p(C_k | C_{k-l}, \ldots, C_{k+l}) \tag{3}$$

The prediction task is typically done via a multi-class classifier, such as softmax. There, we have

$$p(C_k | C_{k-l}, \ldots, C_{k+l}) = \frac{e^{y_{C_{k-l}}}}{\sum_j e^{y_j}} \tag{4}$$

Each of y_j is un-normalized log-probability for each selected association attribute j, computed as

$$y = b + \mathrm{Uh}(C_{k-l}, \ldots, C_{k+l}; C) \tag{5}$$

where U, b are the softmax parameters. h is constructed by a concatenation of intermediate vectors extracted from the associated context series L^r in order r and the index vector extracted from I. In addition, we take stochastic gradient decent (SGD) to train the CCAE and the gradient obtained by back propagation can be used to update

parameters in our model.

Algorithm 1. CCAE (C, I, ω, d, τ, k)

Input:
C: Categorical attribute set, I: Index of each attribute, ω: window size, d: embedding size, τ: training epochs, K: Length of selected association attribute list
Output:
Categorical embedded representation in multi-order:
$e^r(C_i) \in \mathbb{R}^d (r = 1, 2, 3;\ i = 1, 2, ..., N_A + N_B + N_C)$
for $C_i \in C$, $i \in I$ **do**
  $\{\text{idf}_i\} \longleftarrow \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$
  // Find co-occurrence combinations by orders
  $\{L_i^r | r = 1, 2, 3\} \longleftarrow \text{GetCoOccur}(C_i, C)$
  for $L_i^r \in L^r$ **do**
// Select TopK combinations by idf_i value
$L_i^r \longleftarrow \text{RankTopK}(L_i^r, K, \{\text{idf}_i\})$
  end for
end for
while $iter = 1 < \tau$ **do**
  Initialization: Sample Θ and Φ from L^r, I
  PV-DM(Θ, Φ, ω, d) [21]
end while

3.2. Order-dependency recurrent networks

Actually, RNN is a neural network in which nodes are connected in a loop, and the internal state of the network can exhibit dynamic timing behavior and is commonly used for temporal modeling. Its advantage is that it can deal with the dependence between sequence variables. For this reason, we transform the order-dependency modeling into a sequence learning task in an ingenious way and adopt RNN to capture the semantic relevance among multi-order representations.

Specifically, we can learn the embedded-based representation from first to third order for each attribute in EHR by the proposed CCAE and the corresponding representation in patient-level are also available. Furthermore, to better model order-dependency among multi-order representations, an recurrent neural network is introduced. We define the process of modeling order-dependency as a generalized time series problem, and the series of r-order ($r = 3$) embedded representation can be defined as:

$$e = (e^1, e^2, e^3) \in \mathbb{R}^{(N_A+N_B+N_C)\times d\times 3} \tag{6}$$

Traditionally, there is a limitation of the fully connected DNNs that the signals of each neurons layer can only be propagated to the next layer, but it is independent between time steps. However, with the feedback loops, recurrent neural networks (RNN) can produce the recurrent connection and model the contextual information for a time series. Then, (e^1, e^2, e^3) can be fed into a standard RNN which computes the hidden vector sequence $h = (h_1, h_2, h_3)$ and output of a single hidden layer RNN $y = (y_1, y_2, y_3)$ by iterating the following equations from $r = 1$ to 3 as follows:

$$h_r = H(W_{\text{ih}} e^r + W_{\text{hh}} h_{r-1} + b_h) \tag{7}$$

$$y_r = O(W_{\text{ho}} h_r + b_o) \tag{8}$$

where the $H(\cdot)$ and $O(\cdot)$ are the activation functions in the hidden layer and the output layer and W_{ih} denotes the input-hidden weight matrix, W_{hh} the hidden-hidden weight matrix, W_{ho} the hidden-output weight matrix, and the b_h denotes hidden bias vector, b_o the output bias vector.

Table 1
Details of SEER research data (1973–2015).

Fields	Attributes	Description
Target value	Survival months	[0, 71]
Demographics	Patient ID	5399 cases
	Sex	Female/male
	Race recode	Black/White/other
	Year of birth	[1917, 2013]
	State	California, etc.
	Marital status	Married, etc.
	Insurance	Insured, etc.
	Age at Dx.	[0, 95]
	Year of Dx.	[2010, 2015]
	Month of Dx.	[1, 12]
Tumor	Site recode	Stomach, etc. (WHO 2008)
	Behavior recode	8340/3, etc. (ICD-O-3)
	Primary site	[1, 750] (ICD-O-3)
	Grade	Grade I, II, etc. (ICD-O-3)
	Derived stage	I, IA, etc. (AJCC)
	Laterality	Left, right, etc.
	CS tumor size	[0,999] (AJCC)
	CS extension	[10,999] (AJCC)
	CS lymph nodes	[0,999] (AJCC)
	CS mets at dx	[0,99] (AJCC)
	Site rec KM	Accidents, etc.
Treatments	Surg Prim Site	[0, 99]
	Surg Reg LN Sur	1–3, etc.
	Surg Oth Reg/Dis	Non-primary, etc.
	RT. sequence	Post-surgical, etc.
	RT. Recode	Beam, etc.
	Chemotherapy	Yes/no (unknown)

4. Experiments

4.1. Data description and settings

Collected by the National Cancer Institute of U.S., the SEER-Medicare[1] (Surveillance, Epidemiology, and End Results), is one of the most representative large-scale cancer registration databases and provides systematic evidence support and valuable first-hand information for clinicians' evidence-based practice and clinical medical research [22,23]. The SEER research data include incidence and population data associated by age, sex, race, year of diagnosis, and geographic areas. SEER collects tumor data on anatomic site, laterality for paired organs, size, and histopathological type which is based on the 2000 International Classification of Diseases for Oncology version 3 or ICD-O-3 [24]. Moreover, SEER also collects surgical management, radiation therapy and chemotherapy data relating to the first course of treatment are extracted from health records. Specifically, the program records the type of radiation therapy and whether delivery was neoadjuvant, adjuvant or intraoperative and data on chemotherapy use (yes, no or unknown) may also be assessed with a specific request. It releases new research data every Spring based on the previous November's submission of data [25] and we use the 1973–2015 SEER research data with additional treatment fields as experimental data.

1

It should be noted that the data we selected only contains non-autopsy records with positive histology, complete survival month and valid follow-up and the CONSORT diagram is shown in Fig. 3. In addition, since the data records the time span of the patient from diagnosis to death in months, this paper achieves a clinical endpoint prediction with monthly granularity and the detailed statistical items for continuous variable of input and target values are given in Tables 2–4. Furthermore, to better analyze the distribution of survival months corresponding to different site and AJCC stage groups, some examples are also shown in Figs. 4–8.

[1] https://seer.cancer.gov/data/

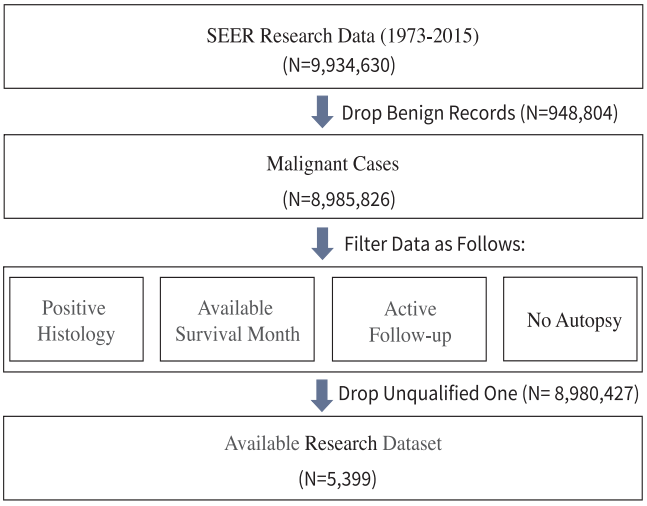

Fig. 3. CONSORT diagram.

Table 2
Detailed statistical items for continuous variable of input for training set.

Statistical items	CS tumor size (/mm)	CS extension (/mm)	CS lymph nodes	CS mets at dx	Age at Dx.
Count	4319	4319	4319	4319	4319
Mean	185.95	442.86	119.32	13.26	61.30
Std	327.82	205.90	167.10	19.90	13.37
Min	0.00	10.00	0.00	0.00	0.00
25%	30.00	300.00	0.00	0.00	52.00
50%	50.00	450.00	100.00	0.00	62.00
75%	90.00	600.00	200.00	26.00	72.00
Max	999.00	999.00	999.00	99.00	95.00

Table 3
Detailed statistical items for continuous variable of input for testing set.

Statistical items	CS tumor size (/mm)	CS extension (/mm)	CS lymph nodes	CS mets at dx	Age at Dx.
Count	1080	1080	1080	1080	1080
Mean	195.86	439.57	125.34	13.16	61.20
Std	339.74	209.14	178.58	19.76	13.57
Min	0.00	10.00	0.00	0.00	0.00
25%	30.00	300.00	0.00	0.00	52.00
50%	50.00	450.00	100.00	0.00	62.00
75%	90.00	595.00	200.00	26.00	72.00
Max	999.00	999.00	999.00	99.00	95.00

Table 4
Detailed statistical items for target values (survival months) in dataset.

Statistical items	Train/valid	Test	Variation
Count	4319	1080	–
Mean	21.14	21.50	0.36
Std	16.15	16.14	0.01
Min	0.00	0.00	0.00
25%	7.00	8.00	1.00
50%	18.00	18.00	0.00
75%	33.00	33.00	0.00
Max	71.00	71.00	0.00

4.2. Parameters settings

There are some parameters in CCAE, i.e., length of selected association attribute list K, embedding dimension d, sampling window size ω and epochs τ. Taking into account efficiency and performance, the setting is: $K = 100$, $d = 3$, $\omega = 3$, $\tau = 50$. In addition, we transform the

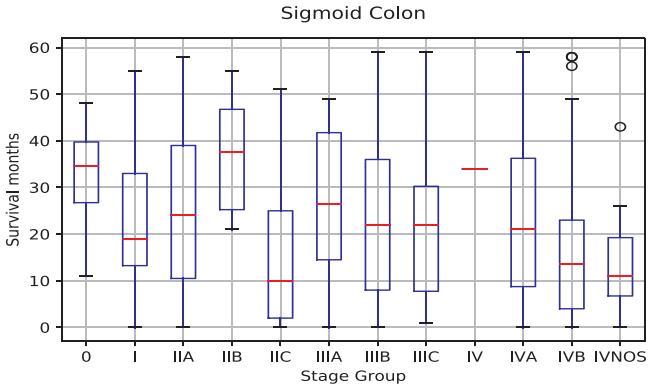

Fig. 4. The boxplot of distribution of survival months (sigmoid colon).

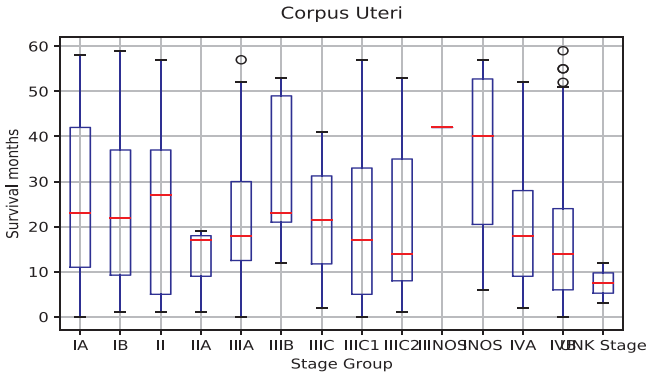

Fig. 5. The boxplot of distribution of survival months (corpus uteri).

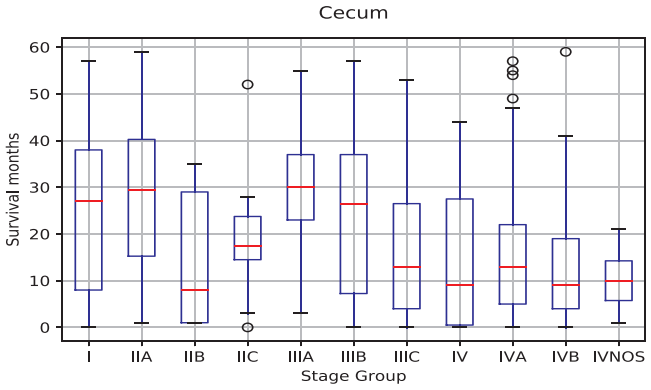

Fig. 6. The boxplot of distribution of survival months (cecum).

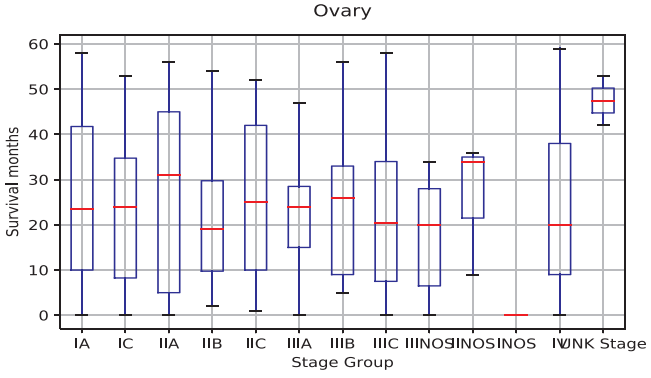

Fig. 7. The boxplot of distribution of survival months (ovary).

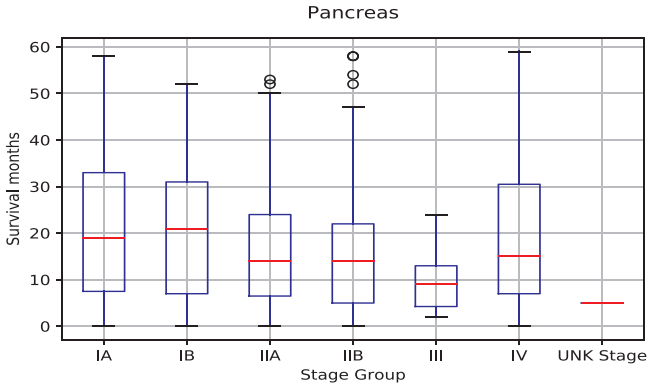

Fig. 8. The boxplot of distribution of survival months (pancreas).

associated attributes of the TopK selected by the ranking of *idf* into three-dimensional vector. The implementation of PV-DM [21] for embedding followed the default setting in the open source framework Gensim (3.4.0). Furthermore, we take a single-layered RNN with size of hidden units: $h = 512$, batchsize $b = 64$. As for setting of dataset, we randomly select 80% of EHR samples for training and validating, the remaining for testing.

4.3. Evaluation metric

Three commonly used metrics: mean absolute errors (MAE), the root mean squared errors (RMSE) and symmetric mean absolute percentage error (SMAPE) [26] are adopted to evaluate the performance of all compared models as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\tilde{y}_t^i - y_t^i| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_t^i - y_t^i)^2} \tag{10}$$

$$\text{SMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\tilde{y}_t^i - y_t^i|}{(\tilde{y}_t^i + y_t^i)/2} \tag{11}$$

where $\tilde{y}_t^i$ is prediction, y_t^i is real value and N is the number of testing samples.

4.4. Baseline methods and settings

Followings are the description on the characteristics and hyperparameter setting of baselines.

- Encoding-based methods: There are three popular unsupervised categorical data encoding methods: One-Hot Encoding, Label Encoding, and IDF Encoding which differentiates values with regard to frequency. we will test these encoding-based methods with completely discretized and partially discretized feature respectively.
- Entity embedding [17]: As an embedded-based representation learning method for structured data, it can map categorical variables in a function approximation problem into Euclidean spaces and be learned by a neural network during the standard supervised training process.
- CatBoost [18]: CatBoost can obtain the new numeric features based on the frequency of occurrence of a certain category. In addition, different combinations of categorical features are also considered to mine the semantic relationships among variables.
- CCAE: Multiple combinations of representations learned from first, second and third order by CCAE will be used for performance comparisons.

The gradient boosting regression tree (GBRT) [27] and random

Table 5
Averaged performance (MAE and SMAPE) comparison of CCAE and baselines on cancer clinical endpoint prediction task (mean ± std).

Model	Type	Encoding methods	MAE	SMAPE
GBRT	Cate. + Num.	One-Hot Encoding	4.6553 ± 0.0122	0.2663 ± 0.0020
		Label Encoding	4.4578 ± 0.0009	0.2610 ± 0.0001
		IDF Encoding	4.5155 ± 0.0351	0.2501 ± 0.0014
	Discretized	One-Hot Encoding	4.4403 ± 0.0317	0.2589 ± 0.0021
		Label Encoding	4.3344 ± 0.0007	0.2606 ± 0.0001
		IDF Encoding	4.4755 ± 0.0065	0.2522 ± 0.0029
	Entity Embedding [17]		4.2653 ± 0.0155	0.2473 ± 0.0028
	CCAE (1-Order)		4.1095 ± 0.0184	0.2384 ± 0.0064
	CCAE (2-Order)		4.1043 ± 0.0171	0.2440 ± 0.0018
	CCAE (3-Order)		4.1105 ± 0.0068	0.2433 ± 0.0016
	CCAE (1-Order + 2-Order)		4.1198 ± 0.0263	0.2429 ± 0.0034
	CCAE (1-Order + 3-Order)		4.1072 ± 0.0249	0.2424 ± 0.0030
	CCAE (2-Order + 3-Order)		4.1100 ± 0.0207	0.2450 ± 0.0018
	CCAE (1-Order + 2-Order + 3-Order)		4.0823 ± 0.0115	0.2405 ± 0.0022
RF	Cate. + Num.	One-Hot Encoding	4.7903 ± 0.0076	0.2616 ± 0.0002
		Label Encoding	4.6391 ± 0.0066	0.2640 ± 0.0003
		IDF Encoding	4.6201 ± 0.0109	0.2368 ± 0.0013
	Discretized	One-Hot Encoding	4.7205 ± 0.0309	0.2575 ± 0.0011
		Label Encoding	4.6023 ± 0.0209	0.2623 ± 0.0009
		IDF Encoding	4.6636 ± 0.0110	0.2358 ± 0.0003
	Entity Embedding [17]		4.5708 ± 0.0132	0.2345 ± 0.0014
	CCAE (1-Order)		4.5277 ± 0.0131	0.2312 ± 0.0008
	CCAE (2-Order)		4.5064 ± 0.0312	0.2306 ± 0.0008
	CCAE (3-Order)		4.5132 ± 0.0117	0.2315 ± 0.0014
	CCAE (1-Order + 2-Order)		4.3569 ± 0.0236	0.2267 ± 0.0017
	CCAE (1-Order + 3-Order)		4.2766 ± 0.0135	0.2230 ± 0.0009
	CCAE (2-Order + 3-Order)		4.3238 ± 0.0156	0.2251 ± 0.0011
	CCAE (1-Order + 2-Order + 3-Order)		4.2363 ± 0.0343	0.2215 ± 0.0014
CatBoost [18]			4.4203 ± 0.0367	0.2382 ± 0.0071
CCAE (1-Order + 2-Order + 3-Order) + RNN			**4.0238 ± 0.0084**	**0.2186 ± 0.0003**

Table 6
Averaged performance (RMSE) comparison of CCAE and baselines on cancer clinical endpoint prediction task (mean ± std).

Model	Type	Encoding methods	RMSE
GBRT	Cate. + Num.	One-Hot Encoding	8.3987 ± 0.0006
		Label Encoding	8.5427 ± 0.0015
		IDF Encoding	8.3862 ± 0.0020
	Discretized	One-Hot Encoding	8.4116 ± 0.0001
		Label Encoding	8.5309 ± 0.0027
		IDF Encoding	8.3506 ± 0.0030
	Entity Embedding [17]		8.5675 ± 0.0033
	CCAE (1-Order)		8.3876 ± 0.0015
	CCAE (2-Order)		8.3117 ± 0.0017
	CCAE (3-Order)		8.2997 ± 0.0025
	CCAE (1-Order + 2-Order)		8.3259 ± 0.0004
	CCAE (1-Order + 3-Order)		8.3660 ± 0.0026
	CCAE (2-Order + 3-Order)		8.3840 ± 0.0019
	CCAE (1-Order + 2-Order + 3-Order)		8.2866 ± 0.0028
RF	Cate. + Num.	One-Hot Encoding	8.3654 ± 0.0009
		Label Encoding	8.2741 ± 0.0035
		IDF Encoding	8.5446 ± 0.0011
	Discretized	One-Hot Encoding	8.4125 ± 0.0008
		Label Encoding	8.2871 ± 0.0004
		IDF Encoding	8.5821 ± 0.0011
	Entity Embedding [17]		8.4750 ± 0.0019
	CCAE (1-Order)		8.3216 ± 0.0015
	CCAE (2-Order)		8.2521 ± 0.0004
	CCAE (3-Order)		8.2648 ± 0.0007
	CCAE (1-Order + 2-Order)		8.2580 ± 0.0011
	CCAE (1-Order + 3-Order)		8.2567 ± 0.0011
	CCAE (2-Order + 3-Order)		8.2491 ± 0.0003
	CCAE (1-Order + 2-Order + 3-Order)		8.2227 ± 0.0004
	CatBoost [18]		8.2241 ± 0.0008
	CCAE (1-Order + 2-Order + 3-Order) + RNN		**8.0015 ± 0.0052**

forest regression (RF) [28] are used to test the performance of different categorical variable encoding or embedding methods. For fairness, we set fixed parameters and the same loss function of mean absolute errors for both model. Specifically, the setting for GBRT is: n_estimators = 150, max_depth = 10 and for RF is: n_estimators = 200, max_depth = 15. As for CatBoost, we set the number of iterations to 200, the learning rate to 0.75, the depth to 12, and the loss function to MAE. In addition, we define the embedding dimension for entity embedding as follows:

$$e_i = \begin{cases} (c_i + 1)/2 & c_i < 50 \\ 50 & \text{else} \end{cases}, \tag{12}$$

where e_i is embedding size, c_i is the amount of categories per feature. Furthermore, we carefully tune each model respectively and tested for five times to reduce random errors and the final results are showed in Table 5.

4.5. Performance comparison

Comparison of our proposed method with other baseline approaches on cancer clinical endpoint prediction task are reported in Tables 5 and 6 and the best performance is in bold. In general, we can see that the proposed method has achieved an effective performance improvement and showed better stability over encoded-based and embedding-based methods under both error evaluation metric. In addition, the key parameter of CCAE: the length of associated attribute list K and the embedding size d which are confirmed by grid searching and showed in Fig. 9.

We can see that the proposed method significantly outperforms traditional encoding-based categorical attribute representation methods. For example, the proposed CCAE with 1–3 ordered representation achieves a relative increase rate of up to 12.31% (from 4.6553 to 4.0823) and 11.57% (from 4.7903 to 4.2363) using the GBRT and RF in MAE respectively. In addition, entity embedding method achieves better performance but such improvement is limited in capturing with cross-domain correlation in EHR data.

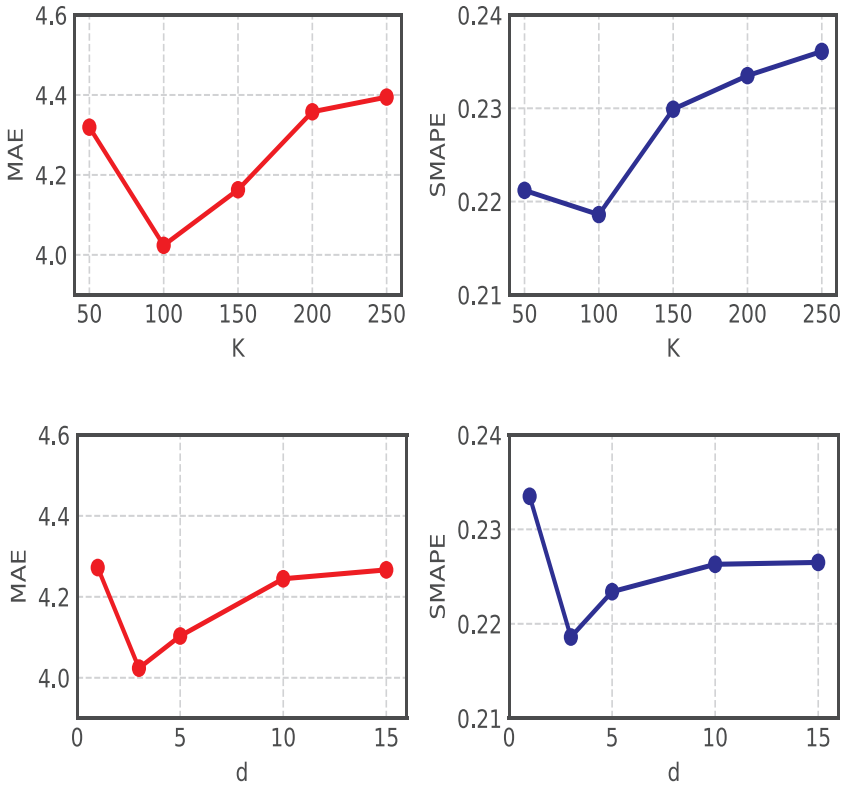

Fig. 9. Parameter sensitivity of the proposed CCAE model.

More specifically, our approach has also achieved a satisfactory performance improvement compared to CatBoost which is the art-of-the-state model for processing categorical data. Furthermore, we can also see that more different ordered representation can get better performance and the introduced recurrent neural networks can better model order-dependency among multi-order representations by combining multi-order representation.

5. Conclusion and future work

In this paper, we introduced an attribute-level representation learning method to obtain a vectorized representation of cancer patients for cancer clinical endpoint prediction and further provide better palliative care for patients with advanced cancer. Specifically, a novel model named CCAE is proposed to respectively learn the embedding-based representation from cross-field categorical variables with strong semantic coupling between each other in different orders. Furthermore, a recurrent neural network is introduced to better model order-dependency among multi-order representations. Experimental results with real-world EHR dataset show the effectiveness of our proposed method over other competitive baselines.

For future work, we will explore personalized treatments recommendation for cancer patients and more relevant applications.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This work was supported in part by Science and Technology Innovation 2030 – “New Generation Artificially Intelligence” Major Project under Grant 2018AAA0102101, in part by the National Natural Science Foundation of China under Grant 61976018 and Grant61532005.

References

[1] Edmondson D, Park CL, Blank TO, Fenster JR, Mills MA. Deconstructing spiritual well-being: existential well-being and HRQOL in cancer survivors. Psycho-Oncology 2008;17(2):161–9.
[2] Sepúlveda C, Marlin A, Yoshida T, Ullrich A. Palliative care: the World Health Organization’s global perspective. J Pain Symptom Manag 2002;24(2):91–6.
[3] Navari RM, Stocking CB, Siegler M. Preferences of patients with advanced cancer for hospice care. JAMA 2000;284(19):2449.
[4] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[6] Avati A, Jung K, Harman S, Downing L, Ng AY, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018;18(S-4):55–64.
[7] Wang L, Zhang W, He X, Zha H. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. KDD 2018:2447–56.
[8] Scarpi E, Maltoni M, Miceli R, Mariani L, Caraceni A, Amadori D, et al. Survival prediction for terminally ill cancer patients: revision of the palliative prognostic score with incorporation of delirium. Oncologist 2011;16(12):1793–9.
[9] Jian S, Cao L, Pang G, Lu K, Gao H. Embedding-based representation of categorical data by hierarchical value coupling learning. IJCAI 2017:1937–43.
[10] Kim Y. Convolutional neural networks for sentence classification. EMNLP 2014:1746–51.
[11] Liu L, Shen J, Zhang M, Wang Z, Tang J. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. AAAI 2018:109–16.
[12] Cao L. Coupling learning of complex interactions. Inf Process Manag 2015;51(2):167–86.
[13] Yoshua B, Aaron C, Pascal V. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798–828.
[14] Pineau E, Lelarge M. InFoCatVAE: Representation learning with categorical variational autoencoders. 2018. arXiv: 1806.08240.
[15] Pang G, Ting KM, Albrecht DW, Jin H. ZERO++: harnessing the power of zero appearances to detect anomalies in large-scale data sets. J Artif Intell Res 2016;57:593–620.
[16] Aizawa A. An information-theoretic perspective of tf–idf measures. Inf Process Manag 2003;39(1):45–65.
[17] Guo C, Berkhahn F. Entity embeddings of categorical variables. 2016. arXiv: 1604.06737.
[18] Prokhorenkova LO, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. NeurIPS 2018:6639–49.
[19] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. KDD 2016:785–94.
[20] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. NIPS 2017:3149–57.
[21] Le Q, Mikolov T. Distributed representations of sentences and documents. Proceeding of the 31st international conference on machine learning 2014:1188–96.
[22] Ferrell B, Connor SR, Cordes A, Dahlin CM, Fine PG, Hutton N, et al. The national agenda for quality palliative care: the National Consensus Project and the National Quality Forum. J Pain Symptom Manag 2007;33(6):737–44.
[23] De Gonzalez AB, Curtis RE, Gilbert E, Berg CD, Smith SA, Stovall M, et al. Second solid cancers after radiotherapy for breast cancer in SEER cancer registries. Br J Cancer 2010;102(1):220–6.
[24] Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The surveillance, epidemiology and end results (SEER) program and pathology: towards strengthening the critical relationship. Am J Surg Pathol 2016;40(12):e94.
[25] Noone AM, Lund JL, Mariotto A, Cronin K, McNeel T, Deapen D, et al. Comparison of SEER treatment data with Medicare claims. Med Care 2016;54(9):e55–64.
[26] Makridakis S, Hibon M. The M3-Competition: results, conclusions and implications. Int J Forecast 2000;16(4):451–76.
[27] Xie J, Rojkova V, Pal S, Coggeshall S. A combination of boosting and bagging for KDD cup 2009 – fast scoring on a large database. Proceedings of KDD-cup competition 2009:35–43.
[28] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 2003;43(6):1947.