Modeling Clinical Decisions with Multinomial Hierarchical Classification

Yinchong Yang [1 2] **Peter A. Fasching** [3] **Volker Tresp** [1 2]

1. Introduction

With the introduction of the Electronic Health Records, a growing amount of available digital information is expected to encourage more personal and precise healthcare services, and thus improve patients experience in clinics (Tresp et al., 2016; Rahman & Reddy, 2015). On the other hand, the physicians are also supposed to consult a large variety and volume of data in order to perform diagnosis and treatment decisions. Such data may include the patients' background information, medical images, genetic profiles and the their entire medical history. The decision making process, therefore, could become increasingly complex in connection with the growing amounts of information collected on each patient. Machine learning based Clinical Decision Support could provide a solution to such data challenges (Choi et al., 2015; Esteban et al., 2016). These works have shown that machine learning models are able to profit from the large amount of data in high dimensional space applying, e.g., various deep neural network architectures. The major advantage of these models lie in their capability of consuming patient features in a variety of forms, and construct more compact and informative latent representations.

However, these models tend to oversimplify the clinical decision process by merely predicting the probability of each decision using plain logistic regression as output layer, thus assuming all possible classes to be unclustered and flat. On the contrary, a clinical decision process is in reality often clustered and hierarchical. For instance, given a patient with breast cancer, a physician has to firstly choose one out of multiple therapy clusters, such as radiotherapy, systemic therapy and surgery. Only then is the physician supposed to further specify the chosen therapy plan. In case of, e.g., radiotherapy, it could be either a curative or a palliative one with respect to therapy intention; and either a Brachytherapy or a percutaneous regarding the therapy type.

Following the concept of Encoder-Decoder Framework (Sutskever et al., 2014), we propose to first encode the patients' features into a latent representation vector using Multilayer Perceptron (MLP) and Recurrent Neural Networks (RNN); and then, more importantly, deploy on top of it a hierarchical classification model, which functions as a decoder that predicts the clustered and hierarchical decision process. It can be seen as a generic form of the so-called Hierarchical Response model in (Tutz, 2011). (Morin & Bengio, 2005) introduced a technically similar architecture to factorize a large softmax layer into a hierarchy. The purpose was to accelerate the calculation of the softmax, which in natural language processing often has the size of the entire vocabulary.

We conduct experiments on a large and up-to-date dataset consisting of almost three thousand metastatic breast cancer patients in Germany. In addition to the advantage of being more realistic, we also show empirically that our proposed architecture improves the prediction quality in term of multiple evaluation metrics.

2. Metastatic Breast Cancer Data

The dataset was provided by the PRAEGNANT study network (Fasching et al., 2015), which has been recruiting patients of metastatic breast cancer since 2014. The original data are warehoused in the secuTrial® database. After exporting and pre-processing, we could extract information on 2,869 valid patients.

There are two classes of patient information that are potentially relevant for modeling the therapy decisions: First the *static* information includes 1) basic patient properties, 2) information on the primary tumor and 3) information on the history of metastasis before entering the study. After dummy-coding we could extract for each patient i a static feature vector denoted with $\boldsymbol{m}_i \in \mathbb{R}^{118}$.

The *sequential* information includes data on 4) local recurrences, 5) metastasis and 6) clinical visits. These are time-stamped clinical events observed on each patient throughout time, and at each time step there can be more than one type of events recorded. All these sequential features are of binary or categorical nature and are also dummy-coded, yielding for patient i at time step t a feature vector $\boldsymbol{x}_i^{[t]} \in \{0,1\}^{189}$. We denote the whole sequence of events

[1]Ludwig Maximilian University of Munich, Germany [2]Siemens AG, Corporate Technology, Germany [3]Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany. Correspondence to: Yinchong Yang <yinchong.yang@siemens.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

for this patient i up to time T_i using a set of $\{\boldsymbol{x}_i^{[t]}\}_{t=1}^{T_i}$.

We attempt to model the therapy decisions concerning 7) radiotherapies, 8) systemic therapies and 9) surgeries, based on the input features as well as former therapy prescriptions.

We extract from the medical history of each patient all possible sub-sequences where the last event consists of one of the three therapies. Therefore in each of these sub-sequences, the last event serves as the target, which the model is trained to predict based on all previous events and the static information.

From the 2,869 patients we could extract in total 16,314 sequences (i.e. 5.7 sequence per patient on average). The length of the sequence before a therapy prescription varies from 0 to 35 and is on average 4.1.

Every time a physician is supposed to prescribe a treatment, he or she is first supposed to choose one of the three *therapy clusters* of radiotherapy, systemic therapy and surgery. For each chosen therapy cluster the physician will then decide the *therapy features*. For radiotherapy there are two 3-dimensional multinomial distributed features: the radiotherapy intention being either curative, palliative or unknown; and the radiotherapy's type being either percutaneous, Brachytherapy or others. For systemic therapy there are three multinomial distributed features. The first one describes 6 types of systemic therapy such as antihormone therapy, chemotherapy, anti-HER2 therapy etc.; the second feature documents the therapy's intention, namely an argument based on the 13 different stagings of the cancer; the third four-dimensional feature records whether the therapy prescription is related to a surgery or is unknown. The last cluster is composed of 10 Bernoulli distributed variables that describe the surgery, such as breast conservation surgery, mastectomy, etc..

3. A Predictive Model of Therapy Decisions

With an encoder RNN we could extract from such sequential input a compact and fixed-size vector representing the entire history of the patient up to a specific time step. For the sake of simplicity, we denote such an RNN –either GRU or LSTM– using a function: $\boldsymbol{h}_i^{[t^*]} = \omega(\{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*})$, where $\boldsymbol{h}_i^{[t^*]}$ is the last hidden state.

In order to also take into account the static features, we concatenate the output of the RNN with the latent representation learned from the static features: $\boldsymbol{z}_i^{[t^*]} = (\boldsymbol{h}_i^{[t^*]}, \boldsymbol{q}_i)$ with $\boldsymbol{q}_i = \sigma(\boldsymbol{H}^T \boldsymbol{m}_i)$. Therefore, the vector $\boldsymbol{z}_i^{[t^*]}$ is expected to represent all available information on patient i up to time t^* in a latent space.

We attempt to model the therapy decisions in a similar fashion as the physicians' prediction procedure: clustered and hierarchical. A physician first has to choose one therapy cluster, and then to specify for the chosen cluster its features. We propose a Multinomial Hierarchical Regression (MHR) to model this procedure.

In the first step we model the probability that each of the three therapy clusters is chosen at time step t^* for patient i using a multinomial variable $C_i^{[t^*]}$ with a softmax activation:

$$\begin{aligned} &\mathbb{P}(C_i^{[t^*]} = k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) \\ =&Softmax\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\gamma}^k\right). \end{aligned} \tag{1}$$

Here $\boldsymbol{z}_i^{[t^*]}$ is the latent representation for the patient up to this time step and $\boldsymbol{\gamma}^k$ serves as the cluster-specific parameter vector.

Then in the second step, given a specific therapy cluster k, we denote the number of therapy features in this cluster with L^k and model the l^k-th multinomial distributed feature variable F_{k,l^k}, whose conditional probability can be modeled with

$$\begin{aligned} &\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \mid C_i^{[t^*]} = k, \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) = \\ =&Softmax\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\beta}_{k,l^k,r}\right), \end{aligned} \tag{2}$$

if k=1 or k=2, i.e. in case of radiotherapy or systemic therapy where therapy features in each cluster are multiple multinomial distributed. Therefore one would need the Softmax function to model the probabilities that the therapy feature takes one specific value r. We denote the parameter vector $\boldsymbol{\beta}_{k,l^k,r}$ with three levels of subscripts: k suggests the cluster of the therapy, l^k selects one specific multinomial feature from this cluster, and r denotes the r-th possible outcome of this feature.

If the therapy cluster suggests the surgery, i.e. k=3, whose features consist of L_k=10 Bernoulli variable, we would have instead of Eq. (2) the following formulation:

$$\begin{aligned} &\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \mid C_i^{[t^*]} = k, \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) \\ =&\sigma\left((\boldsymbol{z}_i^{[t^*]})^T \boldsymbol{\beta}_{k,l^k,r}\right), \end{aligned} \tag{3}$$

with $r = 1$ in all cases, because a Bernoulli variable has an one-dimensional outcome.

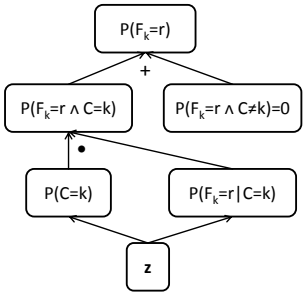
The product of Eq. (1) and (2) as well as that of Eq. (1) and (3) yields the joint probability of both therapy feature and cluster as

$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \wedge C_i^{[t^*]} = k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}). \tag{4}$$

But due to the fact that

$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r \wedge C_i^{[t^*]} \neq k \mid \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}) = 0, \tag{5}$$

Figure 1. A simplified illustration of deriving the marginal probability of the therapy feature. $\boldsymbol{z}$ is the latent representation of a patient.



in all cases, this joint probability of Eq. (4) is equal to

$$\mathbb{P}(F_{i,k,l^k}^{[t^*]} = r | \boldsymbol{m}_i, \{\boldsymbol{x}_i^{[t]}\}_{t=1}^{t^*}), \tag{6}$$

applying the law of total probability, yielding the marginal prediction and allowing us to perform the optimization against the target vector. A simplified illustration of this calculation can be found in Fig. 1.

The major difference between our model and the one in (Tutz, 2011) lies in the fact that the latter one only defines one multinomial response on the second level, which is linked with each cluster on the first level. Our model is more generic in that it allows multiple multinomial or binary responses to be linked to each cluster.

Finally we illustrate the complete model architecture in Fig. 2. There the RNN encoder outputs its last hidden state that represents the whole sequence and is concatenated with the latent representation mapped from the static patient information. This concatenated vector forms the input to the hierarchical model, which in the first step calculates the therapy cluster probabilities and in the second step the therapy feature probabilities conditioned on corresponding cluster. These two levels of probabilities are multiplied, giving the joint probabilities of cluster and feature, which are equivalent to marginal feature probabilities as proven in Eq. (5).

4. Experiment

We conduct cross-validation by splitting the 2,869 patients into 5 disjoint sets, and then query their corresponding sequences to form the training and test sets.

We present two classes of evaluation metrics. First, column-wise average Area Under ROC (AUROC) and Area Under Precision-Recall-Curve (AUPRC), which are well-known metrics applied to measure the classification quality, should indicate the models' capability to assign patients to the correct therapy features. Secondly, we report multi-label ranking-based metrics of Coverage Error (CE) (Tsoumakas et al., 2009) and Label Ranking Average Precision (LRAP) (Madjarov et al., 2012) in the scikit-learn library (Pedregosa et al., 2011). In contrast to precision and recall based metrics, they are calculated row-wise and thus evaluate for each patient how many recommended therapies were actually prescribed. LRAP ranges between 0 and 1 just as AUROC and AUPRC. CE describes how many steps one has to go in a ranked list of recommendations till one covers all ground truth labels. In our case, the average number of labels in each patient case is 4.4 and the total number of possible labels is 39. The CE shall therefore be ideally 4.4, suggesting a perfect prediction, and be 39 in worst case scenario (Tab. 1).

Table 1. Results of experiments with two weak baselines: Random Prediction and Constant Most Popular Prediction.

Weak Baselines	AUROC	AUPRC	CE	LRAP
Random	49.7%	9.4%	38.2	11.2%
Most Popular	50.0%	21.3%	13.9	38.6%

In Tab. 2 we show experiments with three encoders and two decoders. The baseline encoder is a simple Feed-Forward Layer (FFL) consuming the raw sequential information that is aggregated with respect to time. Then the aggregated feature vector is concatenated with the static feature vector for each patient case. Such aggregation can be interpreted as a hand-engineered feature processing, where each feature represents the total number of observed feature values. It also corresponds to the bag-of-words approach (Harris, 1954) in Natural Language modeling, this approach completely neglects the *order* in which the feature values are observed. As a more advanced solution we apply GRU and LSTM as RNN encoders, which are expected to capture the information regarding the events order as well.

The baseline decoder is a single-layered logistic regression, which is a popular choice in multi-class multi-label classification tasks in machine learning. For instance, a therapy *feature* variable is multinomially distributed, implying the mutual exclusiveness of the probable outcomes of the feature values and this aspect cannot be taken into account with a flat logistic regression. For instance, a physician is only supposed to prescribe one medication from a class of medications of similar function.

Both decoders on top of the baseline FFL encoder show suboptimal results compared with those on top of RNN encoders, i.e., GRU and LSTM encoders significantly improves the prediction quality even with a mere logistic regression as decoder. In comparison with the baseline logistic regression as last layer, the MHR model is shown to further improve the prediction in term of all evaluation metrics.

Figure 2. Our proposed model architecture. The radiotherapy features consist of two 3D multinomial variables (red-colored). The systemic therapies consist of on 4D, one 16D and one 3D multinomial variables (orange-colored). The surgery feature consists of 10 Bernoulli variables (purple-colored).

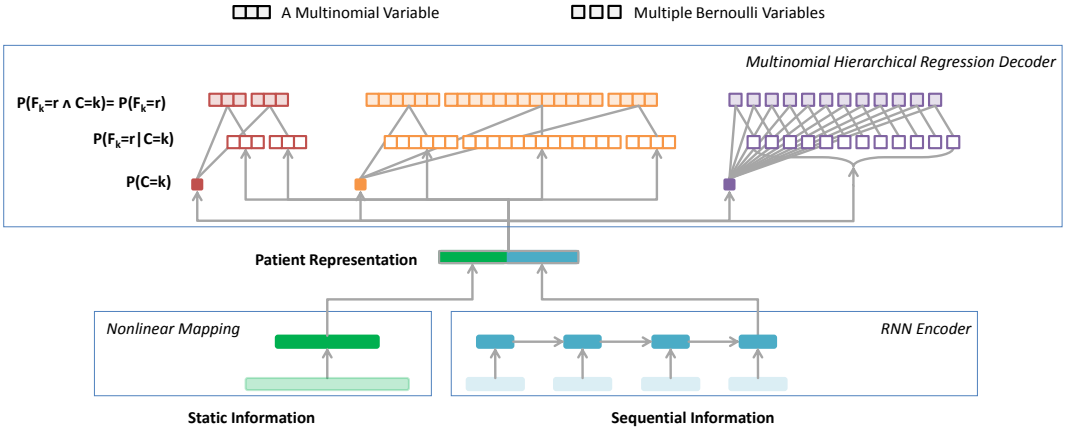


Table 2. Average Results of Experiments with Different Encoders and Decoders, with $\boldsymbol{q}_i \in \mathbb{R}^{128}$ and $\boldsymbol{h}_i^{[t^*]} \in \mathbb{R}^{256}$

Enc.	Dec.	AUROC	AUPRC	CE	LRAP
FFL	Logistic	69.4%	13.4%	12.61	48.6%
	MHR	70.3%	13.9%	11.79	49.3%
GRU	Logistic	81.8%	28.8%	8.57	61.3%
	MHR	**82.1%**	**31.2%**	**8.26**	**62.3%**
LSTM	Logistic	79.6%	24.7%	9.47	57.9%
	MHR	81.9%	30.2%	8.53	61.4%

References

Choi, Edward, Bahadori, Mohammad Taha, and Sun, Jimeng. Doctor ai: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*, 2015.

Esteban, Cristóbal, Staeck, Oliver, Yang, Yinchong, and Tresp, Volker. Predicting clinical events by combining static and dynamic information using recurrent neural networks. *arXiv preprint arXiv:1602.02685*, 2016.

Fasching, P.A., Brucker, S.Y., Fehm, T.N., Overkamp, F., Janni, W., Wallwiener, M., Hadji, P., Belleville, E., Häberle, L., Taran, F.A., Luftner, D., Lux, M.P., Ettl, J., Muller, V., Tesch, H., Wallwiener, D., and Schneeweiss, A. Biomarkers in patients with metastatic breast cancer and the praegnant study network. *Geburtshilfe Frauenheilkunde*, 75(01):41–50, 2015. URL http://www.praegnant.org/.

Harris, Zellig S. Distributional structure. *Word*, 10(2-3): 146–162, 1954.

Madjarov, Gjorgji, Kocev, Dragi, Gjorgjevikj, Dejan, and Džeroski, Sašo. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.

Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pp. 246–252. Citeseer, 2005.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rahman, Rajiur and Reddy, Chandan K. Electronic health records: a survey. *Healthcare Data Analytics*, 36:21, 2015.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Tresp, Volker, Overhage, Marc, Bundschus, Markus, Rabizadeh, Shahrooz, Fasching, Peter, and Yu, Shipeng. Going digital: A survey on digitalization and large scale data analytics in healthcare. *arXiv preprint arXiv:1606.08075*, 2016.

Tsoumakas, Grigorios, Katakis, Ioannis, and Vlahavas, Ioannis. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pp. 667–685. Springer, 2009.

Tutz, Gerhard. *Regression for Categorical Data:*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2011. doi: 10.1017/CBO9780511842061.