Performance of Probabilistic Approach and Artificial Neural Network on Questionnaire Data Concerning Taiwanese Ecotourism

Vladislav Bína[1,2], Václav Kratochvíl[2,1], Lucie Váchová[1,2], Radim Jiroušek[2,1], and Tzong-Ru Lee[3]

[1] University of Economics, Prague
Faculty of Management, Czech Republic
[2] Czech Academy of Sciences
Institute of Information Theory and Automation, Czech Republic
[3] National Chung Hsing University
Department of Marketing, Taichung City, Taiwan
bina@vse.cz, velorex@utia.cas.cz, vachova@vse.cz, radim@utia.cas.cz, trlee@dragon.nchu.edu.tw

Abstract. This paper aims to perform modeling of Taiwanese farm and ecotourism data using compositional models as a probabilistic approach and to compare its results with the performance of an artificial neural network approach. Authors use probabilistic compositional models together with the artificial neural network as a classifier and compare the accuracy of both approaches. The probabilistic model structure is learned using hill climbing algorithm, the weights of multi-layer feedforward artificial neural network are learned using an R implementation of H2O library for deep learning. In case of both approaches, we employ a non-exhaustive cross-validation method and compare the models. The comparison is augmented by the structure of the compositional model and basic characterization of artificial neural network. As expected, the compositional models show significant advantages in interpretability of results and (probabilistic) relations between variables, whereas the artificial neural network provides more accurate yet "black-box" model.

Keywords: compositional models, artificial neural network, model comparison, Taiwanese ecotourism dataset

1 Introduction

When analyzing a sample data set, we cannot infer any results concerning some population with certainty. Therefore, during last more than one hundred years, many tools were developed which are able in a way to handle data having the uncertain nature. They start from the basic test methods of mathematical statistics and so far arrived at the probabilistic and other alternative algorithms and systems more or less capable of serving as a basis for artificial intelligence approaches.

The probabilistic graphical models were considered as a standard tool for support of decision-making under uncertainty and become popular as a tool interconnecting probabilistic description and graphical presentation. The approaches for decision-support initially started from the simple managerial tools assessing causes of problematical status. Then it continued as qualitative schemes of influence diagrams and evolved into the tools of flowcharts, causal loop diagrams, stock and flow diagrams and in the last three decades developed influence diagrams based on Bayesian networks. However, the usability and local computations of Bayesian networks (see, e.g., Jensen [1]) brink together one important source of confusion, namely its graphical representation using directed acyclic graphs. The arrows used in such graphical tools resembles less experienced users direction from causes to effects which is rather intuitive and usually accepted but not correct. Moreover, this graphical presentation is of only little use in case of large diagrams having higher tens or even hundreds of nodes (as we show below). In these situations, one may acknowledge an alternative, yet equivalent algebraic approach based on compositional models Jiroušek [2] and Jiroušek and Kratochvíl [3].

In recent year, based on better algorithms of deep learning and higher performance of computers the approach of artificial neural networks become successful in a wide variety of tasks including computer vision, text and opinion mining, machine translation, image and video processing, etc. It is sometimes referred as a "deep learning revolution" (see, e.g., Sejnowski [4]). Similarly to the comparison of probabilistic and neural network approaches by Tavana et al. [5] or by Simfukwe et al. [6], the aim of this paper is to compare the accuracy of the powerful and well-developed methodology of artificial neural networks with a (yet developing) probabilistic approach of compositional models. Of course, the main disadvantage of compositional models is rather the current state of the art which (unlike the case of neural networks) lack a professional and user-friendly implementation of algorithms. But still, compositional models represent an easy to interpret "white-box" approach and have thus a great advantage in comparison with neural networks. On a simple (but not toy) example data set concerning Taiwanese farm and ecotourism, we present that under certain circumstances, the performance of compositional models is comparable to the performance of artificial neural networks.

2 Brief Summary of Theoretical Background

Throughout the paper, we use two modeling approaches. Compositional models as a probabilistic model structure a multi-layer feedforward artificial neural network. In this section, we will briefly characterize both approaches and set a basis for their comparison using a Taiwanese farm and ecotourism data.

2.1 Compositional Models

The theory of compositional models was summarized in Jiroušek [2] with the important structural properties summarized in Jiroušek and Kratochvíl [3]. Sim-

ilarly, we will adopt the notation in the following sense. Throughout this paper we analyze a set of n finite valued *variables* $\{X_1, X_2, \ldots, X_n\}$. Subsets of variables are denoted by lower-case Roman alphabets (e.g., x, y, and z). $\langle X_i \rangle$ denotes the set of values (states) of variable X_i. Analogously, for sets of variables x, y the respective Cartesian products of all combinations of their values are denoted by $\langle x \rangle, \langle y \rangle$, respectively. Elements of these sets, i.e., the (combinations of) values of variables will be denoted by lower-case boldface Roman characters (e.g., $\mathbf{a} \in \langle x \rangle$).

Conditional distributions will be denoted using a standard notation, e.g., $\pi(y \,|\, X)$. In case that we consider conditioning by a specific value of variable X by $\pi(y|X = \mathbf{a})$. Let us stress that since we deal with finite valued variables, the conditional distribution $\pi(y|X)$ is represented by a table where $\pi(y|X = \mathbf{a})$ is a probability distribution for each $\mathbf{a} \in \langle x \rangle$.

For a probability distribution $\pi(x)$ its *marginal* distribution for $y \subset x$ is denoted either by $\pi(y)$, or by $\pi^{\downarrow y}$. Under a notion of extension (in a way opposite to marginalization) we understand any distribution κ defined for a superset of variables, i.e., $\kappa(z)$ for $z \supset x$, such that $\kappa(x) = \pi(x)$. The set of all extensions of distribution $\pi(x)$ for variables $z \supset x$ will be denoted by $\Psi[\pi; z]$. The symbol $\pi(x \cap y) \ll \kappa(x \cap y)$ denotes that $\kappa(x \cap y)$ *dominates* $\pi(x \cap y)$. This holds (in the considered finite setting) when

$$\kappa^{\downarrow x \cap y}(\mathbf{b}) = 0 \Longrightarrow \pi^{\downarrow x \cap y}(\mathbf{b}) = 0$$

for all $\mathbf{b} \in \langle x \cap y \rangle$.

Now, consider two distributions $\pi(x)$ and $\kappa(y)$. Obviously, there exists their joint extension if and only if they are *consistent*, i.e., if $\pi(x \cap y) = \kappa(x \cap y)$. In case that they are not consistent then one can be interested in getting an extension of π containing from κ *as much information as possible*. Speaking more precisely, one can look for a distribution $\mu(x \cup y)$ that is a *projection* of κ into the set of all extensions $\Psi[\pi; x \cup y]$:

$$\mu(x \cup y) = \arg \min_{\lambda \in \Psi[\pi; x \cup y]} Div(\lambda(y); \kappa(y)).$$

If we consider a Kullback-Leibler divergence the Theorem 6.2 in [2] states that this type of projection can be got as a composition of π and κ. The composition is defined only if $\pi(x \cap y) \ll \kappa(x \cap y)$ by the formula

$$\mu(x \cup y) = \pi(x) \rhd \kappa(y) = \frac{\pi(x) \cdot \kappa(y)}{\kappa(x \cap y)}.$$

The use of the operator of composition can be iterated. The result of the repeated application to the sequence of low-dimensional distributions is (if defined) a multidimensional distribution which can be written in the following way:

$$\kappa_1 \rhd \kappa_2 \rhd \kappa_3 \rhd \ldots \rhd \kappa_n := (\ldots((\kappa_1 \rhd \kappa_2) \rhd \kappa_3) \rhd \ldots) \rhd \kappa_n.$$

Throughout this paper we will focus on the models composed from the marginals of one data distribution, thus there are no inconsistent distributions.

2.2 Artificial Neural Networks

An approach of Artificial Neural Networks (ANN) had roots in 1940s when the first computational model for neural networks was developed (see McCulloch and Pitts[7]). The methodology of ANN became useful and extensively employed after the development of backpropagation algorithms (see Schmidhuber [8]). But in the case of networks with higher number of hidden layers tended to give worse results than the shallow networks (see, e.g., Alom et al. [9]). The success of deep-learning approaches together with improved computational capacities (use of GPUs and employment of its vector computation features) resulted in last ten years in a so-called deep learning revolution, i.e. a radical change of artificial intelligence industry and massive use of ANNs (this breakthrough resulted in awarding of Turing Award in March 2019).

Let us clarify that under the notion of deep network we understand an artificial neural network with at least three hidden layers, oppositely, the ANN with one or two hidden layers is called shallow (this is quite frequent classification (see, e.g., [9]).

One of the modern and most successful open-source systems for artificial intelligence is a H2O platform capable of analyzing (using in-memory compression) huge data samples. Moreover, it has a linear scalability and is able to interconnect with, e.g., R, Python and Hadoop (see [10]). We used it as an implementation of artificial neural network and its learning under R software (we used a version 3.6.1 [11]) augmented by a H2O package version 3.26.0.2 [12].

In our case, we use the artificial neural network as a model of dependencies among the set of categorical variables (most of them binomial, two multinomial variables). Because of the architecture of ANN layers, we need to choose one of the variables as a variable in an output layer. The H2O implementation does not provide a possibility to handle multiple output variables. The nonlinear character of activation functions provides a possibility to perform a classification task.

The considered artificial neural network has a structure of a multilayer feed-forward (nodes do not form a cycle) neural network (perceptron). The network is trained with stochastic gradient descent algorithm based on backpropagation and it is necessary to specify the number of network layers and number of neurons in each layer, see [10]. The binary character of variables allowed to choose among different activation functions (the Tanh activation function performed with the analyzed data set better than ReLU and Maxout). For the classification into multiple classes, i.e., for the multinomial output layer, a softmax activation function can be used. See, e.g., Glorot et al. [13].

Though the H2O package is capable to efficiently handle huge data sets thanks to the parallelization of its procedures, this was not our case. However, relatively small Taiwanese data were divided one hundred times in order to perform multiple training and validation cycles.

2.3 Measures Based on Confusion Matrix

The paper employs a basic set of measures derived from the confusion matrix (see, e.g., Fawcett [14]). The confusion matrix visualizes the correspondence of

predicted class based on model and class observed in data in a contingency table with the setting according to Table 1.

Table 1. Confusion matrix

	Observed class	
Predicted class	True positive (TP) False negative (FN)	False positive (FP) True negative (TN)

Sensitivity (or true positive rate) is defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

specificity (or true negative rate) is given by

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

precision (or positive predictive value) is defined as

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

under *accuracy* we understand

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

and *F1 score* is a harmonic mean of precision and sensitivity which result into

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

3 Data Set and Preprocessing

The Taiwanese farm and ecotourism data set contains answers of 1235 respondents who filled the questionnaire in the period from 2015 to 2017. The answers to the questionnaire in case of the first ten respondents are shown in Table 2 as an illustration of sample data set.

The answers of six main multiple-choice questions were converted into 51 binary variables containing answers accompanied by respondent's gender and age category (0–18; 19–25; 26–35; 36–45; 46–55; 56–65 and 66–). The six main questions are:

Q1 Reasons why you would like to focus on agricultural information.

Table 2. Record of answers of the first ten respondents in the Taiwanese farm and ecotourism data set.

Que. 1	Que. 2	Que. 3	Que. 4	Que. 5	Que. 6	Age	Gender
3, 4	1, 2, 8	F, I, L	1, 3	1, 2	1, 2, 3	36–45	female
3, 4, 5	1, 2, 3, 4, 5, 6, 7, 8	4, 8, 9, A, G, H, K, L	1, 2	1, 2	5	19–25	female
3, 4	3, 4, 8	4	1	1, 2	5	36–45	male
1	1, 6, 7, 8	4, A	5	1	1, 2	19–25	female
5	8	3, 8, D, K, L	3	1	5	36–45	female
4	7	6, A, B, F, I	1, 2, 3, 4	3	5	36–45	male
2, 3	1, 7	C, D	1, 3, 4	1, 2	1, 2, 4	46–55	female
3, 4, 5	4, 5, 7, 8	2, D, I, O	1, 3	2, 3	1, 3	26–35	male
3, 5	1, 8	3, 4, A, B, K	3	1	1, 2	36–45	female
3, 4	6, 7, 8	4, 7, B, L	3, 4	1	2	26–35	male

Q2 What kinds of instant message you would like to see.
Q3 What kinds of products or stories that you are interested in.
Q4 Reasons why you are interested in participating in work exchange.
Q5 What kinds of workshop you are interested in.
Q6 What did you experience from the ecotourism.

Since all the particular questions are not significant for the sake of comparison of two modeling approaches, we will not describe in detail all possible answers to the six main questions. Let us only mention that there were 5 possible answers for the question Q1, 8 answers for the Q2, 24 answers for the Q3, 6 answers for the Q4 and 3 answers for Q5. Now let us focus on the answer to the 6th question which will be analyzed as predicted variables. The possible answers to a multiple-choice question 6 were:

Q6.1 Agricultural experience and understand planting methods.
Q6.2 Enjoy local natural food.
Q6.3 Special festivals participation.
Q6.4 Local culture exchange.
Q6.5 Not yet experienced.

4 Models

The main result of the presented paper includes two types of different modeling approaches describing the Taiwanese farm and ecotourism data. The first approach uses probabilistic compositional models, whereas the second approach employs the artificial neural network methodology.

4.1 Resulting Compositional Model

The structure of a compositional model is learned from the data set using a structural EM algorithm (see Friedman [15]) where a maximization step is performed using a tabu search generalization of hill climbing greedy approach in the

space of models structures (see Russell and Norvig [16]). The resulting compositional model can be (obviously) written as a model formula. However, because of its rather complex structure and long formula let us only take a taste on the shortened expression, i.e.

$$\hat{\mu} = \pi(x_{3.4}) \triangleright \pi(x_{3.O}) \triangleright \pi(x_{3.4}, x_{3.A}) \triangleright \pi(x_{3.A}, x_{3.B}) \triangleright \pi(x_{3.4}, x_{3.5}, x_{3.B}) \triangleright \cdots .$$

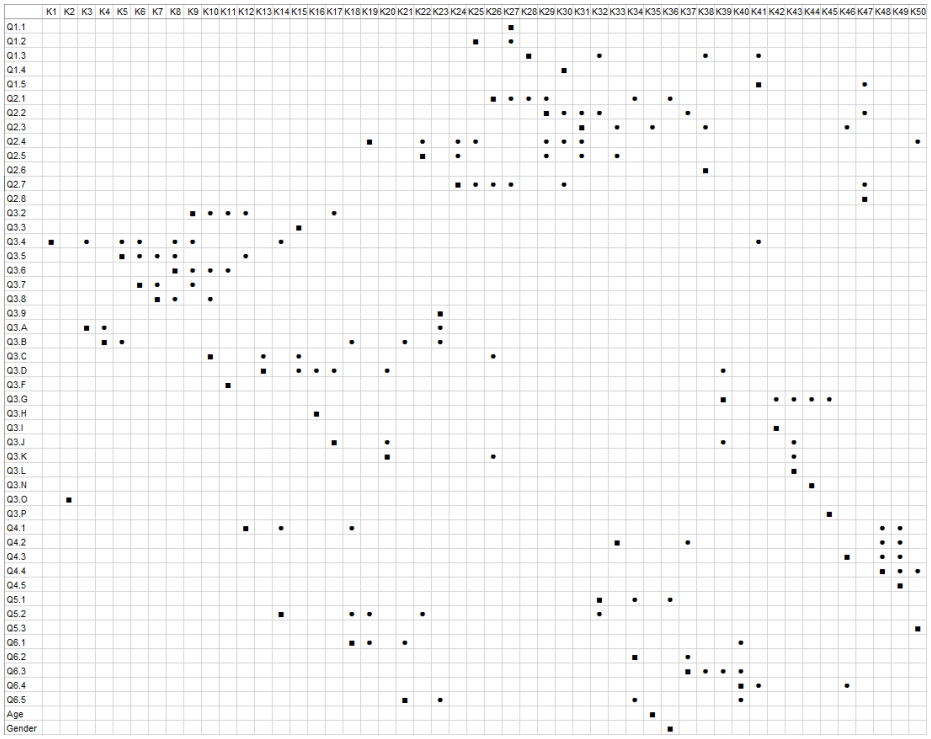

Fig. 1. Persegram of a compositional model structure learned from the Taiwanese farm and ecotourism data.

As the kind author can imagine, because of the big number of variables the complete compositional model formula would be very long and not transparent. Instead of this, we use an equivalent representation, i.e. a structure visualization using a graphical tool of persegram (introduced by Jiroušek [17], for an application, see, e.g., Kratochvíl [18]). This tool describes a dependency structure of considered variables[4] and is capable to clearly present both all the particular

[4] In Jiroušek [17] the following assertion is formulated: "Every independence statement read from the structure (or its persegram) of a compositional model corresponds to probability independence statement valid for every multidimensional probability distribution represented by a compositional model with this structure."

distributions composed in the model (as columns in the table) and occurrence of a variable throughout the distributions within a model (as rows of the table) where the first occurrence is marked by a square and all other occurrences are marked using a bullet. The model learned from an example data set is depicted in Figure 1 and provides an easy insight into the dependence structure of considered variables. Namely, the answers to question 4.1 to 4.5 shows an apparent interdependency, or similarly, the answers to questions 3.2, 3.4 to 3.8.

Probably, the readers are more accustomed to the expression of probabilistic models in the form of a directed acyclic graph (DAG). The Figure 2 shows the above-described dependence structure in the form of DAG, which is obviously rather hard to read and to search the particular (conditional) dependencies. Moreover, the arrows in the graph might be misleading and can lead to an incorrect interpretation as causal relations.

Particular distributions and their conditional variants usable for the process of composition can be easily computed from the data. E.g., conditional distribution $\pi(x_{6.1} \,|\, x_{3.B}, x_{4.1}, x_{5.2})$ can be summarized in a form of Table 3.

Table 3. An example of one probability distribution from compositional model $\hat{\mu}$ in a form of conditional probability table, namely $\pi(x_{6.1} \,|\, x_{3.B}, x_{4.1}, x_{5.2})$.

		Q5.2=yes				Q5.2=no			
		Q4.1=yes		Q4.1=no		Q4.1=yes		Q4.1=no	
		Q3.B		Q3.B		Q3.B		Q3.B	
		yes	no	yes	no	yes	no	yes	no
Q6.1	yes	0.639	0.909	0.500	0.634	0.231	0.671	0.258	0.365
	no	0.361	0.091	0.500	0.366	0.769	0.329	0.742	0.635

4.2 Resulting Artificial Neural Network

Deep learning algorithm was used 5 times to build five models. Each of the models has one of the questions 6.1 to 6.5 as an output node. The method in each case learned the weights of 5 layers multilayer perceptron with 162 neurons in an input layer, 100, 40 and 6 neurons in hidden layers with tanh activation function and an output layer with two nodes uses a softmax activation function. The algorithm in each case learned values of 20600 weights and biases.

The creation of five models was necessary since the current implementation of H2O does not support multiple response columns. The authors suggest to train a new model for each response variable. An example of model metrics achieved for the model with an output being a Question 6.1 summarized in Table 4. Metrics of all particular models and their comparison is performed in the next subsection.

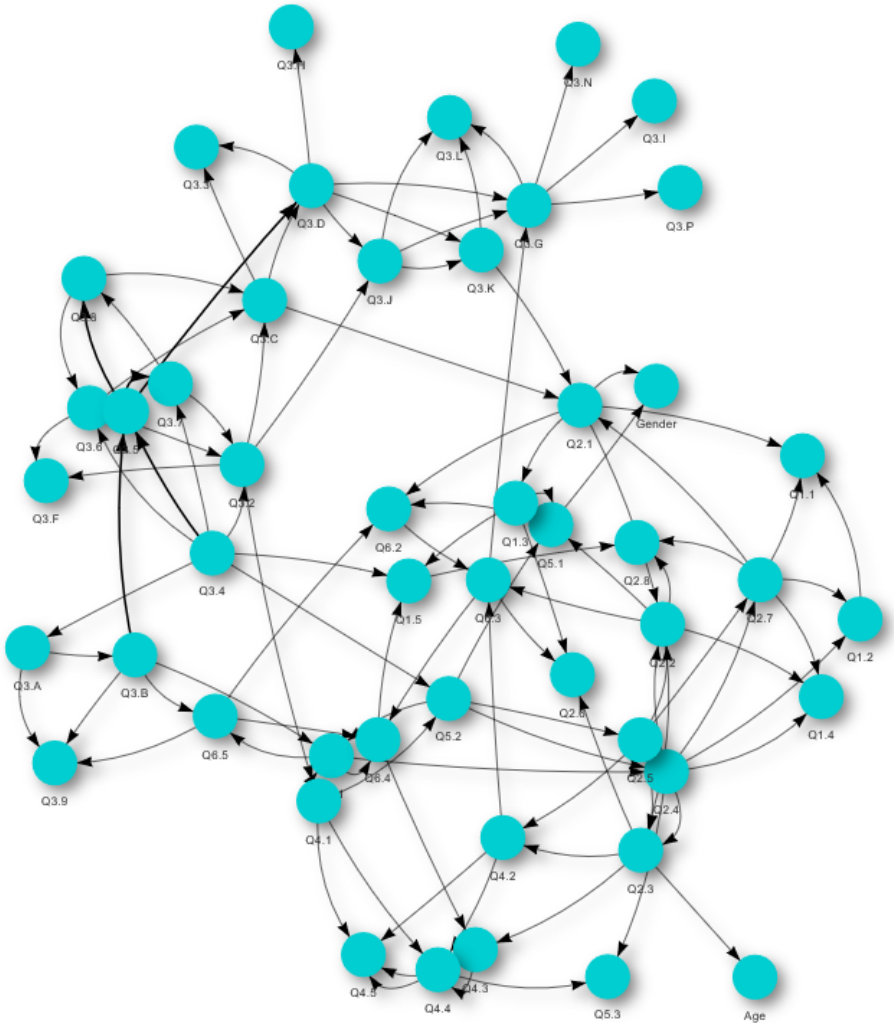

Fig. 2. A directed acyclic graph corresponding to a compositional model structure learned from the Taiwanese farm and ecotourism data.

Table 4. An example of basic maximum metrics and their indices in case of model for Q6.1.

Metric	Value	Index
Max F1	0.867	95
Max accuracy	0.798	79
Max precision	0.966	28
Max sensitivity	1.000	101
Max specificity	0.973	0

4.3 Comparison of Both Types of Models

In each case, the data set was divided into a training frame and a validation frame in the ratio of 80% to 20%. In other words, the training process was conducted on a sample of 1121, and validation was conducted on 114 statistical units. For both approaches, a non-exhaustive cross-validation method was used which do not compute all possible ways of the splitting of the original sample being thus an approximation of leave-p-out cross-validation approach. In each case, we used 100 iterations of the cross-validation.

The results of the comparison are summarized in Table 5 where the metrics of sensitivity, specificity, precision, accuracy and F1 score are provided for both types of models and for each of the five questions from 6.1 to 6.5. This numerical comparison is augmented by a graphical presentation in Figure 3 of ROC space graphs for each of the five questions.

Table 5. The set of measures based on confusion matrix of artificial neural network (ANN) and compositional model (CM) for particular questions from 6.1 to 6.5.

Question	Model	Sensitivity	Specificity	Precision	Accuracy	F1 score
6.1	ANN	0.835	0.887	0.969	0.841	0.897
	CM	0.921	0.358	0.785	0.759	0.844
6.2	ANN	0.801	0.867	0.959	0.810	0.872
	CM	0.959	0.413	0.776	0.782	0.856
6.3	ANN	0.646	0.852	0.759	0.758	0.692
	CM	0.585	0.771	0.598	0.704	0.581
6.4	ANN	0.692	0.821	0.851	0.736	0.760
	CM	0.646	0.852	0.759	0.758	0.692
6.5	ANN	0.969	0.993	0.947	0.990	0.957
	CM	0.145	0.983	0.538	0.886	0.159

5 Conclusion

The paper presented two model approaches for modeling of categorical data. The compositional model approach was applied to build and to use one model approximating the whole data set. The approach of artificial neural networks was employed in order to create five particular models with an output variable of each of five questions from 6.1 to 6.5.

The comparison of both approaches showed that in the case of Questions 6.1 and 6.2, both approaches provided more or less similar quality of models. In case of questions 6.3 and 6.4 the approach of artificial neural networks provided a model of higher quality, and finally, in case of question 6.5, the approach of compositional models failed to provide reasonable predictions, whereas the artificial neural network approach was very successful. This was caused by an

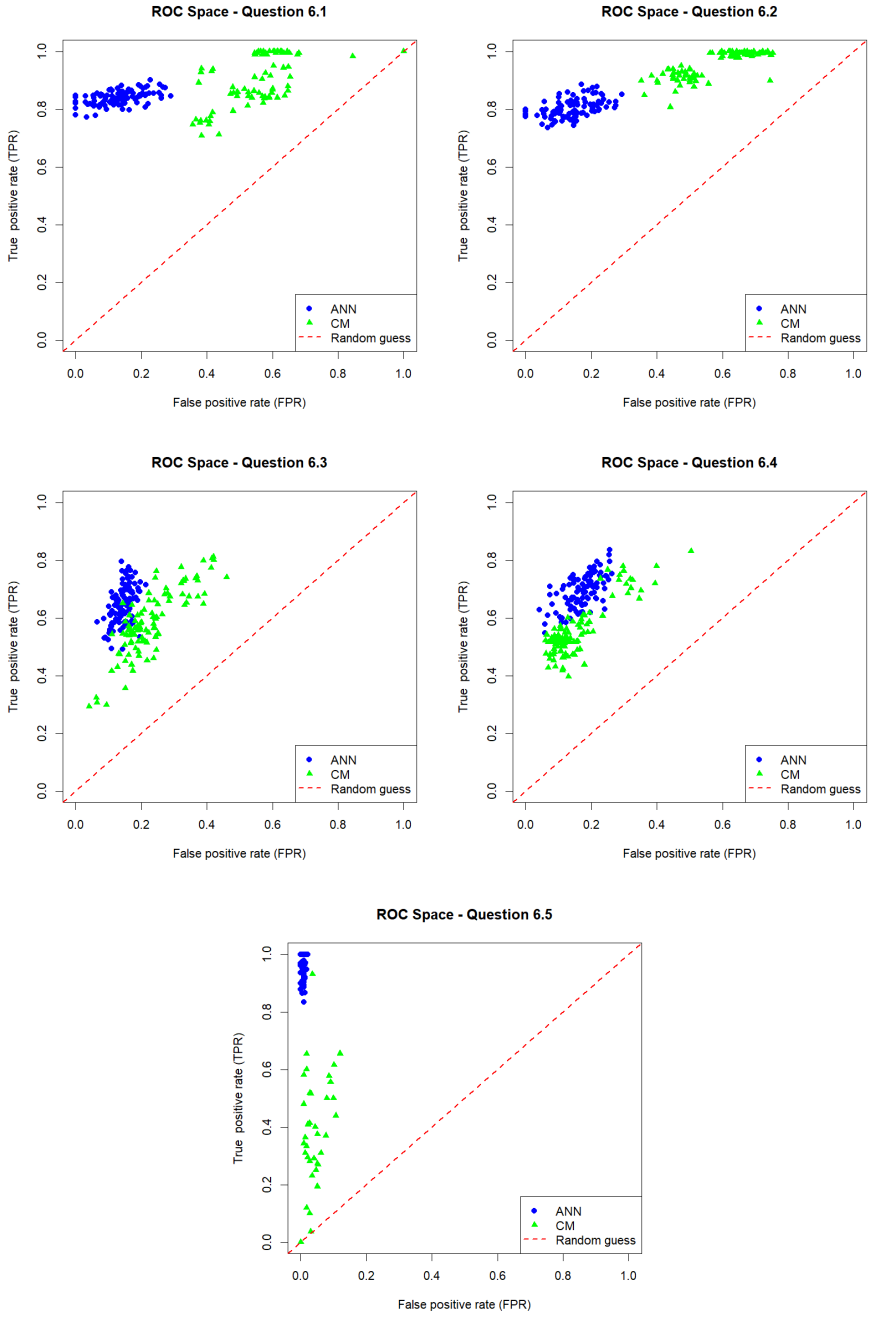

Fig. 3. ROC space of artificial neural network and compositional model for non-exhaustive cross-validation of prediction for particular questions from 6.1 to 6.5.

answer collecting in a way the rest of respondents and having an unbalanced ratio of answers to question 6.5 (142 positive answers and 1093 negative answers) in contrary to more or less comparable frequencies of both answers in case of other answers.

This documents the most serious limitation of the approach of compositional models. But let us mention an important advantage that the compositional models represent a white-box approach, i.e., the possibility to analyze and interpret its building blocks (the low-dimensional distributions to be composed) as probabilities usual for description of uncertainty. Moreover, similarly to the approach of compositional models, the user is able to insert evidence into the model and to analyze interesting marginal distributions which can be calculated more or less easily from the compositional model. The artificial neural network approach comprises in a way a black-box. It is theoretically possible to look at the weights of each neuron, but its possible interpretation is very limited.

Acknowledgement

The research was financially supported by grants GAČR no. 19-06569S (first, second and fourth author) and AVČR no. MOST-04-18 (remaining authors).

References

1. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs. Information Science and Statistics series. New York: Springer-Verlag, 2nd edition (2007).
2. Jiroušek, R.: Foundations of compositional model theory. Int. J. of General Systems **40** (6), 623–678 (2011). doi:10.1080/03081079.2011.562627
3. Jiroušek, R., Kratochvíl, V.: Foundations of compositional models: structural properties. Int. J. of General Systems **44** (1), 2–25 (2015). doi:10.1080/03081079.2014.934370
4. Sejnowski, T.J.: The deep learning revolution. MIT Press (2018).
5. Tavana, M., Abtahi, A.R., Di Caprio, D., Poortarigh, M.: An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking. Neurocomputing **275**, 2525–2554 (2018). doi:10.1016/j.neucom.2017.11.034
6. Simfukwe, M., Kunda, D., Chembe, C.: Comparing Naive Bayes Method and Artificial Neural Network for Semen Quality Categorization. International Journal of Innovative Science, Engineering & Technology 2(7), 689–694 (2015).
7. McCulloch, W., Pitts, W.: A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. **5** (4): 115–133 (1943). doi:10.1007/BF02478259
8. Schmidhuber, J.: Learning complex, extended sequences using the principle of history compression. Neural Computation **4**, 234–242 (1992).
9. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A., Asari, V.K.: A state-of-the-art survey on deep learning theory and architectures. Electronics **8**(3), 292 (2019).
10. Candel, A., Parmar, V., LeDell, E., Arora, A.: Deep Learning with H2O. 6th Edition. H2O.ai, Inc. (2019). http://h2o.ai/resources

11. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2019). https://www.R-project.org/
12. LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava M.: h2o: R Interface for 'H2O'. R package version 3.26.0.2. (2019). https://CRAN.R-project.org/package=h2o
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323 (2011).
14. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters **27**(8), 861–874 (2006). doi:10.1016/j.patrec.2005.10.010
15. Friedman, N.: Learning Belief Networks in the Presence of Missing Values and Hidden Variables. In: Proceedings of the 14th International Conference on Machine Learning, pp. 125–133 (1997).
16. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition (2009).
17. Jiroušek, R.: Persegrams of Compositional Models Revisited: conditional independence. In: Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, pp. 915–922 (2008).
18. Kratochvíl, V.: Probabilistic Compositional Models: solution of an equivalence problem. Int. J. Approx. Reason. **54** (5), 590–601 (2013). doi:10.1016/j.ijar.2013.01.002