Conceptual Alignment Method

Completed Research Full Paper

Wolfgang Maass
German Research Center for Artificial Intelligence (DFKI) and Saarland University, Saarbruecken, Germany
wolfgang.maass@dfki.de

Arturo Castellanos
The College of William and Mary, Williamsburg, Virginia, United States
aacastellanosb@wm.edu

Monica Tremblay
The College of William and Mary, Williamsburg, Virginia, United States
monica.tremblay@mason.wm.edu

Roman Lukyanenko
University of Virginia, Charlottesville, Virginia, United States
romanl@virginia.edu

Veda C. Storey
Department of Computer Information Systems, Georgia State University, Atlanta, Georgia, United States
vstorey@gsu.edu

Jonas S. Almeida
National Cancer Institute, National Institutes of Health, Rockville, Maryland, United States
jonas.dealmeida@nih.gov

Abstract

This paper proposes a Conceptual Alignment (CA) Method for conceptual modeling and machine learning. The model consists of a three-step cycle that selects an initial conceptual model, aligns it with machine learning models, and refines both models to reach predictive consistency. Alignment is based on composition methods that can be instantiated by methods that satisfy contribution properties. The Conceptual Alignment Method is applied to a healthcare use case on hospital inpatient discharges. The machine learning model trained for total cost predictions is aligned with a conceptual model. We show how this refined conceptual model is used for explaining the machine learning model for a very large healthcare dataset.

Keywords

machine learning, conceptual modeling, explainable artificial intelligence, concept alignment.

Introduction

Without models, humans cannot exist within environments (Moore and Golledge, 1976). Conceptual models in information systems are used by information system designers for making sense of complex socio-economic environments that partially reside in physical and social environments but also reside in digital environments. In this sense, conceptual models are bridging both types of environments. In mathematics and statistics, analytical models are used for representing exact equations between variables whereas probabilistic models are approximations of data. Machine learning models are probabilistic models with very high capabilities for complex data, such as visual data or language data. The quality of data predetermines the quality of machine learning models (Sheng et al. 2008).

As our society's dependence on machine learning grows, it is important to ensure that machine learning models perform well, are compliant with legal and ethical requirements, and are interpretable and transparent for different types of users (Maass et al. 2022). This trade-off is often exacerbated by opaque transformations in the input data (feature engineering), making it challenging to assess the effectiveness of the input data on the outcome. Large deep learning models train billions of parameters for finetuning the performance on predictive tasks. Such models can only be interpreted by technical experts; not by decision makers and other users. Therefore, users do not have proper models that allow them to make sense of the environments for which machine learning models are used. It creates a situation in which predictions made by machine learning models are perceived as scientific phenomena, such as molecular behavior in Chemistry, DNA behavior in Biology or behavior of atoms in Physics. This is astonishing because, although machine learning models are developed by humans, we have reached a point of complexity that exceeds human understanding. This "tipping point" results in the persistence of numerous challenges when using machine learning models, including biases, discrimination, lower performance, lack of transparency, interpretability, and explainability (Arrieta et al. 2020).

Given this interpretability gap between decision makers' understanding and machine learning models, conceptual models are well-placed to serve as useful mediators. Traditionally, conceptual models are created by humans based on conceptual modeling grammar (Wand and Weber 2002). Conceptual modeling grammars provide a set of constructs and a set of production rules for representing perceptions for a given domain negotiated between design teams (Burton-Jones et al. 2009). The open issue lies in the lack of transparency of machine learning models which poses a challenge for building a conceptual model on a machine learning model. Conceptual models that are derived from input-output behavior of machine learning models are notoriously weak because they do not represent proper conceptual definitions and relationships between input data and output data. We propose an iterative approach, called *Conceptual Alignment (CA) Method*, that consists of a three-step cycle: (1) *selection of initial conceptual models*, (2) *alignment with machine learning models*, and (3) *refinement of conceptual models and machine learning models*. The iterations stop when a fixpoint has been reached, based on the measurement of *predictive consistency*.

To appreciate the situation more tangibly, consider the following example. A financial investor group hired a leading AI company *DL Force* for building a prediction system for their trading department. The trade department has given schema information about available input data and targeted output data to DL Force. The input data has been selected from the information system's database that has been used thus far. The designers assume that the conceptual model used for the previous information system has similarities with the information system that uses machine learning technologies. Therefore, they provide this as input to the Conceptual Alignment Method. The first iteration shows that the conceptual model in inconsistent in some input concepts with respect to the prediction (binary buy or sell trade decision for simplicity). The CA method indicates which attributes of the conceptual model are predictive inconsistent by measuring the predictive inconsistency of all concepts and attributes. The designers make decision on redesigning the conceptual model according to the recommendations given by the CA method. After several iterations the general predictive consistency score (GPCS) cannot be improved which indicates a maximum. The revised conceptual model is optimally aligned with the machine learning model and provides a proximate explanation for the machine learning model. It includes prediction support values (PSV) between input concepts and the predicted attribute of the outcome concept, i.e., trade decision. Therefore, the revised conceptual model is an improved representation of the ML-based information system that can be in discussions with developers of DL Force and in discussions with superior management.

This example describes the utility of conceptual models as an interpretive framework for machine learning models. By alignment of conceptual models with machine learning models, similarities, and differences between the two can be identified. If there is strong predictive consistency, there is a high probability that the predictive behavior of the machine learning model will conform to the conceptual model. If there is a low level of predictive consistency, the machine learning model behaves significantly differently than assumed by domain experts. This is either an indication that the conceptual model is not properly reflecting the domain or that the machine learning model provides some unintended or novel facts about the domain (Storey et al. 2022).

Machine Learning and Conceptual Modeling

Conceptual models are shared representations expressed in various forms, such as texts and graphics. According to model theory, conceptual models are projected and abbreviated representations made by human experts and used for a purpose (Stachowiak 1973). In contrast, machine learning models are statistical abstractions derived by algorithmic fitting mathematical functions to data according to a purpose given by an objective function (Hastie et al. 2009). Data used for model fitting and also for construction of conceptual models are representations of domains themselves. In essence, conceptual models are socially constructed, and machine learning models are algorithmically constructed. Mental models and machine learning models are equally difficult to access. Therefore, conceptual models occupy a mediating position between the two.

Conceptual models are understood as a means for designing and implementing better information systems (Wand and Weber 2002). Wand and Weber identified the evaluation of ``competing scripts generated via the same grammar to describe some phenomenon'' as a key element for investigation of conceptual-modeling scripts (Wand and Weber 2002). Because information systems increasingly use data analytical models as core components, conceptual models provide a means for designing machine learning models as well.

So far, conceptual-modeling scripts (cf. Wand and Weber 2002) are designed according to syntactic, semantic, and pragmatic properties (Lindland et al. 1994). It does not matter if the attributes serve as input for the prediction of other attributes. Therefore, concepts are designed independent of predictive consistency of attributes. For information systems that use machine learning models, conceptual models can also support the design and implementation of better machine learning models. This means that inconsistencies of concept definitions with respect to predictions are to be minimized so that they better reflect the behavior of intended machine learning models. A conceptual model provides higher predictive consistency if attributes are homogeneous in predicting attributes of concepts to which they are connected by conceptual relationships.

Supervised learning, a subtype of machine learning, and the focus of our work, guides the learner in acquiring knowledge in a domain through examples, so new cases can be handled based on the knowledge learned from similar cases. Modern supervised machine learning efforts have been rapidly increasing due to the availability of training data, growing computational capabilities, and the availability of new methods and techniques (e.g., deep learning neural networks, reinforcement learning).

The increase in the use of complex machine learning models has revealed challenges in explaining the decision logic of these models. Transparency research in AI is a growing societal concern and an emergent research area (Arrieta et al. 2020). For example, constrained models and post-hoc explanation techniques can help in building responsible AI systems (Arrieta et al. 2020). In our work we seek to leverage knowledge of domain experts, externalized as conceptual models, to allow for detection of biases or even unintended behavior of machine learning models.

Conceptual modeling formally describes “some aspects of the physical and social world around us for the purposes of understanding and communication” (Mylopoulos 1992). Conceptual models have been used extensively in information technology development, especially for database design and process engineering, and to facilitate communication and domain understanding (Recker et al. 2021, Wand and Weber 2002). Some ideas exist for the use of conceptual models on machine learning models. For instance, Recker et al. envision digital agents that generate conceptual models from machine learning models (Recker et al. 2021). Conceptual models are characterized by various traits, including simplicity, abstraction, communication, flexibility, and modifiability (Mylopoulos 1992, Wand and Weber 2002). However, none of these traits are related to predictability and performance of machine learning models. This might explain why data scientists rarely use conceptual models during the design of data analytical models and pipelines.

Recent research has proposed combining conceptual modeling with artificial intelligence or, specifically machine learning (Maass et al. 2022, Bork et al. 2020, Lukyanenko et al. 2020). Doing so can provide reliable rules about the domain without being dependent on extracting them from the data. Despite these efforts, conceptual models are rarely used in the process of building machine learning models or to increase machine learning model transparency and interpretability. At the same time, machine learning

invariably relies on understanding of reality in the minds of data scientists or users of machine learning models, who either develop or interpret machine learning solutions, in light of their individual mental models. This inevitably leads to differences between the shared understanding of the information system design team and the data scientist team responsible for the machine learning model. This, in turn, will lead to inconsistent decision behavior of the information system that uses machine learning models.

Research hypothesis: Conceptual models are aligned with machine learning models provide conceptual explanations for machine learning models.

Explanation Models

Humans are good in developing and using linear prediction models. If the stock market went up the past three days by 1%, we easily extend this into a prediction of 1% today. The same holds for weather, grades, or inflation rates. But reality shows that linear models are naïve and result in large prediction errors because linear models are not specific enough, i.e., they cannot capture functional behavior of non-linear signals. For modeling non-linear phenomena, science has developed different approaches, such as differential equations, Taylor series, Interpolating Polynomial and Splines in calculus and statistical models, such as autoregressive models, support vector machines, decision trees, and neural networks. These approaches generally lead to tradeoff between accuracy and human interpretability.

Explanation models are simpler models reversing this tradeoff for better human interpretability by optimizing only locally. Explanation models, such as Lime and SHAP are linear models that approximate point estimates for given input values. Lundberg and Lee proposed SHAP (Shapley Additive Explanation) values as a generalization of additive feature attribution models, i.e., models that are sums over weighted input values plus mean (Lundberg and Lee 2017). SHAP values are Shapley values of a conditional expectation function of the machine learning model, i.e., the fitted model is used for determining local contribution of single features to an outcome.

Shapley values formalize coalition games and determine additive marginal contributions of single players to an overall payoff of the coalition of players. They are defined by an operator ϕ that assigns for each game v a vector of payoffs $\phi(v) = (\phi_1, ..., \phi_n)$. $\phi_i(v)$ is player i's marginal and additive contribution to the outcome of a game over all permutations with all other players (Shapley 1952). Shapley values are locally accurate, i.e., match the original model *f(x)*, are not affected by missing values, and are consistent wrt. inequality relation between two models *f(x)* and *f'(x)* (Lundberg and Lee 2017). The Shapley value of a feature *i* is a weighted mean of its marginal value of feature *i*, averaged over all possible subsets of features:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!v} f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$$

with *F* the set of all features and $f_{S \cup \{i\}}$ with a model trained with feature *i* and $f_S(x_S)$without. SHAP values determine additive contributions of input features *X* on the prediction of an output feature *y*, i.e., *y=f(X)*, agnostic to the machine learning model used with *f* the original model and ϕ a vector of all Shapley values ϕ_i for all independent features. Input feature $x \in X$ is transformed into a simplified vector $z \in Z$ with $z \in \{0, 1\}$, i.e., SHAP value $\phi_i z_i$ is zero if the feature $\phi_i z_i$ is missing and ϕ_i carries the whole *feature contribution* of a feature to an outcome otherwise.

Conceptual Alignment Method

Following the theory-grounded arguments for incorporating conceptual models for explainability, we advance a new method, the conceptual alignment (CA) method. The CA method iteratively aligns conceptual models with machine learning models. The proposed Model Embedding Method consists of three steps that are iterated until a stopping criterium has been reached:

(1) Selection of initial conceptual models

(2) Alignment with machine learning models

(3) Refinement of the conceptual models and machine learning models

The objective of the CA method is the transformation of an initial conceptual model in a way that is best aligned with a machine learning model. Because CA iterations can also lead to redesigning the machine learning model, both models are moving toward each other.

Selection of initial conceptual models

Data is the starting point for the analysis of machine learning models but is also a crucial element of conceptual models. Data schema for training machine learning models is often flat with numerical data only, i.e. numerical tabular data with columns for features and rows for samples. The semantics of features is generally well-defined with a small focus, such as age or income. For supervised learning models, samples also carry values, aka labels, for outcome variables. The goal is the prediction of labels y for known input data X ($f(X) = \hat{y}$). Conceptual model group data attributes into classes and connect concepts by relationships with cardinalities.

The CA method requires a mapping between attributes of conceptual models and features of machine learning models. Explanability is best supported if a one-to-one mapping is used but more complex one-to-many mappings might be used as well. N-to-m mappings diminish the possibility of clearly attributing contributions of features to attributes in conceptual models. For simplicity, we assume a one-to-one mapping in the following with identical value ranges.

Class definitions are key decisions made by conceptual modelers. The CA method uses schema information on classes, i.e., class definitions with associated attributes. A key requirement is that all attributes are mapped with all features used by the machine learning model. Conceptual models that satisfy this requirement are valid candidates for initial conceptual models.

Composition method

The alignment of conceptual models with machine learning models requires a composition mechanism that is comparable to the composition of measurement models and structural models in structural equation models (Bollen 2014). The CA method uses explanation models on machine learning models for extracting concept contributions that are used for determining the predictive consistency score of the conceptual model depending on the machine learning model. This approach builds on the superimposition approach proposed by Lukyanenko et al. (2020) that graphically imposes feature weights onto conceptual models to show which entities (concepts) the weights belong to.

Any composition method has the following *contribution properties* (Shapley 1952):

(1) Contribution values are *superadditive*.

(2) Contribution values satisfy local accuracy, missingness and consistency properties (Lundberg and Lee 2017).

(3) Contribution values can be injectively mapped onto attributes of conceptual models (feature mapping).

(4) Feature mappings are semantically sound.

The *superadditivity* property means that for two machine learning models S and T the following holds for their contribution values $CV(C \cup T) \geq CS(C) \cup C(T)$ given feature sets of C and D are disjoint. This means that composition methods can be applied to ensembles of disjoint machine learning models. The second property of composition methods is derived from Shapley's theory as discussed in Lundberg and Lee (2017). The third property requires that a feature can only be mapped onto one attribute of the conceptual model.

In the following, we use a composition method that uses SHAP Values (Lundberg and Lee 2017) with a one-to-one mapping between features and attributes. SHAP Values satisfy property (1) and (2) and the one-to-one mapping satisfies property (3).

Alignment process

With the selected composition method, we can determine the contribution values of a machine learning model (MLM), map these onto attributes and integrate these into concept contributions relative to a given prediction task. The alignment process with the SHAP-based composition method consists of three steps:

(1) *Feature Contributions*: determination of contribution values for all input features of an MLM for predicting an outcome feature *p*.

(2) *Mapping Contribution Values*: application of the mapping function that maps feature contributions values onto attribute contribution values.

(3) *Concept Contributions*: identification of concept contributions on prediction features.

Note that attributes are defined in the context of conceptual models and features in the context of machine learning models.

Feature Contribution: The composition method calculates contribution values. In our example, we are using SHAP values represented by vector Φ (Lundberg and Lee 2017). Feature contributions are irrespective of the association of features to attributes and concepts, i.e., feature contributions are determined by machine learning models independent of conceptual models.

Mapping Feature Contributions: Despite the simplicity of standard mappings between features of the feature set *F* and attributes of the attribute set *A*, mapping contribution values F_{CV} of features in *F* onto attribute contribution values A_{CF} is an important conceptual step: map(F_{CV} (F) → A_{CV}(A)). F_{CV}(F) and A_{CV}(A)are required to satisfy the properties of contribution values. It requires an in-depth analysis of the semantics underlying features and attributes and their functional relations. The selected mapping function *map* needs to be consistent with the functional relation between features and attributes. A one-to-one mapping fulfills properties 3 and 4 for contribution methods. A feature mapping is semantically sound if the semantic of a feature *f element of F* has an interpretation in the domain of a conceptual model *CM* that is a subset of the corresponding attribute of an *element of A*.

Concept Contribution: After feature mapping feature contributions onto attribute contributions, concept contributions are determined by aggregating attribute contributions. For each concept *c* in *CM*, a n-ary *concept contribution* vector g_c^o is constructed by selecting only attribute contribution values associated to a concept *c*. Selection is controlled by masking vectors δ:$g_c^o = A_{CF} \circ \delta$.

For the SHAP-based composition method with one-to-one mapping, g_c^o is the sign-preserving, minmax-scaled mean of Shapley values for a feature *f* mapped onto *a*. This allows to normalize the attribute contribution values for the absolute largest value, i.e., the normalization is driven by the largest positive or negative attribute contribution value.

Refinement of the conceptual models and machine learning models

Resulting concept contributions g_c^o represent the relative positive or negative contribution of features *f* that is mapped onto *a* for predicting an outcome feature *o*. Therefore, concept contributions are governed by associated attribute contributions. Concept contributions can be purely positive, purely negative, close to zero or multimodal. Pure concept contributions only contain attribute contributions that are all positive or negative, i.e., pointing in the same predictive direction. This can be seen as support for a clear conceptual relationship between both concepts from a data analytical perspective in addition to syntactic, semantic, and pragmatic qualities (Lindland et al. 1994). Concept contributions close to zero indicate that a concept is irrelevant for the prediction of an outcome feature *o*. Multimodal concept contributions contain negative and positive attribute contributions, i.e., attributes are predicted in both directions. This could lead to situations in that strong attribute contribution values cancel out each other. This would mean that strong predictive effects are not transparent on the conceptual level. Multimodal concept contributions are an indicator of misalignment between conceptual models and machine learning models. Therefore, multimodal concept contributions require *conceptual separation* for making prediction effects transparent. First, all attribute contributions below a threshold (e.g., <0.1) are neglected. Second, attributes with negative and positive attribute contribution values are separated into distinct concepts. By

using a heuristic, the larger set is kept in the original concept while a new concept is generated for the smaller set.

After conceptual separation, the alignment process is re-iterated until the number of multimodal concept contributions increases. Whether this is a global or local optimum depends on the initial conceptual model. In general, concept separation is a NP-hard problem that is a subclass of the knapsack problem.

Example

Domain, prediction task, data, and pre-processing

Data analytics in healthcare on patient data is the basis for personalized medicine (Fröhlich et al. 2018). The Office of Quality and Patient Safety of the New York State Department of Health has established a Statewide Planning and Research Cooperative System (SPARCS) that provides a comprehensive all-payer data collection system. SPARCS currently collects patient-level details on patient characteristics, diagnoses and treatments, services, and charges for inpatient and outpatient services (ambulatory surgery, emergency department, and outpatient services) (https://www.health.ny.gov/statistics/sparcs/training/docs/sparcs_dgc_manual.pdf). Cost control is an important aspect of modern healthcare systems. Therefore, predicting individual total costs based on key factors is used in the following.

The dataset Hospital Inpatient Discharges contains 2.3M sample with about 600MB, each representing one patient incident in 2017. The dataset contains 34 features including "*total_costs*". All features are transformed into numerical data. For our evaluation, we only used 1% from the dataset so that it could be tested on a Macbook Air (M1, 16GB, 8 cores, 2020). The dataset has few high total cost values with a mean of 16,708 USD and a maximum value of 5,544,736 USD for the subsampled dataset.

Selection of initial CM

As an initial conceptual model, we defined a class model without relationships. The assignment of attributes to classes was constructed based on discussions with domain experts, semantic analysis of the features, and evaluation of the SPARCS DGC technical documentation.

- Hospital: 'Hospital Service Area', 'Hospital County', 'Operating Certificate Number', 'Permanent Facility Id', 'Facility Name'
- Patient: 'Age Group', 'Zip Code - 3 digits', 'Gender', 'Race', 'Ethnicity', 'Patient Disposition', 'Payment Typology 1', 'Payment Typology 2', 'Payment Typology 3'
- Admission: 'Length of Stay', 'Type of Admission', 'Discharge Year', 'Emergency Department Indicator', 'Birth Weight', 'Abortion Edit Indicator'
- Illness: 'CCS Diagnosis Code','CCS Diagnosis Description', 'CCS Procedure Code', 'CCS Procedure Description', 'APR DRG Code', 'APR DRG Description', 'APR MDC Code', 'APR MDC Description', 'APR Risk of Mortality','APR Medical Surgical Description', 'APR Severity of Illness Code', 'APR Severity of Illness Description'

Definition and application of the composition method

The composition method is based on SHAP values (Lundberg and Lee 2017). A one-to-one mapping between features and attributes which guarantees semantic soundness. Therefore, contributions properties are satisfied. Additionally, we set the multimodality threshold to 0.05.

Application of the alignment method

Contribution values show that hospital, patient, and illness are multimodal. Table 1 shows the feature values for concept *illness* after the initial iteration. *APR MDC Code* and *APR DRG Code* clearly point in different directions so that the concept contribution of illness becomes close to zero. By conceptual separation, three concepts (c1, c2, c3) are constructed that capture diverging features. After the second iteration, '*APR MDC Code*' and '*APR Medical Surgical Description*' were identified as root causes for strong multimodality in *illness*. Another concept could be automatically generated but we opted for

integration of these two features into *c3* due to semantic similarity of features. This is an intervention option that is prompted to conceptual modelers and domain experts. It shall be noted that features *Discharge Year* and *Abortion Edit Indicator* are kept as control variables whose feature contribution should always be zero because they are fixed to value *2017* and *0*, respectively.

Concept: Illness	mean	range
CCS Diagnosis Code	-0.005557	[1...2617]
CCS Diagnosis Description	-0.000823	[0...262]
CCS Procedure Code	0.055127	[0...231]
CCS Procedure Description	-0.09329	[0...222]
APR DRG Code	0.973663	[1...956]
APR DRG Description	0.025985	[0...319]
APR MDC Code	-1.0	[0...25]
APR MDC Description	0.003322	[0...25]
APR Risk of Mortality	0.0089	[0 1 2 3 4]
APR Medical Surgical Description	-0.080236	[0 1]
APR Severity of Illness Code	-0.051258	[0 1 2 3 4]
APR Severity of Illness Description	-0.001722	[0 1 2 3 4]
Concept contribution	*-0.0138*	

Table 1. Feature contributions and concept contributions for *illness*

	mean	ranges
c1		
Hospital County	-0.034587	[0...56]
Operating Certificate Number	-0.035327	[101000.0...7004010.0]
Facility Name	0.000249	[0...209]
gc_bar:	*-0.0232*	
c2		
Age Group	-0.005381	[0 1 2 3 4]
Race	-0.002907	[0 1 2 3]
Patient Disposition	-0.010671	[0...19]
Payment Typology 3	-0.000267	[0...10]
Concept contribution	*0.0048*	
c3		
CCS Procedure Description	-0.0901	[0...222]
APR Severity of Illness Code	-0.003219	[0 1 2 3 4]
APR MDC Code	-1.0	[0...25]

APR Medical Surgical Description	-0.078736	[0 1]
Concept contribution	*-0.293*	

Table 2. Feature contributions for added concepts c1, c2, and c3 (iteration 3)

Concept Contributions

The final concept contribution models in graph representation (Figure 2b) shows that all added concepts are negative while all others are positive. None is multimodal beyond the multimodality threshold. Therefore, the algorithm stops after iteration 3. The conceptual model (Iteration 3) is a valid proxy for the machine learning model (34 features and 600MB data). This conceptual model can be scrutinized by domain experts and information system designers. For example, it is worth noting that zip code, gender, and ethnicity have neither positive nor negative effects on total costs, suggesting that the machine learning model and predictions have a higher likelihood of accounting for associated biases.

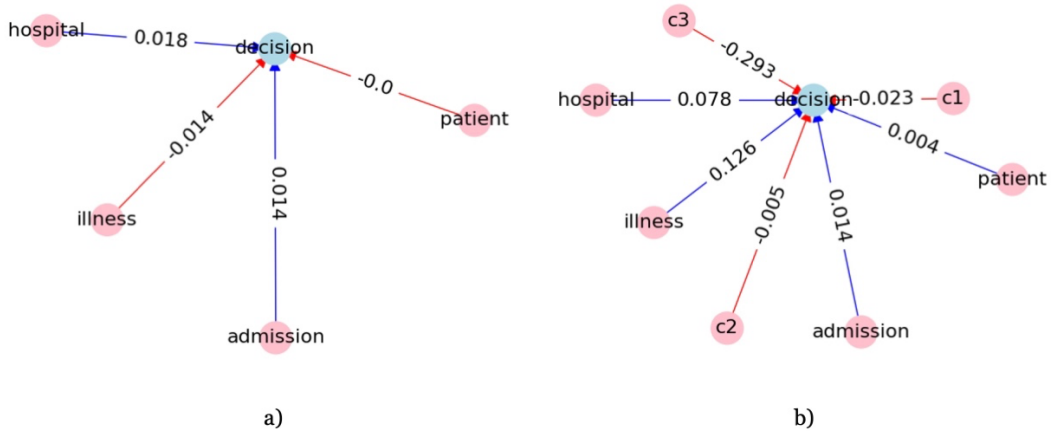

Figure 1: Concept contributions for decision with feature total costs: a) iteration 1 and b) iteration 3

In contrast to well-known explainable AI methods, conceptual alignment determines directed dependencies between input and output concepts instead of features only. Therefore, the conceptual alignment method abstracts from features to concepts and provides conceptual information to domain experts. With this information, domain experts can explore conceptual dependencies that are implicit to machine learning models and training datasets.

Conclusion

This paper has proposed a *Conceptual Alignment (CA) Method* for conceptual modeling and machine learning. The model consists of a three-step cycle that selects an initial conceptual model, aligns it with machine learning models, and refines both models to reach predictive consistency. Future work is needed to apply this method to multiple applications to assess its feasibility and effectiveness.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Burton-Jones, A., Wand, Y., & Weber, R. (2009). Guidelines for empirical evaluations of conceptual modeling grammars. *Journal of the Association for Information Systems*, *10*(6), 1.

Bork, D., Garmendia, A., & Wimmer, M. (2020, November). Towards a Multi-Objective Modularization Approach for Entity-Relationship Models. In *ER Forum/Posters/Demos* (pp. 45-58).

Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., ... & Zupan, B. (2018). From hype to reality: data science enabling personalized medicine. *BMC medicine*, *16*(1), 1-15.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Lindland, O. I., Sindre, G., & Solvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE software*, *11*(2), 42-49.

Lukyanenko, R., Castellanos, A., Storey, V. C., Castillo, A., Tremblay, M. C., & Parsons, J. (2020). Superimposition: augmenting machine learning outputs with conceptual models for explainable AI. In *Advances in Conceptual Modeling: ER 2020 Workshops CMAI, CMLS, CMOMM4FAIR, CoMoNoS, EmpER, Vienna, Austria, November 3–6, 2020, Proceedings 39* (pp. 26-34). Springer International Publishing.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Maass, Wolfgang, Arturo Castellanos, M. Tremblay, Roman Lukyanenko, and Veda C. Storey. "AI Explainability: A conceptual model embedding." In *International conference on information systems*, pp. 1-8. 2022.

Moore, G. T., & Golledge, R. G. (1976). *Environmental knowing: Theories, research and methods*. Dowden.

Mylopoulos, J. (1992). Conceptual modelling and Telos. *Conceptual modelling, databases, and CASE: An integrated view of information system development*, 49-68.

Recker, J. C., Lukyanenko, R., Jabbari Sabegh, M., Samuel, B., & Castellanos, A. (2021). From representation to mediation: a new agenda for conceptual modeling research in a digital world. *MIS Quarterly: Management Information Systems*, *45*(1), 269-300.

Shapley, Lloyd S., A Value for N-Person Games. Santa Monica, CA: RAND Corporation, 1952.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 614-622).

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Springer.

Storey, V. C., Lukyanenko, R., Maass, W., & Parsons, J. (2022). Explainable AI. *Communications of the ACM*, *65*(4), 27-29.

Wand, Y., & Weber, R. (2002). Research commentary: information systems and conceptual modeling—a research agenda. *Information systems research*, *13*(4), 363-376.