Global and Local Interpretability of Belief Rule Base

DOI:
10.1142/9789811223334_0009

Document Version
Accepted author manuscript

Global and Local Interpretability of Belief Rule Base*

S. Sachan†, J. B. Yang and D. L. Xu
Decision and Cognitive Science Research Centre, University of Manchester, Manchester, M15 6PB, United Kingdom
†*E-mail: swati.sachan@manchester.ac.uk*
www.alliancembs.manchester.ac.uk

The rapid adoption of artificial intelligence in automating human-centred tasks has accentuated the importance of interpretable decisions. The Belief-Rule-Base (BRB) is a hybrid expert system that can accommodate human knowledge and capture nonlinear causal relationships as well as uncertainty. This paper presents the strategy to interpret BRB locally for a single instance to understand the decision-making process by the importance of activated rules and attributes and globally to understand most important rules and attributes in an entire rule base.

Keywords: Interpretability; Decision; Rule-Base

1. Introduction

1.1. *Background*

The widespread adoption of artificial intelligence (AI) for smart automation of subset of tasks in various domains, specifically in sensitive areas such as health care and finance has forced research community to focus on develop of interpretable models and methods to understand black-box machine learning (ML) models. The model-agnostic approach Local Interpretable Model-agnostic Explainer (LIME)[1] and Shapley Additive Explainer (SHAP)[2] are presented to understand the contribution of important features in predicting a single instance. Both LIME and SHAP utilize an interpretable ML as a surrogate models for a black-box model to provide an explanation of a specific decision. These model-agnostic methods can be applied to any black-box ML model. Layer-wise Relevance Propagation[3] and Taylor decomposition[4] technique are used to interpret neural networks by decomposing the output into the sum of the relevance

* Work is funded by Alliance Strategic Research Investment Award - AA15032 of the Alliance Manchester Business School, University of Manchester.

of neurons in preceding layers. Some ML/AI model such as decision tree and generalized linear models - logistic regression and quantile regression fall into the class of interpretable models; however, these model lack accuracies compared to complex black-box models. The decision trees can be unstable, a small variation in the dataset could generate a completely different tree. In artificial intelligence, an expert system is an interpretable computer system that emulates the decision-making ability of a human expert in a problem domain. It has three main components – knowledge base, rule inference engine and user interface. The Belief-Rule-Base (BRB)[5] is a hybrid expert system based on the D–S theory of evidence, decision theory, and fuzzy set theory. It can accommodate the knowledge of human experts and capture nonlinear causal relationships as well as uncertainty. Recently its adoption in the development of a transparent system for automation for loan underwriting was presented[6]. This paper presents the strategy to interpret BRB by leveraging a simple example on mortgage loan lending decision.

1.2. *Local and Global Interpretability*

Understanding the data is an initial step towards the interpretability of the ML model to grasping the most relevant features in a dataset. After training, a model can be interpreted globally by judging the overall importance of features in a training dataset and locally by the importance of features in providing a decision for a single instance, shown in Figure 1.

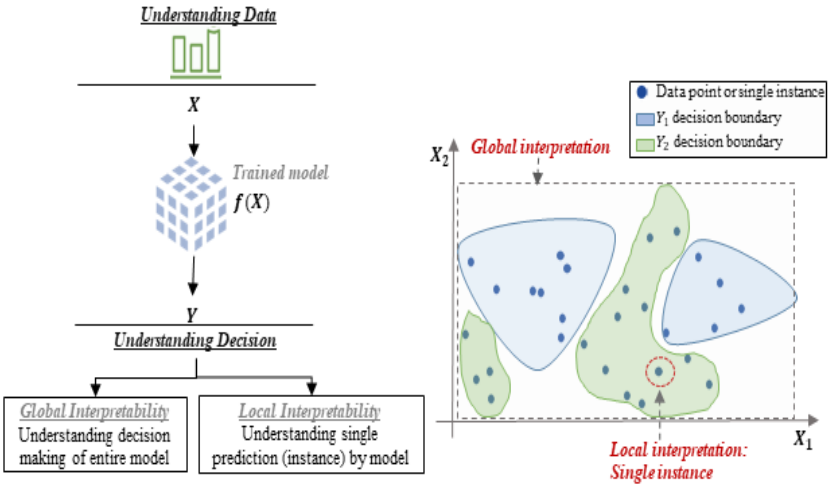

Fig. 1. Local and global interpretability

2. Belief-Rule-Base

In past belief-rule-base (BRB) was adopted to model the decision making for customer preferences[7], loan underwriting[6], pipeline leak detection[8], and trauma outcome[9]. In this paper, strategy to interpret decisions by BRB is demonstrated through a simple example based on mortgage loan lending.

2.1. *Referential values of antecedent and consequence attributes*

Suppose, a decision on the mortgage loan application is given by considering three antecedent attributes A_1: credit score, A_2: affordability, and A_3: property type. Both affordability and property type are evaluated externally by a affordability calculator and property evaluation methods, respectively. The number of antecedent attribute is denoted by I, $i \in \{1, \dots I\}$. The consequence attribute is Y: Decision (decision to fund or reject). The input data is shown in Table 1. The structure of BRB is shown in Figure 2. The qualitative and quantitative antecedent attributes has set of categorical and continuous referential values, respectively. The v^{th} referential value of i^{th} antecedent attribute is denoted by $A_{v,i}$, $v = 1, \dots V_i$. The training of continuous referential values of quantitative antecedent attributes is part of training process of BRB. The referential value for antecedent and consequence attribute is shown below:

$$Referential\ Value = \begin{cases} \mathrm{A_1}: \text{credit score} = \{0,150,300,600\} \\ \mathrm{A_2}: \text{affordability} = \{\text{Yes, No}\} \\ \mathrm{A_3}: \text{Property type} = \{\text{Good, Average, Poor}\} \\ \mathrm{Y}: \text{Decision} = \{\text{Fund (F), Reject (R)}\} \end{cases} \quad (1)$$

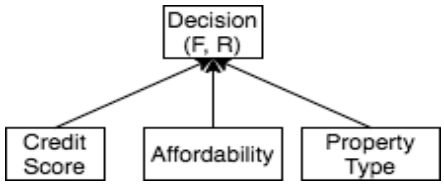

Fig. 2. BRB Structure

Table 1. Dataset

Data point	A_1	A_2	A_3	Y
1	600	Yes	Good	F
⋮	⋮	⋮	⋮	⋮
m	300	No	Good	R
⋮	⋮	⋮	⋮	⋮
M	600	Yes	Poor	R

2.2. *Rule in Rule-base and activation of rules*

The human knowledge in BRB is represented by IF-THEN rules. The rules in rule-base are exhaustive combination of referential values of antecedent attributes to match referential value of attributes and rules. The number of rules in a rule-base is the Cartesian products of the number of referential values in antecedent attributes:

$$K = V_1 \times V_2 \times V_3 = \prod_{i=1}^{I=3} V_i = 24 \tag{2}$$

In Eq. (2), K ($k \in \{1, \dots, K\}$) is numbers of rules. The IF-THEN rules in rule-base are illustrated in Table 2. It has 24 rules. The degrees of belief for each rule can be given by experts initially. The weight of k^{th} rule (θ_k), weight of i^{th} attribute in k^{th} rule (δ_i^k), and n^{th} belief degree of k^{th} rule ($\beta_{n,k}$) are trained by data. The initial value of θ_k and δ_i^k are set equal to 1.

Table 2. Initial untrained rule-base

Rule Number	Rule Weight	Rule	Attribute weight			Belief Degree (F,R)
1	1	0 and Yes and Good	1	1	1	(0.80, 0.20)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	1	150 and Yes and Poor	1	1	1	(0.90, 0.10)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$K = 24$	1	600 and No and Poor	1	1	1	(0.00, 1.00)

Each attribute in dataset for BRB are transformed in form of a distribution to measure the matching degree to its set of referential values. The matching degree of the m^{th} data point to the v^{th} referential value of the i^{th} antecedent attribute is denoted by $\alpha_{v,i}(x_{m,i})$. For example, a customer x_m (a data point) have following referential values in each attribute:

$$x_m = [\mathrm{A}_1 = 532, \mathrm{A}_2 = \mathrm{Yes}, \mathrm{A}_3 = \mathrm{Average}] \tag{3}$$

The vector of transformed data point x_m is:

$$S(x_m) = [\{(\mathbf{0}, 0), (\mathbf{150}, 0), (\mathbf{300}, 0.226), (\mathbf{600}, 0.774)\}, \{(\mathbf{Yes}, 1), (\mathbf{No}, 0)\}, \{(\mathbf{Good}, 0), (\mathbf{Average}, 1), (\mathbf{Poor}, 0)\}] \tag{4}$$

In Eq. (4), the credit score 532 lies between 300 and 600. The matching degree for 600 is 0.887 (77.4%) and 356 is 0.226 (22.6%). It is closer to 600 than 300. Similarly, for qualitative attributes affordability = Yes and property type =

Average has 100% membership towards one referential value. The number of rules activated by a transformed data point $S(x_m)$ is equal to Cartesian product of number of non-zero matching degree in each attribute:

$$number\ of\ activated\ rules = \prod_{i=1}^{I} \sum_{v=1}^{V_i} \mathbb{I}(\alpha_{v,i}(x_{m,i}) \neq 0) \quad (5)$$

In Eq. (5), $\mathbb{I}(\alpha_{v,i}(x_{m,i}) \neq 0)$ is indication for non-zero matching degree for v^{th} referential value of the i^{th} antecedent attribute in m^{th} data point. The values of $(\alpha_{v,i}(x_{m,i}) \neq 0) = 1$ for non-zero matching degree, else it is ignored in counting of number of non-zero matching degrees in transformed data of an attribute. In Eq. (4), the transformed data point $S(x_m)$ will activate two rules $(2 \times 1 \times 1)$ out of 24 rules in the rule-base shown in Table 2.

2.3. *Aggregation of activated rules and training BRB*

The activation weight of k^{th} rule by a input data point x_m is given by

$$W_k(x_m) = \frac{\theta_k \prod_{i=1}^{I} (\alpha_{v,i})^{\overline{\delta^k}_i}}{\sum_{k=1}^{K} \left[\theta_k \prod_{i=1}^{I} (\alpha_{v,i})^{\delta_i^k}\right]} \text{ and } \overline{\delta^k}_i = \frac{\delta_i^k}{max_{i \in \{1,\dots,I\}} \{\delta_i^k\}} \quad \forall k \in \{1, \dots, K\} \quad (7)$$

where $\theta_k \in [0,1]$ $(k \in \{1, \dots, K\})$ is rule weight. $\delta_i^k \in [0,1]$ $(i \in \{1, \dots, I\})$ is the weight of antecedent attribute in the k^{th} rule. The term $\prod_{i=1}^{I} (\alpha_{v,i})^{\overline{\delta^k}_i}$ is the combined matching degree. The final inference output $o(\widehat{y_m})$ is generated by aggregating all the rules activated by the transformed data $S(x_m)$. The inference output $o(\widehat{y_m})$ generated by x_m can be represented in the following way:

$$o(\widehat{y_m}) = \{(h_n, \beta_n(x_m)), n \in \{1, \dots, N\}\} \quad (8)$$

The analytical evidential reasoning (ER) approach is used for the inference of output. The activated rules are combined together by ER. The rule weight, attribute weight, and belief degree are trained. The detailed methodology for ER approach and training of BRB can be seen here[5,6].

2.4. *Global Interpretability*

The decision making process of a BRB system can be interpreted globally by analysing the trained rule-base, as shown in Table 3. The trained continuous referential values of credit score is A_3: $credit\ score = [0,100,356,600]$. The attribute weight (δ_i^k) in each rule provides the global importance of attribute in the entire dataset. The rule weight (θ_k) represents overall importance of rule in a rule-base. Affordability is most important attribute in all the rules, Figure 3.

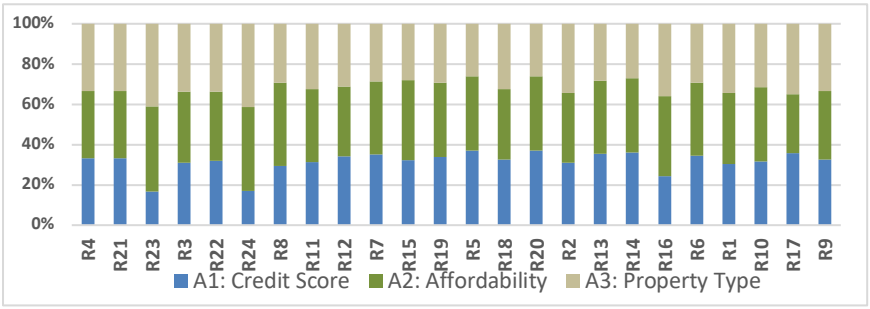

Fig. 3. Rules in trained rule-base arranged in descending order of their importance and global feature importance

Table 3. Trained rule-base

Rule Number	Rule Weight	Rule	Attribute weight			Belief Degree (F,R)
1	0.80	0 and Yes and Good	0.852	0.99	0.97	(0.70, 0.30)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	0.90	100 and Yes and Poor	0.89	0.95	0.89	(0.85, 0.15)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$K = 24$	1	600 and No and Poor	0.4	0.99	0.98	(0, 1.0)

2.5. *Local Interpretability*

A data point (i) represent an loan application or customer. It activates rules in the rule-base at varying degree of importance. The number of rules activated by a data point is given by Eq. (5). A decision by BRB can be explained locally for a single instance by measuring importance of activated rules and importance of attributes (features) in activated rules.

(i) Importance of activated rules ($\mathcal{I}_k(x_m)$): The importance of activated rules in a rule-base is measured by activation weight of each rule.

$$\mathcal{I}_k(x_m) = W_k(x_m) \tag{9}$$

In Eq. (9), $\mathcal{I}_k(x_m)$ is importance of k^{th} rule activated by data point x_m. The activation weight is zero for inactive rules. A high matching degree of a referential value in transformed data $S(x_m)$ has high importance of the rule and vice versa.

(ii) Importance of attributes in an activated rule ($\hat{\Delta}_{v,i}^k(x_m)$): The importance of v^{th} referential value of i^{th} attribute in k^{th} rule $\hat{\Delta}_{v,i}^k$ is measured by attribute weight (δ_i^k) and its matching degree ($\alpha_{,i}$) in transformed data.

$$\Delta_{v,i}^k(x_m) = attribute\ weight\ \times\ matching\ degree \tag{10}$$

The normalized $\Delta_{v,i}^{k}(x_m)$ for all attributes in a activated rules is:

$$\hat{\Delta}_{v,i}^{k}(x_m) = \frac{\Delta_{v,i}^{k}(x_m)}{\sum_{i=1}^{I} \Delta_{v,i}^{k}(x_m)} \tag{11}$$

The importance of rules and its attribute for two loan applications $x_1 = [\mathrm{A}_1 = 581, \mathrm{A}_2 = \mathrm{Yes}, \mathrm{A}_3 = \mathrm{Average}]$ and $x_2 = [\mathrm{A}_1 = 61, \mathrm{A}_2 = \mathrm{Yes}, \mathrm{A}_3 = \mathrm{Average}]$ is shown in Figure 4. Both applications activated 2 rule in the rule-base. The loan application x_1 activated rules R11 = [356 and Yes and Average] and R12 = [600 and Yes and Average] and x_2 activated rules R1= [0 and Yes and Good] and R2 = [100 and Yes and Good]. The aggregated decision in form of belief degree obtained by ER approach is {(F,0.94),(R,0.06)} and{(F,0.11),(R,0.89)} for x_1 and x_2, respectively. The ability to afford and average property type has high importance towards the decision to fund loan application x_1. Inability to afford the loan has highest importance towards rejection of loan application. For loan application x_2. The 3-Fold cross-validation was performed for BRB. The average accuracy of 3-fold cross-validation set was 0.95. A detailed comparison of accuracy of BRB and other machine learning methods for such system can be seen here[5,6].

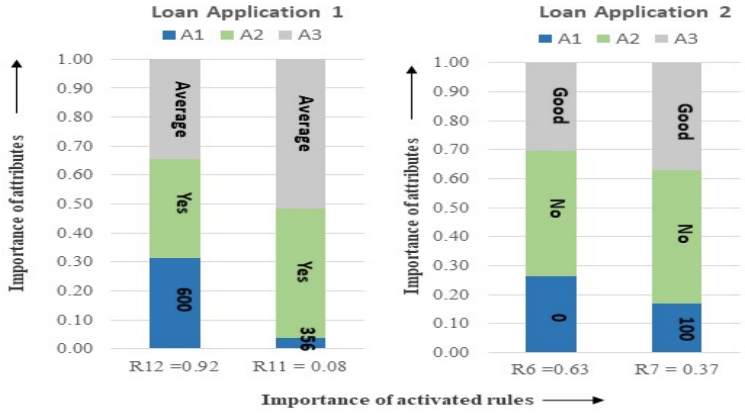

Fig. 4. Local interpretation (single instance) of loan application 1 and 2

3. Conclusion

In this paper, we have presented the global and local interpretability of BRB. The significance of rules and its attributes in a rule-base reveals overall (global) decision-making process by BRB. The importance of rules activated by a data point and importance of referential values in attributes of activated rules provide

local interpretation for single instance. We have demonstrated that BRB has a good trade-off between prediction accuracy and interpretability.

References

1. M. T. Ribeiro, S. Singh and C. Guestrin, Model-agnostic interpretability of machine learning. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), arXiv preprint arXiv:1606.05386.* (New-York, 2016).
2. S. M. Lundberg and S.I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing,* (NIPS2017), (Long Beach, CA, USA, 2017).
3. S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller and W. Samek, W, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PloS one*, ***10*** (2015).
4. G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K. R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, *Pattern Recognition*, ***65***, 211 (2017).
5. J. B. Yang, J. Liu, J. Wang, H. S. Sii and H. W. Wang, Belief rule-base inference methodology using the evidential reasoning approach-RIMER, *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, ***36***, 266 (2006).
6. S. Sachan, J. B. Yang, D. L. Xu, D. E. Benavides and Y. Li, An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, ***144***, 113100 (2020).
7. J. B. Yang, Y. M. Wang, D. L. Xu, K. S. Chin and L. Chatton, Belief rule-based methodology for mapping consumer preferences and setting product targets, *Expert Systems with Applications*, **39**, 4749 (2012).
8. D. L. Xu, J. Liu, J. B. Yang, G. P. Liu, J. Wang, I. Jenkinson, J. Ren, Inference and learning methodology of belief-rule-based expert system for pipeline leak detection, *Expert Systems with Applications*, **32**, 103 (2007).
9. G. L. Kong, D. L. Xu, J. B. Yang, X. F. Yin, T. B. Wang, B. G. Jiang and Y. H. Hu, Belief rule-based inference to predict trauma outcome, *Knowledge-based Systems*, **95**, 35 (2016).