Efficiently Explaining Decisions of Probabilistic RBF Classification Networks

Marko Robnik-Šikonja[1], Aristidis Likas[2], Constantinos Constantinopoulos[3], Igor Kononenko[1], and Erik Strumbelj[1]

[1] University of Ljubljana, Faculty of Computer and Information Science,
Tržaška 25, 1001 Ljubljana, Slovenia
{marko.robnik,igor.kononenko,erik.strumbelj}@fri.uni-lj.si

[2] University of Ioannina, Department of Computer Science,
GR 45110 Ioannina, Greece
arly@cs.uoi.gr

[3] Barcelona Media - Centre d'Innovació, Av. Diagonal, 177, planta 9,
08018 Barcelona, Spain
konstantinos.konstantinopoulos@barcelonamedia.org

Abstract. For many important practical applications model transparency is an important requirement. A probabilistic radial basis function (PRBF) network is an effective non-linear classifier, but similarly to most other neural network models it is not straightforward to obtain explanations for its decisions. Recently two general methods for explaining of a model's decisions for individual instances have been introduced which are based on the decomposition of a model's prediction into contributions of each attribute. By exploiting the marginalization property of the Gaussian distribution, we show that PRBF is especially suitable for these explanation techniques. By explaining the PRBF's decisions for new unlabeled cases we demonstrate resulting methods and accompany presentation with visualization technique that works both for single instances as well as for the attributes and their values, thus providing a valuable tool for inspection of the otherwise opaque models.

Keywords: classification explanation, model explanation, comprehensibility, probabilistic RBF networks, model visualization, game theory.

1 Introduction

In many areas where machine learning and data mining models are applied, their transparency is of crucial importance. For example, in medicine the practitioners are just as interested in the comprehension of the decision process, explanation of the model's behavior for a given new case, and importance of the diagnostic features, as in the classification accuracy of the model. The same is true for other areas where knowledge discovery dominates prediction accuracy.

Research in statistics, data mining, pattern recognition and machine learning is mostly focused on prediction accuracy. As a result we have many excellent

A. Dobnikar, U. Lotrič, and B. Šter (Eds.): ICANNGA 2011, Part I, LNCS 6593, pp. 169–179, 2011.

prediction methods, which are approaching the theoretically achievable prediction accuracy. Some of the most successful and popular approaches are based on support vector machines (SVM), artificial neural networks (ANN), and on ensemble methods (e.g., boosting and random forests). Regrettably none of these approaches offer an intrinsic introspection into their decision process or an explanation for labeling a new instance.

The probabilistic radial basis function network (PRBF) classifier [9, 10] is also a very effective black box classifier [9, 2]. PRBF is a special case of the RBF network [1] that computes at each output unit the density function of a class. It adopts a cluster interpretation of the basis functions, where each cluster can generate observations of any class. This is a generalization of a Gaussian mixture model [4, 1], where each cluster generates observations of only one class. In [2] an incremental learning method based on Expectation-Maximization (EM) for supervised learning is proposed that provides classification performance comparable to SVM classifiers. Unfortunately, like SVM, PRBF also lacks explanation ability.

Recently, in [6, 13, 12] general explanation methods have been presented that are in principle independent of the model (the model can be either transparent, e.g., decision trees and rules, or a black box, e.g., SVM, ANN and classifier ensembles) and can be used with all classification models that output probabilities. These explanation methods decompose the model's predictions into individual contributions of each attribute. Generated explanations closely follow the learned model and enable its visualization for each instance separately. The methods work by contrasting a model's output with the output obtained using only a subset of features. This demands either retraining of the model for several feature subsets which is computationally costly or using a simulation (e.g., averaging over all possible feature's values). For the Naive Bayesian (NB) classifier it was shown in [6] that due to the independence assumption of the classifier, one can simply omit attributes from classification and get classification with feature subsets without retraining or approximation.

In this paper we present an efficient decomposition for PRBF, which is also exact and avoids the approximation techniques thereby efficiently transforming PRBF into a white box non-linear classifier. We demonstrate how explanation methods from [6, 13] can explain the PRBF model's decisions for new unlabeled cases and present a graphical description of the otherwise opaque model. As a result, a highly effective classifier such as PRBF becomes transparent and can explain its individual decisions as well as its general behavior.

The aim of the paper is twofold, first to present a specific exact solution needed for PRBF explanation, and second to show how a general explanation technique can be applied to an opaque model to explain and visualize its decisions.

For a good review of neural network explanation methods we refer the reader to [3]. Techniques for extracting explicit (if-then) rules from black box models (such as ANN) are described in [11, 7]. Some less complex non-symbolic models enable the explanation of their decisions in the form of weights associated with each attribute. A weight can be interpreted as the proportion of the information contributed by the corresponding attribute value to the final prediction. Such explanations can be easily visualized. In [5] nomograms were developed for visualization of NB decisions. In [6] the NB list of information gains was generalized to a list of attribute weights for any prediction model.

Throughout the paper we use the notation where each of the n learning instances is represented by an ordered pair (x, y); each attribute vector x consists of individual values of attributes A_i, $i = 1, ..., a$ (a is the number of attributes), and is labeled with y representing one of the discrete class values y_j, $j = 1, ..., c$ (c is the number of class values). We write $p(y_j)$ for the probability of the class value y_j.

In Sect. 2 we present the explanation principle of [6] and [13]. In Sect. 3 we describe PRBF and the marginalization property of the Gaussian distribution which is exploited for explanation. In Section 4 we demonstrate instance and model based explanation possible with PRBF. Section 5 summarizes the properties of explanation and provides some directions for further work.

2 Explanation Methods

Following [6] we identify two levels of explanation: the *domain level* and the *model level*. The domain level tries to find the true causal relationship between the dependent and independent variables. Typically this level is unreachable unless we are dealing with artificial domains where all the relations as well as the probability distributions are known in advance. The model level explanation aims to make transparent the prediction process of a particular model. The prediction accuracy and the correctness of explanation at the model level are orthogonal: the correctness of the explanation is independent of the correctness of the prediction. However, empirical observation shows [13] that better models (with higher prediction accuracy) enable better explanation at the domain level.

We present two recently developed general explanation methods which are based on decomposition of a model's predictions into contributions of each attribute. Both methods work by explaining decisions taken by the model in classifying individual instances. To get a broader picture both methods combine these individual explanations and thereby provide also a better view of the model and problem. We first present the simpler of the two [6] (called EXPLAIN in the reminder of the text), which computes the influence of the feature value by changing this value and observing its impact on the model's output. The EXPLAIN assumes that the larger the changes in the output, the more important role the feature value plays is in the model. The shortcomings of this approach is that it takes into account only a single feature at time, therefore it cannot detect certain higher order dependencies (in particular disjunctions) and redundancies in the model. The method which solves this problem is called IME [13] and is presented in the second subsection.

2.1 Method EXPLAIN

The idea of the explanations proposed in [6] is to observe the relationship between the features and the predicted value by monitoring the effect on classification caused by the lack of knowledge of a feature's value. To monitor the effect that the attributes' values have on the prediction of an instance, the EXPLAIN method decomposes the prediction into individual attributes' values and define the model's probability $p(y|x\backslash A_i)$, as the model's probability of class y for instance x without the knowledge of event $A_i = a_k$ (marginal prediction), where a_k is the value of A_i for observed instance x. The comparison of the values $p(y|x)$ and $p(y|x\backslash A_i)$ provides insight into the importance of event $A_i = a_k$. If the difference between $p(y|x)$ and $p(y|x\backslash A_i)$ is large, the event $A_i = a_k$ plays an important role in the model; if this difference is small, the influence of $A_i = a_k$ in the model is minor. Due to its favorable properties the weight of evidence is an appropriate way how to evaluate the prediction difference, so we assume its use in this work. The odds of event z is defined as the ratio of the probability of event z and its negation: $\mathrm{odds}(z) = \frac{p(z)}{p(\overline{z})} = \frac{p(z)}{1-p(z)}$. The weight of evidence of attribute A_i for class value y is defined as the log odds of the class value y with the knowledge about the value of A_i and without it:

$$\mathrm{WE}_i(y|x) = \log_2(\mathrm{odds}(y|x)) - \log_2(\mathrm{odds}(y|x\backslash A_i)) \quad [bit] \tag{1}$$

To get the explanation factors an evaluation of (1) is needed. To compute factor $p(y|x)$ one just classifies the instance x with the model. To compute the factors $p(y|x\backslash A_i)$ the simplest, but not always the best option, is to replace the value of attribute A_i with a special unknown value (NA, don't know, don't care, etc.) which does not contain any information about A_i. This method is appropriate only for modeling techniques which handle unknown values naturally (e.g., the Naive Bayesian classifier just omits the attribute with unknown value from the computation). For other models we have to bear in mind that while this approach is simple and seemingly correct, each method has its own internal mechanism for handling unknown values. The techniques for handling unknown values are very different: from replacement with the most frequent value for nominal attributes and with median for numerical attributes to complex model-based implantations. To avoid the model dependent treatment of unknown values, a technique is suggested in [6] that simulates the lack of information about A_i with several predictions. For nominal attributes the actual value $A_i = a_k$ is replaced with all possible m_i values of A_i, and each prediction is weighted by the prior probability of the value resulting in the following equation

$$p(y|x\backslash A_i) = \sum_{s=1}^{m_i} p(A_i = a_s|x\backslash A_i)p(y|x \leftarrow A_i = a_s) \tag{2}$$

which can be approximated as

$$p(y|x\backslash A_i) \doteq \sum_{s=1}^{m_i} p(A_i = a_s)p(y|x \leftarrow A_i = a_s) \tag{3}$$

The term $p(y|x \leftarrow A_i = a_s)$ represents the probability for y when in x the value of A_i is replaced with a_s. The simplification is used for the prior probability $p(A_i = a_s)$ which implies that (3) is only an approximation.

2.2 Method IME

The main shortcoming of EXPLAIN method described above is that it observes only a change of a single feature at a time and therefore cannot detect disjunctive or redundant concepts expressed in a model. The solution to this is the method IME (Interactions-based Method for Explanation) proposed in [13] where all attribute interactions are scanned and the difference in predictions caused by each interaction is assigned to attributes taking part in each interaction. Such a procedure demands the generation of 2^a attribute subsets and is therefore limited to data sets with a relatively low number of features. Fortunately, this problem can be viewed from the point of coalitional game theory [12]. Within this framework the contributions assigned to individual feature values by IME method correspond to the Shapley value [8] and are therefore fair according to all interactions in which they are taking part. Furthermore a sampling approximation algorithm is presented in [12] which drops the requirement for all 2^a subsets and makes the method practically much more attractive.

Formally, for a given instance x the IME method introduces the notion of the prediction using the set of all features $\{1, 2, ...a\}$, the prediction using only a subset of features Q, and the prediction using the empty set of features $\{\}$. Let these predictions be $h(x_{\{1,2,...a\}})$, $h(x_Q)$, and $h(x_{\{\}})$, respectively. Since an instance x is fixed in explanation, for readability sake we omit it from expressions below, but remain aware that the dependence exists. The basis for explanation is therefore Δ_Q, the difference in predictions using the subset of features Q and the empty set

$$\Delta_Q = h(x_Q) - h(x_{\{\}}). \tag{4}$$

This difference in prediction is a result of an influence individual features may have, as well as the influence of any feature interactions $\mathscr{I}_Q$. The interaction contribution for the subset Q is defined recursively as $\mathscr{I}_Q = \Delta_Q - \sum_{W \subset Q} \mathscr{I}_W$ where $\mathscr{I}_{\{\}} = 0$. This definition takes into account that due to interactions of the features in the set Q the prediction may be different from prediction using any proper subset of Q.

The contribution π_i of the ith feature in classification of instance x is therefore defined as the sum of contributions of all relevant interactions

$$\pi_i = \sum_{Q \subseteq \{1,2,...,a\} \wedge i \in Q} \frac{\mathscr{I}_Q}{|Q|} \tag{5}$$

where $|Q|$ is the power of the subset Q. Equivalently, we can express this sum only with the differences in predictions [13].

$$\pi_i = \sum_{Q \subseteq \{1,2,\ldots,a\} - \{i\}} \frac{1}{a\binom{a-1}{a-|Q|-1}} (\Delta_{Q \cup \{i\}} - \Delta_Q) \tag{6}$$

Both, the complete algorithm and the sampling described in [12] require classification with a subsample of attributes. In the next Section we show that an efficient solution exists for PRBF, which does not require retraining of the classifiers for each feature subset, but still provides exactly the same classification.

3 Probabilistic RBF Networks

Consider a classification problem with c classes y_k $(k = 1, \ldots, c)$ and input instances $x = (A_1, \ldots, A_a)$. For this problem, the corresponding PRBF classifier has a inputs and c outputs, one for each class. Each output provides and estimate of the probability density $p(x|y_k)$ of the corresponding class y_k.

Assume that we have M components (hidden units), each one computing a probability density value $f_j(x)$ of the input x. In the PRBF network all component density functions $f_j(x)$ are utilized for estimating the conditional densities of all classes by considering the components as a common pool [9]. Thus, for each class a conditional density function $p(x|y_k)$ is modeled as a mixture model of the form:

$$p(x|y_k) = \sum_{j=1}^{M} \pi_{jk} f_j(x), \qquad k = 1, \ldots, c, \tag{7}$$

where the mixing coefficients π_{jk} are probability vectors; they take positive values and satisfy the following constraint:

$$\sum_{j=1}^{M} \pi_{jk} = 1, \qquad k = 1, \ldots, c. \tag{8}$$

Once the outputs $p(x|y_k)$ have been computed, the class of data point x is determined using the Bayes rule, i.e. x is assigned to the class with maximum posterior $p(y_k|x)$ computed by

$$p(y_k|x) = \frac{p(x|y_k)P_k}{\sum_{\ell=1}^{c} p(x|y_\ell)P_\ell} \tag{9}$$

The class priors P_k are usually computed as the percentage of training instances belonging to class y_k.

In the following, we assume the Gaussian component densities of the general form:

$$f_j(x) = \frac{1}{(2\pi)^{a/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\} \tag{10}$$

where $\mu_j \in \Re^a$ represents the mean of component j, while Σ_j represents the corresponding $a \times a$ covariance matrix. The whole adjustable parameter vector of the model consists of the mixing coefficients π_{jk} and the component parameters (means μ_j and covariances Σ_j).

It is apparent that the PRBF model is a special case of the RBF network, where the outputs correspond to probability density functions and the output layer weights are constrained to represent the prior probabilities π_{jk}. Furthermore, the separate mixtures model [4] can be derived as a special case of PRBF by setting $\pi_{jk} = 0$ for all classes k, except for the class that the component j belongs to. Training of the PRBF network is simple and efficient using the EM algorithm for likelihood maximization [9, 2].

3.1 Marginalization of PRBF

A notable convenient characteristic of the Gaussian distribution is the *marginalization property*: if the joint distribution of a set of random variables $S = \{A_1, \ldots, A_a\}$ is Gaussian with mean μ and covariance matrix Σ, then for any subset A of these variables, the joint distribution of the subset $Q = S - A$ of the remaining variables is also a Gaussian. The mean $\mu_{\backslash A}$ of this Gaussian is obtained by removing from μ the components corresponding to the variables in subset A and covariance matrix $\Sigma_{\backslash A}$ is obtained by removing the rows and columns of Σ corresponding to the variables in subset A. Therefore, if we know the mean and covariance of the joint distribution of a set of variables, we can immediately obtain the distribution of any subset of these variables.

For an input $x = (A_1 = v_1, \ldots, A_a = v_a)$ each output $p(x|y_k)$, $k = 1, \ldots, c$ of the PRBF is computed as a mixture of Gaussians:

$$p(x|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x; \mu_j, \Sigma_j) \tag{11}$$

Consequently, based on the marginalization property of the Gaussian distribution, it is straightforward to analytically compute $p(x\backslash\{A\}|y_k)$ obtained by excluding subset A of the attributes:

$$p(x\backslash\{A\}|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x\backslash\{A\}; \mu_{j\backslash\{A\}}, \Sigma_{j\backslash\{A\}}) \tag{12}$$

where $\mu_{j\backslash A}$ and $\Sigma_{j\backslash\{A\}}$ are obtained by removing the corresponding elements from μ_j and Σ_j. Then we can directly obtain $p(y_k|x\backslash\{A\})$ as

$$p(y_k|x\backslash\{A\}) = \frac{p(x\backslash\{A\}|y_k)P_k}{\sum_{\ell=1}^{c} p(x\backslash\{A\}|y_\ell)P_\ell} \tag{13}$$

and use it as a replacement for (2) and (3) in EXPLAIN method.

We use the same property in IME method and for a subset of features Q we get

$$h(x_Q|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x_Q; \mu_{jQ}, \Sigma_{jQ}) \tag{14}$$

where μ_{jQ} and Σ_{jQ} are obtained by retaining only the elements from Q in μ_j and Σ_j. We obtain $h(y_k|x_Q)$ as

$$h(y_k|x_Q) = \frac{h(x_Q|y_k)P_k}{\sum_{\ell=1}^{c} h(x_Q|y_\ell)P_\ell} \tag{15}$$

and use it in (4) and (6).

As a consequence, with PRBFS models EXPLAIN and IME explanation becomes much more efficient and exact as no approximation is needed. Classification with a subset of features requires only a mask which selects appropriate elements from μ and appropriate rows and columns from Σ matrix.

4 Instance Level and Model Level Explanation of PRBF

To show the practical utility of the proposed marginalization we demonstrate it by a visualization on a well-known Titanic data set. As there are no strong interactions in this data set we can use EXPLAIN method. IME methods produces similar graphs, and its details can be seen in [13].

The learning task is to predict the survival of a passenger in the disaster of the Titanic ship. The three attributes report the passenger's status during travel (first, second, or third class, or crew), age (adult or child), and gender of the passenger. Left-hand graph in Fig. 1 shows an example of an explanation for the decision of the PRBF network for an instance concerning a first class adult male passenger.

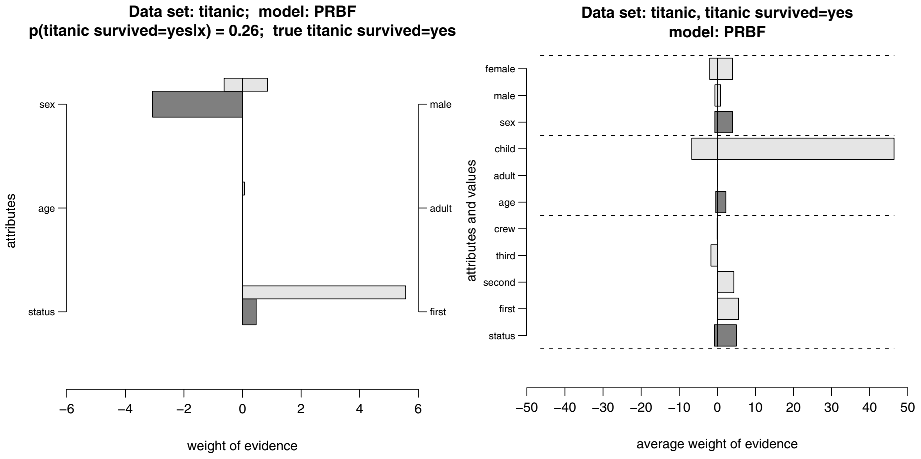

Fig. 1. Instance explanation (left-hand side) for one of the instances in the Titanic data set classified with PRBF model. The explanation of PRBF model is presented on right-hand side.

The weight of evidence is shown on the horizontal axis. The vertical axis contains names of the attributes on the left-hand side and their values for the chosen instance on the right-hand side. The probability $p(y|x)$ for class "survived=yes" returned by the PRBF model for the given instance x is reported on the top (0.26). The lengths of the thicker bars correspond to the influences of the given attribute values in the model, expressed by (1) and using (13). The positive weight of evidence is given on the right-hand side and the negative one is on the left-hand side. Thinner bars above the explanation bars indicate the average value of the weight of evidence over all training instances for the corresponding attribute value. For the given instance we observe that "sex = male" speaks strongly against the survival and "status=first class" is favorable for survival, while being adult has no influence. The exact probabilities of $p(y_k|x\backslash A)$ caused by the decomposition (13) are 0.73 for "sex = male", 0.19 for "status=first class", and 0.26 for "age=adult". Thinner average bars mostly agree with the effects expressed for the particular instance (being male is on average less dangerous than in the selected case, and being in the first class is even more beneficial).

To get a more general view of the model we can use explanations for the training data and visualize them in a summary form, which shows average importance of each feature and its values. An example of such visualization for titanic data set is presented in right-hand graph on Fig. 1.

On the left-hand side on the vertical axis, all the attributes and their values are listed (each attribute and its values are separated by dashed lines). For each of them the average negative and the average positive weights of evidence are presented with the horizontal bar. For each attribute (a darker shade) the average positive and negative influences of its values are given. For titanic problem status and sex have approximately the same effect, and age plays less important role in PRBF model. Particular values give even more precise picture: first and second class is perceived as undoubtedly advantageous, being a child or female has greater positive than negative effect, traveling in third class or being male is considered as a disadvantage, while a status of crew or being adult plays only a minor negative role.

5 Conclusions

We presented a decomposition for PRBF classifier by exploiting the marginalization property of the Gaussian distribution and applied this decomposition inside two general methods for explaining predictions for individual instances.

We showed how we can explain and visualize the classifications of unlabeled cases provided by the otherwise opaque PRBF model. We demonstrated a visualization technique which uses explanations of the training instances to describe the effects of all the attributes and their values at the model level.

The explanation methods EXPLAIN and IME exhibit the following properties:

- Model dependency: the decision process is taking place inside the model, so if the model is wrong for a given problem, explanation will reflect that and will be correct for the model, therefore wrong for the problem.

- Instance dependency: different instances are predicted differently, so the explanations will also be different.
- Class dependency: explanations for different classes are different, different attributes may have different influence on different classes (for two-class problems, the effect is complementary).
- Capability to detect strong conditional dependencies: if the model captures strong conditional dependency the explanations will also reflect that.
- EXPLAIN method is unable to detect and correctly evaluate the utility of attributes' values in instances where the change in more than one attribute value at once is needed to affect the predicted value. IME method samples the space of feature interactions and therefore avoids this problem. In PRBF it is straightforward and at no extra cost to marginalize any number of features, which is a useful property for sampling subsets of features.
- Visualization ability: the generated explanations can be graphically presented in terms of the positive/negative effect each attribute and its value have on the selected class.

While explanation methodology presented has been successfully used in medical application [13], in future work we plan to use also PRBF and use the feedback provided by the experts to further improve the visualization tool.

References

[1] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)

[2] Constantinopoulos, C., Likas, A.: An incremental training method for the probabilistic RBF network. IEEE Trans. Neural Networks 17(4), 966–974 (2006)

[3] Jacobsson, H.: Rule extraction from recurrent neural networks: A taxonomy and review. Neural Computation 17(6), 1223–1263 (2005)

[4] McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley & Sons, Chichester (2000)

[5] Možina, M., Demšar, J., Kattan, M.W., Zupan, B.: Nomograms for visualization of naive bayesian classifier. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 337–348. Springer, Heidelberg (2004)

[6] Robnik Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering 20(5), 589–600 (2008)

[7] Setiono, R., Liu, H.: Understanding neural networks via rule extraction. In: Proceedings of IJCAI 1995, pp. 480–487 (1995)

[8] Shapley, L.S.: A value for n-person games. In: Contributions to the Theory of Games, vol. II. Princeton University Press, Princeton (1953)

[9] Titsias, M.K., Likas, A.: Shared kernel models for class conditional density estimation. IEEE Trans. Neural Networks 12(5), 987–997 (2001)

[10] Titsias, M.K., Likas, A.: Class conditional density estimation using mixtures with constrained component sharing. IEEE Trans. Pattern Anal. and Machine Intell. 25(7), 924–928 (2003)

[11] Towell, G.G., Shavlik, J.W.: Extracting refined rules from knowledge-based neural networks. Machine Learning 13(1), 71–101 (1993)

[12] Štrumbelj, E., Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory. Journal of Machine Learning Research 11, 1–18 (2010)

[13] Štrumbelj, E., Kononenko, I., Robnik-Šikonja, M.: Explaining instance classifications with interactions of subsets of feature values. Data & Knowledge Engineering 68(10), 886–904 (2009)