Neurocomputing (2012) 96:37-46

Quality of Classification Explanations with PRBF

Marko Robnik-Šikonja, Igor Kononenko, Erik Štrumbelj
University of Ljubljana,
Faculty of Computer and Information Science,
Tržaška 25, 1001 Ljubljana, Slovenia
e-mails: {marko.robnik,igor.kononenko,erik.strumbelj}@fri.uni-lj.si

March 20, 2013

Abstract

Recently two general methods for explaining classification models and their predictions have been introduced. Both methods are based on an idea that importance of a feature or a group of features in a specific model can be estimated by simulating lack of knowledge about the values of the feature(s). For the majority of models this requires an approximation by averaging over all possible feature values. A probabilistic radial basis function network (PRBF) is one of the models where such approximation is not necessary and therefore offers a chance to evaluate the quality of approximation by comparing it to the exact solution.

We present both explanation methods and demonstrate their behavior with PRBF. The explanations make individual decisions of classifiers transparent and allow inspection and visualization of otherwise opaque models.

We empirically compare the quality of explanations based on marginalization of the Gaussian distribution (the exact method) and explanation with averaging over all feature values (the approximation). The results show that the approximation method and the exact solution give very similar results, which increases the confidence in the explanation methodology also for other classification models.

Keywords: classification explanation, model explanation, quality of explanation, model comprehensibility, probabilistic RBF networks, model visualization

1 Introduction

Researchers in statistics, data mining, pattern recognition and machine learning are mostly focused on prediction accuracy. As a result we have many excellent prediction methods, which are approaching the theoretically achievable prediction accuracy. Some of the most successful approaches are support vector machines (SVM), artificial neural networks (ANN), and ensemble methods (e.g., boosting and random forests). Regrettably none of these approaches offers an intrinsic introspection into its decision process or an explanation of predictions.

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

The probabilistic radial basis function network (PRBF) classifier [34, 35] is an effective black box classifier [34, 5]. PRBF is a special case of the RBF network [3] 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 [22, 3], where each cluster generates observations of only one class. In [5] 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 [28, 38, 37] two general explanation methods IME and EXPLAIN have evolved that are in principle independent of the prediction model and can be used with all classification models that output probabilities. The model can be either transparent, e.g., decision trees and rules, or a black box, e.g., SVM, ANN and classifier ensembles. 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 an approximation (e.g., averaging over all possible feature's values). The quality of this approximation is one of the topics of this work.

We present an efficient decomposition for PRBF, which allows exact explanations thereby efficiently transforming PRBF into a white box non-linear classifier. We demonstrate how explanation methods from [28, 38] 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.

To our knowledge there are only two methods which do not need approximations in the explanation procedure: Naive Bayesian classifier (NB) and (P)RBF networks. For the NB classifier it was shown in [28] that due to the independence assumption of the classifier, the exact explanation method (omitting the attribute from the classifier) is exactly equivalent to the approximation (averaging over all feature's values). Since approximation is possible but not needed for PRBF explanations, PRBF offers a unique chance to evaluate the quality of the approximation procedure by comparing it to the exact method.

The aim of the paper is threefold, first to present a specific exact solution needed for PRBF explanation, second to demonstrate how a general explanation technique can be applied to an opaque model to explain and visualize its decisions, and third, in contrast to [27], to evaluate the approximation procedure used by the explanation methods.

1.1 Related work

We first discuss some representative explanation methodologies specific to the neural computing, while for a more complete review we refer the reader to [14]. Techniques for extracting explicit (if-then) rules from black box models (such as ANN) are described in [36, 7, 30, 33]. In [8] a black box model is approximated with a symbolic model, such as decision tree, by sampling additional training instances from the model. Extraction of fuzzy rules from trained ANNs using interval propagation is suggested in [24], while an approach which can provably extract sound and non-monotonic rules was presented in [9]. In [36] a search-based method has been proposed suitable for ANNs with binary units only, while the application of the method proposed in [30] requires discretization of the hidden unit activations and pruning of the ANN architecture. By extracting these models from trained ANNs the knowledge stored in the ANN is represented in a symbolic form. To obtain the explanation of a prediction, one simply has to read the corresponding rule in the extracted symbolic model (rules, trees, ...). Whether such explanation is comprehensive in the case of large trees or rule sets is questionable. Framling [11] explains neural network behavior by using contextual importance and utility of attributes which are defined as the range of changes of attribute and class values. Importance and utility can be efficiently computed only in a special INKA network. In contrast, IME and EXPLAIN methods described here are based on marginalization and can be efficiently computed for any probabilistic classifier.

The works summarized next are not aimed directly at neural network models but are relevant to our discussion as they aim to provide explanation and visualization principles and applications to specific methods. 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. For logistic regression a well known approach is to use nomograms, first proposed in [20]. In [15] nomograms were developed for SVM, but they work only for a restricted class of kernels and cannot be used for general non-linear kernels. SVMs can also be visualized using projections into lower dimensional subspaces [4, 25] or using self-organizing maps from unsupervised learning [13]. These methods concentrate on visualization of the separating hyperplane and decision surface and do not provide explanations for individual decisions. The Naive Bayesian classifier is able to explain its decisions as the sum of information gains [17]. A straightforward visualization of NB was used in [2] while in [23] nomograms were developed for visualization of NB decisions. In [28] the NB list of information gains was generalized to a list of attribute weights for any prediction model.

Madigan [21] consider the case of belief networks and use each (binary or multi-valued discrete) attribute as a node in the graph. By computing “evidence flows” in the network it is possible to explain its decisions. This approach is not general and cannot be applied to PRBF because for it the whole input vector (containing all attributes) is considered a single variable that is assumed to be sampled from multivariate Gaussian distributions. Also PRBF assumes continuous attributes as inputs. The negation of continuous attributes is meaningless, but would be needed to compute weights of evidence as defined by [21].

In ExplainD framework [26] weights of evidence and visualizations similar to EXPLAIN are used, but the explanation approach is limited to (linear) additive models, while EXPLAIN can be used for all probabilistic models. As PRBF is not linear in the space of features, but rather in the space of Gaussian components, where each component presents multivariate Gaussian kernel, it cannot be visualized by ExplainD. Additionally, EXPLAIN is not restricted to the weight of evidence, but can use also other interpretations of difference of probabilities (e.g., information difference).

In the context of feature subset selection, attributes are evaluated in [18] as the difference between the correct and perturbed output, which is similar to EXPLAIN approach to the model level explanation [28]. Lemaire [19] has extended their approach to instance level explanations and applied it to a customer relationship management system in telecommunications industry. The marginalization of PRBF presented here, which allows to omit output perturbations would benefit also the approach of Lemaire. However both EXPLAIN and the method of Lamaire are limited as they cannot detect disjunctive concepts and redundancies in the model. These deficiencies are ameliorated in the method IME which is described in Section 2.2.

Finally, similar to most other explanation methods, presented explanations are related to statistical sensitivity analysis and uncertainty analysis [29]. In that methodology sensitivity of models is analyzed with respect to models' input which is what we call model level explanation. Presented visualization of averaged explanations can therefore be viewed as a form of sensitivity analysis.

1.2 Taxonomy

Andrews et al. [1] introduced a taxonomy and the criteria for evaluation of rule extraction methods from ANN. Our approach for PRBF via EXPLAIN and IME is graphical, does not produce rules, and is oriented towards explaining effects for individual learning instances. It is therefore not directly comparable, but some aspects of the taxonomy are relevant to it, so we put it into perspective. According to [1, 14] the explanation techniques for neural networks can be classified according to:

1. *expressive power* describing the language of extracted knowledge: propositional logic (i.e.. if-then rules), nonconventional logic (e.g., fuzzy logic), first-order logic, and finite state machines (deterministic, nondeterministic, stochastic, e.g., Moore and Mealy types).

2. *translucency* describing the degree to which an explanation method looks inside the ANN: decompositional (combining rules for individual neurons), pedagogical (treating ANN as a black box), and eclectic methods (combining both compositional and pedagogical types),

3. *portability* describes how well the technique covers the set of available ANN architectures,

4. *quality* of the extracted rules, which can be further split into several criteria: accuracy (generalization ability), fidelity (how well the rule reflects behavior of ANN), consistency (similarity of behavior for different ANN trained on the same task), and comprehensibility (readability and size of the rules).

5. *algorithmic complexity.*

By imbedding EXPLAIN and IME into the taxonomy above, we can classify them as follows:

1. Expressive power: extracted explanations are in the form of contributions of individual attributes (and for IME also their combinations), which speak in favor/against the given class; these contributions can be seen as a limited form of propositional logic.

2. Translucency: both methods are pedagogical (the model is treated as a black-box, only the causal relationship between input and output is considered).

3. Portability: the approaches are general and applicable to arbitrary ANN architecture (nevertheless, the contribution of this paper is based on the specific property of PRBF networks, i.e., marginalization of Gaussians, see Sect. 3.1).

4. Quality of extracted knowledge: while accuracy and consistency are not an appropriate criteria as they apply only to rule based methods, the produced explanations are fidel to the learned model [28]. Concerning comprehensibility EXPLAIN and IME can be considered very good due to their simplicity and graphical visualization.

5. Algorithmic complexity: the marginalization of PRBF with Eqs. (16) and (18) can be efficiently computed by masking or selecting appropriate elements from vector μ and matrix Σ. The generation of explanations does not affect the learning phase for PRBF. If a is a number of attributes, than for a single instance and for a particular model we need $O(a)$ classifications for EXPLAIN (one for each attribute). For IME we need 2^a classifications with a model using all subsets of features, but a sampling methods reduces this to $O(a)$ classifications. In practice, the time required to generate individual explanations is negligible for EXPLAIN, and reasonable for IME.

1.3 Notation and Organization

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 [28] and [38]. 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 with PRBF. In Section 5 we compare approximate and exact explanations generated for PRBF models. Section 6 summarizes the contributions of the paper and provides some directions for further work.

2 Explanation Methods

Following [28] 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 aim is 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 observations show that better models (with higher prediction accuracy) enable better explanation at the domain level [38].

In this paper we use the following terminology:

- *instance explanation*: explanation of a classification of a single instance with given model i.e., at the model level,
- *model explanation*: average of explanations over many training instances at the model level, which provide more general explanations of features and their values. This averaging over many instances enables identification of different roles attributes may play in the classifications of instances. To avoid loss of information due to averaging in our visualization we collect evidence for and against each class separately. In this way we can, for example see that a particular value of an attribute may support specific class but it may not always be so.
- *domain explanation*: by definition this level is unreachable, however, if the accuracy of the model is high, it should be quite similar to the model explanation.

We present two general explanation methods which are based on decomposition of a model's predictions into contributions of each attribute. Both methods work by explaining decisions of the model in classifying individual instances. To get a broader picture both methods combine the individual explanations and thereby provide a better view of the model and problem. We first present the simpler of the two [28] (called EXPLAIN), which computes the influence of the feature value by 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 [38] and is presented in the second subsection.

2.1 Method EXPLAIN

The idea of the explanations proposed in [28] 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 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.

For the evaluation of the prediction difference there are several options: information difference, the weight of evidence (or equivalently log-odds ratio or Kullback-Leibler divergence) [31, 12, 6], or using difference of probabilities directly.

$$\mathrm{probDiff}_i(y|x) = p(y|x) - p(y|x\backslash A_i) \tag{1}$$

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: $\mathrm{odds}(z) = \frac{p(z)}{p(\bar{z})} = \frac{p(z)}{1-p(z)}$. 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:

$$\mathrm{WE}_i(y|x) = \log_2(\mathrm{odds}(y|x)) - \log_2(\mathrm{odds}(y|x\backslash A_i)) \;\; [bit] \tag{2}$$

By restricting the reasoning to the model, the explanations are provided also for multi-attribute dependencies (interactions), where each of the dependent attributes A_i affects the prediction and so also the score $\mathrm{probDiff}_i(y|x)$, therefore the explanations for all A_i are nonzero. In this way the generated explanations not only provide information about simple one-attribute-at-a-time dependencies but also about complex multi-attribute dependencies, as long they are expressed in a given model.

A weakness of this methodology is that it is not able to 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 (disjunctions). For example, in the model expressing $C = A_1 \vee A_2$, for instance $x = (A_1 = 1,\ A_2 = 1)$ the probability of class value 1 is $p(C = 1|x) = 1$. Decompositions on any single attribute would make no difference in probability $p(C = 1|x\backslash A_1) = p(C = 1|x\backslash A_2) = 1$ and $\mathrm{probDiff}_1(C = 0|x) = \mathrm{probDiff}_2(C = 0|x) = 0$, and therefore both A_1 and A_2 are incorrectly considered irrelevant for classification of this instance. However, for this model other combinations of values for A_1 and A_2 are explained correctly. This problem is overcome with IME method described below, which takes into account interactions of attributes.

To get the explanation factors an evaluation of (2) 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 imputations. To avoid the model dependent treatment of unknown values, a technique is suggested in [28] 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

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

which can be approximated as

$$p(y|x\backslash A_i) \doteq \sum_{s=1}^{m_i} p(A_i = a_s)p(y|x \leftarrow A_i = a_s) \tag{4}$$

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 (4) is only an approximation.

For numerical attributes the procedure is similar. A discretization splits the values of A_i into sub-intervals. The middle points of these sub-intervals are taken as the representative replacement values in (4) for which predictions $p(y|x \leftarrow A_i = a_s)$ are computed. Instead of prior probabilities of single values $p(A_i = a_s)$, the probabilities of the sub-intervals are used for weighting the predictions.

The generation of explanations does not affect the learning phase; for a single instance and for a particular model we need $O(a)$ predictions (at least one prediction for each attribute).

In Sect. 3 we show that in the case of PRBF networks, the approximation (4) is not needed and the lack of information about attribute values can be handled in a way which is computationally efficient (does not require relearning of the model or several predictions) and sound (we get exactly the same results as by learning without the knowledge of the given attribute).

Next we present a method which takes into account interactions between attributes.

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 [38] 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 [37]. Within this framework the contributions assigned to individual feature values by IME method correspond to the Shapley value [32] and are therefore fair according to all interactions in which they are taking part. Furthermore a sampling approximation algorithm is presented in [37] which drops the requirement for all 2^a subsets. Experiments have shown that the improved method is practically usable for data sets with up to 100 attributes.

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

$$\Delta_Q = h(x_Q) - h(x_{\{\}}). \tag{5}$$

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

$$\pi_i = \sum_{Q \subseteq \{1,2,...,a\} \wedge i \in Q} \frac{\mathscr{I}_Q}{|Q|} \tag{6}$$

where $|Q|$ is the power of the subset Q. Equivalently, we can express this sum only with the differences in predictions [38].

$$\pi_i = \sum_{Q \subseteq \{1,2,...,a\} - \{i\}} \frac{1}{a\binom{a-1}{a-|Q|-1}} (\Delta_{Q \cup \{i\}} - \Delta_Q) \tag{7}$$

Viewed from the point of coalitional game theory, Eq. (7) corresponds to the Shapley value for the coalitional game of a players with Δ as the characteristic function. For larger a, the computation of Eq. (7) becomes infeasible, so an alternative formulation is used:

$$\begin{aligned} \pi_i &= \frac{1}{a!} \sum_{\mathscr{O} \in \pi(a)} \left(\Delta_{Pre^i(\mathscr{O}} \cup \{i\}) - \Delta_{Pre^i(\mathscr{O})} \right) = \\ &= \frac{1}{a!} \sum_{\mathscr{O} \in \pi(a)} \left(h(x_{Pre^i(\mathscr{O}) \cup \{i\}}) - h(x_{Pre^i(\mathscr{O})}) \right), \end{aligned} \tag{8}$$

where $R(a)$ is the set of all permutations of a elements and $Pre^i(\mathscr{O})$ is the set of all input variables which precede the i-th variable in permutation $\mathscr{O} \in R(a)$.

The conditional expectations in Eq. (8) are difficult to compute for an arbitrary model, so in [37] an additional step is taken to simplify the approximation. However, because we can easily compute the conditional predictions $h(x_C)$ for the PRBF model, we can use the following unbiased estimator:

$$\hat{\pi}_i = \frac{1}{m}\sum_{j=1}^{m}\left(h(x_{Pre^i(\mathcal{O}_|)\cup\{i\}}) - h(x_{Pre^i(\mathcal{O}_|)})\right), \tag{9}$$

where $\mathcal{O}_\infty, ..., \mathcal{O}_\Updownarrow$ is a sequence of permutations randomly drawn from $R(a)$.

The number of samples needed to achieve a certain degree of confidence about the approximation error of an attribute's contribution is proportionate to the variance of the estimator in Eq.(9). In Section 5.2 we show that the variance of the terms in Eq.(9) is lower than the variance of the terms in the originally proposed estimator, which makes the computation of the explanation for the PRBF classifier more efficient.

3 Probabilistic RBF Networks

Consider a classification problem with c classes y_k $(k = 1, \dots, c)$ and input instances $x = (A_1, \dots, A_a)$. 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 [34]. Thus, for each class a conditional density function $p(x|y_k)$ is modeled as a mixture model of the form:

$$p(x|y_k) = \sum_{j=1}^{M} \pi_{jk} f_j(x), \quad k = 1, \dots, c, \tag{10}$$

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

$$\sum_{j=1}^{M} \pi_{jk} = 1, \quad k = 1, \dots, c. \tag{11}$$

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

$$p(y_k|x) = \frac{p(x|y_k)P_k}{\sum_{\ell=1}^{c} p(x|y_\ell)P_\ell} \tag{12}$$

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:

$$f_j(x) = \frac{1}{(2\pi)^{a/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\} \tag{13}$$

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 [22] 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 [34, 5].

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 $S = \{A_1, \dots, A_a\}$ 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 $x = (A_1 = v_1, \dots, A_a = v_a)$ each output $p(x|y_k)$, $k = 1, \dots, c$ of the PRBF is computed as a mixture of Gaussians:

$$p(x|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x; \mu_j, \Sigma_j) \tag{14}$$

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:

$$p(x \backslash \{A\}|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x \backslash \{A\}; \mu_{j \backslash \{A\}}, \Sigma_{j \backslash \{A\}}) \tag{15}$$

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

$$p(y_k|x \backslash \{A\}) = \frac{p(x \backslash \{A\}|y_k) P_k}{\sum_{\ell=1}^{c} p(x \backslash \{A\}|y_\ell) P_\ell} \tag{16}$$

and use it as a replacement for (3) and (4) in EXPLAIN method.

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

$$h(x_Q|y_k) = \sum_{j=1}^{M} \pi_{jk} N(x_Q; \mu_{jQ}, \Sigma_{jQ}) \tag{17}$$

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

$$h(y_k|x_Q) = \frac{h(x_Q|y_k)P_k}{\sum_{\ell=1}^{c} h(x_Q|y_\ell)P_\ell} \tag{18}$$

and use it in (5) and (7).

As a consequence, with PRBF models EXPLAIN and IME explanation becomes much more efficient and exact as no approximation of the conditional predictions 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 use EXPLAIN method. IME method produces similar graphs, and its details can be seen in [38].

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.

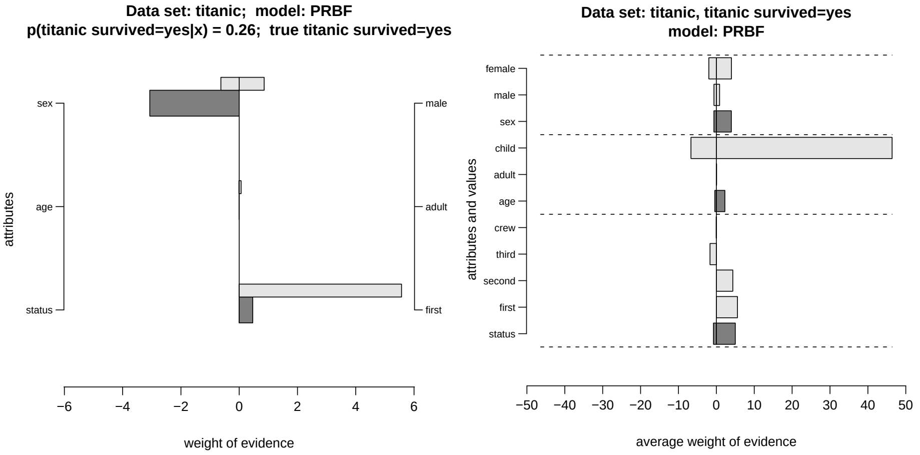


Figure 1: Instance explanation (left-hand side) for one of the instances in the Titanic data set classified with PRBF model. The explanation of PRBF model as a whole is presented on right-hand side.

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 (2) and using (16). 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 (16) 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).

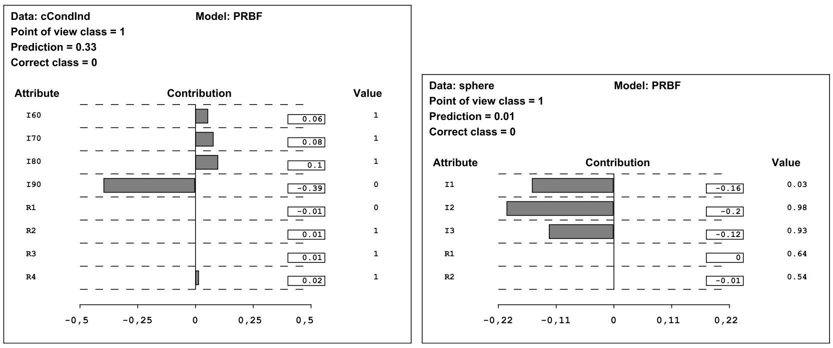
While in our simple example we present all the attributes, for domains with many attributes a threshold parameter can select only attributes with sufficient impact for presentation. In some application domains (for example in medicine) experts interpret similar graphs as points in favor/against the decision and expect that the total sum of points will be 1 (or 100), which is also possibility with our approach.

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 of Fig. 1.

On the left-hand side on the vertical axis, all the attributes and their values are listed (attributes 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 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.

Figures 2(a) and 2(b) show two instance level explanations computed using IME method (for description of data sets used see Section 5). The visualization and its interpretation are the same as those for Fig. 1, but without the bars that indicate the average contribution of the feature. For example, Fig. 2(a) shows how the first three important attributes, whose values are 1, contribute towards the class value 1, while the most important attribute is 0 and strongly contributes against class value 1. The latter attribute is the main contributor towards the models' prediction of 0.33 probability of the class value being 1. In this case, the classification is correct, because the actual class value was indeed 0.

The advantage of the IME method is that it takes into account the interactions between attributes when computing their contribution. For example, the instance from the sphere data set (see Fig. 2(b)) is a good example of attribute value redundancy. All three important coordinates are far enough from the center of the sphere so that knowing just two of them would be sufficient to predict that the point lies outside the sphere. Therefore, when using a one-attribute-at-a-time EXPLAIN approach, all three



(a) condInd data set

(b) sphere data set

Figure 2: Two instance level explanations for the PRBF classifier using the IME method.

important attributes would be marked as unimportant (that is, they would get a zero contribution), which is incorrect and misleading. On the other hand, IME method captures the interactions and correctly assigns a non-zero contribution to each important attribute.

5 Quality of Approximation

Methods EXPLAIN and IME can explain PRBF classifications using approximation with (4) or with the exact marginalized solution given by (16) and (18). These two possibilities allow us to test the quality of the approximation.

To make results comparable with previous work we use the following measure for closeness of explanations. Let a distance between two explanations (e and f) for an instance x be an average distance over all the attributes:

$$d_{e,f}(x) = \frac{1}{a}\left(\sum_{i=1}^{a}\left(\frac{1}{2}\left(\frac{e_i(x)}{\sum_{j=1}^{a}|e_i(x)|} - \frac{f_i(x)}{\sum_{j=1}^{a}|f_i(x)|}\right)\right)^2\right)^{\frac{1}{2}} \tag{19}$$

Here $e_i(x)$ and $f_i(x)$ represent the explanations (i.e., the contributions) of the attribute A_i for the prediction of the instance x. For attribute A_i the $e_i(x)$ and $f_i(x)$ can be either positive (if the attribute's value supports the selected class) or negative (if the attribute's value supports some other class). In the computation of the average explanation distance over all the instances, we want all instance explanations to be equally represented, therefore the absolute values of the model's explanations are normalized to 1 for each instance: we divide the $e_i(x)$ and $f_i(x)$ with the sum of their respective absolute values, and multiply them by $\frac{1}{2}$ (the explanations are thereafter normalized to $[0,1]$).

Let $\overline{d_{e,f}}$ be an average of (19) over all testing instances

$$\overline{d_{e,f}} = \frac{1}{n}\sum_{i=1}^{n} d_{e,f}(x_i). \tag{20}$$

The $\overline{d_{e,f}}$ is an indicator of how close are the model's explanations for two explanation techniques.

Following [28] and [38] we use the following artificial data sets.

condInd The class value is a boolean variable with 50% probability of 1. The four important conditionally independent attributes have the same value as class in 90, 80, 70 and 60% of the cases, respectively. There are also 4 attributes unrelated to the class (random) in this data set.

xor The data set consists of three important attributes describing a parity problem of order 3. We added noise to the class value by reverting it in 10% of the cases. There are also 3 attributes unrelated to the class.

groups Two important attributes I_1 and I_2 define three class values scattered around group centers. Two attributes are unrelated to the class.

cross This problem has two important attributes I_1 and I_2 defining a cross like distribution of values with separated class values. There are also 4 attributes unrelated to the class.

chess Two important attributes define a 4x4 chessboard with class values being the color (black or white). Two attributes are unrelated to the class.

sphere The data can be described as a sphere inscribed into the unit cube. The points within the sphere belong to one class and the remaining points to the other. The three important attributes correspond to three dimensions. There are two additional irrelevant attributes.

For all data sets we generated 1000 training instances and another 1000 to test explanation methods.

5.1 Method EXPLAIN

Method EXPLAIN assigns contributions to individual attributes based on their effect on classification. For PRBF the contributions can be computed with the marginalization (16) or with the approximation given by (4). The explanations using marginalization (16) are denoted e_{marg} in the results below. We measure the quality of approximation by comparing exact explanations (16) with approximations (4), namely we compute the average distance between the explanations (20). Note that for (4) numerical attributes have to be discretized and the discretization technique used also affects the quality of explanations. For the artificial data sets defined above the optimal discretization bounds for all the numerical attributes are known. The explanations using this optimal discretization are labeled with d_o. We tried also several well-known discretization techniques, but since they consider only class dependencies of single attributes, they are not appropriate for data sets where class is conditionally dependent on several attributes. Some discretization techniques returned no intervals (for example [10]), and others returned hundred of intervals (for example [16]). To observe the effect of discretization we therefore decided to use simple equal-width discretization, but increased the number of intervals beyond the known optimal number by 2 or 10. We label the explanations using these discretizations d_{o+2} and d_{o+10}, respectively.

In [28] the "true explanations" for the above data sets and method EXPLAIN were suggested based on the known properties of the distributions and the optimal models. For example, in *chess* the two important attributes with coordinate values equally contribute to the determination of the class value, so for the selected class their true explanations is 0.5 (-0.5 for the opposite class), and for irrelevant attributes it is 0. The true explanations for the data set *sphere* were not defined. The average distances between different explanations are collected in Table 1.

Table 1: The classification accuracy and average distance between different explanations with EXPLAIN method. For sphere data set the "true explanation" is not defined. Attributes in condInd and xor data sets are already discrete, so the $\sim$ sign indicates no change from ideal discretization.

		distance to "true explanation"				distance to e_{marg}		
data set	ca%	e_{marg}	d_o	d_{o+2}	d_{o+10}	d_o	d_{o+2}	d_{o+10}
condInd	89	0.07	0.07	$\sim$	$\sim$	0.01	$\sim$	$\sim$
groups	99	0.03	0.03	0.04	0.06	0.01	0.01	0.11
xor	90	0.12	0.12	$\sim$	$\sim$	0.07	$\sim$	$\sim$
cross	98	0.13	0.13	0.25	0.15	0.01	0.02	0.11
chess	54	0.19	0.11	0.50	0.19	0.10	0.39	0.10
sphere	84	-	-	-	-	0.01	0.01	0.01

The left-hand side of the table shows the average difference of given explanations to the "true explanation". Comparison between the exact method (column e_{marg}) and approximation with ideal discretization (column d_o) shows that there is practically no difference in the quality of explanation. Lower difference for d_o in *chess* is surprising only at first sight as the model in this data set is practically unusable (54% classification accuracy in balanced two class problem), which means that the explanations are unreliable as observed also by [28]. Using imperfect discretizations (columns d_{o+2} and d_{o+10}) increases the distance to "true explanation", but the difference is small.

The right-hand side of the Table 1 presents the differences between exact explanation and approximations. The differences are small, but larger for imperfect discretizations. For *chess* where the learned model is unreliable the differences are larger and vary more.

The conclusion based on both sides of Table 1 is optimistic for EXPLAIN method. The differences between approximation and exact method are very small, indicating that classification models for which no exact explanations exist will still get reliable explanations using the approximation.

5.2 Method IME

As mentioned in Section 2.2, the number of samples needed to compute the IME contribution for an attribute to a certain degree of accuracy is proportionate to the variance of the estimator. Therefore, the total number of samples to compute the contribution of all a attributes for an instance is proportionate to the mean variance across all attributes.

We compared the original algorithm and the improvement using the marginalization property by training the PRBF and estimating the mean variance within a test instance across several test instances. We used 1000 samples per feature across 50 instances for each data set for a total of $50 \times 1000 \times a$ samples for each data set, where a is the number of features. For comparison, we also included the Naive Bayes classifier, for which the marginalization property also holds [28].

Table 2: Mean variance of the original algorithm $\overline{\sigma^2}$ and the proposed marginalization-based improvement $\overline{\sigma^2_{marg}}$.

	PRBF			Naive Bayes		
data set	ca%	$\overline{\sigma^2}$	$\overline{\sigma^2_{marg}}$	ca%	$\overline{\sigma^2}$	$\overline{\sigma^2_{marg}}$
condInd	89	$2.67 \cdot 10^{-2}$	$8.66 \cdot 10^{-3}$	91	$3.22 \cdot 10^{-2}$	$3.52 \cdot 10^{-3}$
groups	99	$1.91 \cdot 10^{-1}$	$4.16 \cdot 10^{-2}$	32	$5.91 \cdot 10^{-5}$	$2.71 \cdot 10^{-9}$
xor	90	$1.59 \cdot 10^{-1}$	$5.41 \cdot 10^{-2}$	50	$3.76 \cdot 10^{-4}$	$3.52 \cdot 10^{-9}$
cross	98	$6.95 \cdot 10^{-2}$	$1.41 \cdot 10^{-2}$	53	$2.35 \cdot 10^{-3}$	$1.19 \cdot 10^{-7}$
chess	54	$2.26 \cdot 10^{-1}$	$7.73 \cdot 10^{-2}$	49	$3.11 \cdot 10^{-4}$	$5.34 \cdot 10^{-10}$
sphere	84	$1.00 \cdot 10^{-2}$	$1.01 \cdot 10^{-3}$	92	$1.09 \cdot 10^{-2}$	$1.92 \cdot 10^{-5}$

Table 2 shows that the proposed improvement consistently has a lower variance across all data sets and both models and therefore improves the efficiency of the explanation computation for both models. The absolute improvement is, on average, larger for the more complex PRBF model. It is also larger for data sets where the models learn the underlying concepts, because the model's predictions have a higher variance. For example, for the sphere and condInd data sets both models have a similar accuracy and subsequently similar mean variance and absolute improvement. On the other hand, when the model learns little, the variance is low and so is the absolute improvement in variance. For example, one can observe this with the Naive Bayes model on all data sets except the previously mentioned sphere and condInd data sets.

6 Conclusions

We presented the decomposition of PRBF classifier which exploits the marginalization property of the Gaussian distribution and applied this decomposition inside two general methods for explaining predictions for individual instances. We demonstrated how one can explain and visualize the classifications of otherwise opaque PRBF model. Explanations of the training instances can capture the impact 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 values have on the selected class.

We compared the distance between PRBF explanations using decomposition, which is an exact method, and averaging over all possible feature values (an approximation). The results show that the differences are small which is an indication that for other learning algorithms without exact methods, the approximation gives reliable explanations.

The presented explanation methodology has been successfully used in medical application [38]. In future work we plan to include also PRBF and use the feedback provided by the experts to further improve the visualization tool. Within this tool we want to control a reliability of classifications and warn a user if a whole model or a single classification is not reliable enough to allow explanations.

References

[1] R. Andrews, J. Diederich, and A. B. Tickle. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6):373–384, 1995.

[2] B. Becker, R. Kohavi, and D. Sommereld. Visualizing the simple bayesian classier. In *KDD Workshop on Issues in the Integration of Data Mining and Data Visualization*, 1997.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[4] D. Caragea and V. Cook, Dianne Honavar. Towards simple, easy-to-understand, yet accurate classifiers. In *Third IEEE International Conference on Data Mining, ICDM 2003.*, pages 497– 500, 2003.

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

[6] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[7] M. Craven and J. W. Shavlik. Using sampling and queries to extract rules from trained neural networks. In *International Conference on Machine Learning*, pages 37–45, 1994.

[8] M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 24–30. The MIT Press, 1996.

[9] A. S. d'Avila Garcez, K. Broda, and D. M. Gabbay. Symbolic knowledge extraction from trained neural networks: a sound approach. *Artificial Intelligence*, 125 (1-2):155–207, 2001.

[10] U. M. Fayad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Aartificial Intelligence (IJCAI'93)*, pages 1022–1027. Morgan Kaufmann, San Francisco, 1993.

[11] K. Främling. Explaining results of neural networks by contextual importance and utility. In *Proceedings of the AISB'96 conference*, 1996.

[12] I. J. Good. *Probability and the weighing of evidence*. C. Griffin London, 1950.

[13] L. Hamel. Visualization of support vector machines with unsupervised learning. In *Proceedings of 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2006.

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

[15] A. Jakulin, M. Možina, J. Demšar, I. Bratko, and B. Zupan. Nomograms for visualizing support vector machines. In R. Grossman, R. Bayardo, and K. P. Bennett, editors, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 108–117. ACM, 2005.

[16] R. Kerber. ChiMerge: Discretization od numeric attributes. In *Proceedings of AAAI'92*, 1992.

[17] I. Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7(4):317–337, 1993.

[18] V. Lemaire and F. Clérot. An input variable importance definition based on empirical data probability and its use in variable selection. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1375–1380 vol.2, 2004. doi: 10.1109/IJCNN.2004.1380149.

[19] V. Lemaire, R. Féraud, and N. Voisine. Contact personalization using a score understanding method. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2008.

[20] J. Lubsen, J. Pool, and E. van der Does. A practical device for the application of a diagnostic or prognostic function. *Methods of Information in Medicine*, 17: 127–129, 1978.

[21] D. Madigan, K. Mosurski, and R. G. Almond. Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics*, 6(2):160–181, 1997.

[22] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[23] M. Možina, J. Demšar, M. W. Kattan, and B. Zupan. Nomograms for visualization of naive bayesian classifier. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004*, pages 337–348. Springer, 2004.

[24] V. Palade, C.-D. Neagu, and R. J. Patton. Interpretation of trained neural networks by rule extraction. In *Proceedings of the International Conference, 7th Fuzzy Days on Computational Intelligence, Theory and Applications*, pages 152–161, London, UK, 2001. Springer-Verlag. ISBN 3-540-42732-5.

[25] F. Poulet. Svm and graphical algorithms: A cooperative approach. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 499–502, 2004.

[26] B. Poulin, R. Eisner, D. Szafron, P. Lu, R. Greiner, D. S. Wishart, A. Fyshe, B. Pearcy, C. Macdonell, and J. Anvik. Visual explanation of evidence with additive classifiers. In *Proceedings of AAAI'06*. AAAI Press, 2006.

[27] M. Robnik-Šikonja, A. Likas, C. Constantinopoulos, I. Kononenko, and E. Štrumbelj. Efficiently explaining decisions of probabilistic RBF classification networks. In A. Dobnikar, U. Lotrič, and B. Šter, editors, *Adaptive and Natural Computing Algorithms, Proceedings of 10th International Conference, ICANNGA 2011, Part I, LNCS 6593*, pages 169–179. Springer, 2011.

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

[29] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity analysis*. Wiley, Chichester; New York, 2000.

[30] R. Setiono and H. Liu. Understanding neural networks via rule extraction. In *Proceedings of IJCAI'95*, pages 480–487, 1995.

[31] C. E. Shannon. A mathematical theory of communications. *The Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

[32] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, volume II*. Princeton University Press, 1953.

[33] S. Thrun. Extracting rules from artifical neural networks with distributed representations. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 505–512. The MIT Press, 1995.

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

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

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

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

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