Knowl Inf Syst (2010) 24:305–324
DOI 10.1007/s10115-009-0244-9

REGULAR PAPER

Explanation and reliability of prediction models: the case of breast cancer recurrence

Erik Štrumbelj · Zoran Bosnić · Igor Kononenko · Branko Zakotnik · Cvetka Grašič Kuhar

Received: 25 October 2008 / Revised: 12 June 2009 / Accepted: 27 June 2009 /
Published online: 20 August 2009

Abstract In this paper, we describe the first practical application of two methods, which bridge the gap between the non-expert user and machine learning models. The first is a method for explaining classifiers' predictions, which provides the user with additional information about the decision-making process of a classifier. The second is a reliability estimation methodology for regression predictions, which helps the users to decide to what extent to trust a particular prediction. Both methods are successfully applied to a novel breast cancer recurrence prediction data set and the results are evaluated by expert oncologists.

Keywords Data mining · Machine learning · Breast cancer · Classification explanation · Prediction reliability

1 Introduction

Machine learning has over the years become more present in medicine as Kononenko summarizes in his overview of machine learning in medical diagnosis [18]. However, Kononenko states that machine learning has not been fully accepted in the medical community and suggests that medical practitioners are reluctant to accept such tools because the tools further complicate their work. Similarly, users from other fields may also be reluctant or may require additional insight into the model's decision. Our long-term goal is to fully utilize the available machine learning techniques while still producing easy to use tools and results that are easy to understand. In this paper we apply standard and innovative machine learning approaches to a data set provided by the Institute of Oncology, Ljubljana, which focuses

E. Štrumbelj (✉) · Z. Bosnić · I. Kononenko
Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
e-mail: erik.strumbelj@fri.uni-lj.si

B. Zakotnik · C. Grašič Kuhar
Institute of Oncology, Ljubljana, Slovenia

on breast cancer recurrence. The contribution of the paper is twofold. First, we successfully apply a method for the explanation of classifiers' predictions. Second, we implement a methodology for estimating the reliability of individual regression predictions. Both methods bridge the gap between machine learning approaches and users, such as oncologists, who are not machine learning experts. To further emphasize and evaluate the benefits of the new approaches we apply them without incorporating prior medical knowledge and results are independently evaluated by oncologists.

1.1 Explaining individual classifications

A lot of related work deals with model-specific explanation methods. For example, naive Bayes and decision trees already have relatively simple but model-specific way of explaining their decisions. A prediction from a naive Bayes model can be explained by information gains of individual attributes [17] and a prediction from a decision tree can be explained with the logical rule that was used to get from the root to the leaf for that particular instance. Bayesian networks are also an example of a highly interpretable model and their interpretability can be further improved [11]. Tools for explaining the importance of individual features for Random Forests were provided by Breiman [7], but these tools are also model-specific. Extracting rules from neural networks has received a lot of attention [1,8,29] and several methods for the explanation of Support Vector machines [14,15,24] and naive Bayes [2,17,23] have been developed. Evolutionary algorithms can also be adapted to produce comprehensible classification rules [10]. The *ExplainD* method [27], which explains the influence of attribute values by assigning a score to each attribute value, uses a visualization similar to our own, with bars that correspond to the size of the contribution. However, their approach is limited to linear additive models. The problem with model-specific methods is that we cannot provide a uniform explanation across different models. Whenever we change the classifier or add a new classifier, the user has to adapt to a new explanation method. This requires extra time and effort from the user, which is exactly what medical practitioners dislike. The advantage of our method over these model-specific methods is that it can be applied to any classifier, regardless of its type. As far as we know there exist two other such methods. The first is a method for explanation proposed by Lemaire et al. [21]. The second is proposed by Robnik Š and Kononenko [25]. However, both of these methods work by observing the sensitivity of the model by changing the value of a single attribute at a time. Therefore, both methods are unable to handle disjunctive concepts as Štrumbelj and Kononenko have shown [28]. We show that this results in less informative explanations. Our explanation can handle such concepts, which results in more informative explanations.

1.2 Reliability of individual predictions

An important aspect of analyzing the quality of induced knowledge is evaluating the accuracy of computed predictions. In contrast to evaluating model accuracy using the averaged accuracy measures, reliability estimates for individual predictions [6] can provide indispensable information, especially in decision-critical prediction fields, such as medicine. Based on the estimates of the individual prediction reliability, the experts can decide to what extent to trust that particular prediction and whether to perform the necessary actions. In previous work, an extensive comparison of reliability estimates for individual predictions was performed [3,5]. The work presented nine reliability estimates, which were based on five different approaches: sensitivity analysis, variance of bagged models, density estimation, local modeling of error, and local cross-validation. In subsequent work [4], a methodology

for the automatic selection of the best performing reliability estimate for a given domain and regression model was developed. The evaluation of former approaches on standard benchmark domains showed a success of both methods and a potential for their usage on practical application domains of machine learning. In this paper we present the first implementation of the former methodology in practice and evaluate its performance on a real medical prognostic problem.

1.3 Organization of the paper

The paper is divided into six sections. In Sect. 2 we present the breast cancer recurrence data set. The prediction of breast cancer recurrence is addressed in Sect. 3. In Sect. 4 we use a method for the explanation of individual decisions to explain the predictions of several different models. The generated explanations are evaluated by oncologists and the results are discussed. In Sect. 5 we apply a reliability estimation method, evaluate it, and analyze its relation with the oncologists' subjective estimation of reliability. We conclude the paper with Sect. 6.

2 Description of the oncologyBC data set

The initial data set was provided by the Institute of Oncology Ljubljana and contains data for 1,035 breast cancer patients. Each patient is described with 22 medical features recorded after breast cancer surgery and 10 features recorded through patient follow-up. The latter reveal whether the patient had a recurrence of breast cancer and when or, in the case of no recurrence, the duration of the follow-up and last recorded state of the patient.

Some of the 22 features recorded at the time of surgery are redundant. For several features both the numerical and the discretized versions were recorded. Note that the features were discretized by oncologists, based on how they use the features in everyday medical practice. Furthermore, preliminary analysis has shown that no significant difference in prediction quality can be achieved using any combination of the numerical versions instead of the discretized versions of the features. Therefore, the numerical versions of redundant features were removed. The remaining features form the *oncologyBC* data set, which is described in Table 1. Note that feature values for *grade*, *PgR*, *famHist*, and *ER* were not determined for every instance and in those cases the missing values are treated as a separate *not determined* feature value. The *not applicable* value of feature *grade* indicates a special type of tumor where tumor grade does not apply. We are primarily interested in the prognosis of breast cancer recurrence. Therefore, the follow-up features were reduced to two features: a binary feature that indicates whether a recurrence has occurred and a numerical feature that indicates either the time of recurrence (if the cancer recurred) or the time of the last follow-up (if the cancer did not recur).

In this paper, we use two variations of the *oncologyBC* data set, each representing a different formulation of the recurrence prediction problem. The *oncologyBC10* is a binary classification problem where the class value indicates whether or not there was a recurrence within 10 years after surgery. The *oncologyBCR* data set is a regression problem where the class value is continuous and indicates the time to recurrence. Patients with class value of more than 10 years are considered to have no recurrence. To correctly assume that there was no recurrence within 10 years after surgery we have to observe a patient at least 10 years, therefore 154 patients that did not have a recurrence but were observed for less than 10 years were removed from *oncologyBCR* and *oncologyBC10*.

Table 1 A detailed description of the features of the *oncologyBC* cancer data set and their values

Feature name	Feature description
menop	Binary feature indicating menopausal status
stage	Tumor stage
	1: less than 20 mm, 2: between 20 and 50 mm, 3: over 50 mm
grade	Tumor grade
	1: good, 2: medium, 3: poor, 4: not applicable, 9: not determined
histType	Histological type of the tumor
	1: ductal, 2: lobular, 3: other
PgR	Level of progesterone receptors in tumor (in fmol per mg of protein)
	0: less than 10, 1: more than 10, 9: unknown
invasive	Invasiveness of the tumor
	0: no, 1: invades the skin, 2: the mamilla, 3: skin and mamilla, 4: wall or muscle
nLymph	Number of involved lymph nodes
	0: 0, 1: between 1 and 3, 2: between 4 and 9, 3: 10 or more
famHist	Medical history
	0: no cancer, 1: 1st generation breast, ovarian or prostate cancer
	2: 2nd generation breast, ovarian or prostate cancer, 3: unknown gynecological cancer
	4: colon or pancreas cancer, 5: other or unknown cancers, 9: not determined
LVI	Binary feature indicating lymphatic or vascular invasion
ER	Level of estrogen receptors in tumor (in fmol per mg of protein)
	1: less than 5, 2: 5 to 10, 3: 10 to 30, 4: more than 30, 9: not determined
maxNode	Diameter of the largest removed lymph node
	1: less than 15 mm, 2: between 15 and 20 mm, 3: more than 20 mm
posRatio	Ratio between involved and total lymph nodes removed
	1: 0, 2: less that 10%, 3: between 10 and 30%, 4: over 30%
age	Patient age group
	1: under 40, 2: 40-50, 3: 50-60, 4: 60-70, 5: over 70 years

Note that not all features are considered relevant for breast cancer recurrence

3 Are classifiers good at predicting breast cancer recurrence?

Our goal is to analyze whether our explanation method for classifiers can provide uniform explanations to non-machine learning expert users. In our case, these users are oncologists. However, the oncologists cannot benefit from the explanation of a classifier if the classifier does not learn any concepts behind breast cancer recurrence. Therefore, we first have to answer the question: *Are classifiers better than default predictions?* Several well-known classifiers were used and the first three are considered to be among the most influential data mining algorithms [30]: a naive Bayes classifier (M_{NB}), a decision tree (M_{DT}) , a tuned SVM with a polynomial kernel (M_{SVM}), a Random Forests classifier (M_{RF}), a multi-layer perceptron artificial neural network (M_{ANN}), and bagging with naive Bayes as the base classifier (M_{bag}) (see for example [19] for a more detailed description of these classifiers). Note that we also used SVM with linear and RBF kernel. However, the results were not significantly different from using a polynomial kernel and were therefore omitted. As a baseline for comparison, we use M_{def}, a default classifier, which always predicts the majority class. Note that

Table 2 Models' mean classification accuracies on the *oncologyBC10* data set

	M_{def}	M_{NB}	M_{DT}	M_{SVM}	M_{RF}	M_{ANN}	M_{bag}
Accuracy	0.490	0.678	0.674	0.599	0.676	0.608	0.680
p-value	–	5.62×10^{-18}	4.78×10^{-17}	1.55×10^{-5}	2.35×10^{-16}	5.01×10^{-7}	4.83×10^{-18}

Results were obtained using tenfold cross-validation. The p-values are for the Wilcoxon signed rank test with the null-hypothesis that the classifier's accuracy equals the default accuracy. The alternative hypothesis is that the classifier's accuracy is higher

Table 3 Models' and oncologists' mean classification accuracies across 100 test instances

	M_{def}	M_{NB}	M_{DT}	M_{RF}	M_{bag}	$\mathcal{O}_1$	$\mathcal{O}_2$
Accuracy	0.48	0.70	0.67	0.68	0.70	0.65	0.63
p-value	0.97	0.20	0.37	0.31	0.21	–	–

The p-values are for the Wilcoxon signed rank test with the null-hypothesis that the classifier's accuracy equals the accuracy of oncologist $\mathcal{O}_1$. The alternative hypothesis is that the classifier's accuracy is higher

the distribution of class values is approximately 51% (recurrence) and 49% (no recurrence), so unbalanced class values are not an issue.

To evaluate the classifiers, we use classification accuracy. Table 2 shows the mean accuracies obtained with tenfold cross-validation. For each classifier a Wilcoxon signed rank test was used to test the significance of the improvement in accuracy. The null-hypothesis is that the accuracy of the classifier equals the default classifier's accuracy. The alternative hypothesis is that the classifier has a higher accuracy than the default classifier. p-values for these tests are provided in Table 2. It is clear that the accuracy of classifiers is significantly higher than the default accuracy.

The classifiers obviously learn some of the concepts of breast cancer recurrence, but how good are their predictions compared to expert oncologists? For validation, we randomly chose 100 instances from the *oncologyBC10* data set. The remaining 781 instances were used to train and tune the classifiers. Table 3 shows the results on the validation set and p-values for the Wilcoxon signed rank tests. The last two columns show the accuracies of two expert oncologists on those 100 examples. We cannot conclude that the classifiers have a significantly higher accuracy than oncologists, but the results suggest that the classifiers' predictions are at least comparable with those of expert oncologists.

4 Enhancing classifier decisions with an explanation

Results from the previous section show that classifiers are able to learn at least some of the concepts behind breast cancer prediction. Now we can apply our explanation method to provide the user with additional information about the classifiers' decision making process. A classifier's decision is explained in the form of an *instance explanation*. The explanation method assigns to each feature value a contribution, which can be positive, negative, or zero. A positive contribution means that the feature value speaks in favor of the class value we are explaining, a negative contribution means that the feature speaks against the class value, and a zero contribution means that the feature value does not influence the decision. The

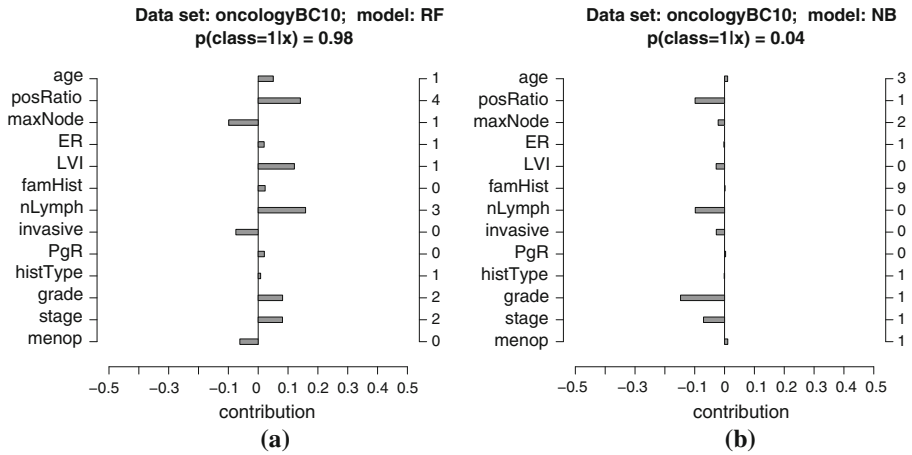

Fig. 1 Visualization of instance explanations for two M_{RF} model decisions. In the *left-hand side* are the names of the features and on the *right-hand side* their values for the instance. The *bars* represent the contributions of features to the model's decision. Information about the used data set and model, and the model's prediction are also provided. The true outcome for the first instance is 1 (recurrence) and for the second instance 2 (no recurrence). **a** *Instance explanation* 1. **b** *Instance explanation* 2

definition of a feature value's contribution and an illustrative example can be found in the Appendix and further information about the method can be found in [28].

Figure 1 shows two instance explanations. The first was generated to explain the decision of the M_{RF} and the second to explain the decision of the M_{NB} model. The instance in Fig. 1a describes a patient that had a recurrence of breast cancer and the model correctly predicted that the recurrence will recur (assigning a 0.98 probability to that outcome). The generated explanation suggests that the three feature value that had the largest contribution towards the prediction of a recurrence are: (*posRatio* = 4), (*LVI* = 1), and (*nLymph* = 3). Oncologists confirmed that this combination of features indeed results in one of the worst possible prognoses for a patient. Figure 1b describes a patient that did not have a recurrence of breast cancer and the decision of a different model is being explained. Again, the model (M_{NB}) correctly predicted that recurrence is not probable (assigning a 0.04 probability to that outcome). Now the generated explanation suggests that (*grade* = 1), (*stage* = 1) and (*nLymph*=0) were the features that most influenced the classifier to make such a prediction. Note that (posRatio = 1) and (*nLymph*=0) give the same information and the explanation correctly assigns a similar contribution to both. Indeed, oncologists also predict a non-recurrence for this instance and are most convinced by the absence of positive lymph nodes (*nLymph*=0), tumor grade, and tumor stage. These two examples show how the instance explanations provide insight into the classifier's decision in a form that the user can relate to.

4.1 Evaluation of individual instance explanations

To evaluate the instance explanations generated by our explanation method we chose 20 from the 100 test instances of the *oncologyBC10* data set. To ensure more diversity, the instances were chosen in a semi-random way, so that half of the instances had no recurrence and half had a recurrence. In each half there were five correctly classified instances and five misclassified instances. Using our method, we generated instance explanations that explain the decisions of the M_{RF} model for these 20 instances.

Table 4 Class value, predicted value, and the number and position of disagreements (×) for the 20 instances used for evaluation

#	class	pred.	age	ER	LVI	nLy.	inv.	PgR	grade	stage	menop
1	2	2	✓	✓	✓	✓	✓	✓	✓	✓	×
2	2	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	2	2	✓	✓	✓	✓	✓	✓	×	✓	✓
4	2	1	×	✓	✓	✓	✓	✓	✓	✓	×
5	2	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	2	2	×	✓	✓	✓	✓	✓	✓	✓	×
7	2	1	×	×	✓	×	×	✓	✓	✓	✓
8	2	2	×	✓	✓	✓	✓	✓	✓	✓	✓
9	2	2	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	2	1	✓	✓	✓	✓	✓	✓	×	✓	✓
11	1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	1	1	✓	×	✓	×	✓	✓	×	✓	✓
13	1	2	✓	✓	✓	✓	✓	✓	✓	✓	✓
14	1	2	×	✓	✓	✓	✓	✓	✓	✓	✓
15	1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
16	1	2	✓	✓	✓	✓	✓	✓	✓	✓	✓
17	1	2	×	✓	✓	✓	✓	×	×	✓	×
18	1	1	×	✓	✓	✓	✓	✓	✓	✓	✓
19	1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓
20	1	2	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sum			7	2	0	2	1	1	4	0	4

For each instance we used the following evaluation procedure: For each feature of the instance the oncologist had to either agree with the generated contribution or disagree (due to the size of the contribution, the direction of the contribution, or both). Note that out of the 13 features, the oncologist use only 9 in every-day medical practice. Therefore the contributions of 4 features (histType, famHis, maxNode and posRatio) could not be evaluated and our evaluation produced a total of 180 agreements/disagreements. The results are shown in Table 4 and we can see that in total there were 21 disagreements, i.e. 12% of all contributions.

Once the contributions were evaluated, we, together with oncologists, made in-depth analysis of the disagreements. Eleven of the disagreements are on the features *age* and *menop*, which is also an age related feature. These contributions describe young age as a factor that speaks in favor of a non-recurrence but the consensus amongst medical specialists is that young age speaks in favor of a recurrence. We have established that these contributions in fact reflect the data and their incorrectness is due to a bias in the data. As oncologists explained, a higher than usual amount of young patients in our data was treated with therapies, because youth was, and still is, considered a negative factor. Consequently, young age sometimes moves the model's decision towards non-recurrence, because it implies a higher probability of therapy, which in turn reduces the chance of recurrence. The cause of the remaining disagreements can be either the classifiers inability to correctly learn the concepts or an incorrect explanation. However, even if all the remaining disagreements are counted as a mistake of the explanation method, we still achieve a 10/180 disagreement ratio (approximately 95%).

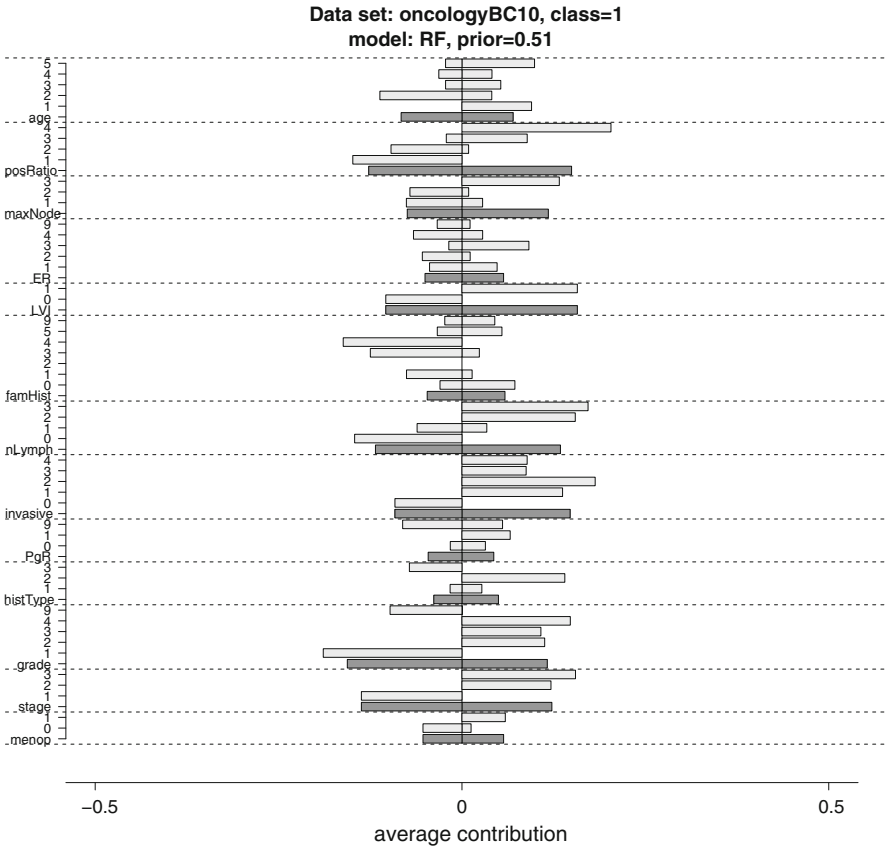

Fig. 2 Visualization of the model explanation for the M_{RF} model. On the vertical axis we have the attributes and their respective values. The *light bars* represent the mean positive and mean negative contribution of each particular value across 100 test cases. The *dark bars* represent the mean positive and mean negative contribution of the whole features

This is very good, considering that each agreement means that the explanation method extracted from the classifier knowledge that the expert oncologist can relate to.

4.2 Model explanations

The contributions generated by our method can be combined over several instance explanations to produce a more general *model explanation* which describes how much features and their individual values, on average, influence the model's predictions. Figure 2 shows a model explanation for the M_{RF} model on the *oncologyBC10* data set. From this model explanation we can extract a lay explanation of breast cancer recurrence. The classifier we are explaining is relatively successful on this data set, so we can view the model as an oncologist with limited knowledge. Therefore, if the explanation method is also successful at explaining the decision-making process of this "oncologist", we should get an explanation that is close to the actual concepts behind breast cancer recurrence. The overall contributions of individual features (dark bars) suggest that *nLymph*, *LVI*, *invasive*, *grade*, *stage*, and *posRatio* are the features which on average have the largest influence on the classifier's decision.

On the other hand, the features *menop*, *famHist*, *histType*, *PgR*, and *ER* have the smallest influence on the classifier. We can also explain the influence of individual feature values. For example, if *LVI* equals 1 then the prognosis for that patient is much worse. It is similar with *age*, where young age results in a worse prognosis, while older age does not have a clearly positive or negative influence. The higher the number of positive lymph nodes (*nLymph* and *posRatio*), the worse the prognosis, etc... These conclusions were made just by observing this model explanation, without prior medical knowledge, yet oncologists confirm that these conclusions indeed reflect current medical knowledge about breast cancer recurrence. This suggests that model explanations can provide the user with useful information and make the model's predictions easier to trust.

In Fig. 3, we have four model explanations, including the one we have already discussed, and all are across the same 100 instances. The visualizations have been cropped and minimized, but the contributions (i.e., the bars) have not been scaled in any way, to facilitate comparison. Figure 3b is the model explanation for M_{NB} and, we can see that the same features and feature values are important both for M_{NB} and for M_{RF}. However, unlike for M_{RF}, for M_{NB} each feature value has either a completely positive or a completely negative contribution. This is due to the naive Bayes' assumption of conditional independence. On the other hand, M_{RF} considers interactions between features and the same feature value can have a positive contribution in one context and a negative contribution in another context. Figure 3c is a model explanation for M_{SVM}, which is not as successful on this data set as the previous two models. From the explanation we can see that the contributions of complete features (dark bars) are smaller for M_{SVM}, compared to M_{RF} and M_{NB}. This implies that this model is less successful because it takes into account all the features and does not capture in full the most important features. Finally, Fig. 3b is a model explanation for M_{RF} generated using the method proposed by Robnik Šikonja and Kononenko [25]. The feature contributions are not as clearly expressed as in Fig. 3, because this explanation method has difficulties with feature interactions, especially redundancy (see [28]). Subsequently, this explanation is not as informative. On the other hand, our explanation method generates more informative explanations. And, as the examples show, the explanations reflect what the classifier learned from the data set.

5 Estimating the reliability of individual predictions

Besides explaining the classifier's decisions, the oncological problem also requires very sensitive handling of the prediction outcomes for individual patients. Based on the predictive model's outputs, the medical experts can take preventive or therapeutical actions, which, if based on inaccurate model predictions, can cause harm. This makes the cancer recurrence problem an opportunity to implement and evaluate the methodology for estimating reliability of individual predictions.

Reliability is generally defined as the ability of a system to perform and maintain its functions in routine or unexpected circumstances [6,9]. It can be evaluated either with positive performance indicators (in terms of *the greater value the better*, e.g. accuracy, availability, responsiveness, etc.) or with negative performance indicators (in terms of *the less is better*, e.g. inaccuracy, downtime rate, latency, etc.). Since reliability is in the most cases defined qualitatively, the *reliability estimate* is therefore an estimate for quantitative measuring of reliability (e.g. an accuracy estimate, error estimate, availability estimate, etc.).

Generally, reliability estimates can be implemented either as *model-dependent reliability estimates* (by exploiting the properties of a particular regression model, e.g. using number

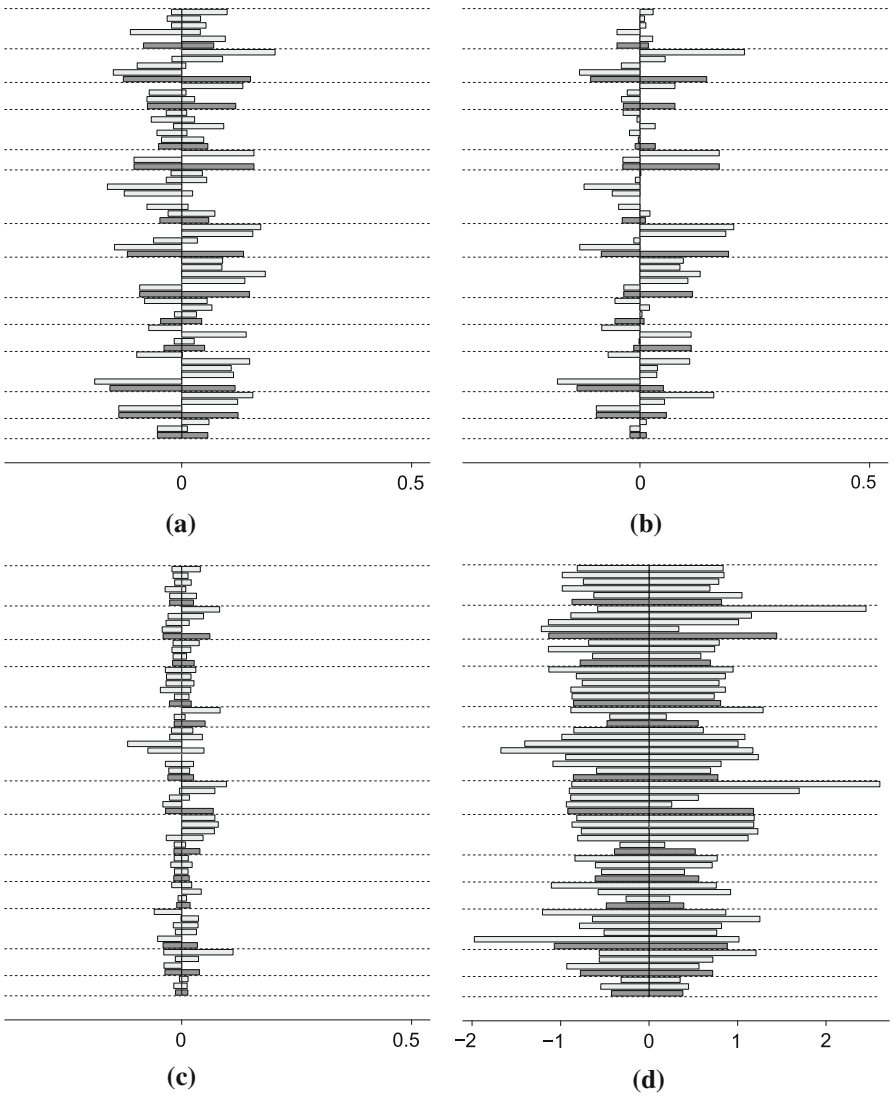

Fig. 3 Four model explanations for the same data set and different models. The explanation in the *bottom-right* was generated using the explanation method proposed by Robnik Šikonja and Kononenko [25]. **a** M_{RF}. **b** M_{NB}. **c** M_{SVM}. **d** M_{RF} with another method

of support vectors [12], Lagrange multipliers in the SVM optimization procedure [20,26], splits in a regression tree, etc.) or as *model independent reliability estimates* (by exploiting general properties of the supervised learning framework, e.g. changing the learning set, etc.). In our previous work, we focused on developing model-independent reliability estimates for individual predictions, which are implemented as *estimates of the prediction error*. As such, these estimates are defined as metrics, of which higher values represent higher estimated prediction error and vice versa. Value zero accordingly represents the estimated reliability of the most accurate prediction. In the following, we present the used reliability estimates and evaluate them in the context of our oncological problem.

5.1 Methodology for the reliability estimation

In our previous work [3,5], we presented nine reliability estimates, named *SAvar*, *SAbias-s*, *SAbias-a*, *CNK-s*, *CNK-a*, *LCV*, *BAGV*, *DENS* and *BVCK*, which are intended to estimate prediction errors of individual predictions, computed by various regression models. In our past experimental work, these reliability estimates were evaluated using the eight regression models, which are listed in Table 5, and showed the promising results for estimation of individual prediction reliability in regression. We summarize the core ideas of the reliability estimates' design as follows:

SAvar: Reliability estimate, based on sensitivity-analysis (an approach that observes how model outputs, i.e. predictions, change with respect to changing its inputs, i.e. the data set) measures the prediction variance. The variance is measured using various predictions, output from the regression model after changing a training data set in a controlled way [3]. The testing results showed good performance of this reliability estimate with linear regression and generalized additive model.

SAbias-s: Estimate, based on sensitivity-analysis, which measures the local prediction bias. The bias is computed by observing whether the model, being influenced by changes in the data set, tends more to increase or decrease its initial predictions. Since values of this reliability estimate can also be negative (suffix -s denotes *signed*), the estimate provides the additional information about the error direction (whether the value of prediction was too high or too low). The estimate achieved outstanding results with the regression trees.

SAbias-a: The absolute version of SAbias-s (suffix -a denotes *absolute*), which is tested for correlation with the absolute prediction error only.

CNK-s: Reliability estimate which models the prediction error locally as the difference between averaged nearest neighbors' label and the prediction of the example in question (similarly to *SAbias-s*, the estimate is signed, hence the suffix *-s*). It achieved the best performance using the regression trees.

CNK-a: The absolute version of CNK-s, which was tested for correlation with the absolute prediction error only. The estimate achieved the best results using the linear regression and the generalized additive model.

LCV: The estimate which locally models the prediction error by applying the cross-validation procedure in the local area of problem space. The estimate is computed as the weighted average of leave-one-out prediction errors, obtained by applying the leave-one-out cross-validation procedure only to the subspace defined by the nearest neighbors of the particular example (for which we are estimating the prediction reliability). The evaluation revealed that *LCV* is the most appropriate estimate for the usage with the support vector regression, locally weighted regression and random forests.

BAGV: Estimate which measures the prediction reliability by creating the bootstrapped replicas of the original model and computes the variance of the bootstrapped models' predictions. Can be used with an arbitrary regression model, in our experiments it is used with the regression trees. Besides achieving the best average performance, the evaluation showed that this estimate is the most appropriate for the usage with locally weighted regression.

DENS: Reliability estimate, based on the distribution of learning examples in the input space. Its design is based on the assumption that the reliability of predictions, made in denser problem space is more reliable than the reliability of predictions

for examples where the information for the neighboring examples is sparse. The estimate did not achieve noticeable results with any of the testing regression models.

BVCK: Linear combination of estimates *BAGV* and *CNK-a*, which outperformed all of the above individual estimates. The estimate achieved the best results with neural networks and with bagging.

Testing of the reliability estimates was performed by observing and statistically evaluating their correlation to the prediction error. The evaluation revealed that the estimates exhibit different magnitudes of the correlation coefficients in different domains and with different regression models. To tackle this challenge, in the following work [4], a methodology for the automatic selection of the best performing reliability estimate for a given domain and regression model was developed. Two approaches for the automatic selection were developed, based on meta-learning and internal cross-validation approach. The evaluation of former approaches showed a success of both methods and a potential for their usage on practical application domains of the machine learning. In the following sections, we evaluate the performance of the reliability estimation methodology on the *oncologyBCR* problem and analyze the benefits of its implementation.

5.2 Selection of the regression model

From the initial *oncologyBCR* data set of 881 examples, 100 examples were randomly selected as test examples and 781 examples remained as the learning examples. The purpose of this separation was to provide an independent test set, which can be used to evaluate the reliability estimate, automatically selected on the learning data. In our experimental work we computed eight different regression models for predicting the time of cancer recurrence. The evaluation of the regression models was performed using the tenfold cross-validation procedure on the learning examples and computation of the models relative mean squared errors (RMSE):

$$\mathrm{RMSE} = \frac{\sum_{(\mathbf{x_i}, T_i) \in E} (T_i - P_i)^2}{\sum_{(\mathbf{x_i}, T_i) \in E} (T_i - \overline{T'})^2}, \text{ where } \overline{T'} = \frac{1}{|E'|} \sum_{(\mathbf{x'_i}, T'_i) \in E'} T'_i$$

and T_i and P_i denote the target and the predicted regression value, respectively, for an example $(\mathbf{x_i}, T_i)$ from the test fold E. T'_i denotes the target value of an example from the training folds, $(\mathbf{x'_i}, T'_i) \in E'$. In other words, the RMSE is the models mean squared error, relative to the mean squared error of predicting according to the mean target value on the training set. The errors for each of the models are shown in Table 5. Since bagging with regression trees achieved the lowest error, we decided to use this model for the implementation of the cancer recurrence predictive system (see Sect. 5.5).

5.3 Evaluation of the reliability estimates

The empirical analysis of the reliability estimates in previous work showed that the strength of reliability estimates' values to the prediction error depends on the used regression model and on a particular problem domain. The analysis also showed that a different number of estimates may significantly correlate with the prediction error in each of the testing domains, which enables to rank the domains according to their difficulty (as *difficult* we denote those domains, in which none or a only few of nine reliability estimates significantly positively correlated to prediction error, and vice versa).

Table 5 Ranking of the regression models in the decreasing order of their performance, evaluated on the subset of oncologyBCR data set using the tenfold cross-validation

Table shows the relative mean squared error (RMSE) for each of the regression models. Note that the default predictor which always predicts mean recurrence time has RMSE = 1.0

Model	RMSE
Bagging with regression trees	0.790
Generalized additive model	0.791
Random forests	0.801
Linear regression	0.806
Support vector machine	0.838
Regression tree	0.852
Locally weighted regression	0.862
Neural network	0.955

Table 6 Correlation coefficients between the reliability estimates and the prediction errors of regression models

	SAvar	SAbias-s	SAbias-a	CNK-s	CNK-a	LCV	BAGV	DENS	BVCK
Learning examples	**0.274**	**0.076**	0.035	0.036	0.019	0.030	**0.274**	**−0.155**	**0.140**
	+1	+5	7	6	9	8	+2	−3	+4
Test examples	**0.234**	0.174	−0.055	0.032	−0.085	0.013	**0.226**	−0.124	0.039
	+1	3	6	8	5	9	+2	4	7

The significant coefficients ($\alpha < 0.05$) are denoted with the boldface. The numbers beneath the coefficients denote the ranking of the coefficient's absolute value (+ denotes significant positive correlation and − denotes significant negative correlation)

To evaluate the reliability estimation difficulty for the oncological problem, we first tested the performance of nine reliability estimates using the selected regression model (bagging with regression trees) on the learning data set. The reliability estimates and the prediction errors were computed using the leave-one-out procedure for all learning examples. Afterwards, the correlation coefficients between the estimates and the prediction errors were computed. The computed coefficients are shown in Table 6 (the upper half of the table). The table also shows the rankings of correlation coefficients and their statistical significance.

We also applied the procedures for automatic selection of the best performing reliability estimate in this regression domain. The meta-classifier, induced using the meta-learning approach [4] predicted that the most suitable estimate is *BAGV*. The internal cross-validation procedure selected the estimate *SAvar*, which also achieved the greatest positive correlation to the prediction error on the learning examples. The results of these two estimates in the upper half of Table 6 show that the correlation coefficients are significant as well.

5.4 Evaluation of the reliability estimates on the test examples

The performance of all reliability estimates was tested on the test examples, similarly as on the learning examples. The prediction errors and the reliability estimates were for each testing example computed using models, induced on the learning examples. The testing results are shown in Table 6 (the bottom half of the table).

By comparing results on the learning and test examples we can see that a lower number of reliability estimates were successful (i.e. achieved significant positive correlation to the prediction error) on the test examples than on the learning examples. Two estimates, which

significantly correlated with the error (*SAvar* and *BAGV*) are a subset of the estimates that were successful on the learning examples, showing that the reliability estimation difficulty is greater for the test examples than for the learning examples. Although the low number of successful reliability estimates indicates that the task of reliability estimation is feasible with this model, it also indicates a challenge for the automatic procedures for selection of the reliability estimates, which therefore have to select one of the non-numerous successful estimates.

Evaluation of the results, achieved on the test examples also shows that the both procedures for automatic selection of the most suitable estimate correctly selected the estimates which perform well on the test examples (*SAvar* and *BAGV*). We can see, that the selected estimates for bagging were also the best two ranked successful reliability estimates for that model. In addition, their correlation coefficients are statistically significant as well, indicating the potential to estimate the prediction error with these two estimates. Based on these results we can evaluate both procedures for automatic selection of estimates as successful. In the next section, we describe the implementation of the oncological regression prediction system.

5.5 Implementation of the reliability estimates with the oncological prediction system

The evaluation showed that bagging and the estimate *SAvar* are, respectively, the most suitable regression model and reliability estimate for the oncological cancer recurrence problem. We supplemented the bare predictions of bagged regression trees with values of the reliability estimate *SAvar*. Since values of *SAvar* belong to an arbitrary, but domain-specific interval of numbers, we transformed these values to follow the distribution of the learning examples' prediction errors, enabling *SAvar* to approximately indicate the magnitude of the prediction error on test examples. This also enabled us to graphically present the predictions with their adjoined reliabilities, as shown in Fig. 4. In the figures, the Gaussian symbolically presents the reliability, with its width defined using the transformed value of *SAvar*.

5.6 Evaluation of the oncological prediction system

To evaluate how the oncological experts benefit from the prediction system, we compared its performance to the manual predictions of two oncologists, made for the same 100 test examples. The results in Table 7 show that the predictions of the regression model highly and significantly correlate with the predictions of both experts (the first table row), which indicates that the model is consistent with the experts' knowledge. The same, although in the lesser extent, is true also for the correlation of our reliability estimate (*SAvar*) to the reliability estimates of the oncologists (the second table row). Namely, we can see that *SAvar* significantly negatively correlates to the reliability estimate of the second oncologist, while its correlation to the estimate of the first oncologist remains statistically insignificant. Note, that the values of estimate *SAvar* are expected to negatively correlate with the values on oncologists' estimate, since the higher values of the former denote higher prediction *errors* and higher values of the latter higher prediction *confidence*.

Analyzing the correlation of the oncologists' prediction error to their reliability estimate (the third table row) we can see that, although correlating negatively, as expected, the coefficients' magnitudes are not significant. By comparing this result to the correlation coefficient of the estimate *SAvar* to the prediction error of bagging (Table 6, value of the coefficient is 0.234 and is significant) we can see that the experts can benefit from the implemented system. Namely, the evaluation of the system showed that it offers a comparable prediction

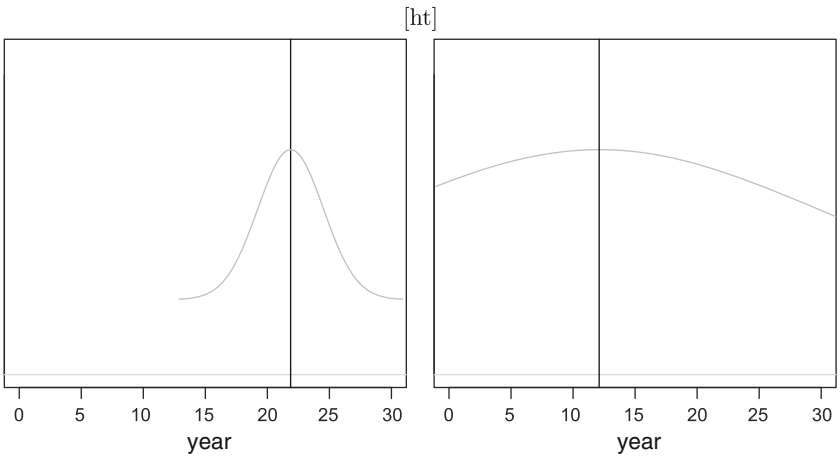

Fig. 4 Graphical representation of two recurrence predictions (*vertical lines*) and their reliability (denoted by the width of the Gaussian, surrounding the vertical line). Gaussians are only symbolic representations of the reliability (their width is a parameter of the value of the reliability estimate) and denote that the real prediction can either be close to the predicted value (high reliability is represented by narrow Gaussian) or on some less determined interval (lower reliability is therefore represented by wide Gaussian). The figure illustrates an example of the high prediction reliability (*left*), and the low prediction reliability (*right*)

Table 7 Evaluation of the oncological prediction system using the manual predictions of two oncologists

Correlation between		Oncologist 1	Oncologist 2
Bagging predictions	Oncologists' predictions	**0.509**	**0.752**
SAvar estimate	Oncologists' reliability estimate	0.171	**−0.217**
Oncologists' prediction error	Oncologists' reliability estimate	−0.115	−0.014

accuracy (for the discretized two-class yes/no recurrence problem) while also providing the informative estimate of prediction reliability.

6 Conclusion

As is the case with most medical prediction problems, standard machine learning approaches achieve results comparable to oncologists in breast cancer recurrence prediction as well. However, achieving optimal performance was not the goal of this paper. Instead, we applied for the first time two approaches that bring models and non-expert users closer together: a classifier explanation method, which can be used on any classifier type, and a reliability estimation methodology.

The model explanations generated by our method reflect the decision-making process of the model and provide the user with additional insight. Oncologists confirm that the generated contributions of individual features and their values reflect current medical knowledge. Oncologists also evaluated the instance explanations of one of the models and agreed with approximately 90% of the contributions in the generated explanations. Only 5% of the disagreements could not be explained by a known bias in the data. We conclude that the

explanation method was successfully applied in practice and the results are encouraging. The explanation method is especially useful when we need a uniform explanation across different classifier types and it does not have the flaws of other such methods.

For the breast cancer recurrence problem, we also evaluated the possibility of implementing the reliability estimates for individual regression predictions. We evaluated the performance of eight regression models and tested the reliability estimates with the best performing model—bagging with regression trees. The evaluation of the procedures for automatic selection of the best performing estimate showed high utility of this approach for practice, since they selected the first and the second most successful reliability estimate for this problem. By implementing the prediction system with the reliability estimate *SAvar* and comparing its predictions to the predictions of the experts, we concluded that our predictions significantly correlate with the predictions of the experts. However, our reliability estimate additionally provided the experts with the validation information of the predictions' accuracies, which they were not able to do before, since the experts' subjective reliability measures do not significantly correlate with their prediction errors. To conclude, the implementation of the oncological prediction system is the first implementation of our reliability estimation methodology in practice. Based on the performed analysis we can evaluate it as successful and promising for implementation in further problems.

Appendix A. Definition of the contribution of a feature value

Let set $S = \{1, 2, ..., n\}$ represent the n features that describe an instance from our data set. Let $\Delta_W = p_W - p_{\text{prior}}$, where p_W is the model's probabilistic prediction for the class value when features not in W are omitted from the instance, and p_{prior} is the prior probability of the class value. Note that W is an arbitrary subset of S and p_W is obtained by retraining the model on a data set with omitted features (see [28] for details) and classifying the instance with omitted feature values. This is similar to the wrapper approach in feature selection [16]. As long as this approach is used to approximate p_W the method works uniformly both for discrete and continuous features. However, continuous features have to be discretized to combine individual instance explanation into a model explanation. Therefore, for visualization purposes only. The major limitation of the explanation method is its exponential time complexity, which makes it unfeasible on data sets with a larger number of features. In such cases we can either to limit the number of subsets we examine (usually by limiting the depth) or by using feature selection to reduce the number of features [13,22]. The development of an approximation method is a part of our future work. However, for the purposes of this paper the full method was used (i.e., all the subsets were examined) as it takes only a few seconds to generate instance explanations on features oncologists use for breast cancer recurrence prognosis.

When W equals the full set of features S, we get $\Delta_S = p_S - p_{\text{prior}}$. This is, in other words, the difference in the classifiers prediction, which is caused by the knowledge of all the feature values for the instance we are trying to explain. How much each feature value contributed to this difference is exactly what we are trying to explain. Next, we define the contribution of the interaction between the features in an arbitrary subset $Q \subseteq S$:

$$\mathcal{I}_Q = \sum_{W \subseteq Q} \left((-1)^{|Q|-|W|} \Delta_W \right) \tag{1}$$

Table 8 A simple data set with 8 instances, 3 features, and a binary class value

A_1	A_2	A_3	C
0	0	0	0
0	1	0	0
0	0	1	0
0	1	1	0
1	0	0	0
1	1	0	1
1	0	1	0
1	1	1	1

The rationale behind (1) starts with the decomposition $\Delta_S = \sum_{W \subseteq S} \mathcal{I}_W$. In other words, the difference we are trying to explain is the sum of *interaction contributions* of all the subsets' of feature values. The interaction contribution of the feature values in set W, $\mathcal{I}_W$, is an abstract notion of the change in prediction that is caused by observing these feature values together but cannot be observed on any subset of those feature values. To explicitly define the interaction contribution, we use the generalization $\Delta_Q = \sum_{W \subseteq Q} \mathcal{I}_W$. By assuming $\mathcal{I}_W = 0$ (i.e., knowing no feature values contributes nothing), we get a recursive definition of an interaction contribution $\mathcal{I}_Q = \Delta_Q - \sum_{W \subset Q} \mathcal{I}_W$, which can be transformed into its explicit form (1). The interaction contribution and its transformation from recursive to explicit form (1) is very similar the inclusion/exclusion principle from set theory. Therefore, the proof is omitted.

We can now define π_i, the contribution of the ith feature, by dividing the contributions of interactions among the involved feature values:

$$\pi_i = \sum_{W \subseteq S \wedge i \in W} \frac{\mathcal{I}_W}{|W|} \tag{2}$$

The division of each interaction into equal parts (2) is justified by the fact that it is the only symmetrical (i.e., fair) division without using background knowledge.

An illustrative example offers more insight into the workings of the explanation method. Table 8 shows a simple data set with 3 features and a binary class value. The concept behind the data set is $C = A_1 \wedge A_2$, so feature A_3 is irrelevant. In this case there are 3 features, so $S = \{1, 2, 3\}$. For instance $(A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 1)$ the Bayesian classifier (i.e., we estimate all the conditional probabilities directly from the data set) would predict class value 1 with probability 1: $P\,(C = 1 | A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 1) = 1$. We can use our explanation method to explain this decision. The prior class probability is $p_{\text{prior}} = P(C = 1) = \frac{1}{4}$, so $\Delta_S = P\,(C = 1 | A_1 = 1 \wedge A_2 = 1 \wedge A_3 = 1) - P(C = 1) = 1 - \frac{1}{4} = \frac{3}{4}$. Therefore, the increase in predicted probability when the feature values are known is $\frac{3}{4}$. To calculate the remaining Δ-terms, we need the conditional probabilities with omitted features. This is done by ignoring the columns of omitted features and recalculating the predictions (in practice, this means retraining the classifier on the data set with omitted features or use an alternative method for simulating such predictions). For example, if we ignore feature A_2, then $P\,(C = 1 | A_1 = 1 \wedge A_3 = 1) = \frac{1}{2} \Rightarrow \Delta_{\{1,3\}} = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$.

Once we have all the Δ-terms, we can use (1) to calculate the interaction contributions. The contribution of the interaction of all three feature values is $I_{\{1,2,3\}} = \Delta_{\{1,2,3\}} - \big(\Delta_{\{1,2\}} + \Delta_{\{1,3\}} + \Delta_{\{2,3\}}\big) + \big(\Delta_{\{1\}} + \Delta_{\{2\}} + \Delta_{\{3\}}\big) = \frac{3}{4} - \big(\frac{3}{4} + \frac{1}{4} + \frac{1}{4}\big) + \big(\frac{1}{4} + \frac{1}{4} + 0\big) = 0$.

Therefore, observing all three values together brings nothing new to the classifier's decision. The only non-zero interaction contributions are: $I_{\{1,2\}} = I_{\{1\}} = I_{\{2\}} = \frac{1}{4}$. Using (2) we can now calculate the contributions of individual feature values: $\pi_1 = \pi_2 = \frac{\frac{1}{4}}{2} + \frac{1}{4} = \frac{3}{8}$ and $\pi_3 = 0$, because $A_3 = 1$ is not involved in any non-zero interaction. Both the interactions and the final contributions make sense because $A_3 = 1$ is indeed irrelevant. Feature values $A_1 = 1$ and $A_2 = 1$ are equally important, both values influence the decision individually and additionally influence the decision when observed together. Also, the sum of the contributions equals the initial change in predicted probability (i.e., Δ_S), which is an important property of the method.

References

1. Andrews R, Diederich J, Tickle AB (1996) Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl Based Syst 8:373–389
2. Becker B, Kohavi R, Sommerfield D (1997) Visualizing the simple bayesian classier. In: KDD Workshop on issues in the integration of data mining and data visualization
3. Bosnić Z, Kononenko I (2007) Estimation of individual prediction reliability using the local sensitivity analysis. Appl Intell 29(3):187–203
4. Bosnić Z, Kononenko I (2008) Automatic selection of reliability estimates for individual predictions. Knowl Eng Rev (in press)
5. Bosnić Z, Kononenko I (2008) Comparison of approaches for estimating reliability of individual regression predictions. Data Knowl Eng 67(3):504–516
6. Bosnić Z, Kononenko I (2009) An overview of advances in reliability estimation of individual predictions in machine learning. Intell Data Anal 13(2):385–401
7. Breiman L (2001) Random forests. Mach Learn J 45:5–32
8. Craven MW, Shavlik J (1994) Using sampling and queries to extract rules from trained neural networks. In: Proceedings of international conference on machine learning, pp 37–45
9. Crowder MJ, Kimber A, Smith RL, Sweeting T (1991) Statistical concepts in reliability. Statistical analysis of reliability data. Chapman & Hall, London
10. De Falco I, Della Cioppa A, Iazzetta A, Tarantino E (2006) An evolutionary approach for automatically extracting intelligible classification rules. Knowl Inf Syst 7:179–201
11. de Santana AL, Frances C, Rocha CA, Carvalho SV, Vijaykumar NL, Rego LP, Costa JC (2007) Strategies for improving the modeling and interpretability of bayesian networks. Data Knowl Eng 63(1):91–107
12. Gammerman A, Vovk V, Vapnik V (1998) Learning by transduction. In: Proceedings of the 14th conference on uncertainty in artificial intelligence. Madison, Wisconsin, pp 148–155
13. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
14. Hamel L (2006) Visualization of support vector machines with unsupervised learning. In: Proceedings of 2006 IEEE symposium on computational intelligence in bioinformatics and computational biology. pp 1–8
15. Jakulin A, Možina M, Demšar J, Bratko I, Zupan B (2005) Nomograms for visualizing support vector machines. In: KDD '05: proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp 108–117
16. Kohavi R, John G (1997) Wrappers for feature subset selection. Artif Intell J 97(1–2):273–324
17. Kononenko I (1993) Inductive and bayesian learning in medical diagnosis. Appl Artif Intell 7:317–337
18. Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 23:89–109
19. Kononenko I, Kukar M (2007) Machine learning and data mining: introduction to principles and algorithms. Horwood, New York
20. Kukar M (2004) Quality assessment of individual classifications in machine learning and data mining. Knowl Inf Syst 9(3):364–384
21. Lemaire V, Féraud R, Voisine N (2008) Contact personalization using a score understanding method. In: International joint conference on neural networks (IJCNN)
22. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. Knowl Data Eng 17(4):491–502

23. Možina M, Demšar J, Kattan M, Zupan B (2004) Nomograms for visualization of naive bayesian classifier. In: PKDD '04: proceedings of the 8th European conference on principles and practice of knowledge discovery in databases, New York, NY, USA, 2004. Springer, New York, pp 337–348
24. Poulet F (2004) Svm and graphical algorithms: a cooperative approach. In: Proceedings of fourth IEEE international conference on data mining. pp 499–502
25. Robnik Šikonja M, Kononenko I (2008) Explaining classifications for individual instances. IEEE Trans Knowl Data Eng 20:589–600
26. Saunders C, Gammerman A, Vovk V (1999) Transduction with confidence and credibility. Proc IJCAI' 99(2):722–726
27. Szafron D, Poulin B, Eisner R, Lu P, Greiner R, Wishart D, Fyshe A, Pearcy B, Macdonell C, and Anvik J (2006) Visual explanation of evidence in additive classifiers. In: Proceedings of innovative applications of artificial intelligence
28. Štrumbelj E, Kononenko I (2008) Towards a model independent method for explaining classification for individual instances. In: Proceedings of data warehousing and knowledge discovery. LNCS. Springer, Berlin, pp 273–282
29. Towell G, Shavlik JW (1993) Extracting refined rules from knowledge-based neural networks. Mach Learn 13:71–101
30. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu R, Yu PS, Steinbach ZZM, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37

Author Biographies

Erik Štrumbelj is a Ph.D. student and junior researcher at the University of Ljubljana, Faculty of Computer and Information Science. His research interests include machine learning, data mining, and forecasting.

Zoran Bosnić obtained his Master and Doctor degrees in Computer Science at University of Ljubljana (Slovenia) in 2003 and 2007, respectively. Since 2006 he has been employed at Faculty of Computer and Information Science and currently works as an assistant professor in the Laboratory of Cognitive Modelling. His research interests include artificial intelligence, machine learning, regression, and reliability estimation for individual predictions, as well as applications in these areas.

Igor Kononenko received his Ph.D. in 1990 from University of Ljubljana, Slovenia. He is a professor at the Faculty of Computer and Information Science in Ljubljana and the head of the Laboratory for Cognitive Modeling. His research interests include artificial intelligence, machine learning, neural networks and cognitive modeling. He is a member of the editorial board of Applied Intelligence Journal and informatica Journal. He is the (co)author of 180 papers and 10 textbooks. Recently, he co-authored the book "Machine Learning and Data Mining: Introduction to Principles and Algorithms" (Horwood 2007).

Branko Zakotnik graduated from medicine in 1979 at the University of Ljubljana. He obtained his board certificate of internal medicine in 1988 and Ph.D. degree in 1999. His appointments include President of the institutional Ethics committee at the Institute of Oncology Ljubljana, 1991-2003, President of Cancerologic society of Slovenia, 1999–2007, European Society of Medical Oncology (ESMO) National representative, 1996–2002. He is a specialist of Internal Medicine, Medical Oncology Department, Institute of Oncology, Ljubljana, and associate professor at the Medical faculty Ljubljana. His main research interests include clinical studies in cancer patients and survival of cancer patients in Slovenia. He is the (co)author of 170 papers and several textbooks.

Cvetka Grašič Kuhar is a medical oncologist at the Institute of Oncology, Ljubljana, Slovenia. She received her Ph.D. in 2009 at the University of Ljubljana, Slovenia. Her research interest includes breast cancer, especially the evaluation of prognostic and predicting factors.