EXPLANATION AND RELIABILITY OF INDIVIDUAL PREDICTIONS: RECENT RESEARCH BY LKM

Igor Kononenko, Erik Štrumbelj, Zoran Bosnić, Darko Pevec, Matjaž Kukar, Marko Robnik-Šikonja
Laboratory for Cognitive Modeling (LKM), Faculty of Computer and Information Science, University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
www: lkm.fri.uni-lj.si; e-mail: lkm@fri.uni-lj.si (name.surname@fri.uni-lj.si)

ABSTRACT

We review some recent research topics by Laboratory for Cognitive Modeling (LKM) at Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Classification and regression models, either automatically generated from data by machine learning algorithms, or manually encoded with the help of domain experts, are daily used to predict the labels of new instances. Each such individual prediction, in order to be accepted/trusted by users, should be accompanied by an explanation of the prediction as well as by an estimate of its reliability. In LKM we have recently developed a general methodology for explaining individual predictions as well as for estimating their reliability. Both, explanation and reliability estimation are general techniques, independent of the underlying model and provide on-line (effective and efficient) support to the users of prediction models.

1 INTRODUCTION

In a typical machine learning scenario a machine learning algorithm is used to construct a model of the relationships between the input features and the target variable with the purpose to predict the target variable of new, yet unseen instances. Explaining the learned relationships is also an important part of machine learning. Some models, such as additive models or small decision trees, are inherently transparent and require little or no additional postprocessing (Jakulin et al., 2005; Szafron, 2006). Other, more complex and often better performing models are non-transparent and require additional explanation. Therefore, model-specific explanation methods have been developed for models such as artificial neural networks and SVM.

In practice, dealing with several different explanation methods requires undesirable additional effort and makes it difficult to compare models of different types. To address this issue, general explanation methods are used – methods, which treat each model as a black-box and can be used independent of the model's type. Most general explanation methods are based on marginalization of features (Lemaire, 2008; Zien, 2009). This approach is computationally efficient. It is also effective as long as the model is additive (that is, as long as the features do not interact). However, several widely-used machine learning models are not additive, which leads to misleading and incorrect explanations of the importance and influence of features (Štrumbelj and Kononenko, 2010). Unlike existing general explanation methods, our method, described in Section 2, takes into account not only the marginal effect of single features but also the effect of subsets of features.

In supervised learning, one of the goals is to get the best possible prediction accuracy on new and unknown instances. As current prediction systems do not provide enough information about single predictions, experts find it hard to trust them. Common evaluation methods for classification and regression machine learning models give an averaged accuracy assessment of models, and in general, predictive models do not provide reliability estimates for their individual predictions. In many areas, appropriate reliability estimates may provide additional information about the prediction correctness and can enable the user (e.g. medical doctor) to differentiate between more and less reliable predictions. In Section 3 we describe our approaches to estimating the reliability of individual predictions. Finally, in Section 4 we overview directions of current research, carried out in LKM.

2 EXPLAINING INDIVIDUAL PREDICTIONS

The idea behind our method for explaining individual predictions is to compute the contributions of individual features to the model's prediction for a particular instance by decomposing the difference between the model's prediction for the given instance and the model's expected prediction (i.e., the model's predictions if none of the features' values were known). We adopt, with minor modifications, the notation used in (Štrumbelj and Kononenko, 2011). Let $A = A_1 \times A_2 \times ... \times A_n$ be our feature space, where each feature A_i is a set of values. Let p be the probability mass function defined on the sample space A. Let f_c: $A \rightarrow [0,1]$ describe the classification model's prediction for class value c. Our goal is a general explanation method which can be used with any model, so no other assumptions are made about $f_{c.}$ Therefore, we are limited to changing the inputs of the model and observing the outputs.

Let $S = \{A_1,..., A_n\}$ be the set of all features. The influence of a certain subset Q of S for classification of a given instance $x \in A$ is defined as:

$$\Delta(Q)(x) = \mathrm{E}[f \mid \text{values of features in } Q \text{ for } x] - \mathrm{E}[f] \qquad (1)$$

The value of the above function for the entire set of features S is exactly the difference between the model's prediction for a given instance and the model's expected prediction that we wish to decompose. Note that we omit the class value in the notation of f. Suppose that for every subset of features Q the value of $\Delta(Q)$ is known. The goal is to decompose $\Delta(S)$ in a way that assigns each feature a fair contribution with respect its influence on the model’s prediction. In (Štrumbelj and Kononenko, 2010) a solution is proposed that is equivalent to the Shapley value (Shapley, 1953) for the coalitional game with the *n* features as players and Δ as the characteristic function. The contribution of the *i*-th feature is defined as follows:

$$\varphi_i(x)=\sum_{Q\subseteq S\{i\}}\frac{|Q|!(|S|-|Q|-1)!}{|S|!}(\Delta Q\cup\{i\})(x)-\Delta(Q)(x)). \qquad (2)$$

These contributions have some desirable properties. Their sum for the given instance *x* equals Δ(*S*), which was our initial goal and ensures implicit normalization. A feature that does not influence the prediction will be assigned no contribution. And, features that influence the prediction in a symmetrical way will be assigned equal contributions.

The computation of Eq. (2) is infeasible for large *n* as the computation time grows exponentially with *n*. The approximation algorithm is proposed in (Štrumbelj and Kononenko, 2010; 2011), where we show its efficiency and effectiveness. It is based on the assumption that *p* is such that individual features are mutually independent. With this assumption and using an alternative formulation of the Shapley value we get a formulation which facilitates random sampling. For a global view on features’contributions, we define the contribution of the feature’s value as the expected value of that feature’s contribution for a given value. Again, random sampling can be used to estimate the expected value (Štrumbelj and Kononenko, 2011).

Let us illustrate the use of the features’ local and global contributions using a simple data set with 5 numerical features A_1,..., A_5 with unit domains [0,1]. The binary class value equals 1 if $A_1 > 0.5$ or $A_2 > 0.7$ or $A_3 > 0.5$. Otherwise, the class value is 0. Therefore, only the first three features are relevant for predicting the class value. This problem can be modeled with a decision tree. Figure 1 shows explanations for such a decision tree. The global contributions of each feature’s values are plotted separately. The black line consists of points obtained by running the approximation algorithm for the corresponding feature and its value corresponing the value on the x-axis. The lighter line corresponds to the standard deviation of the samples across all values of that particular feature and can therefore be interpreted as the overall importance of the feature. The lighter lines reveal that only the first three features are important. The black lines reveal the areas where features contribute towards/against class value 1. For example, if the value of feature A_1 is higher than 0.5 it strongly contributes towards class value being 1. If it is lower, it contributes against class value being 1.

For example, the instance *x* = (0.47, 0.82, 0.53, 0.58, 0.59) belongs to class 1, which the decision tree correctly predicts. The visualization on Figure 2 shows the individual features’ contributions for this instance.

The last two features have a 0 contribution. The only feature value that contributes towards class = 1 is A_2 = 0.82, while the remaining two features’ values have a negative contribution.

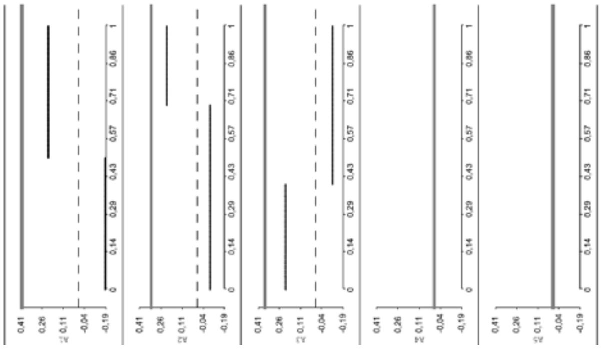

Figure 1: *Explanation of a decision tree*

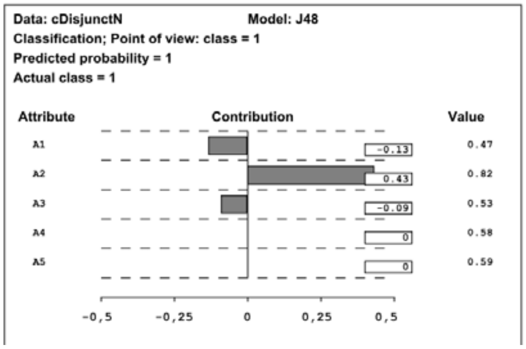

Figure 2: *Visualization of the individual features’ contributions for particular instance*

We have successfully applied this research to post-processing tools for breast cancer recurrence prediction (Štrumbelj et al., 2010), maximum shear stress prediction from hemodynaic simulations (Bosnić et al., 2011), and businesses' econonomic results precition (Pregeljc et al., 2012).

3 RELIABILITY OF INDIVIDUAL PREDICITONS

Because model independent approaches are general, they cannot exploit specific parameters of a given predictive model, but rather focus on influencing the parameters that are available in the standard supervised learning framework (e.g. a learning set and attributes). We expect from reliability estimators to give insight into the prediction error and we expect to find positive correlation between the two. The first two algorithms described in this Section arebased on Reverse transduction and Local sensitivity analysis and follow the same transductive idea, though the first is applicable only to classification, and the second to regression models. Other algorithms are general and need only minor adaptations when converted from regression to classification models, or vice versa.

3.1 Reverse transduction and Local sensitivity analysis

Transduction can be used in the reverse direction, in the sense of observing the model's behavior when inserting modified learning instances of the unseen and unlabeled instance (Kukar and Kononenko, 2002). Let *x* represent an instance and let y be its class label. Let us denote a learning instance with known label y as (x, y) and let $(x, \_)$ be an unseen and unlabeled instance, for which we wish to estimate the reliability of an initial prediction K. It is possible to create a modified learning instance by inserting the unseen instance $(x, \_)$ into the learning set and label it with the same (*first*) or different class (*second* best or *last*) as predicted by the initial model. The distance between the initial probability vector and that from the rebuilt model forms the reliability estimate. The three reliability estimators for classification models derived from three instances are labeled $TRANS_{first}$, $TRANS_{second}$ and $TRANS_{last}$.

In the regression the procedure is similar except that the predicted label is first slightly corrupted: $y = K + \delta$ and then we insert the newly generated instance (*x*, y) into the learning set and rebuild the predictive model. We define $\delta = \varepsilon(l_{\max} - l_{\min})$, where ε expresses the proportion of the distance between largest ($l_{\max}$) and smallest ($l_{\min}$) prediction. In this way we obtain a sensitivity model, which computes a sensitivity estimate $K\varepsilon$ for the instance $(x, \_)$. To widen the observation window in local problem space and make the measures robust to local anomalies, the reliability measures use estimates from the sensitivity models, gained and averaged across different values of $\varepsilon \in E$. For more details see (Bosnić and Kononenko, 2007). Let us assume we have a set of nonnegative ε values $E = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_{|E|}\}$. We define the estimates as follows:

- Estimate *SAvar* (Sensitivity Analysis local variance):

$$SAvar = \frac{\sum_{\varepsilon \in E}(K_\varepsilon - K_{-\varepsilon})}{|E|} \quad (3)$$

- Estimate *SAbias* (Sensitivity Analysis local bias):

$$SAbias = \frac{\sum_{\varepsilon \in E}(K_\varepsilon - K) + (K_{-\varepsilon} - K)}{2|E|} \quad (4)$$

3.2 Bagging variance

In related work, the variance of predictions in the bagged aggregate of artificial neural networks has been used to estimate the reliability of the aggregated prediction (Heskes, 1997; Carney and Cunningham, 1999). The proposed reliability estimate is generalized to other models (Bosnić and Kononenko, 2008).

Let K_{i}, $i = 1\ldots m$, be the predictor's class probability distribution for a given unlabeled example $(x, \_)$. Given a bagged aggregate of m predictive models, where each of the models yields a prediction B_{k}, $k = 1\ldots m$, the reliability estimator *BAGV* is defined as the variance of predictions' class probability distribution:

$$BAGV = \frac{1}{m}\sum_{k=1}^{m}\sum_i (B_{\mathrm{k,i}} - K_{\mathrm{i}})^2. \quad (5)$$

The algorithm uses a bagged aggregate of 50 predictive models as default.

3.3 Local cross-validation

The *LCV* (Local Cross-Validation) reliability estimate is computed using the local leave-one-out (LOO) procedure. Focusing on the subspace defined by k nearest neighbors, we generate k local models, each of them excluding one of the k nearest neighbors. Using the generated models, we compute the leave-one-out predictions K_{i}, $i = 1\ldots k$, for each of the k excluded nearest neighbors. Since the labels C_{i}, $i = 1\ldots k$, of the nearest neighbors are known, we are able to calculate the local leave-one-out prediction error as the average of the nearest neighbors' local errors:

$$LCV = \frac{1}{k}\sum_i |C_{\mathrm{i}} - K_{\mathrm{i}}|. \quad (6)$$

In experiment, the parameter k was assigned to one tenth of the size of the learning set.

3.4 Local error modeling

Given a set of k nearest neighbors, where C_i is the true label of the i-th nearest neighbor, the estimate CNK ($C_{\mathrm{Neighbors}} - K$) is defined as the difference between average label of the k nearest neighbors and the instance's prediction K:

$$CNK = \frac{\sum_i C_{\mathrm{i}}}{k} - K. \quad (7)$$

CNK is not a suitable reliability estimate for the k-nearest neighbors algorithm, as they both work by the same principle. In our experiments we used $k = 5$. In regression tests, *CNK-a* denotes the absolute value of the estimate, whereas *CNK-s* denotes the signed value.

3.5 Density based estimation

This approach assumes that an error is lower for predictions in denser problem subspaces, and higher for predictions in sparser subspaces. Note that it does not consider the learning instances' labels. The reliability estimator *DENS* is a value of the estimated probability density function for a given unlabeled example.

3.6 Empirical evaluation of estimators

For testing, 20 standard benchmark data sets were used for classification and 28 data sets for regression problems, gathered from UCI Machine Learning Repository (Asuncion & Newman, 2007) and from the StatLib DataSets Archive (Department of Statistics at Carnegie Mellon University, 2005). We tested the reliability estimators with eight regression and seven classification models, all implemented in statistical package R (decision and regression trees, linear regression, ANN, SVM, bagging, *k*-nearest neighbors, locally weighted regression, random forests, naive Bayes and generalized additive model). The testing was performed using a leave-one-out cross-validation. For each learning instance left out in each iteration, a prediction and all the reliability estimates were computed. In classification, the performance of the reliability estimates was measured by computing the Spearman correlation coefficient between the reliability estimate and the prediction error. In regression tests, the Pearson correlation coefficient was used. Figure 3 presents the results for regression, showing the average performance of reliability estimates ranked in the decreasing order with respect to the percent of positive correlations (a similar figure was obtained for classification, except that the order of measures was, from the best towards the worse: LCV, BAGV, CNK, *TRANS*$_{first}$, *TRANS*$_{second}$, *TRANS*$_{last}$, *DENS*). The results indicate that the estimators *TRANS*$_{first}$, *SAbias*, *CNK*, *BAGV* and *LCV* have a good potential for estimation of the prediction reliability.

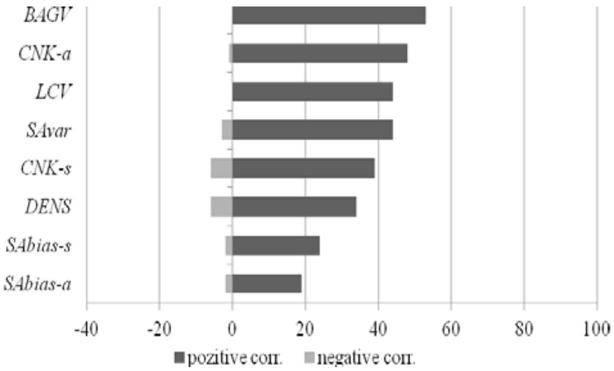

Figure 3: *Ranking of reliability estimators by the average percent of significant positive and negative correlation with the prediction error in regression experiments.*

The proposed reliability estimation methodology has been implemented in several applications of machine learning and data mining, e.g. breast cancer recurrence prediction problem (Štrumbelj et al., 2010), electricity load forecast prediction problem (Bosnić et al., 2011), and predicting maximum wall shear stress magnitude and position in the carotid artery bifurcation (Bosnić et al., 2012).

4 CURRENT LKM RESEARCH DIRECTIONS

Current LKM research is focused on several topics:

- evaluation of ordinal features in the context of surveys and customer satisfaction in marketing;
- learning of imbalanced classification problems;
- applying evolutionary computation to data mining, focused on using ant colony optimization;
- prediction intervals which represent the distribution of individual future points in a more informative manner;
- spatial data mining with multi-level directed graphs;
- employing background knowledge analysis for search space reduction in inductive logic programming;
- profiling web users in an online advertising network;
- employing algebraic methods, particularly matrix factorization for text summarization;
- detection of (non)-ischaemic episodes in ECG signals;
- heuristic search methods in clickstream mining;
- implementation and evaluation of reliability estimators in online learning (data streams);
- adaptation of the explanation methodology for online learning (data streams);
- modelling the progression of team sports matches and evaluation of the individual player contribution.

Acknowledgments

Research of LKM is supported by ARRS (Slovenian Research Agency). We thank all former and current members of LKM for their contributions.

References

Asunction, A., & Newman, D. J. (2007). *UCI ML repository*.

Bosnić, Z., & Kononenko, I. (2007). Estimation of individual prediction reliability using the local sensitivity analysis. *Applied Intelligence*, *29*(3), 187–203.

Bosnić, Z., & Kononenko, I. (2008). Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering*, *67*(3), 504–516.

Bosnić, Z., Rodrigues, P.P., Kononenko, I., Gama, J (2011). Correcting streaming predictions of an electricity load forecast system using a prediction reliability estimate. In: Czachorski, T. et al (eds.): *Man-machine interactions 2 (proceedings)*, Springer, 2011, 343-350.

Bosnić, Z., Vračar, P., Radović, M.D., Devedžić, G., Filipović, N., Kononenko, I. (2012) Mining data from hemodynamic simulations for generating prediction and explanation models. *IEEE trans. inf. technol. biomed.*, vol. 16, no. 2, 248-254.

Carney, J., & Cunningham, P. (1999). Confidence and prediction intervals for neural network ensembles. In *Proceedings of IJCNN'99, The International Joint Conference on Neural Networks*, Washington, USA, (pp. 1215–1218).

Department of Statistics at Carnegie Mellon University. (2005). *Statlib – Data, software and news from the statistics community*. Retrieved from http://lib.stat.cmu.edu/

Heskes, T. (1997). Practical confidence and prediction intervals. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in Neural Information Processing Systems, 9*, 176–182. The MIT Press.

Jakulin, A., Možina, M., Demšar, J., Bratko, I., Zupan, B. (2005). *Nomograms for visualizing support vector machines*. In: KDD '05: ACM SIGKDD, 108–117.

Kukar, M., & Kononenko, I. (2002). Reliable classifications with machine learning. In Elomaa, T., Manilla, H., & Toivonen, H. (Eds.), *Proceedings of Machine Learning: ECML-2002* (pp. 219–231). Helsinki, Finland: Springer.

Lemaire, V., Feraud, V., & Voisine, N. (2008). Contact personalization using a score understanding method. In *International Joint Conference on Neural Networks (IJCNN)*, 649-654.

Pregeljc, M., Štrumbelj, E., Mihelcic, M., & Kononenko, I. (2012). Learning and Explaining the Impact of Enterprises' Organizational Quality on their Economic Results. In R. Magdalena-Benedito, M. Martínez-Sober, J. Martínez-Martínez, J. Vila-Francés, & P. Escandell-Montero (Eds.), Intelligent Data Analysis for Real-Life Applications: Theory and Practice (pp. 228-248).

Shapley, L. S. (1953). *A Value for n-person Games*. volume II of Contributions to the Theory of Games. Princeton University Press.

Štrumbelj, E. & Kononenko, I. (2010). *An Efficient Explanation of Individual Classifications using Game Theory*. Journal of Machine Learning Research 11, 1-18.

Štrumbelj, E. & Kononenko, I. (2011). *A General Method for Visualizing and Explaining Black-Box Regression Models, International Conference on Adaptive and Natural Computing Algorithms.* ICANNGA 2011.

Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B. & Grašič, C. (2010). Explanation and reliability of prediction models: the case of breast cancer recurrence. *Knowledge and information systems*, vol. 24, no. 2, 305-324.

Szafron, D., Poulin, B., Eisner, R., Lu, P., Greiner, R., Wishart, D., Fyshe, A., Pearcy, B., Macdonell, C. , & Anvik, J. (2006). Visual explanation of evidence in additive classifiers. In *Proceedings of Innovative Applications of Artificial Intelligence – Volume 2*. AAAI Press.

Zien, A., Krämer, N., Sonnenburg, S., & Rätsch, G. (2009). The feature importance ranking measure. In *ECML PKDD 2009*, Part II, pages 694–709. Springer-Verlag.