Case-Based Reasoning for Explaining Probabilistic Machine Learning

Tomas Olsson[1, 2], Daniel Gillblad[2], Peter Funk[1], and Ning Xiong[1]

[1]School of Innovation, Design, and Engineering, Mälardalen University, Västerås, Sweden

{tomas.olsson, peter.funk, ning.xiong}@mdh.se,

[2]SICS Swedish ICT, Isafjordsgatan 22, Box 1263, SE-164 29 Kista, Sweden

{tomas.olsson, daniel.gillblad}@sics.se

Abstract

This paper describes a generic framework for explaining the prediction of probabilistic machine learning algorithms using cases. The framework consists of two components: a similarity metric between cases that is defined relative to a probability model and an novel case-based approach to justifying the probabilistic prediction by estimating the prediction error using case-based reasoning. As basis for deriving similarity metrics, we define similarity in terms of the principle of interchangeability that two cases are considered similar or identical if two probability distributions, derived from excluding either one or the other case in the case base, are identical. Lastly, we show the applicability of the proposed approach by deriving a metric for linear regression, and apply the proposed approach for explaining predictions of the energy performance of households.

Keywords

Case-based Reasoning, Case-based Explanation, Artificial Intelligence, Decision Support, Machine Learning

1. Introduction

Explanation has been identified as a key factor for user acceptance of an intelligent system [1–6]. If a user does not understand or trust a decision support system, it will be less likely that the system will be accepted. For instance, in the medical domain it is well-known that a good prediction performance will not automatically mean user acceptance unless the physicians understand the reasoning behind [7]. In this paper, we propose a novel approach to using case-based reasoning (CBR) as an intuitive approach for justifying (explaining) the predictions of a probabilistic model as a complement to traditional statistical measures of uncertainty such as the mean value and the variance.

CBR is a conceptually simple and intuitive, but yet powerful approach for knowledge management and learning [8, 9]. In contrast to model-based approaches in machine learning and statistics, inference in CBR is done directly from a set of cases without generalizing to a model. The fundamental idea in CBR is that similar problems have similar solutions and therefore, new solutions can be created from previous solutions. Traditionally, CBR is not used if there is a sufficiently good model-based solution to a problem. Yet, CBR has some advantages that complement model-based learning approaches. For instance, a probabilistic machine learning model can be hard to understand for non-experts while CBR is conceptually much more intuitive and easy to explain. Therefore, by complementing a probabilistic model with a CBR-based explanation facility, we can make the system more understandable.

Explanation using preceding cases has some advantages compared to other approaches. For instance, it has been shown in a user experiment that users in some domains prefer case-based over rule-based explanation [10]. In [11], Nugent et al. list three advantages of case-based explanation. Firstly, it is a natural form of explanation in many domains where explanation by analogy is common. Second, it uses real evidence in form of a set of cases relevant to the task at hand. This, we argue in this paper, is the key strength of combining probabilistic methods with CBR for explanation. Lastly, it is a fixed and simple form of explanation that is directly related to the problem at hand. Thus, regardless of the complexity of the problem at hand, the content of the explanation can be kept rather constant.

The purpose of explaining a system can vary depending on which type of user the explanation is addressing. For instance, Sørmo et al. list five common types of explanations: transparency, learning, conceptualization, relevance, and justification [12]. The goal of transparency is to explain how the system computed the prediction. This goal is more relevant for expert users that are able to assess the reasoning process by themselves. The goal of learning is to help novice users to learn the application domain. The goal of conceptualization is to help novice users to understand the meaning of concepts used in the system. The explanation goal relevance is about explaining why the system does something such as asking a question for more information. In this paper, we only consider the last type of explanation – justification – where the goal is to support a non-expert user in assessing the reliability of the system predictions using CBR.

The work in the current paper presents developments in line with previous work in knowledge light CBE that uses cases to explain model-based machine learning algorithms [13–15]. However, while previous work define similarity metrics ad-hoc and by intuitive means, we differ by defining similarity metrics with a good theoretical foundation and with a clear meaning. This can be achieved by restricting the application to learning algorithms that result in probability models. In probabilistic machine learning, the inference is visible in terms of probability distributions. So by analyzing the probability distributions from two different probability models, we can draw some conclusions of the relation between them. This probabilistic assumption makes the proposed approach generically applicable to any probabilistic machine learning algorithm.

In addition, we also differ with respect to previous work in how a prediction is explained. Since the probability model is most likely trained on all cases, it would not be sufficient to justify the system performance by only showing a list of the most similar cases. A list of similar cases would lead the user into checking each case and compare its difference or similarity to the new case. This would be very time consuming and would not say much of the system performance at a larger scale. As justification of the system performance, we instead propose using the average prediction error for all similar preceding cases. This is a straightforward application of CBR applied to predicting the probabilistic prediction error.

The rest of the paper is organized as follows. Sect. 2 presents related work in case-based explanation. In Sect. 3, we give background to similarity metrics and statistical metrics. Sect. 4 presents the overall framework for explanation by cases. We present a definition of similarity in terms of probability distributions and our approach to case-based explanation. We also derive a similarity metric for linear regression. Sect. 5 describes the application of the proposed approach to explaining the prediction of the energy performance of a household. In Sect. 6, we make concluding remarks and describe future work.

2. CASE-BASED EXPLANATION

Case-based explanation (CBE) is a research field within CBR that investigates the use of cases for explaining systems [16–18, 5]. CBE can, similarly to CBR, be divided into knowledge intensive and knowledge light CBE where the former makes use of explicit domain knowledge while the latter uses mainly knowledge already contained in the similarity metric and the case base [10]. The main work in knowledge light CBE has been in explaining classification while less work has been invested in explaining regression. The current work is an instance of knowledge light CBE where no explicit explanation mechanisms are modeled and the considered learning task is regression.

Furthermore, CBE research differs in how cases are explained. One type of knowledge light CBE research investigates the use of other types of learning methods for explanation. The ProCon system described in [19, 20] uses a naive Bayes classifier trained on all cases to find which features of a case support or oppose the classification. The author argues that even the most similar case can some times contain information that contradicts the prediction and that must be made explicit in order to keep the confidence of the user. The system presented in [21] by the same author generates rules from the nearest neighbors in order to explain the retrieved cases. Similarly to the previous system, it is ensured that the learned rules subsumes both cases that supports and opposes the classification. Case-based explanation of a lazy learning approach in the same vein for classifying chemical compounds was presented in [23]. The approach lets the user compare the molecular structures using the similarities between cases with respect to cases from different classes. Thus, by showing the user similarities only common to cases of each class the user is also able to understand the difference between classifications. In [22], the authors describes a system that applies logistic regression to a set of retrieved cases and uses the logistic model to explain the importance of features and assign a probabilistic confidence measure of the system's prediction.

A second type of research considers which cases to present to a user for explaining classification of new cases. This research was started when it was noticed that the set of cases used for making the classification is not necessary also the best cases for explaining the classification. Instead of presenting the most similar case, it might be better to show a case close to the decision border between two classes. In [24], the authors compare similarity metrics optimized for explanation with similarity metrics optimized for classification, while in [25], the authors use the same similarity metric as for classification but explore different rules for selecting which case to use as explanation. In [22], logistic regression is used to find cases close to the classification border. A more recent work describing all these three approaches is presented in [11].

A third type of knowledge light CBE research has addressed the problem of explaining model-based machine learning methods using cases, but so far, mainly neural networks have been investigated [13, 14, 26, 15, 27, 28]. The first knowledge light CBE for model-based learning algorithms was presented in [13, 14]. In the first paper, the author sketches ideas on how to use the model of a neural network or a decision tree as a similarity metric. In case of neural networks the activation difference between two cases was proposed as a metric while the leaves in the decision tree naturally contain similar cases. In [26] the authors explain the prediction of an ensemble of neural networks using rules extracted from the network. Then, for a new case, only rules relevant for explaining that case are used. The rules are filtered using heuristic criteria. The work presented in [27, 28] explains the output of an ensemble of neural networks using the most important feature values relative to a case. In [15], a generic knowledge light CBE framework for black-box machine learning algorithms is presented. The authors trained a locally weighted linear model to approximate a neural network using artificial cases generated from the neural network. Then they used the coefficients of the linear model both as feature weights of a similarity metric to retrieve relevant cases and for explaining which features are most import for a prediction.

3. PRELIMINARIES

In this section, we define the notion of a true metric that is important in order to index cases for fast retrieval. In addition, we present the J-divergence that is a statistical measure of similarity between probability distribution that we use in our definition of similarity between cases.

3.1. True Metrics

In order to make fast retrieval of cases similarity metrics should adhere to the axioms of a true metric. Given a true metric, the search space can be partitioned into smaller regions and organized so that there is no need to search through all regions.

In this paper, we use the term metric informally as any function that makes a comparison between two cases, while a true metric is a metric in a mathematical sense. This means that a true metric is a function d that satisfy the following three axioms where X denotes the case base with the set of all cases:

1. $d(x,y) \geq 0$ (non-negative and identity) with $d(x,y) = 0$ if and only if $x = y, \quad \forall x,y \in X$
2. $d(x,y) = d(y,x)$ (symmetric) for all $x,y \in X$
3. $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality) for all $x,y,z \in X$

There is a discussion in the CBR literature whether all of the above axioms are required for useful similarity and distance metrics [29, 30]. A common true metric that we will use in this paper is the Manhattan distance. The definition of the Manhattan distance is

$$d(x,y) = \sum_{k} |x^k - y^k|$$

where $|\ldots|$ denotes the absolute value function.

3.2. Statistical Metrics

A commonly used statistical metric for comparing two probability distributions is the well-known Kullback-Leibler divergence (KL) [31]. KL is also sometimes called the relative entropy or the information gain, since it is closely related to the entropy concept introduced by Shannon [32, 33]. The KL for the two probability distributions p_i, p_j, with parameter θ, is

$$D(p_i \| p_j) = \int \log\left(\frac{p_i(\theta)}{p_j(\theta)}\right) p_i(\theta) d\theta \tag{1}$$

In case of discrete variables, the integral is replaced with a sum.

KL is not symmetric but by computing the KL divergence in both directions and then add them together we get a symmetric metric. This is an important characteristic if we desire a true metric as described in Sect. 3.1. The symmetric KL is often called Jeffreys divergence (J-divergence). The J-divergence will then be:

$$\begin{aligned} J(p_i,p_j) &= D(p_i\|p_j) + D(p_j\|p_i) \\ &= \int \log\left(\frac{p_i(\theta)}{p_j(\theta)}\right) p_i(\theta)d\theta + \int \log\left(\frac{p_j(\theta)}{p_i(\theta)}\right) p_j(\theta)d\theta \\ &= \int \log\left(\frac{p_i(\theta)}{p_j(\theta)}\right) \left(p_i(\theta) - p_j(\theta)\right)d\theta \end{aligned} \tag{2}$$

In this paper, we use the J-divergence as basis for the similarity metrics, because it is a commonly used measure and it has a clear information theoretical interpretation. Other statistical metrics for comparing distributions are also available such as the total variation distance, the Euclidean distance and the Jensen-Shannon divergence [34, 31,35–37].

4. THE CASE-BASED EXPLANATION FRAMEWORK

In this paper, we propose a generic case-based explanation framework that justifies the predictions of an intelligent system by estimating the prediction reliability case by case. We have restricted the approach to probabilistic machine learning methods, since that gives the framework a good theoretical foundation. However, the explanation part can in principle also be used for any learning algorithm.

Assume that we have trained a probability model for predicting an unknown variable, and that we have derived a relevant similarity metric. Then, the framework creates a case-based explanation according to the following pseudo-algorithm:

1. Make prediction for a new case using the probability model
2. Retrieve most similar previous cases
3. For each previous case
    a. Make prediction for the case using the probability model
    b. Compute the absolute prediction error given the ground truth
4. Estimate the prediction error for the new case as the average of previous prediction errors
5. Present predicted value and estimated prediction error to the user

We will apply this framework to a real example in Sect. 5 where we explain the predictions from a linear regression model of the energy performance of a household. However, before that, we will describe the approach in more detail in the rest of this section. First, in Sect. 4.1, we describe a generic approach to defining a similarity metric using methods from statistics and machine learning. Then, Sect 4.2 shows how a similarity metric based on the linear regression model can be derived. Last, Sect. 4.3 describes a CBR approach to estimating the prediction error.

4.1. A Statistical Measure of Similarity

In this section, we present a principled approach to defining similarity metrics using methods from statistics and machine learning.

As a means to relate the similarity between two cases to probability models, we have formulated the principle of interchangeability as a general definition. We define the principle of interchangeability as follows:

Definition 1. Two cases x_i, x_j in case base X are similar if they can be interchanged such that the two probability distributions P_i, P_j inferred from excluding x_i and x_j respectively from the case base – $X \setminus x_i$ and $X \setminus x_j$ – are identical with respect to some parameter(s) of interest.

We have chosen to use the J-divergence presented in Sect. 3.2 for comparing two probability distributions. The J-divergence distance between two cases x_i, x_j in case base X is

$$d(x_i,x_j) = J(p_i,p_j) = \int \log \frac{p_i(\theta)}{p_j(\theta)}(p_i(\theta) - p_j(\theta))d\theta \qquad (3)$$

where $p_i(\theta)$ and $p_j(\theta)$ are probability distributions derived from the case base when excluding the cases x_i and x_j respectively and θ are the parameters that define in what aspects cases are similar.

The resulting measure between two cases can then be interpreted information theoretically as the sum of the information gain from including one case in the case base over the other case and the information gain from including the other case in the case base over the first case. However, the resulting J-divergence distance is not a true metric, so an additional step might be needed that turns it into a final distance that fulfills the axioms of a true metric.

4.2. Derivation of a Similarity Metric for Linear Regression

In this section, we derive a similarity metric for linear regression showing that this simple statistical model leads to a simple metric based on the Manhattan distance. The derived distance metric is the ratio between the distribution parameters plus the square of the sum of the weighted differences between case features.

In linear regression, we assume that each case x_i in case base X can be modeled as a weighted sum of a feature vector:

$$y_i = \omega_0 + \sum_{k=1}^{K} \omega_k x_i^k + \varepsilon_i \tag{4}$$

where ε_i are the residual error and the ω is a weight vector, and an unknown value y for a new case x is estimated by:

$$\hat{y} = \omega_0 + \sum_{k=1}^{K} \omega_k x^k \tag{5}$$

Let ε be normally distributed with mean 0 and standard deviation σ and then, the predictive distribution, conditioned on the weights and the standard deviation from a point estimation will be denoted as:

$$p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}[y-(\omega_0+\sum_{k=1}^{K}\omega_k x^k)]^2} \tag{6}$$

Theorem 1. Let $p_i\ (y|x)$ and $p_j\ (y|x)$ be the predictive probability distributions for linear regression derived from excluding x_i and x_j respectively. Let $\omega_i, \sigma_i, \omega_j, \sigma_j$ be the corresponding point estimations for the parameters of the distributions. Then, it can be shown that the J-divergence distance for two cases x_i, x_j is:

$$d(x_i,x_j) = \left(\frac{\sigma_i^2}{2\sigma_j^2} + \frac{\sigma_j^2}{2\sigma_i^2} - 1\right) + \left(\frac{1}{2\sigma_j^2} + \frac{1}{2\sigma_i^2}\right)\left[\sum_{k=0}^{K}(\omega_i^k x_i^k - \omega_j^k x_j^k)\right]^2 \tag{7}$$

Remark 1. Notice that a special case of Eq. 7 is when $\sigma_i \approx \sigma_j \approx \sigma$ and $\omega_i \approx \omega_j \approx \omega$, in which case the distance metric becomes:

$$d(x_i,x_j) \approx \frac{1}{\sigma^2}\left[\sum_{k=1}^{K}\omega_k(x_i^k - x_j^k)\right]^2 \tag{8}$$

which is approximately true with a large number of cases.

If we ignore the difference between the standard deviations, the resulting metric is square of the difference between the predicted value of $\hat{y}_i$ and $\hat{y}_j$ for case x_i and x_j respectively. The square root of this is a true metric, with respect to the predicted values, but not with respect to the cases. As a consequence, two cases that lead to the same predicted values would be considered similar.

If a true metric is desirable, we have to ensure that different features do not cancel out each other. One way of doing this is by rewriting the distance metric as follows:

$$
\begin{aligned}
d(x_i,x_j) &\propto \left[\sum_{k=1}^{K} \omega_k (x_i^k - x_j^k)\right]^2 \\
&\leq \left[\sum_{k\in P} \omega_k (x_i^k - x_j^k) + \sum_{k\in N} |\omega_k||x_i^k - x_j^k|\right]^2 \\
&= \left[\sum_{k=1}^{K} |\omega_k||x_i^k - x_j^k|\right]^2
\end{aligned}
\tag{9}
$$

where P and N are all features k where the contribution $\omega_k (x_k - x_k)$ to the sum is positive and negative respectively. Then, by then taking the square root of the last metric, we get a true metric that is a weighted version of the Manhattan distance:

$$
d(x_i,x_j) = \sum_{k=1}^{K} |\omega_k||x_i^k - x_j^k| \tag{10}
$$

So, the absolute weights are indicating the importance of each feature, which is similar to how the weights of a locally weighted linear regression model are used in [15]. Thereby, we can theoretically justify the use of the regression weights in a distance metric.

4.3. Estimating the Prediction Error

In this paper, we have formulated case-based explanation as a regression problem that we solve using CBR. Thus, by retrieving a set of similar cases, we can estimate the error of the prediction for a new case. Below, we describe the simple average prediction estimation that we use for estimating the prediction error.

The average prediction error approach is a simple application of the k nearest neighbor algorithm [9]. Thus, given a new case x_i, we can retrieve the set of k most similar cases using the metric in Eq. 10. Thereafter, we can estimate the prediction error e_i as the average prediction error of the retrieved cases:

$$
\hat{e}_i = \frac{1}{k} \sum_{j=1}^{k} e_j \tag{11}
$$

where $e_j = |\hat{y}_j - y_j|$ is the actual prediction error, $\hat{y}_j$ is the predicted value for case x_j from Eq. 12 and y_j is the true value.

5. Explaining Predicted Energy Performance of a Household

In this section, we apply the proposed framework to explaining the prediction of energy performance of households. The energy performance of a household is measured in kilowatt hours per square meter and year, and corresponds to the energy need in a building for space and hot water heating, cooling, ventilation, and lighting based on standard occupancy. The energy performance has been computed by a certified energy and it is intended to make it easy to compare houses when selling and buying a house.

The rest of this section is organized as follows. Sect. 5.1 describes the used data set. Sect. 5.2 presents which features we have used and the result from fitting a linear regression model to the data. Sect. 5.3 describes the implementation and evaluation of the proposed case-based explanation framework for energy performance prediction. Finally, Sect. 5.4 ends this section by showing examples of case-based explanation for two households.

5.1. Energy Performance Data

The data that we have used comes from around 1800 energy reports collected in the ME3Gas project [38,41]. The energy report consists of four attributes related to the building and location of a house, which are shown in the upper part of Table 1, and energy measurements for 12 different heating system types, which are shown in Table 2. Each used energy system type is measured in kWh. In addition, there are also the time period when measuring was done and the energy performance measurement that are shown in the second part of Table 1.

Table 1. The building and location attributes with summary of the data followed by measuring period start and energy performance. Climate zone indicate location in Sweden: zone 1 is in the most northern part and 4 is in the most southern part of Sweden.

House attribute	Summary
Year of construction	mean: 1956, min: 1083, max: 2013
Climate zone 1-4	90, 248, 934, and 515 in each zone 1-4
House type	#Detached: 1727, #Terraced: 57
Size of heated area (m^2)	mean: 177, min: 38, max: 925
Start dates of measuring period (year)	mean: 2011, min: 2007, max: 2013
Energy performance	mean: 111, min: 24, max: 388

Table 2. Heating systems and how many households with each.

Heating System			
District heating	200	Electric direct Acting	738
Heating oil	75	Electric airborne	50
Natural gas	19	Geothermal heat pump electrical	317
Firewood	512	Exhaust air heat pump	147
Wood chips/pellets	133	Heat pump Air Air	417
Electric water-borne	316	Heat pump Air Water	137

5.2. Log-Linear Regression Model Fitting

In this section, we first select which features to use and then fit a linear regression model to the energy performance data. Our first observation is that the energy performance is not normally distributed, but log-normally distributed. This means that the logarithm of the energy performance is normally distributed. In addition, all relations between features and the log of the energy are not linear. Thus, the following set of new features were added that capture non-linear relations: age of the house when the measuring was started, log of age that is the natural logarithm of age, log of climate zone, and log of heated area. For the household heating systems, we assume that it is only known which types of heating systems a new household uses, not how much energy is used by each heating system. Each heating system is therefore represented as 1 if present or 0 if not.

A linear regression model was then fitted to the data using the ridge regularization implementation of the scikit-learn project [39]. This resulted in a log-linear regression model with the weights listed in Table 3. The energy performance can then be predicted as below, using the exponential power of the result from Eq. 5 plus an extra term ($s^2/2$) that is an adjustment for the bias of the log-normal model:

$$\hat{y} = e^{\omega_0 + \Sigma_{k=1}^{K} \omega_k x_i^k + s^2/2} \tag{12}$$

where s^2 is the estimated standard error of the predicted value $\hat{y}$. The distance metric would still be the same as in Eq. 10, but with the extra features added to the cases, and that the distance metric is defined with respect to the distribution of *log(y)* and not *y*.

Table 3. The regression weight of each feature.

House Characteristics	Weight (ω)	Heating Systems	Weight (ω)
Year of construction	0.002	District heating	0.142
Age	0.002	Heating oil	0.190
Log of Age	0.107	Natural gas	0.040
Climate zone	-0.021	Firewood	0.150
Log of Climate zone	-0.064	Wood chips/pellets	0.263
Detached house	0.068	Electric water-borne	0.100
Terraced house	0.064	Electric direct Acting	0.008
Heated area	0.000	Electric airborne	0.036
Log of Heated Area	-0.221	Geothermal heat pump electrical	-0.441
		Exhaust air heat pump	-0.046
		Heat pump Air Air	-0.159
		Heat pump Air Water	-0.197

5.3. Evaluation of the Estimation of the Prediction Error

In this section, we evaluate the estimation approach proposed in Sect 4.3 for estimating the prediction error of the log-linear regression model. We compare the estimated error to the true error.

In the experiment, we split the data set 10 times into three sets: a 60% training set, a 20% validation set and a 20% test set. Each time the log-linear model is trained on the training set. Then, the *k* nearest neighbor algorithm was trained on the training set and configured to use the distance metric from Sect. 5.2 together with average prediction error approaches from Sect. 4.3. Thereafter, the performance of the *k* nearest neighbor algorithm was measured on both the validation set and the test set. The results from all data splits were then averaged. The validation set is used for selecting the algorithm parameters k. The performance is measured using root mean square error (RMSE).

In Figure 1, we show the average RMSE of the estimated prediction error on a validation set and a test set. The figure shows that the performance of both the test and the validation set is quite consistent. The minimum error of the validation set is located somewhere between $k = 25$ and $k = 45$, so we can select $k = 35$ as the size of the case set that will be used for explaining the prediction. However, we should also consider the experience of the users. If the users distrust the explanation because we have selected a too small set of previous cases, we should consider selecting a larger k. For instance, the difference in performance between $k = 35$ and $k = 60$ is quite small, so the latter could be preferred in some situations if it is considered more convincing.

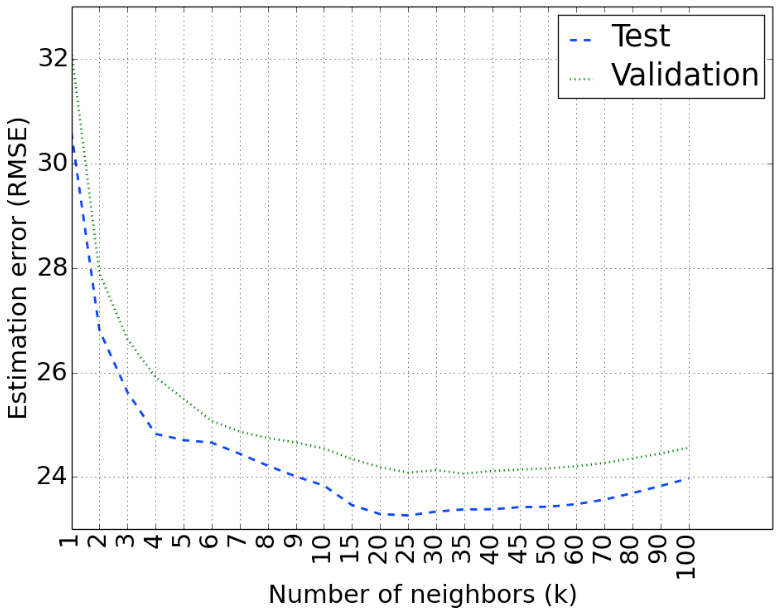

Figure 1. Number of k nearest neighbors (x-axis) versus the root mean square error (y-axis) of the estimated prediction error.

5.4. Case-based Explanation Examples

This section presents two examples of case-based explanations for two different households using the average prediction error as estimation with $k = 60$ neighbors. The two examples are shown in Table 4 and Table 5 respectively. Beginning from top of the tables. First, the characteristics of the house are listed, and then the used heating systems. Thereafter, the predicted value is shown, together with the true value that is assumed to be unknown but is here shown for comparison. Last, we explain the prediction in words, where we classify a estimated prediction error to be low if less than 10% of the predicted value, medium high if less than 20%, quite high if less than 30%, high if less than 40% and very high otherwise. In classifying the estimated prediction error, we use background knowledge in that low values are better than large and that the energy performance should be larger than zero. Thus, a relative value is an intuitive means of assessing the severity of the error. As can be seen from the examples, the explanations show that the prediction errors are medium high or high and that the prediction are not completely reliable, which can be confirmed by looking at the value of the true energy performance compared to the predicted value (especially the second household). However, the reason that the prediction of the second household is so bad is that there are very few houses with that combination of heating systems, especially wood chips/pellets, which can be easily seen by in addition listing the most similar cases. In Table 6 are the three most similar households listed, and as can be seen, none of them are very similar to example 2.

Table 4. Household example 1

Feature	Value
Year of construction	1977
Climate zone 1-4	2
House type	Detached house
Size of heated area	215 m2
Electric direct Acting	Yes
Heat pump Air Air	Yes
Predicted energy performance	84.0 kWh/m2
True energy performance	90 kWh/m2 (Unknown)
Explanation: The predicted energy performance of this house is 84.0 kWh/m2. The average prediction error for the 60 most similar houses is 14.5 kWh/m2. That is about 17.3% of the predicted energy performance, which is medium high.	

Table 5. Household example 2

Year of construction	1961
Climate zone 1-4	1
House type	Detached house
Size of heated area	100
Wood chips/pellets	Yes
Electric water-borne	Yes
Heat pump Air Air	Yes
Predicted energy performance	185.6 kWh/m2
True energy performance	81 kWh/m2 (Unknown)
Explanation: The predicted energy performance of this house is 185.6 kWh/m2. The average prediction error for the 60 most similar houses is 57.5 kWh/m2. That is about 31.0% of the predicted energy performance, which is high.	

Table 6. The three most similar cases to household example 2.

Feature	Case 1	Case 2	Case 3
Year of construction	1968	1909	1987
Heated area	105	107	112
Detached	Yes	Yes	Yes
Climate zone	1	2	3
Heating system	Firewood, Electric direct Acting	Electric water-borne, Heat pump Air Air	District heating
Predicted energy performance	186.0	128.7	137.3
True energy performance	172	132	128

6. Conclusions and Future Work

In this paper, we have proposed a framework for knowledge light case-based explanation of probabilistic machine learning. The first contribution of this work is a principled and theoretically well-founded approach to defining a similarity metric for retrieving cases relative a probability model.

As a second contribution, we have proposed a novel approach to justifying a prediction by estimating the prediction error as the average prediction error of the most similar cases. Since the justification is based on real cases and not merely on the correctness of the probability model, we argue that this is a more intuitive justification of the reliability than traditional statistical measures. However, it should be regarded as a complement to traditional measures rather than a replacement.

The work in this paper can be developed further in many directions. Clearly, we could develop case-based explanation approaches for other types of probability models. Especially, we would like to extend this approach to classification tasks. In addition, in order to evaluate the proposed explanation any further we would need to conduct user studies where we let real users use different versions of explanations and assess the effectiveness of each approaches.

Acknowledgements

The authors gratefully acknowledge the funding from the Swedish Knowledge Foundation (KK-stiftelsen) [40] through ITS-EASY Research School, the European ARTEMIS project ME3Gas (JU Grant Agreement number 100266), the ITEA 2 (ITEA 2 Call 5, 10020) and Swedish Governmental Agency for Innovation Systems (VINNOVA) grant no 10020 and JU grant no 100266 making this research possible. The energy performance data is used by courtesy of CNet Svenska AB [41].

References

[1] Wick, M.R., Thompson, W.B, (1992) “Reconstructive expert system explanation”, Artificial Intelligence 54(1), pp33–70

[2] Ye, L.R., Johnson, P.E. (1995) “The impact of explanation facilities on user acceptance of expert systems advice”, MIS Quarterly, pp157–172

[3] Gregor, S., Benbasat, I. (1999) “Explanations from intelligent systems: Theoretical foundations and implications for practice”, MIS quarterly, pp497–530

[4] Lacave, C., Diez, F.J.) (2004) “A review of explanation methods for heuristic expert systems”, The

Knowledge Engineering Review 19(02, pp133–146

[5] Leake, D., McSherry, D. (2005) “Introduction to the special issue on explanation in case-based reasoning”, Artificial Intelligence Review 24(2), pp103–108

[6] Darlington, K. (2013) “Aspects of intelligent systems explanation”, Universal Journal of Control and Automation 1, pp40 – 51

[7] Langlotz, C.P., Shortliffe, E.H. (1983) “Adapting a consultation system to critique user plans”, International Journal of Man-Machine Studies 19(5), pp479–496

[8] Aamodt, A., Plaza, E. (1994) “Case-based reasoning: Foundational issues, methodological variations, and system approaches”, AI communications 7(1), pp39–59

[9] Aha, D., Kibler, D., Albert, M. (1991) “Instance-based learning algorithms”, Machine learning 6(1), pp37–66

[10] Cunningham, P., Doyle, D., Loughrey, J. (2003) “An evaluation of the usefulness of case-based explanation”, In: Case-Based Reasoning Research and Development. Springer, pp122–130

[11] Nugent, C., Doyle, D., Cunningham, P. (2009) “Gaining insight through case-based explanation”, Journal of Intelligent Information Systems 32(3), pp267–295

[12] Sørmo, F., Cassens, J., Aamodt, A. (2005) “Explanation in case-based reasoning–perspectives and goals”, Artificial Intelligence Review 24(2), 109–143

[13] Caruana, R., Kangarloo, H., Dionisio, J., Sinha, U., Johnson, D. (1999) “Case-based explanation of non-case-based learning methods”, In: Proceedings of the AMIA Symposium, American Medical Informatics Association, pp212

[14] Caruana, R. (2000) “Case-based explanation for artificial neural nets”, In: Artificial Neural Networks in Medicine and Biology. Springer, pp303–308

[15] Nugent, C., Cunningham, P. (2005) “A case-based explanation system for black-box systems”, Artificial Intelligence Review 24(2), pp163–178

[16] Schank, R.C., Leake, D.B. (1989) “Creativity and learning in a case-based explainer”, Artificial Intelligence 40(1), pp353–385

[17] Aamodt, A. (1994) “Explanation-driven case-based reasoning”, In: Topics in case-based reasoning. Springer, pp274–288

[18] Doyle, D., Tsymbal, A., Cunningham, P. (2003) “A review of explanation and explanation in case-based reasoning”, Dublin, Trinity college https://www.cs.tcd.ie/publications/techreports/reports 3

[19] McSherry, D. (2003) “Explanation in case-based reasoning: an evidential approach”, In: Proceedings of the 8th UK Workshop on Case-Based Reasoning, pp47–55

[20] McSherry, D. (2004) “Explaining the pros and cons of conclusions in CBR”, In: Advances in Case-Based Reasoning. Springer, pp317–330

[21] McSherry, D. (2012) “A lazy learning approach to explaining case-based reasoning solutions”, In: Case-Based Reasoning Research and Development. Springer, pp241–254

[22] Nugent, C., Cunningham, P., Doyle, D. (2005) “The best way to instil confidence is by being right”, In: Case-Based Reasoning Research and Development. Springer, pp368–381

[23] Armengol, E. (2007) “Discovering plausible explanations of carcinogenecity in chemical compounds”, In: Machine Learning and Data Mining in Pattern Recognition. Springer, pp756–769

[24] Doyle, D., Cunningham, P., Bridge, D., Rahman, Y. (2004) "Explanation oriented retrieval", In: Advances in Case-Based Reasoning. Springer, pp157–168

[25] Cummins, L., Bridge, D.: Kleor (2006) A knowledge lite approach to explanation oriented retrieval. Computing and Informatics 25(2-3), pp173–193

[26] Wall, R., Cunningham, P., Walsh, P. (2002) "Explaining predictions from a neural network ensemble one at a time", In: Principles of Data Mining and Knowledge Discovery. Springer, pp449–460

[27] Green, M., Ekelund, U., Edenbrandt, L., Björk, J., Hansen, J., Ohlsson, M. (2008) "Explaining artificial neural network ensembles: A case study with electrocardiograms from chest pain patients", In: Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications.

[28] Green, M., Ekelund, U., Edenbrandt, L., Björk, J., Forberg, J.L., Ohlsson, M. (2009) "Exploring new possibilities for case-based explanation of artificial neural network ensembles", Neural Networks 22(1), pp75–81

[29] Burkhard, H.D., Richter, M.M. (2001) "On the notion of similarity in case based reasoning and fuzzy theory", In: Soft computing in case based reasoning. Springer , pp29–45

[30] Burkhard, H.D. (2001) "Similarity and distance in case based reasoning", Fundamenta Informaticae 47(3), pp201 – 215

[31] Kullback, S., Leibler, R.A (1951) "On information and sufficiency", The Annals of Mathematical Statistics 22(1), pp79–86

[32] Ihara, S. (1993) "Information theory for continuous systems", Volume 2. World Scientific

[33] Shannon, C.E. (2001) "A mathematical theory of communication", ACM SIGMOBILE Mobile Computing and Communications Review 5(1), pp3–55

[34] Rachev, S. T., Stoyanov, S. V., & Fabozzi, F. J. (2011) "A Probability Metrics Approach to Financial Risk Measures", John Wiley & Sons.

[35] Lin, J. (1991) "Divergence measures based on the Shannon entropy", Information Theory, IEEE Transactions on 37(1), pp145–151

[36] Cha, S. H. (2007) "Comprehensive survey on distance/similarity measures between probability density functions", City, 1(2), 1.

[37] Dragomir, S.C. (2008) "Some properties for the exponential of the Kullback-Leibler divergence", Tamsui Oxford Journal of Mathematical Sciences 24(2), pp141–151

[38] Me3gas – smart gas meters & middleware for energy efficient embedded services. url: http://www.me3gas.eu

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011) "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research 12, pp2825–2830

[40] KK-Stiftelse: Swedish Knowledge Foundation. http://www.kks.se (Last Accessed: April 2014)

[41] CNet Svenska AB. http://www.cnet.se (Last Accessed: April 2014)