ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Causes of Rejects in Prototype-based Classification Aleatoric vs. Epistemic Uncertainty

Johannes Brinkrolf[1], Valerie Vaquet[1], Fabian Hinder[1], and Barbara Hammer[1] *

1- Bielefeld University - Faculty of Technology
Universitätsstraße 25, 33615 Bielefeld - Germany

Abstract. Prototype-based methods constitute a robust and transparent family of machine-learning models. To increase robustness in real-world applications, they are frequently coupled with reject options. While the state-of-the-art method, *relative similarity*, couples the rejection of samples with high aleatoric and epistemic uncertainty, the technique lacks transparency, i.e., an explanation of why a sample has been rejected. In this work, we analyze the relative similarity analytically and derive an explanation scheme for reject options in prototype-based classification.

1 Introduction

Many real-world applications require accurate and robust machine learning models that can be trained on small datasets with limited computational power. Besides, additional requirements like transparency and fairness need to be met in accordance with the European AI Act [1]. One particularly suitable family of models for such tasks is prototype-based models as they natively provide transparency while being flexible and robust [2]. To further increase the robustness of classification models, frequently the option to reject uncertain samples instead of providing an uncertain potentially erroneous prediction is added [3]. To ensure transparency, some works explore explanation schemes for reject options, e.g. [4, 5]. One particular suitable reject strategy for prototype-based models is *Relative Similarity* (RelSim) [6]. This state-of-the-art method has the advantage of rejecting both samples of high *aleatoric* and high *epistemic* uncertainty. Thereby, aleatoric uncertainty refers to unclear and noisy data, such as data lying in the borderline region of two classes, while epistemic uncertainty refers to insufficient knowledge, e.g. outliers [7]. While having one reject strategy in place for both types is a convenient and efficient way to implement a reject scheme, providing information on why a sample has been rejected to a human user or operator would be desirable and increase the transparency of the overall pipeline. Unlike model explanations, such as counterfactual explanations of learning vector quantization (LVQ) as discussed in [5], we aim for explanations of the model uncertainty, i.e., its reason to reject a prediction, rather than the prediction itself.

Thus, in this paper, we analyze the properties of RelSim. Based on our considerations, we derive an explanation scheme that further categorizes rejects according to the type of uncertainty. Before analyzing the properties of rejects in Section 3, we briefly recall prototype-based models, the concept of reject options, and define RelSim in Section 2. Afterward, in Section 4, we propose a categorization of reasons for rejects to explain why a sample is rejected. We then

*We gratefully acknowledge funding from the European Research Council (ERC) under the ERC Synergy Grant Water-Futures (Grant agreement No. 951424)

evaluate the method on a medical real-world dataset in Section 5 showing that our methods exhibit intuitive behavior and conclude this paper in Section 6.

2 Reject Options for Prototype-Based Models

In this work, we will focus exemplarily on LVQ-based models. They consist of two components, the prototypes $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n \in \mathcal{D}$ with associated class $c(\boldsymbol{w}_i)$ and a semi-metric $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$. Classification is realized by the winner-takes-all principle, i.e., if $d(\boldsymbol{x}, \boldsymbol{w}^+) < d(\boldsymbol{x}, \boldsymbol{w}')$ for all $\boldsymbol{w}' \neq \boldsymbol{w}^+$ then $\boldsymbol{x}$ is assigned the label $c(\boldsymbol{w}^+)$. Note that the proposed methodology also applies to all other prototype-based classification schemes.

A common way to realize reject options is to consider a certainty measure $\mathrm{r}\colon \mathcal{D} \to \mathbb{R}$, $\boldsymbol{x} \mapsto \mathrm{r}(\boldsymbol{x})$. The classification for a point $\boldsymbol{x}$ is rejected if $\mathrm{r}(\boldsymbol{x}) < \theta$ where $\theta \in \mathbb{R}$ characterizes the reject threshold. Non-rejected points are refered to as accepted and we denote the set of all as $\mathrm{Acpt} = \{\boldsymbol{x} \mid \mathrm{r}(\boldsymbol{x}) \geq \theta\}$. Several certainty measures have already been proposed for prototype-based models [6]. A particularly efficient one is called *Relative Similarity* (RelSim) that is geometrically motivated and can be computed efficiently. It is defined as $\mathrm{r}_{\mathrm{RelSim}}(\boldsymbol{x}) = \frac{d(\boldsymbol{x}, \boldsymbol{w}^-) - d(\boldsymbol{x}, \boldsymbol{w}^+)}{d(\boldsymbol{x}, \boldsymbol{w}^+) + d(\boldsymbol{x}, \boldsymbol{w}^-)}$ where $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ are the closest prototype belonging to different classes [6]. Due to construction, $\mathrm{r}_{\mathrm{RelSim}}(\boldsymbol{x}) \in [0, 1]$ where 0 is reached exactly if $\boldsymbol{x}$ lies on the decision boundary and 1 if $\boldsymbol{x}$ and $\boldsymbol{w}^+$ coincide. An additional advantage of RelSim is that the method rejects both borderline samples and outliers. As RelSim is a prototype-based method, we can decompose the set of all accepted points according to the closest prototype $\mathrm{Acpt} = \cup_{\boldsymbol{w}} \mathrm{Acpt}(\boldsymbol{w})$. We will analyze this strategy in the next section. In particular, we will investigate a more refined decomposition of the rejected points $\mathrm{Rej} = \mathcal{D} \setminus \mathrm{Acpt}$.

3 Analyzing Relative Similarity

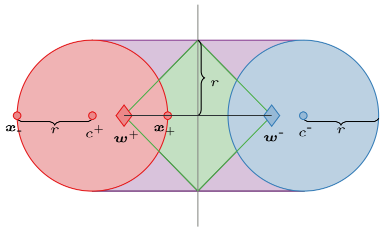
While the RelSim function can be applied to each classified point to decide whether it will be rejected or classified, obtaining a better understanding of this process is valuable. Therefore, we analyze which areas in the data space will be rejected.

Theorem 1. *Assume a binary model with two prototypes $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ with reject option induced by RelSim for threshold $\theta \in (0, 1)$. Then, the accepted points $\boldsymbol{x}$ classified as $c(\boldsymbol{w}^+)$ that are closest to/furthest away from $\boldsymbol{w}^-$ are found by the following optimization problems*

$$\begin{aligned} &\min_{c(\boldsymbol{x})=c(\boldsymbol{w}^+)} \pm d(\boldsymbol{x}, \boldsymbol{w}^-) \\ \text{s.t.} \quad &\mathrm{r}_{RelSim}(\boldsymbol{x}) \geq \theta \end{aligned}$$

which yield the solutions $\boldsymbol{x}_\pm = \boldsymbol{w}^+ + \lambda_\pm(\boldsymbol{w}^- - \boldsymbol{w}^+)$ where $\lambda_\pm = \frac{\theta - 1 \pm \sqrt{1-\theta^2}}{2\theta}$. The accepting area $Acpt(\boldsymbol{w}^+)$ of $\boldsymbol{w}^+$ is the ball with $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ as poles.

Due to space constraints, we refer to [8, Theorem 5] for the proof and skip a generalization to more than two prototypes that is straightforward as in classification, RelSim only considers the two closest prototypes of different classes.



1. $\mathrm{Acpt}(\boldsymbol{w}^+) =$ 🟥, $\mathrm{Acpt}(\boldsymbol{w}^-) =$ 🟦
2. $\overline{\boldsymbol{w}^+\boldsymbol{w}^-} =$ —
3. $\mathrm{conv}(\mathrm{Acpt}(\boldsymbol{w}^+),\, \mathrm{Acpt}(\boldsymbol{w}^-)) =$ 🟥 $\cup$ 🟪 $\cup$ 🟦
4. $\{\boldsymbol{x} \mid d(\boldsymbol{x},\, \boldsymbol{w}^+) = d(\boldsymbol{x},\, \boldsymbol{w}^-)\} =$ —
5. $\mathrm{Brd}(\boldsymbol{w}^+,\, \boldsymbol{w}^-) =$ 🟩

Fig. 1: Construction for borderline samples

Theorem 1 is of high relevance as it provides a closed-form description of the accepted samples. This way, the efficiency of computation of reject explanations, e. g. in [5], can drastically be increased as those are just projections onto the accepting ball. However, we are interested in a more detailed explanation of the reject allowing a deeper understanding. Based on Theorem 1, we propose an explanation scheme that provides a further categorisation of the rejected samples.

4 Explaining Reject Options

While the fact that RelSim rejects points lying in the border of two classes (Brd) and outliers (Out) is very convenient, it neglects valuable information that might be informative for potential downstream tasks and users interacting with the system. Consider for example a medical application: if a sample is rejected, it is of high relevance for a medical practitioner whether this is due to the fact that the sample lies in the borderline area between a healthy state and a medical condition (Brd) or whether it is an outlier (Out), e.g., a state which is not covered by the considered system or contains measurement errors which require running some test again. Thus, we want to differentiate the two types of rejected samples. Here, we will focus on the identification of borderline samples and obtain the outliers as the remaining rejected points.

For simplicity, we will consider the special case of two-classes. This is reasonable as RelSim is based on $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$. Thus, we are in the situation of Theorem 1. We define borderline samples as those that lie between classes (Btw) and are rejected (Rej), i.e., Brd = Btw ∩ Rej. We now define the set Btw by the following axioms:

(i) All data points lying on the line segment between the prototypes are between classes, i.e., $\overline{\boldsymbol{w}^+\boldsymbol{w}^-} \subseteq \mathrm{Btw}(\boldsymbol{w}^+,\, \boldsymbol{w}^-)$.

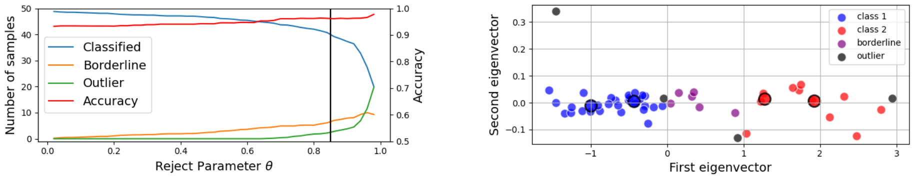
(ii) All points not lying in between any pair of accepted points are outliers, i.e., $\mathrm{Btw}(\boldsymbol{w}^+,\, \boldsymbol{w}^-) \subseteq \mathrm{conv}(\mathrm{Acpt}(\boldsymbol{w}^+),\, \mathrm{Acpt}(\boldsymbol{w}^-))$ where conv is the convex hull.

(iii) If a point lies between any pair of points in Btw, it is also in Btw, i.e., $\mathrm{Btw}(\boldsymbol{w}^+, \boldsymbol{w}^-)$ is convex.

(iv) All points on the decision boundary are either in Btw or in Out according to (ii), i.e., $\mathrm{conv}(\mathrm{Acpt}(\boldsymbol{w}^+),\, \mathrm{Acpt}(\boldsymbol{w}^-)) \cap \{\boldsymbol{x} \mid d(\boldsymbol{x},\, \boldsymbol{w}^+) = d(\boldsymbol{x},\, \boldsymbol{w}^-)\} \subseteq \mathrm{Btw}(\boldsymbol{w}^+,\, \boldsymbol{w}^-)$.

(v) $\mathrm{Btw}(\boldsymbol{w}^+)$ is the smallest set fulfilling (i)-(iv).

The properties (i)-(iii) are rather clear as they correspond to intuitive ideas of what it means to "lie between". However, they do not suffice to give a



(a) Number of samples per reject type and accuracy of classified samples

(b) Visualization of a test set in GMLVQ projection for $\theta = 0.85$.

Fig. 2: Exploration of dataset. Effect of θ on number of accepted and type of rejected samples and accuracy.

unique solution. As can be seen in Fig. 1, both formalize extremal situations, include samples that lie very far from the decision boundary ((ii)/purple) or exclude samples that can be seen as borderline ((i)/black line) and thus, lead to counter-intuitive situations. To resolve this issue, we also include the points on the decision boundary (iv). Uniqueness is then assured by minimality (v).

A great advantage of this description is that it admits a closed-form description and leads to an efficient algorithmic solution.

Lemma 1. *Consider the same setup as in Theorem 1. The constraints (i)-(v) yield a unique solution Btw*$(\boldsymbol{w}^+, \boldsymbol{w}^-)$ *that takes on the shape of a double cone with* $\boldsymbol{w}^+, \boldsymbol{w}^-$ *as tips and the cut of the decision boundary and the convex hull as base. In particular,* $\boldsymbol{x} \in \mathit{Brd}(\boldsymbol{w}^+, \boldsymbol{w}^-)$ *if and only if* $\boldsymbol{x} \in \mathit{Rej}$ *and*

$$\left\|(\boldsymbol{x} - \boldsymbol{w}^+) - \boldsymbol{d}\langle \boldsymbol{d}, \boldsymbol{x} - \boldsymbol{w}^+\rangle\right\| < 2\langle \boldsymbol{d}, \boldsymbol{x} - \boldsymbol{w}^+\rangle r$$

where $\boldsymbol{d} = \frac{\boldsymbol{w}^- - \boldsymbol{w}^+}{\|\boldsymbol{w}^- - \boldsymbol{w}^+\|}$ *and* r *is the radius of Acpt*$(\boldsymbol{w}^+, \boldsymbol{w}^-)$*.*

Proof. By Theorem 1, Acpt$(\boldsymbol{w}^+, \boldsymbol{w}^-)$ is a ball with center c^+ and radius $r = \frac{1}{2}\|\boldsymbol{x}_+ - \boldsymbol{x}_-\|$ and analogous for Acpt$(\boldsymbol{w}^-, \boldsymbol{w}^+)$ by symmetry. Furthermore, the decision boundary is orthogonal to the line spanned by $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ which also contains c^+ and c^-. Thus, by (iv), the intersection of the decision boundary and the ball around $\frac{1}{2}(\boldsymbol{w}^+ + \boldsymbol{w}^-)$ with radius r is contained in Btw$(\boldsymbol{w}^+, \boldsymbol{w}^-)$. Take C as the convex hull of this disc and the line segment $\overline{\boldsymbol{w}^+\boldsymbol{w}^-}$. Then C is a double cone as described and testing whether $\boldsymbol{x} \in C$ can be done as described above. Furthermore, $C \subseteq$ Btw$(\boldsymbol{w}^+, \boldsymbol{w}^-)$ by (i),(iv). Yet, as the convex hull is the smallest such set, we have equality by (v). □

5 Experiments

To evaluate the proposed methodology, we conducted two experiments on the Adrenal dataset [9] which consists of 147 samples from adrenal tumor patients with 32 preselected features of steroid metabolites. The two classes represent a benign and a malignant diagnosis. In the first experiment, we explore the different kinds of rejects for the dataset. In the second, we analyze the effect of different perturbations on the number of outliers and borderline samples[1].

[1] Our code of these experiments is available at `https://github.com/jbrinkro/Causes-of-Rejects-in-Prototype-based-Classification`

Table 1: Ratio of borderline samples (brl)/outliers (out) for different permutations (for line samples we use $\alpha = 0.4 + 0.2k/5$).

	interpolate		crossover		noise ($\sigma = 0.5$)		noise ($\sigma = 1$)		noise ($\sigma = 2$)		noise ($\sigma = 3$)		noise ($\sigma = 5$)	
k	brl	out	brl	out	brl	out	brl	out	brl	out	brl	out	brl	out
1	75.76	0.16	6.94	2.61	1.94	0.68	3.97	1.63	7.62	4.37	9.87	8.28	11.35	17.88
2	79.83	0.12	13.50	3.21	3.15	1.18	6.47	2.83	11.64	8.13	13.71	15.51	12.35	31.97
3	81.14	0.12	18.42	3.99	4.01	1.53	8.20	3.88	13.95	11.44	14.70	21.57	10.91	43.14
4	81.80	0.11	22.54	4.81	4.78	1.79	9.66	4.76	15.22	14.43	14.68	27.01	8.81	52.47
5	82.01	0.12	25.86	5.37	5.57	2.16	11.01	5.79	16.07	17.78	14.02	33.02	6.84	60.75

In the experiments, we apply GMLVQ [10] as a classifier which extends LVQ using metric learning by means of linear transformation. Our approach is applied after the linear transformation.

Exploration of the dataset The dataset is visualized in Fig. 2b. We first analyze the dataset by considering the number of accepted, borderline samples, and outliers for different choices of θ (see Fig. 2a). As can be seen, for small θ values, the number of borderline examples increases fast with increasing θ which corresponds to a growing margin, i. e. only borderline samples are rejected. Notice that the number of borderline examples linearly correlates with the accuracy. This implies that our method successfully identifies and rejects those samples that lie "between classes" and thus regularly misclassified. For large θ values, we observe a growing number of rejects due to outliers with increasing values of θ. While the rejection of the first outliers (for $\theta \approx 0.7$) results in an increase in accuracy rejecting additional samples does not benefit the performance indicating that the dataset does not contain many outliers. In the next experiment, we substantiate this claim using a quantitative evaluation.

Effect of perturbation To quantitatively substantiate the claim that our method assigns the reject type in an intuitive way, we consider different types of perturbations and their effect on the type of reject. We consider (pairs of) accepted samples $\boldsymbol{x}$ (and $\boldsymbol{x}'$) from the different classes to construct perturbed datapoints $\tilde{\boldsymbol{x}}$ using one of the following techniques: (i) linear interpolations, i.e., $\tilde{\boldsymbol{x}} = \alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{x}'$ for $\alpha \in (0,1)$, (ii) crossover, i.e., choose $C \subseteq \{1,\ldots,d\}$ where d is the number of features with $|C| = k$ and set $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}'_i$ if $i \in C$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$ otherwise, (iii) noise, i.e., choose C as before and set $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \mathbf{1}[i \in C]\varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0,\sigma)$. Linear interpolation and crossover can be seen as a way to mix samples. The strength of mixing is determined by α and k, respectively. Both are expected to lead to an increase in the number of borderline samples although crossover can also lead to out-of-manifold samples if the data lives on a non-axis-aligned hyperplane. Noise perturbs the samples in an arbitrary way and thus will likely lead to out-of-manifold samples, i.e. additional outliers.

We apply the perturbations to the dataset and report the number of induced outliers and borderline samples. We repeat the experiment 1000 times for various mixing (α, k) and perturbation strengths (σ). The results are shown in Table 1. As can be seen, linear interpolation as well as crossover lead to a significant increase of borderline samples. For stronger mixing, the ratio changes significantly towards borderline samples. Noise leads to more outliers if the perturbation strength is large enough. This is reasonable as one would expect a certain amount of noise in the data so that the effect only shows if we significantly exceed the data-specific noise level. Thus, the reject types assigned by our methods reflect our intuitive understanding of borderline and outlier cases.

Analyzing how permutation in specific features affects the rejection, we find that we reconstruct the relevance profile of GMLVQ. If we apply our method without the transformation, we see that applying noise to any feature will most likely lead to outliers independent of the selected feature. For crossover, different features are relevant for creating borderline samples (5, 6, 19) and outliers (4, 25, 26, 28). The set of features resulting in borderline samples coincides with the set of the most relevant features found by the model and those reported for the dataset in the literature [9].

6 Conclusion

In this work, we extended RelSim to allow for explanations on why a sample has been rejected by introducing the sub-categories of borderline samples (aleatoric) and outliers (epistemic uncertainty). We provided a closed-form description of the accepted samples which allows for an efficient algorithmic solution. Based thereon, we derived a decomposition of the rejected samples based on their relative position to the closest prototypes. We empirically showed that the method shows intuitive behavior when confronted with perturbed samples.

The proposed methodology provides a better understanding of the model's decision in the rejection case and is therefore of high relevance in domains like medicine enabling informed decision-making. However, future research is necessary: The explanation of whether a sample has been rejected due to aleatoric or epistemic uncertainty could be coupled with feature attributions. As briefly discussed, Theorem 1 provides an efficient way to compute counterfactual examples by projecting samples on the accepting ball. Besides, so far, our method is only designed for two class problems. However, in the case of multi-class problems with samples from several classes occupying neighboring locations in dataspace, a more in-depth analysis is desirable to provide insights into the classification.

References

[1] European Commission and Directorate-General for Communications Networks and Content and Technology. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2021.

[2] Michael Biehl, Barbara Hammer, and Thomas Villmann. Prototype-based models in machine learning. *WIREs Cognitive Science*, 7(2):92–111, 2016.

[3] Kilian Hendrickx, Lorenzo Perini, Dries Van Der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: a survey. *Machine Learning*, 113(5):3073–3110, May 2024.

[4] Andrea Pugnana, Carlos Mougan, and Dan Saattrup Nielsen. Model Agnostic Explainable Selective Regression via Uncertainty Estimation. *CoRR*, abs/2311.09145, 2023.

[5] André Artelt, Johannes Brinkrolf, Roel Visser, and Barbara Hammer. Explaining reject options of learning vector quantization classifiers. In *IJCCI 2022*, pages 249–261, 2022.

[6] Lydia Fischer, Barbara Hammer, and Heiko Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

[7] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021.

[8] Johannes Brinkrolf. *Learning Vector Quantization for the Real-World: Privacy, Robustness, and Sparsity.* PhD thesis, 2023.

[9] Michael Biehl, Petra Schneider, David Smith, Han Stiekema, Angela Taylor, Beverly Hughes, Cedric Shackleton, Paul Stewart, and Wiebke Arlt. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In *ESANN 2012*, 2012.

[10] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.