Balanced Explanation of Predictions with BEEF

The problem of understanding the reasons behind why different machine learning classifiers make specific predictions is very complex. The main reason is that classifiers, especially ones that use deep learning, are often very complex and often use many possible functions that are combined in many different ways to generate a final prediction.

Our computational framework, named Balanced English Explanations of Forecasts (BEEF), can generate balanced explanations for predictions made by any binary classifier and provide these explanations in plain English. A balanced explanation (BE, for short) of a prediction explains not only why the prediction might be true, but also why it might be false. As a result, a decision maker who is presented with a computer-generated prediction can see what the evidence for/against the prediction might be before making a decision based on the prediction.

The Explanation Computation Problem

We assume that a predictive model (or classifier) is given, along with a set $P$ of points that have already been classified either green or red by it. Based on these input points, we identify a set of blue and red clusters. These sets of clusters must jointly satisfy various desirable properties: 1) Blue clusters and red clusters must overlap as little as possible (minimum overlap), 2) each cluster must mostly consist of objects of one color (maximal purity), 3) most points in $P$ must belong to at least one cluster (maximal inclusion), and 4) we shouldn’t have too many clusters (simplicity).

As we are working in an n-dimensional space, we assume an axis-aligned hyper rectangle representation; cluster boundaries are therefore planar surfaces. A clear advantage of this choice is that users can more easily understand their description as compared to other representations, such as spherical-shaped clusters or strange hyper-surfaces. As an example, consider the figure above. Suppose we want to generate an English explanation for a query point $p$ with a blue prediction label falling in the blue cluster. A possible balanced explanation for $p$ can be of the form: “Given that $x_1$ lies between $u^1_1$ and $l^1_1$ and $x_2$ lies between $u^2_2$ and $l^2_2$, the query point is classified as blue with a confidence of 77%.”

We prove that the Explanation Computation Problem is NP-complete.

Balanced Explanations Extraction: The BEE Algorithm

A balanced explanation of a forecast generated by a classifier consists of three parts:

- Primary explanation: describes cluster with highest purity – confidence of the explanation is purity of the cluster.
- Supporting explanation(s): describes other clusters that support the forecast.
- Alternative explanation(s): describes clusters supporting the opposite of the primary explanation.

BEEF extracts high quality balanced explanations within a reasonable amount of time. It starts from the output of a generic clustering algorithm and computes the two sets of hyper rectangles by performing the following two operations: (i) it changes the boundaries of the clusters so that overlap is reduced, with the objective of providing higher-quality explanations that only include supporting and/or opposing parts whenever it is absolutely necessary; and (ii) it selects a set of features/dimensions for each cluster that is not too large, yet is still sufficiently representative—the goal is to keep the explanations bounded to a reasonable size so that they remain human-readable, but without decreasing their quality.

For each feature $f$ in the data, we assume the existence of a set of English-language templates to explain that feature. For instance, suppose we consider a health care application where a feature named $hrm$ measures heart rate per minute. In this case, the set of templates may contain two templates: (i) “The heart rate of the patient lies between $\#L$ and $\#U$”, and (ii) “The patient’s heart rate per minute lies in the $[\#L; \#U]$ interval”. Of course, there can be many other templates as well. In these templates, $\#L$ and $\#U$ are reserved “variable” symbols,
denoting lower/upper bounds, which may be instantiated. Likewise, for the dependent variable, we assume the existence of an English language template.

Evaluation of extracted explanations

We experimentally evaluated BEEF in two stages focusing on: (i) the quality of the explanations obtained and (ii) evaluation of the interpretability of the explanations using human subjects. We used 8 standard datasets from the machine learning repository (UCI).

For the second stage, a survey was conducted on Amazon Mechanical Turk with 100 participants to assess the readability of the generated explanations w.r.t. the explanations generated by an older method called LIME. The MTurk task consisted of 25 questions, each consisting of two parts. The first part asked the respondents to rate explanations (on a 7-point Likert scale) provided by the two different explanation engines for the same dataset. The second part asked which was more usable.

Results show that for 4 out of 5 datasets, a significant number of respondents gave positive ratings to the explanations generated by BEEF.

They also preferred BEEF to LIME; the only case in which LIME performed better is the MAGIC dataset, where the English explanation involves domain-related jargon that is difficult to understand for a nonexpert user. For the second part, more than 60% of respondents preferred BEEF over LIME for the same 4 out of 5 datasets.

---

**PARTICIPANTS**

Lead: V.S. Subrahmanian
Current: Chiara Pulice
Past: Sachin Grover, Gerardo Simari

---

**References**