Summary

The R (R Core Team, 2020) package stabm provides functionality for quantifying the similarity of two or more sets. For example, consider the two sets \( \{ A, B, C, D \} \) and \( \{ A, B, C, E \} \). Intuitively, these sets are quite similar because their overlap is large compared to the cardinality of the two sets. The R package stabm implements functions to express the similarity of sets by a real valued score. Quantifying the similarity of sets is useful for comparing sets of selected features. But also for many other tasks like similarity analyses of gene sets or text corpora, the R package stabm can be employed.

In the context of feature selection, the similarity of sets of selected features is assessed in order to determine the stability of a feature selection algorithm. The stability of a feature selection algorithm is defined as the robustness of the set of selected features towards different data sets from the same data generating distribution (Kalousis et al., 2007). For stability assessment, either \( m \) data sets from the same data generating process are available or \( m \) data sets are created from one data set. The latter is often achieved with subsampling or random perturbations (Awada et al., 2012). Then, the feature selection algorithm of interest is applied to each of the \( m \) data sets, resulting in \( m \) feature sets. To quantify the stability of the feature selection algorithm, the similarity of the \( m \) sets is calculated. In the context of feature selection stability, set similarity measures are called stability measures.

The R package stabm provides an open-source implementation of the 20 stability measures displayed in the table below. Argument checks are performed with checkmate (Lang, 2017) to provide helpful error messages. It is publicly available on CRAN and on Github and it has only a few dependencies.

<table>
<thead>
<tr>
<th>Name</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>stabilityDavis</td>
<td>Davis et al. (2006)</td>
</tr>
<tr>
<td>stabilityDice</td>
<td>Dice (1945)</td>
</tr>
<tr>
<td>stabilityHamming</td>
<td>Dunne et al. (2002)</td>
</tr>
<tr>
<td>stabilityIntersectionCount</td>
<td>Bommert &amp; Rahnenführer (2020)</td>
</tr>
<tr>
<td>stabilityIntersectionGreedy</td>
<td>Bommert &amp; Rahnenführer (2020)</td>
</tr>
<tr>
<td>stabilityIntersectionMBM</td>
<td>Bommert &amp; Rahnenführer (2020)</td>
</tr>
<tr>
<td>stabilityIntersectionMean</td>
<td>Bommert &amp; Rahnenführer (2020)</td>
</tr>
<tr>
<td>stabilityJaccard</td>
<td>Jaccard (1901)</td>
</tr>
<tr>
<td>stabilityKappa</td>
<td>Carletta (1996)</td>
</tr>
<tr>
<td>stabilityLustgarten</td>
<td>Lustgarten et al. (2009)</td>
</tr>
<tr>
<td>stabilityNogueira</td>
<td>Nogueira et al. (2018)</td>
</tr>
<tr>
<td>stabilityNovovicova</td>
<td>Novovičová et al. (2009)</td>
</tr>
<tr>
<td>stabilityOchiai</td>
<td>Ochiai (1957)</td>
</tr>
<tr>
<td>stabilityPhi</td>
<td>Nogueira &amp; Brown (2016)</td>
</tr>
<tr>
<td>stabilitySechidis</td>
<td>Sechidis et al. (2020)</td>
</tr>
</tbody>
</table>
Name | Reference
--- | ---
stabilityUnadjusted | Bommert & Rahnenführer (2020)
stabilityWald | Wald et al. (2013)
stabilityYu | Yu et al. (2012)
stabilityZucknick | Zucknick et al. (2008)

Statement of Need

The R package *stabm* provides an implementation of many stability measures. For theoretical and empirical comparative studies of the stability measures implemented in *stabm*, we refer to Bommert et al. (2017), Bommert & Rahnenführer (2020), Bommert (2020), and Nogueira et al. (2018). It has been demonstrated that considering the feature selection stability when fitting a predictive model often is beneficial for obtaining models with high predictive accuracy (Bommert et al., 2017; Bommert, 2020; Schirra et al., 2016). The stability measures implemented in the R package *stabm* have been employed in Bommert et al. (2017), Bommert et al. (2020), Bommert & Rahnenführer (2020), and Bommert (2020).

Related Software

A subset of the implemented stability measures is also available in other R or Python packages. The R package sets (Meyer & Hornik, 2009) and the Python package *scikit-learn* (Pedregosa et al., 2011) provide an implementation of the Jaccard index (Jaccard, 1901) to assess the similarity of two sets. The Python package *GSimPy* (Zhang & Cao, 2020) implements the Jaccard index, the Dice index (Dice, 1945), and the Ochiai index (Ochiai, 1957). The source code for the publication Nogueira et al. (2018) provides an implementation of their stability measure in R, Python, and Matlab.

Acknowledgements

This work was supported by the German Research Foundation (DFG), Project RA 870/7-1, and Collaborative Research Center SFB 876, A3.

References


