# Feature engineering for clustering student solutions

The MIT Faculty has made this article openly available. *Please share* how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1145/2556325.2567865">http://dx.doi.org/10.1145/2556325.2567865</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Association for Computing Machinery (ACM)</td>
</tr>
<tr>
<td>Version</td>
<td>Author's final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Dec 12 12:55:13 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/90409">http://hdl.handle.net/1721.1/90409</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/4.0/">http://creativecommons.org/licenses/by-nc-sa/4.0/</a></td>
</tr>
</tbody>
</table>
Feature Engineering for Clustering Student Solutions

Elena L. Glassman  Rishabh Singh  Robert C. Miller
MIT CSAIL, 32 Vassar St, Cambridge, MA
{elg,rishabhs,rcm}@mit.edu

ABSTRACT
Open-ended homework problems such as coding assignments give students a broad range of freedom for the design of solutions. We aim to use the diversity in correct solutions to enhance student learning by automatically suggesting alternate solutions. Our approach is to perform a two-level hierarchical clustering of student solutions to first partition them based on the choice of algorithm and then partition solutions implementing the same algorithm based on low-level implementation details. Our initial investigations in domains of introductory programming and computer architecture demonstrate that we need two different classes of features to perform effective clustering at the two levels, namely abstract features and concrete features.

RELATED WORK
A common goal of the prior work cited here is to help teachers monitor the state of their class, or provide solution-specific feedback to many students. However, the techniques for analyzing solutions have not converged on a particular method. Huang et al. [1] use unit test results and AST edit-distance algorithms to identify clusters of submissions that could potentially receive the same custom feedback message. Taherkhani et al. [4] identify which sorting algorithm a student implemented using supervised machine learning methods. Each solution is represented by statistics about language constructs, measures of complexity, and detected roles of variables.

OUR APPROACH
We are pursuing a two-level hierarchical clustering methodology. The high-level clusters are intended to partition solutions along the separation dimension, where each cluster represents a particular algorithm. We have used k-means to create these high-level clusters of solutions based on abstract features. The abstract features for Python programs consist of 12 features that include the position of conditional statements relative to the loop statements (before, after, or inside), the depth of nested loops, number of AST nodes, return statements, loops, comparisons, etc.

The sub-clusters within each high-level cluster are intended to capture the generalization dimension, where the only dif-
We use the adjusted mutual information (AMI) metric to compare TAs’ clusterings with each other and with our k-means clustering. We plan to use k-means again on solutions within each high-level cluster, based on low-level, concrete features. The concrete features for Python programs consist of 48 low-level features that include the number of specific types of operators (add, subtract, etc.), comparisons (<, >, etc.), loops (while or for), library functions, and statements (assignments, conditional, or loop), number of program variables, constant values, etc.

**PRELIMINARY RESULTS**

We use abstract features for k-means clustering of student solutions for the separation dimension, which partitions the solutions into k clusters. We compute clusterings for different k values, and then compare these clusterings to those created by two course teaching assistants (TAs). The TAs were given 50 randomly chosen student solutions as a clustering task. We did not give them specific directions for clustering, in order to better understand how the TAs naturally group solutions. We observed that they ignored low-level features, e.g., they clustered together solutions implementing the same algorithm but using different functions such as pop, list slicing, and delete.

We use the adjusted mutual information (AMI) metric to compare TAs’ clusterings with each other and with our k-means clustering. An AMI value of 0 indicates purely independent clusterings, whereas a value of 1 indicates perfect agreement between the clusterings. The agreement of the two TAs’ clusterings, referred to here as the inter-TA AMI, is only 0.3275. When k was sufficiently high, i.e., at least 15, the k-means-produced clusterings agreed, as measured by AMI, with each TA’s clusterings as much or more than the TAs’ clusterings agreed with each other. We found high agreement between our k-means and TA-produced clusterings on two additional coding assignments as well.

**FUTURE WORK**

We are generalizing this approach to two additional domains. The Mathworks runs an online game, Cody. Users submitted 218,000 Matlab functions as solutions to 1000 or so problems. We hope to categorize software metrics, library functions, and language constructs within Matlab functions as abstract features, differentiating algorithms, or concrete features, distinguishing implementations of the same algorithm. The second domain is code written by MIT students in a hardware description language. Students define their own library of circuits, from which larger circuits are composed. Within each high-level cluster based on overall structure, we could cluster based on low-level library circuit implementation.

**REFERENCES**