Search-based Diverse Sampling from Real-world Software Product Lines

Real-world software product lines (SPLs) often encompass enormous valid configurations that are impossible to enumerate. To understand properties of the space formed by all valid configurations, a feasible way is to select a small, valid and representative sample set. Even though a number of sampling strategies have been proposed, they either fail to produce diverse samples with respect to the number of selected features (an important property to characterize behaviors of configurations),or achieve diverse sampling but with limited scalability (the handleable configuration space size is limited to 1013). To resolve this dilemma, we propose a scalable diverse sampling strategy, which uses a distance metric in combination with the novelty search algorithm to produce diverse samples in an incremental way. The distance metric is carefully designed to measure similarities between configurations, and further diversity of a sample set. The novelty search incrementally improves diversity of samples through the search for novel configurations. We evaluate our sampling algorithm on 39 real-world SPLs. It is able to generate the required number of samples for all the SPLs, including those which cannot be counted by sharpSAT, a state-of-the-art model counting solver. Moreover, it performs better than or at least competitively to some state-of-the-art samplers with respect to the diversity of the sample sets. Our results suggest that only the proposed sampler (among all tested ones) achieves scalable diverse sampling.

Full camera ready version to come soon.