Regression techniques can be used not only for legitimate data analysis, but also to infer private information about individuals. In this paper, we demonstrate that regression trees, a popular data-analysis and data-mining technique, can be used to effectively reveal individuals’ sensitive data. This problem, which we call a “regression attack,” has not been addressed in the data privacy literature, and existing privacy-preserving techniques are not appropriate in coping with this problem. We propose a new approach to counter regression attacks. To protect against privacy disclosure, our approach introduces a novel measure, called digression, which assesses the sensitive value disclosure risk in the process of building a regression tree model. Specifically, we develop an algorithm that uses the measure for pruning the tree to limit disclosure of sensitive data. We also propose a dynamic value-concatenation method for anonymizing data, which better preserves data utility than a user-defined generalization scheme commonly used in existing approaches. Our approach can be used for anonymizing both numeric and categorical data. An experimental study is conducted using real-world financial, economic, and healthcare data. The results of the experiments demonstrate that the proposed approach is very effective in protecting data privacy while preserving data quality for research and analysis.

Keywords: Privacy, data analytics, data mining, regression, regression trees, anonymization