Zusammenfassung

Nach einer Beschreibung der Wichtigkeit von Datenqualität im Kontext von *Data Science* Projekten und maschinellen Lernens, werden Grundlagen zu Datenfehlern und Methoden des maschinellen Lernens erläutert. Anschließend werden aktuelle Fehlerdetektionsmethoden vorgestellt. Keine der beschriebenen Methoden nutzt Datensätze, für welche die Fehler bereits bekannt sind, um Datenfehler in neuen Datensätzen zu detektieren. Die entwickelte Fehlerdetektionsmethode verwendet *Meta Learning* um dies umzusetzen. Durch diese Systematik kann die Laufzeit unter bestimmten Voraussetzungen reduziert werden. In der Evaluation der implementierten Methode werden unter Variation von verwendeten Parameter diese Situationen aufgezeigt. Auswirkungen auf die Leistung der Fehlerdetektion werden ebenfalls dargestellt. Des Weiteren werden die Ergebnisse mit aktuellen Fehlerdetektionsmethoden verglichen. Zum Abschluss folgt eine Zusammenfassung sowie ein Ausblick auf weitere Optimierungen und Forschungsmöglichkeiten.

Schlagworte: Fehlererkennung, Meta Learning, Datenaufbereitung, Datenbereinigung, Clustering

Abstract

Upon a description of the importance of data quality in the context of data science projects and machine learning, basics about data errors and methods of machine learning are explained. Current error detection methods are then presented. None of the described methods use datasets, for which the errors are already known, to detect data errors in new datasets. The developed error detection method uses meta learning to implement this. Through this systematic approach, the runtime can be reduced under certain conditions. In the evaluation of the implemented method these conditions are shown under variation of used parameters. Effects on the error detection performance are also presented. Furthermore, the results are compared with current error detection methods. Conclusively, this report provides a summary of the findings and an outlook on further optimizations and research possibilities.

Keywords: error detection, meta learning, data preparation, data cleaning, clustering