ABSTRACT

Over the last decades many mechanical tasks have been automated to a certain degree by machines. In recent years, the trend has shifted to automate more complex tasks, while imitating the human ability to make intuitive decisions using Machine Learning (ML). Many ML algorithms are known to need a lot of curated and labeled data to be trained. The industry can often easily provide raw data, but lacks the knowledge needed to apply it to ML. This poses the challenge of acquiring relevant data and processing it to a degree where it becomes usable for ML tasks. For the industry, labeling is much harder than just providing the raw data. The reason for this is, that labeling in the industry often requires human experts, who are always scarce and expensive. This gives an incentive to improve the labeling process so that it can be done by non-experts or machines, with minimal expert supervision. This work proposes a concept, namely the Data Refinery (DR), to efficiently label data by exploring and exploiting semantic information, based on Deep Metric Learning (DMeL). This is not achieved without human expertise, but instead with a decreased effort, thus simplifying an quickening the process of acquiring relevant data. For this purpose two DRs were implemented which only differ in their selection strategy. One DR selected data samples to show to a human based on a score (DR-S) and the other randomly without considering the score at all (DR-R). The idea behind the score was to know in advance which data samples are valuable for training. The results indicate that DR-S, in contrast to DR-R, specifically adds new data samples for training in classes where it underperformed. Contrary to the expectations, this did not lead to an overall improvement of DR-S when compared with DR-R, but to a similar performance for a classification task. Regardless, the goal of the case study was not to train a state of the art classification model for a given task, but to efficiently mine data for such a model. Even though the success of the case study is only moderate performance wise, the new approach of the DR as a concept proves to be a promising way to acquire specific data that can be used to create curated data sets. Such curated data sets, are not only valuable for the industry but can be used for all kinds of ML tasks throughout all industries.

ZUSAMMENFASSUNG

In den letzten Jahrzehnten wurden viele mechanische Aufgaben bis zu einem gewissen Grad durch Maschinen automatisiert. In den letzten Jahren hat sich der Trend dahingehend verschoben, komplexere Aufgaben zu automatisieren und dabei die menschliche Fähigkeit, intuitive Entscheidungen zu treffen, mit Machine Learning (ML) zu imitieren. Es ist bekannt, dass viele ML Algorithmen eine Menge kuratierter und gelabelte Daten benötigen, um trainiert zu werden. Die Industrie kann meist leicht Rohdaten zur Verfügung stellen, aber es fehlt ihr an Wissen, um diese auf ML anzuwenden. Daraus ergibt sich die Herausforderung, relevante Daten zu beschaffen und sie so weit zu verarbeiten, dass sie für ML Aufgaben nutzbar werden. Für die Industrie ist das labeln viel schwieriger als die Bereitstellung der Rohdaten. Der Grund dafür ist, dass das labeln in der Industrie oft menschliche Experten erfordert, die immer knapp und teuer sind. Dies gibt einen Anreiz, den Labelings-prozess so zu verbessern, dass er von Nicht-Experten oder Maschinen mit minimaler Expertenaufsicht durchgeführt werden kann. Diese Arbeit schlägt ein Konzept vor, die sogenannte Data Refinery (DR), um Daten effizient zu labeln, indem semantische Informationen erforscht und genutzt werden, basierend auf Deep Metric Learning (DMeL). Dies geschieht nicht ohne menschliches Fachwissen, sondern mit einem verringerten Aufwand, wodurch der Prozess relevante Daten zu beschaffen, vereinfacht und beschleunigt wird. Zu diesem Zweck wurden zwei DR implementiert, die sich nur in ihrer Selektionsstrategie unterscheiden. Die eine DR wählt Datenmuster, die einem Menschen gezeigt werden sollen, anhand eines Scores aus (DR-S), die andere zufällig, ohne den Score überhaupt zu berücksichtigen (DR-R). Die Idee hinter dem Score war, im Voraus zu wissen, welche Datenproben für das Training wertvoll sind. Die Ergebnisse zeigen, dass DR-S im Gegensatz zu DR-R gezielt neue Datenproben für das Training in Klassen hinzufügt, in denen es unterdurchschnittlich abschneidet. Entgegen den Erwartungen führte dies nicht zu einer Gesamtverbesserung von DR-S im Vergleich zu DR-R, sondern zu einer ähnlichen Leistung bei einer Klassifikationsaufgabe. Unabhängig davon war das Ziel der Fallstudie nicht das Trainieren eines State-of-the-Art-Klassifikationsmodells für eine bestimmte Aufgabe, sondern das effiziente Mining von Daten für ein solches Modell. Auch wenn der Erfolg der Fallstudie in Bezug auf die Performance nur mäßig ist, erweist sich der neue Ansatz des DR als Konzept als vielversprechender Weg, um spezifische Daten zu gewinnen, die zur Erstellung kuratierter Datensätze verwendet werden können. Solche kuratierten Datensätze sind nicht nur für die Industrie wertvoll, sondern können für alle Arten von ML Aufgaben in allen Branchen verwendet werden.