Maschinelles Lernen in Datenstrommanagementsystemen

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften

vorgelegt von
Dipl.-Inform. Dennis Geesen

Gutachter:
Prof. Dr. Dr. h.c. H.-Jürgen Appelrath
Prof. Dr. Otthein Herzog
Jun.-Prof. Dr. Daniela Nicklas

Tag der Disputation: 29. November 2013

Extended Abstract
Sowohl im privaten als auch im wirtschaftlichen und öffentlichen Umfeld werden kontinuierliche Sensordaten zusammen mit weiteren Informationen häufig zu Überwachungszwecken verwendet, aber vermehrt auch als Grundlage zum selbstständigen Handeln, sodass Anwendungen situativ über Aktoren reagieren oder sich an bestimmte Situationen anpassen können. Beispielsweise kann Temperatur oder Licht in einem Smart Home automatisch in Abhängigkeit von Strompreis, Wetter und Verhalten der Bewohner geregelt werden, was neben einer Arbeitserleichterung auch zur Senkung des Energiebedarfs führen kann.

Um hierbei mit individuellen Bewohnern, heterogenen Umgebungen oder auch verrauschten Sensordaten umgehen zu können, wird meist eine Sensordatenfusion durchgeführt, deren Ergebnisse als Eingabe für anschließende Verfahren aus dem Machine Learning (ML), wie Clustering, Klassifikation oder Assoziationsanalyse, dienen. Technologisch bieten Datenstrommanagementsysteme (DSMS) eine ideale Grundlage für die Umsetzung, da sie unter anderem eine universelle, flexible, deterministische und optimierbare Sensordatenfusion erlauben. Da ein DSMS nicht über Konzepte zum ML verfügt, werden die entsprechenden Teile der Anwendung typischerweise jedoch monolithisch umgesetzt. So sind diese Teile bspw. nicht universell einsetzbar oder bieten nur wenige Optimierungsmöglichkeiten. Hieraus ergibt sich die Fragestellung, ob Verfahren des ML, ebenso wie die Sensordatenfusion in einem DSMS, universell und flexibel zur Verfügung gestellt werden können und eine integrierte Verarbeitung zusätzliche Optimierungsmöglichkeiten bietet.

Ziel der Arbeit ist die Beantwortung dieser Fragestellung, indem Verfahren des ML in ein DSMS integriert werden, um dadurch eine gemeinsame Verarbeitung und ggf. Optimierungen zu ermöglichen. Die Semantik vorhandener Verarbeitungsschritte wie Selektion oder Verbund werden in einem DSMS über logische Operatoren beschrieben, um dadurch unter anderem deterministische Ergebnisse zu zusichern. Durch Übersetzung in physische Operatoren, welche die eigentlichen Implementierungen beinhalten, können die Operatoren ausgeführt werden. Analog zu diesem Konzept werden in der Arbeit zunächst logische Operatoren zum ML für die formale Beschreibung der Semantik entworfen. Anschließend werden physische Operatoren entwickelt. Die Implementierung ist dabei unabhängig von konkreten ML-Algorithmen, um auf existierende und

etablierte Verfahren zurückgreifen zu können. Sind mehrere logische Operatoren zu einem Plan verknüpft, kann das DSMS semantisch äquivalente Optimierungen durchführen, indem bspw. durch Äquivalenzregeln effizientere Reihenfolgen gewählt werden. Hierzu wurde untersucht, ob es durch die neuen Operatoren zusätzliche Äquivalenzregeln gibt. Durch Anwendung dieser Regeln können dann Anwendungen, die auf eine integrierte Sensordatenfusion und ML beruhen, optimiert werden.

Die Demonstration zur Umsetzbarkeit des Ansatzes erfolgt anhand des Prototypen OdysseusML, indem die Konzepte in das DSMS Odysseus integriert wurden und zwei Anwendungen für ein Smart Home und für eine Windenergieprognose umgesetzt werden. Die Evaluation untersucht ergänzend Effektivität und Effizienz des Ansatzes. Zum einen wird evaluiert, ob die Integration generell ein ML für Datenströme durch DSMS ermöglicht. Zum anderen wird evaluiert, ob die entwickelten Optimierungen die Latenz bzw. den Durchsatz verbessern und wo die Grenzen liegen.