Open-Source-Data-Mining-Werkzeuge im Vergleich

Asa Toure Kwonang

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
asa.toure.kwonang@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Frank Kammer

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
frank.kammer@mni.thm.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Big Data, Data Analytics, Data Mining, Software Tools, Data-Mining Werkzeuge, Daten, Wissen, Informationen, Klassifikation, Entscheidungsbaum

Zusammenfassung

Im Hinblick auf Herausforderungen wie Big Data in der Wirtschaft oder Gensequenzierung in der Biowissenschaft ist Data Mining sowohl für alltägliche Probleme als auch für Spezialgebiete wichtig. Im Bereich der Finanzdatenanalyse wird Data Mining zum Beispiel für die Analyse der Kreditwürdigkeit, die Vorhersage der Wahrscheinlichkeit einer Kreditrückzahlung, die Klassifizierung und das Clustering von Kunden für gezieltes Marketing und die Aufdeckung von Geldwäsche und anderen Finanzdelikten genutzt. Im Marketing wird Data Mining genutzt um große Datenmengen aus den Bereichen Verkauf, Einkaufshistorie, Warentransport, Konsum und Dienstleistungen auszuwerten.

Der Unternehmer kann die gewonnenen Informationen nutzen, um Entscheidungen über die richtigen Strategien und Taktiken zu treffen, die den Bedürfnissen des Kunden entsprechen. Die Fähigkeit von Data Mining, aus riesigen Datenmengen gewonnene Informationen zu liefern, wurde zu einem effektiven Werkzeug für Unternehmen und Einzelpersonen. Mit der zunehmenden Bedeutung dieser Wissenschaft stieg auch die Zahl der Open-Source- Werkzeuge, die zur Umsetzung ihrer Konzepte entwickelt wurden. Die Entwicklung und Anwendung von Data Mining-Algorithmen setzen den Einsatz leistungsfähiger Softwaretools voraus.

Diese Bachelorarbeit gibt einen Überblick über die historische Entwicklung von Data-Mining, die vier wichtigsten Data-Mining-Verfahren und stellt vier der verbreitetsten Open-Source-Data-Mining-Werkzeuge vor. Vorgestellt werden R, RapidMiner, WEKA und KNIME. Obwohl diese Werkzeuge komplexen Algorithmen zur Extraktion von Mustern aus Datenbeständen verwenden, sind sie auch für Nichtfachleute leicht zu bedienen und plattformübergreifend.

Welches Werkzeug die gewünschte Aufgabe besser erfüllt, war nicht einfach zu beurteilen. Hinzukommt, dass man sich nicht nur auf die Beschreibung der jeweiligen Anbieter verlassen kann. Anhand Kriterien wie die genutzte Plattform, Format der zu verarbeitenden Daten, die benötigte Datenvisualisierungsform, die Leistungsfähigkeit und der Absicht, neue Funktionen zu entwickeln, konnte eine Bewertung erfolgen.

Zur Prüfung der Leistung der Werkzeuge, werden in dieser Arbeit drei Klassifizierungsalgorithmen ausgewählt, nämlich der Naive Bayes-, der Entscheidungsbaum- und der K-nächster Nachbar-Algorithmus. Für Testzwecke werden sechs verschiedene Datensätze, die sich in ihrem Bereich, Anzahl der Instanzen, Attribute und Klassen-beschriftungen unterscheiden aus dem UCI Repository verwendet. Das Experiment wurde auf

einem Laptop mit bestimmten Spezifikationen durchgeführt.

Anhand dieser Bewertungen konnte festgestellt werden, dass R eine größere Anzahl von Ein-/ Ausgabeformaten und Visualisierungsarten unterstützt. In Bezug auf die Anwendbarkeit der Klassifikatoren ist WEKA das beste Werkzeug, um die ausgewählten Klassifikatoren auszuführen, gefolgt von R, RapidMiner und schließlich KNIME.

Diese Studie ist zu dem Schluss gekommen, dass kein Werkzeug besser ist als das andere, da die Leistung der Werkzeuge für Klassifikationsaufgabe von der Art des verwendeten Datensatzes und der Art und Weise, wie die Klassifikationsalgorithmen in den Werkzeugen implementiert wurden, beeinflusst wird.

Das es auf dem Markt sehr viele Anbieter von Data-Mining-Werkzeuge gibt, ist die Suche nach einem neuen und vor allem anforderungsgerechten Werkzeug nicht trivial. Bevor man sich für ein Werkzeug entscheidet, sollte man einen Softwareauswahlprozess durchführen. Grundlage hierfür bildet die Strategie, da eine Software, die genau zu dem Unternehmen passt, ein Schlüssel zum Unternehmenserfolg ist.

Literatur

Cleve, Jürgen / Lämmel, Uwe, Data Mining, München 2014

Fayyad, Usama / Piatetsky-Shapiro, Gregory / Smyth, Padhraic, From Data Mining to Knowledge Discovery in Databases, AI Mag. 1996, 37–37

Bissantz, Nicolas / Hagedorn, Jürgen, Data Mining (Datenmustererkennung), in: WIRTSCHAFTS-INFORMATIK 2009, 139–144

Jovic, A. / Brkic, K. / Bogunovic, N., An overview of free software tools for general data mining, 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) 2014, 1112–1117

Wuttke, Laurenz, Data Mining: Algorithmen, Definition, Methoden und Anwendungsbeispiele, https://datasolut.com/was-ist-data-mining/

o.V. Klassifikationsverfahren im Data Mining | Data Mining Grundlagen, https://www.daten-banken-verstehen.de/business-intelligence/data-mining-grundlagen/klassifikationsverfahren/