

Sebastian Wuttke

Optimierung von aktiven Klassifikatoren zur Reduktion des Ressourcenaufwands in der Trainingsphase

Heutzutage ist maschinelles Lernen in vielen Bereichen der Fernerkundung im Einsatz. Einen großen Teil der dort behandelten Aufgabenstellungen lösen Methoden der überwachten Klassifikation. Dabei wird ein Klassifikator mit Hilfe von Beispielen trainiert, die bereits mit dem korrekten Klassifikationsergebnis versehen sind (mit der sogenannten Ground Truth). Meist ist das Beschaffen dieser Ground Truth sehr ressourcenaufwendig, aber unabdingbar.

Die aktive Klassifikation stellt eine Möglichkeit dar, die benötigten Ressourcen zu verringern. Ein aktiver Klassifikator ist eine Kombination aus einer Selektionsstrategie und einem überwachten Klassifikationsalgorithmus. Die Selektionsstrategie wählt aus den noch unmarkierten Beispielen eine kleine Auswahl und nur von dieser wird die Ground Truth beschafft. Anschließend kann der Klassifikator wie gewohnt trainiert werden. Im Gegensatz zu einer zufälligen (passiven) Auswahl reduziert eine passende aktive Selektionsstrategie den Ressourceneinsatz für das Training des Klassifikators ohne die Klassifikationsgüte zu verschlechtern. Auf diese Art wird trotz geringerem Ressourceneinsatz ein größerer Trainingserfolg erreicht.

In den letzten zehn Jahren wurden verschiedene aktive Klassifikatoren entwickelt. Jedoch ist sowohl die Auswahl der anzuwendenden Selektionsstrategie, als auch des zu verwendenden Klassifikationsalgorithmus oft unklar. Um eine effiziente Kombination zu bestimmen, wird heutzutage meist die Ground Truth für einige Beispiele bestimmt und anschließend verschiedene Kombinationen ausprobiert. Dieses Vorgehen basiert auf mangelnder Kenntnis über das Zusammenspiel zwischen Selektionsstrategie und Klassifikator in Abhängigkeit von den neuen noch unmarkierten Daten.

In dieser Arbeit soll ein Verfahren zur Auswahl des aktiven Klassifikators entwickelt und optimiert werden, um so schneller eine effiziente Kombination für einen neuen Datensatz zu finden. Dazu soll zunächst ein flexibles Framework entworfen werden, welches die Bewertung von verschiedenen Kombinationen aus Datensatz, Selektionsstrategie und Klassifikator einfach ermöglicht. So soll vor allem das Zusammenspiel zwischen Klassifikator und Selektionsstrategie untersucht werden. Die Untersuchung soll dabei datengetrieben geschehen, das heißt basierend auf den zu untersuchenden Datensätzen. Dies soll zeigen, ob Datensätze spezielle Strukturmerkmale aufweisen müssen, um mit bestimmten Klassifikatoren eine hohe Klassifikationsgüte erreichen zu können.

Ziel dieser Arbeit ist die Auswahl der Kombination aus Selektionsstrategie und Klassifikator zu vereinfachen und so den Ressourcenaufwand für das Training zu reduzieren. Die erwarteten Ergebnisse sollen

- das Verständnis des Zusammenspiels zwischen Selektionsstrategie und Klassifikationsalgorithmus verbessern.
- die Vorhersage des Verhaltens und die Bewertung von aktiven Klassifikatoren basierend auf den noch unmarkierten Daten ermöglichen.
- die Auswahl eines aktiven Klassifikators vereinfachen, ohne dass Ressourcenaufwand für das Beschaffen der Ground Truth erforderlich ist.