Zusammenfassung

Die Fähigkeit aus Erfahrung zu lernen ist ein faszinierendes und charakteristisches Kennzeichen intelligenten Lebens. Im Verlauf der Evolution sind immer fortschrittlichere Strategien zur Verhaltensanpassung an die speziellen Umgebungsbedingungen entstanden. Durch die Effizienz von natürlichen Lernvorgängen motiviert, betrachtet das überwachte maschinelle Lernen ein formales Modell des Erlernens von spezifischen Eingabe/Ausgabe-Abbildungen und übernimmt das Paradigma des *Lernens aus Beispielen* um Funktionen zu induzieren, welche die Zielobjekte zu gegebenen Merkmalsvektoren vorhersagen können. Im konventionellen maschinellen Lernen ist der Lernprozess durch einen einseitig ausgerichteten Informationsfluss zwischen den beiden idealisierten Protagonisten charakterisiert. Hierbei werden Informationen nur vom Lehrer zum Schüler übermittelt. Im Gegensatz hierzu basieren natürliche Lernvorgänge typischerweise auf komplexen Interaktionsprozessen zwischen dem Schüler und dem Lehrer. Bidirektionale Kommunikation - vom Schüler zum Lehrer und in entgegengesetzter Richtung - stellt eine fundamentale Komponente des Lernens dar. Wie durch die Natur motiviert, substituiert man im aktiven maschinellen Lernen das konventionelle passive Modell durch ein erweitertes Modell, welches eine einfache Form von Interaktion ermöglicht. Die Zielsetzung ist hierbei, den künstlichen Lernprozess zu beschleunigen. Wir untersuchen das Konzept des aktiven Lernens, welches den Lernprozess durch eine Reduzierung der Anzahl an repräsentativen Beispielen effizienter gestaltet, im Forschungsfeld der kernbasierten Maschinen für verschiedene Kategorien von Lernproblemen.

Ein umfassender Korpus an flexiblen Techniken zur Konstruktion von Vorhersagesystemen mit exzellenten Generalisierungsfähigkeiten ist als konstruktives Ergebnis aus der inspirierenden Synergie von statistischer Lerntheorie, Optimierungstheorie und anwendungsorientierteren Forschungsbereichen des maschinellen Lernens in den vergangenen Jahren entstanden. Unter der Vielfalt der entwickelten Methoden gehören Support Vektor Maschinen zu den bekanntesten. Mit Zunahme der heute zur Verfügung stehenden Rechenkapazitäten sind optimierte Trainingsalgorithmen in der Lage, Probleme von großem Umfang zu bewältigen. Jedoch lösen Fortschritte bei der Rechengeschwindigkeit und effizientere Trainingsalgorithmen nicht das inhärente Problem, dass konventionelles überwachtes maschinelles Lernen auf eine Menge von Merkmalsvektoren respektive Beispielen angewiesen ist, die den korrespondierenden Zielobjekten zugeordnet werden müssen. In vielen Anwendungsbereichen kann die Aufgabe der Zuordnung von Zielobjekten nicht automatisch durchgeführt werden, sondern beruht auf zeit- und kostenaufwändigen Ressourcen, wie zum Beispiel komplexen Experimenten oder von Menschen zu treffenden Entscheidungen. Somit berücksich-

tigt die zu Grunde liegende Annahme, dass eine den Zielobjekten zugeordnete Menge von Beispielen dem Lernalgorithmus zur Verfügung gestellt wird, den in vielen Fällen notwendigen Zuordnungsaufwand nicht.

Der Oberbegriff des aktiven Lernens umfasst verschiedene Ansätze, die darauf ausgerichtet sind, den Zuordnungsaufwand beim überwachten maschinellen Lernen zu reduzieren. Wir betrachten das so genannte poolbasierte Modell des aktiven Lernens. Der Grundgedanke hinter diesem Modell ist es, aus einer vorgegebenen endlichen Menge viel versprechende noch nicht zugeordnete Beispiele in dem Sinne auszuwählen, dass die zugehörigen Zielobjekte zu einer Verbesserung der Genauigkeit der Vorhersagefunktion beitragen. Im Gegensatz zum konventionellen überwachten Lernen betrachtet man im poolbasierten Modell des aktiven Lernens ein erweitertes Lernmodell, in dem dem Lernalgorithmus Zugang zu einer Menge zunächst nicht zugeordneter Beispiele gewährt wird und er mit der Möglichkeit ausgestattet wird, über die Reihenfolge des Zuordnungsprozesses zu entscheiden. Hierbei steht die Zielsetzung im Mittelpunkt ein hohes Maß an Genauigkeit zu erreichen, ohne die Gesamtmenge an zugehörigen Zielobjekten anzufordern.

Die vorliegende Dissertation verfolgt als generelle Zielsetzung die Verbesserung und Verallgemeinerung bestehender Ansätze im poolbasierten aktiven Lernen mit kernbasierten Maschinen auf weitere Kategorien von Lernproblemen und die eingehende Analyse zu Grunde liegender theoretischer Gesichtspunkte. Die wesentlichen Beiträge können wie folgt zusammengefasst werden:

- *Effizienzverbesserung:* Im Bereich des binären Klassifikationslernens stellen wir eine Strategie zur Auswahl einer Menge von Beispielen vor. Als grundlegende Komponente berücksichtigt diese Strategie ein Diversifitätsmaß, um eine effizientere Strategie bezüglich der Anzahl der zum Erreichen eines bestimmten Genauigkeitsniveaus notwendigen zugeordneten Beispiele zur Verfügung zu stellen. Darüber hinaus entwickeln wir modifizierte Auswahlstrategien für Klassifikationsprobleme mit einer beliebigen Klassenanzahl auf Basis verschiedener binärer Dekompositionsverfahren. Diese Strategien erzielen substantielle Verbesserungen gegenüber vorhergehenden Ergebnissen.

- *Verallgemeinerung:* Das Erlernen von Rankfunktionen gehört zur Oberkategorie des Präferenzlernens und wurde im Forschungsbereich des aktiven Lernens noch nicht untersucht, obwohl der Zuordnungsaufwand einen Gesichtspunkt von noch größerer Relevanz darstellt als beim Klassifikationslernen. Auf Basis des nebenbedingungsbeschränkten Klassifizierens und einer paarbasierten Methode stellen wir Verallgemeinerungen des poolbasierten aktiven Lernens vor und erzielen substantielle Verbesserungen beim Lernvorgang.

- *Theoretische Analyse:* Wir betrachten ein lineares Lernszenario, um Schwachpunkte volumenbasierter Auswahlstrategien aus theoretischer Sicht zu untersuchen und zeigen, dass die Minimierung des Volumens des Versionsraumes eine notwendige Voraussetzung für die Minimierung des vorgestellten verbesserten Auswahlkriteriums darstellt. Darüber hinaus beweisen wir ein neues Konvergenztheorem für die volumenbasierte SIMPLE Auswahlstrategie für den Fall der Approximation mittels einer Kugel mit maximalem Radius.

Hierzu stellen wir einen generellen Überblick über das Konzept des überwachten Lernens dar und formalisieren das betrachtete Lernmodell, um eine allgemeine Notationsbasis zu schaffen. Der Schwerpunkt wird auf die Hypothesenklasse der linearen Klassifikatoren gelegt, die auf elegante Art und Weise durch das flexible Konzept der Kerne erweitert werden kann. Unter der Vielfalt der kernbasierten Maschinen stellen Support Vektor Maschinen und Bayes-Punkt Maschinen eingehend untersuchte Beispiele für lineare Lernalgorithmen dar, die im so genannten Versionsraum-Modell eine besondere geometrische Interpretation besitzen. Darüber hinaus dienen uns lineare Klassifikatoren als grundlegende Komponenten, um komplexere Problemkategorien des überwachten Lernens zu bearbeiten. Wir stellen Veränderungen für einen kernbasierten Billard-Algorithmus zur Approximation des Bayes-Punktes vor, um dessen numerische Stabilität und Genauigkeit zu verbessern.

Die Klasse aktiver Lernstrategien, die darauf abzielt, das Volumen des Versionsraumes durch Approximation des Massenzentrums zu reduzieren, wird detailliert erörtert und die zu Grunde liegenden theoretischen Überlegungen werden erläutert. Darüber hinaus analysieren wir ein lineares Lernszenario, in dem volumenbasierte Strategien gewisse Schwachpunkte offenbaren und entwickeln eine verbesserte binäre Auswahlstrategie zur Bewältigung dieses Problems. Zur Reduzierung der Berechnungskomplexität der vorgeschlagenen Auswahlstrategie diskutieren wir eine fortschrittliche Vorauswahltechnik, die eine Teilmenge der noch nicht zugeordneten Beispiele gemäß einer weniger aufwändigen volumenbasierten Strategie als Ausgangsbasis für unsere neu entwickelte Auswahlstrategie bestimmt. Aus theoretischer Sicht stellt die Minimierung des Volumens des Versionsraumes eine *notwendige* Voraussetzung für die Minimierung des verbesserten Auswahlkriteriums dar. Wie durch die theoretische Analyse begründet, zeigen die experimentellen Ergebnisse, dass der zweischichtige Vorauswahlansatz die Berechnungskomplexität substantiell verringern kann, ohne dass hierbei die Klassifizierungsgenauigkeit reduziert wird. Unabhängig von praktischen Gesichtspunkten beleuchtet das betrachtete Szenario potentielle Schwachpunkte volumenbasierter Strategien aus theoretischer Perspektive.

Der grundlegende Ablauf des poolbasierten aktiven Lernens ist durch den sequentiellen Prozess des Auswählens von *einzelnen* Beispielen aus der Menge der nicht zugeordneten Beispiele und des Anforderns der zugehörigen Zielobjekte charakterisiert. Jedoch ist es sowohl mit Blick auf den Berechnungsaufwand, als auch im Hinblick auf häufig gegebene Charakteristika von Lernproblemen in der Praxis vorteilhaft, dieses Schema insoweit zu verallgemeinern, als dass *Mengen* von noch nicht zugeordneten Beispielen ausgewählt werden. Wir entwickeln eine verallgemeinerte Auswahlstrategie zur aktiven Auswahl von Mengen mehrerer Beispiele im Bereich des binären Klassifikationslernens, die ein Diversifitätsmaß mit einbezieht, und präsentieren experimentelle Ergebnisse, die zeigen, dass dieser neue Ansatz eine effizientere Methode bezüglich der Anzahl des zum Erreichen eines bestimmten Genauigkeitsniveaus notwendigen zugeordneten Beispiele als eine häufig verwendete Erweiterung einer volumenbasierten Auswahlstrategie darstellt.

Während der wesentliche Teil der Forschung im Bereich des aktiven Lernens mit kernbasierten Maschinen auf dem Gebiet der binären Klassifikationsprobleme geleistet wurde, ist dem Problem des Klassifikationslernens im Falle mehrerer Klassen weni-

ger Aufmerksamkeit gewidmet worden. Wir betrachten drei häufig verwendete Dekompositionsmethoden zur Formulierung von Multiklassenproblemen als Mengen von binären Klassifikationsproblemen und entwickeln neue aktive Auswahlstrategien, um den Zuordnungsaufwand zu reduzieren. Eine Vielzahl von Experimenten, die auf praxisorientierten Datensätzen durchgeführt wurden, demonstrieren die Vorteile unseres Ansatzes im Vergleich zu vorhergehenden Ergebnissen.

Der Aufwand, der in einem überwachten Lernszenario notwendig ist, um Mengen von den Zielobjekten zugeordneten Beispielen zu erstellen, wird oft vernachlässigt, obwohl dieser in vielen Anwendungen einen zeitaufwändigen und kostenintensiven Arbeitsschritt darstellt. Während der Zuordnungsprozess bereits beim Lernen von Klassifikationsfunktionen einen wesentlichen Gesichtspunkt darstellt, wird dieser zu einem Gegenstand von noch substantiellerer Relevanz im Bereich des Lernens von Rankfunktionen, das sich mit totalen Ordnungen über einer vorgegebenen Menge von Alternativen als komplexerem Raum für Zielobjekte beschäftigt. Wir stellen eine neue Verallgemeinerung des poolbasierten aktiven Lernens zur Reduzierung des Zuordnungsaufwandes vor, die sowohl auf dem Ansatz des paarbasierten als auch dem Ansatz des nebenbedingungsbeschränkten Klassifizierens zur Darstellung von Rankfunktionen basiert.

Der abschließende Teil dieser Dissertation beschäftigt sich mit der eingehenden Analyse volumenbasierter Auswahlstrategien im Bereich des binären Klassifikationslernens. Wir leiten ein Konvergenztheorem für eine häufig verwendete volumenbasierte Auswahlstrategie für den Fall der Approximation des Massenzentrums durch eine Kugel mit maximalem Radius her und präsentieren einen Gesamtüberblick weiterer Ergebnisse dieses Themenkomplexes.