Aktives Lernen für Klassifikationsprobleme unter der Nutzung von Strukturinformationen

Tobias Reitmaier[1]

Abstract: Heutzutage werden mediale, kommerzielle und auch persönliche Inhalte immer mehr in der digitalen Welt konsumiert, ausgetauscht und somit gespeichert. Diese Daten versuchen IT-Unternehmen mittels Methoden des Data Mining oder des maschinellen Lernens verstärkt wirtschaftlich zu nutzen, wobei in der Regel eine zeit- und kostenintensive Kategorisierung bzw. Klassifikation dieser Daten stattfindet. Ein effizienter Ansatz, diese Kosten zu senken, ist aktives Lernen (AL), da AL den Trainingsprozess eines Klassifikators durch gezieltes Anfragen einzelner Datenpunkte steuert, die daraufhin durch Experten mit einer Klassenzugehörigkeit versehen werden. Jedoch zeigt eine Analyse aktueller Verfahren, dass AL nach wie vor Defizite aufweist. Insbesondere wird Strukturinformation, die durch die räumliche Anordnung der (un-)gelabelten Daten gegeben ist, unzureichend genutzt. Außerdem wird bei vielen bisherigen AL-Techniken noch zu wenig auf ihre praktische Einsatzfähigkeit geachtet. Um diesen Herausforderungen zu begegnen, werden in der diesem Beitrag zugrundeliegenden Dissertation mehrere aufeinander aufbauende Lösungsansätze präsentiert: Zunächst wird mit probabilistischen, generativen Modellen die Struktur der Daten erfasst und die selbstadaptive, (fast) parameterfreie Selektionsstrategie 4DS (Distance-Density-Distribution-Diversity Sampling) entwickelt, die zur Musterauswahl Strukturinformation nutzt. Anschließend wird der AL-Prozess um einem transduktiven Lernprozess erweitert, um die Datenmodellierung während des Lernvorgangs anhand der bekanntwerdenden Klasseninformationen iterativ zu verfeinern. Darauf aufbauend wird für das AL-Training einer Support Vector Machine (SVM) der neue datenabhängige Kernel RWM (Responsibility Weighted Mahalanobis) definiert.

1 Einführung

In unseren heutigen Informationsgesellschaft wächst die Anzahl der (mobilen) Internet-Teilnehmer und das durch sie erzeugte Datenvolumen rasant. Aktuelle Studien belegen, dass fast 40% der Weltbevölkerung einen Internetzugang besitzen, wobei die durch Social Media, mobile Apps und Clouds erzeugten Daten im Jahr 2019 ein Volumen von ca. 10,4 Zettabytes erreichen werden. Diese riesigen Datenmengen versuchen IT-Unternehmen wie beispielsweise Amazon, Google oder Yahoo wirtschaftlich zu nutzen, z. B. für Empfehlungsdienste oder personalisierte Werbung. Hierbei stellen diese Daten aufgrund ihres enormen Volumens, ihrer Heterogenität und ihrer Schnelllebigkeit eine große Herausforderung dar, die unter dem Begriff *Big Data* zusammengefasst wird. Aber auch aktuelle Bestrebungen wie *Internet of Things* (IoT), *Machine-to-Machine* (M2M) oder *Industrie 4.0* führen zu einem weiteren Kommunikations- und Datenanstieg.

Neue technische Entwicklungen (wie Hadoop, Spark oder Storm) ermöglichen es immer besser, die resultierenden Daten zu strukturieren, sodass das nächste erklärte Ziel darin besteht, die richtigen Fragen zu stellen, um Erkenntnisse bzw. Zusammenhänge aus diesen

[1] Universität Kassel, Fachbereich Elektrotechnik/Informatik, Fachgebiet Intelligente Eingebettete Systeme, tobias.reitmaier@uni-kassel.de

Daten zu extrahieren. Hierzu kommen statistische Verfahren, Optimierungsalgorithmen und Methoden des Data Mining oder des maschinellen Lernens zum Einsatz. Maschinelles Lernen verfolgt i. A. das Ziel der *Generalisierung*, d. h., ausgehend von Beispieldaten Muster oder Gesetzmäßigkeiten einer Problemstellung zu erlernen bzw. zu *verallgemeinern*, um somit auch unbekannte Daten der gleichen Problemstellung beurteilen zu können. Sind Beispieldaten bereits mit entsprechenden *Zielwerten* (Klassenzugehörigkeiten, Labels) versehen, so kann auf Basis dieser Daten ein *Lernmodell* (Klassifikator) erzeugt bzw. trainiert werden, das die Eingabedaten (Trainingsdaten) geeignet auf die gegebenen Zielwerte abbildet. In den meisten Fällen jedoch stehen die Daten ungelabelt oder nur teilgelabelt zur Verfügung, wodurch im Bereich des maschinellen Lernens zwischen *überwachtem*, *unüberwachtem* und *halbüberwachtem* Lernen unterschieden wird.

Sei ein Klassifikationsproblem gegeben, so hängt die Güte des resultierenden Klassifikators hauptsächlich von den gelabelten Trainingsdaten ab. Weitere Einflussfaktoren sind der verwendete Klassifikatortyp (Modellierung), der Lernalgorithmus und auch die gewählte Parametrierung. Heute ist es oft einfach, große Mengen ungelabelter Daten zu sammeln; doch diese mit den passenden Klassen zu versehen, ist meist schwierig und mit hohen zeitlichen und finanziellen Kosten verbunden. Eine effiziente Lösung um diese Kosten zu senken, stellt *aktives Lernen* (AL) dar, da AL durch Stellen der „richtigen“ Fragen die Musterauswahl und somit das Training eines Klassifikators gezielt steuert.

Was ist AL? AL basiert auf der Annahme, dass wenn es dem Lernmodell erlaubt ist, die Daten zu wählen, von denen es lernen will, d. h., aktiv, neugierig und erkundend zu sein, so erzielt dieses trotz einer kleineren Anzahl an Trainingsdaten eine gleichbleibende oder verbesserte Leistung [Se12]. Hierzu nimmt AL an, dass zu Beginn des Lernprozesses eine Menge (Pool) *U ungelabelter Muster* (Instanzen oder Beobachtungen) und eine (meist) kleine Menge *L gelabelter Daten* mit $|L| \ll |U|$ zur Verfügung stehen. Daraufhin erhöht AL iterativ die Anzahl der gelabelten Muster, wobei in jeder *Anfragerunde i* eine *Anfragemenge* $\mathcal{S}_i \subseteq U_{i-1}$ hoch informativer[2] Muster mit Hilfe einer *Selektionsstrategie* $\mathcal{Q}$ unter Berücksichtigung des aktuellen „Wissensstands“ des aktiv zu trainierenden *Klassifikators* $\mathbf{G}_{i-1}$ ausgewählt wird. Die Muster der Menge $\mathcal{S}_i$ werden daraufhin durch ein *Orakel* $\mathcal{O}$ (bzw. einen menschlichen Experten) mit einer Klasse $c \in C$ versehen (gelabelt), von U_{i-1} entfernt und zur Trainingsmenge L_{i-1} hinzugefügt. Anhand von $L_i = \mathcal{S}_i \cup L_{i-1}$ wird $\mathbf{G}_{i-1}$ „neu“ trainiert. Dies wird so oft wiederholt, bis ein zuvor festgelegtes Abbruchkriterium, wie z. B. eine bestimmte Anzahl an Anfragerunden, erfüllt ist. Im Bereich des AL stellen *poolbasiertes aktives Lernen* (PAL), *streambasiertes aktives Lernen* (SAL) und *Membership Query Learning* (MQL), die wichtigsten Lernszenarien dar, wobei die beiden zuletzt genannten Szenarien nicht weiter betrachtet werden.

Wer nutzt AL und für welche Problemstellungen? Anhand vieler aktueller Veröffentlichungen ist das wachsende Interesse namhafter Unternehmen wie AT&T, IBM, Microsoft, Mitsubishi oder Yahoo an AL zu erkennen, die bereits heute AL in verschiedensten Anwendungsgebieten wie z. B. Bildklassifikation, Spracherkennung oder Wirkstoffdesign erfolgreich einsetzen. Aber auch für IoT, M2M und Industrie 4.0 wird AL eine wichtige

[2] Ein Muster wird als *informativ* angesehen, wenn es mit großer Wahrscheinlichkeit die Güte des Lernmodells erhöht, sobald dieses gelabelt für dessen Training berücksichtigt wird.

Rolle spielen, da im Zuge dieser Entwicklungen immer mehr technische Systeme miteinander verknüpft werden, wobei diese unterschiedlichste Quellen, wie Sensoren, das Internet, Datenbanken, etc., als Informationslieferanten nutzen werden, um autonom zu lernen, sich zu konfigurieren oder an Ihre Umgebung anzupassen. Die Relevanz von AL in der Forschung ist auch an verschiedenen aktuellen DFG-Schwerpunktprogrammen zu erkennen, wie Organic Computing, Autonomous Learning oder Kooperativ Interagierender Automobile, die zur Erhöhung der Autonomie technischer Lernsystemen u. a. AL nutzen.

Eine Analyse aktueller Verfahren (Übersichten siehe [RS13, RCS14, Se12]) zeigt, dass AL nach wie vor diverse Defizite aufweist und hinsichtlich der zu erzielenden Klassifikationsgüte und der praktischen Einsetzbarkeit weiterhin Verbesserungspotential besitzt. Insbesondere wird diejenige Information (Strukturinformation), die durch die räumliche Anordnung der (ungelabelten) Daten im Eingaberaum eines Klassifikators gegeben ist, unzureichend berücksichtigt. Außerdem wird bei vielen bisherigen Ansätze noch zu wenig auf eine leichte Handhabung im praktischen Einsatz geachtet (wie auf eine möglichst kleine Anzahl initial gelabelter Muster oder wenig einzustellende Parameter). Um diesen Herausforderungen zu begegnen, werden in der diesem Beitrag zugrundeliegenden Dissertation mehrere Ziele verfolgt und jeweils neuartige, aufeinander aufbauende Lösungsansätze präsentiert. Das dabei angestrebte Ziel besteht in der Entwicklung eines effektiveren, effizienteren und praxistauglicheren AL-Ansatzes.

2 Ziele, Lösungsansätze und innovative Aspekte

In jedem der folgenden Abschnitte werden zunächst die gemachten Annahmen, das zu erreichende Ziel und daraufhin dessen entsprechender Lösungsansatz beschrieben.

Nutzung von Strukturinformation zur aktiven Musterauswahl: Maschinelles Lernen ist grundsätzlich anwendbar, wenn „Strukturen“ oder „Regelmäßigkeiten“ in einer Menge von Beispieldaten erkannt und ausgenutzt werden können. Für Klassifikationsprobleme bedeutet dies, dass die Daten meist Cluster mit willkürlichen Formen bilden, wobei derartige Strukturen für AL zu erfassen bzw. modellieren sind, um sie für die aktive Musterauswahl und damit für das aktive Training eines Klassifikators zu berücksichtigen. In einer realen Anwendung muss ein AL-Ansatz „from scratch“ starten können, d. h., ohne jegliche Klasseninformation ($L_0 = \emptyset$), sodass die räumliche Anordnung der Daten nur mit Hilfe eines unüberwachten Verfahrens modelliert werden kann. Zudem gilt für einen erfolgreichen Einsatz von AL, dass je „einfacher“ die Selektionsstrategie $\mathcal{Q}$ ist, desto mehr initial gelabelte Muster sollten verteilt im Eingaberaum zur Verfügung stehen, da L_0 dem „initialen Wissen“ des Klassifikators $\mathbf{G}$ entspricht und folglich einen großen Einfluss auf den aktiven Lernverlauf von $\mathbf{G}$ besitzt. In der Literatur wird dieses Problem jedoch kaum beachtet [Re15], sodass viele AL-Ansätze von einer unrealistisch großen Anzahl initial gelabelter Muster ausgehen. Überdies zeigt die aktuelle Forschung, dass eine effiziente Selektionsstrategie einen Kompromiss zwischen Exploration und Exploitation eingehen muss. Zum Beispiel sollten zu Beginn des AL-Prozesses Muster aus allen Regionen des Eingaberaums gewählt werden, in denen sich Daten befinden (Explorationsphase), und zum Ende sollte durch die Auswahl von Mustern nahe der Entscheidungsgrenze des Klas-

sifikators diese feinabgestimmt werden (Exploitationsphase). Somit ist das „Stellen der richtigen Fragen“, d. h. die Auswahl informativer Muster, ein facettenreiches Problem, das i. A. zu Mehr-Kriterien-Strategien führt, die in der Praxis kaum oder nur bedingt einsetzbar sind. Der Hauptgrund hierfür ist, dass derartige Selektionsstrategien meist viele Parametern besitzen, die in der Praxis nicht akkurat einzustellen sind, da im praktischen Einsatz von AL nur ein Versuch möglich ist. Das erste Ziel besteht folglich darin, (1) die Struktur der Daten hinreichend zu modellieren, (2) den AL-Prozess ohne initial gelabelten Muster zu starten ($L_0 = \emptyset$) und (3) eine selbstadaptive, (möglichst) parameterfreie Mehr-Kriterien-Strategie zu entwickeln, die Muster unter Berücksichtigung der Datenstruktur wählt.

In einem ersten Schritt wurden in der Dissertation probabilistische, generative Modelle genutzt, um die Struktur der ungelabelten Daten zu erfassen. Dabei handelt es sich um probabilistische Mischmodelle, die unüberwacht (d. h. ohne Klasseninformationen) mit Hilfe von *Variationaler Bayes'scher Inferenz* (VI) *einmalig* vor Beginn des AL-Prozesses (offline) geschätzt werden. Wurde das Dichtemodell $\mathbf{M}$ bestimmt, kann dieses anhand gelabelter Muster, von denen während des AL-Prozesses eine immer größere Anzahl zur Verfügung steht, zu einem Klassifikator (CMM: Classifier based on Mixture Modells) erweitert werden. Der CMM-Klassifikator ordnet hierbei den Komponenten von $\mathbf{M}$ die Klasseninformationen graduell zu. Dies bedeutet, dass sich die Modellkomponenten die Klassen „teilen“ (share), wodurch dieser Klassifikator auch als CMM_{sha} bezeichnet wird. Wird der AL-Prozess ohne jegliche Klasseninformation gestartet ($L_0 = \emptyset$), so steht in der ersten Anfragerunde kein trainierter Klassifikator $\mathbf{G}_0$ zur Verfügung, dessen Wissen von der Selektionsstrategie $\mathcal{Q}$ berücksichtigt werden kann, um Muster aktiv zu wählen. Daher wurde der „konventionelle“ PAL-Zyklus erweitert, sodass in der ersten Anfragerunde mittels eines dichtebasierten Ansatzes (basierend auf $\mathbf{M}$) Muster aus Regionen des Eingaberaums mit hoher Dichte gewählt werden. Derartige Muster befinden sich i. A. in der Nähe der Zentren der Modellkomponenten und stellen daher Prototypen für ihre Nachbarmuster dar. Des Weiteren wurde in der Dissertation eine neue Selektionsstrategie namens *Distance-Density-Diversity-Distribution Sampling* (4DS) entwickelt, die vier verschiedene Selektionskriterien für ihre aktive Musterauswahl berücksichtigt. Drei dieser Kriterien (Distanz, Dichte und Verteilung) werden von 4DS *selbstadaptiv*, d. h., während des AL-Prozesses in Abhängigkeit des zu trainierenden Klassifikators gewichtet, um einen guten Kompromiss zwischen einer Exploration und Exploitation zu finden. Außerdem ist 4DS *parameterfrei*, falls auf die Auswahl mehrerer Muster pro Anfragezyklus verzichtet werden kann (Diversität). Ein weiterer innovativer Aspekt der Strategie 4DS ist, dass 4DS mittels des Verteilungskriteriums Muster derart aktiv auswählen kann, dass aus jedem Anfragezyklus eine gelabelte Mustermenge resultiert, deren Klassenverteilung die „wahre“, unbekannte Klassenverteilung der Gesamtdaten bestmöglich approximiert.

Aktive Verfeinerung der Strukturinformation: Bisher wurde angenommen, dass das Mischmodell $\mathbf{M}$ einmalig vor Beginn des AL-Prozesses nur auf Basis der ungelabelten Daten unüberwacht bestimmt wird. Demzufolge wird jegliche Klasseninformation, die während des AL-Prozesses zur Verfügung steht, nicht für eine Feinabstimmung von $\mathbf{M}$ verwendet. Allerdings kann unter Umständen nur anhand der Klassenlabels erkannt werden, ob datengenerierende Prozesse unterschiedlicher Klassen Muster erzeugen, die im Eingaberaum sich überlappende Cluster bilden. Angenommen, es stünden für alle Trai-

ningsmuster ihre Klassenlabels zur Verfügung, so könnten anhand eines überwachten Modellierungsansatzes *separate* Mischmodelle für jede Klasse $c \in C$ geschätzt werden, die daraufhin zu einem Mischmodell kombiniert werden. Diese Vorgehensweise führt zu einem Klassifikator, der CMM_{sep} genannt wird. Dabei ist anzunehmen, dass dieser Klassifikator aufgrund der überwachten Modellierung eine höhere Klassifikationsgüte erzielt als der Klassifikator CMM_{sha}. Jedoch stehen in einem AL-Prozess initial keine oder bestenfalls nur sehr wenige, gelabelte Muster zur Verfügung, sodass das zweite Ziel darin besteht, die Strukturinformationen anhand der Klasseninformationen, die im Verlauf des AL-Prozesses bekannt werden, iterativ zu verfeinern.

Für die Schätzung des Mischmodells, das dem Klassifikator CMM_{sep} zugrunde liegt, müssen bereits alle Muster der Problemstellung gelabelt sein, da anderenfalls die Varianzen der Modellkomponenten unterschätzt werden. Aufgrund dessen kann ein CMM_{sep}-Klassifikator *nicht direkt* aktiv trainiert werden. Um dennoch den Klassifikator CMM_{sep} aktiv zu trainieren, wird in der vorliegenden Arbeit der AL-Prozess um einen *transduktiven Lernprozess* erweitert. Dieser transduktive Lernprozess nutzt einerseits einen zusätzlichen CMM_{sha}-Klassifikator, um in jeder Iteration i des AL-Prozesses alle noch ungelabelten Muster transduktiv (kostenfrei) zu klassifizieren. Folglich steht nach jeder Anfragerunde eine vollständig gelabelte Mustermenge $(U_i \cup L_i)$ zur Verfügung, auf der der Klassifikator CMM_{sep} „aktiv“ trainiert werden kann. Andererseits adaptiert der transduktive Lernprozess auf Basis der angefragten, gelabelten Muster das zugrunde liegende Mischmodell des CMM_{sha}-Klassifikators iterativ durch lokale Änderungen mit dem Ziel, Modellkomponenten zu erhalten, die möglichst nur Muster modellieren, die derselben Klasse angehören. Folglich überführt der transduktive Lernprozess den zusätzlichen CMM_{sha}-Klassifikator iterativ in Richtung eines CMM_{sep}-Klassifikators. Die innovativen Aspekte dieses erweiterten AL-Prozesses sind, dass er eine Adaption der Datenmodellierung auf Basis der bekannt werdenden Klasseninformationen ermöglicht, vollständig probabilistisch ist und prinzipiell mit jeder beliebigen Selektionsstrategie kombiniert werden kann. Zudem kann anstelle des Klassifikators CMM_{sep} auch jeder beliebige andere Klassifikator mit diesem neuen Ansatz aktiv trainiert werden.

Nutzung von Strukturinformation für das aktive Training von SVM: Generell ist bei der Erzeugung eines Klassifikators zwischen *generativen* und *diskriminativen* Ansätzen zu unterschieden. Generative Ansätze modellieren die datengenerierenden Prozesse, z. B. anhand einer Wahrscheinlichkeitsverteilung, und nutzen diese Verteilung meist, um mit Hilfe des Bayes'schen Theorems einen Klassifikator zu erzeugen. Diskriminative Ansätze bestimmen stattdessen direkt aus den Daten eine Diskriminanzfunktion, um die Klassen der Eingabedaten bestmöglich zu separieren. Aus AL-Sicht besitzen beide Ansätze Vor- und Nachteile. Generative Ansätze haben den Vorteil, eine Modellierung der Daten bereitzustellen, die für eine aktive Musterauswahl nützliche Informationen liefert. Zudem können generative Klassifikatoren auch auf Basis teilgelabelter Daten trainiert werden. Diskriminative Ansätze hingegen erreichen meist höhere Klassifikationsgüten als generative Ansätze. Doch wegen ihres „Black-Box“-Verhaltens stellen sie weniger Informationen für eine aktive Musterauswahl bereit. Aufgrund dessen besteht das dritte Ziel darin, Strukturinformation für das (aktive) Training eines diskriminativen Klassifikators wie z. B. *Support Vector Machines* (SVM) zu nutzen. Da sich dadurch die Vorteile beider Ansätze kombinie-

ren lassen, d. h., die Klassifikationsfähigkeit einer SVM, als Stand der Technik im Bereich der Musterklassifikation, mit der Modellierungsfähigkeit eines generativen Ansatzes.

Grundsätzlich steht in jeder Anfragerunde $i > 0$ eines AL-Prozesses ein großer Pool U_i ungelabelter Daten und eine Menge L_i gelabelter Muster zur Verfügung. Jedoch werden für das (überwachte) Training einer SVM, d. h., für die Bestimmung der Support Vektoren, nur Informationen der gelabelten, nicht aber die der ungelabelten Muster berücksichtigt. Aufgrund dessen wurde in der Dissertation ein neuer, datenabhängiger Kernel definiert, der es ermöglicht, in jeder AL-Iteration i eine SVM halbüberwacht, d. h. basierend auf den Informationen beider Mengen U_i und L_i, zu trainieren. Dieser Kernel wird (RWM: *Responsibility Weighted Mahalanobis*)-Kernel genannt, da er die Ähnlichkeit zweier Muster auf Basis von Mahalanobis-Distanzen bewertet, die in den jeweiligen Komponenten des Mischmodells $\mathbf{M}$ enthalten sind. Hierbei werden die Mahalanobis-Distanzen umso stärker berücksichtigt, je „verantwortlicher“ deren Modellkomponenten für die Generierung der betrachteten Muster sind. Die innovativen Aspekte des neuen RWM-Kernels sind, dass er ohne algorithmische Anpassungen mit jeder Standardimplementierung einer SVM und auch mit jedem Standardalgorithmus der quadratischen Optimierung für SVM, wie Sequential Minimal Optimization (SMO), verwendet werden kann. Zudem ist eine SVM mit RWM-Kernel einfach zu parametrieren, da sich effiziente Suchheuristiken einer C-SVM mit (RBF: Radiale Basisfunktionen)-Kernel auf den neuen Kernel übertragen lassen.

Teile der Dissertation wurden bereits in Zeitschriften sowie auf internationalen Konferenzen veröffentlicht. In [RS11] wird die Selektionsstrategie 3DS vorgestellt, wobei 3DS die Informativität eines Muster mit Hilfe einer Linearkombination, bestehend aus einem Distanz-, Dichte- und Diversitätskriterium, bestimmt. Die Gewichte der beiden zuerst genannten Kriterien bestimmt 3DS adaptiv. Eine Erweiterung von 3DS beschriebt [RS13], die 4DS genannt wird, da 4DS ein zusätzliches Verteilungskriterium (Distribution) verwendet. In [RCS14] wird der erweiterte, transduktiver AL-Ansatz vorgestellt, der es einerseits ermöglicht, in jeder Anfragerunde einen beliebigen Klassifikator anhand einer vollständig gelabelten Datenmenge zu trainieren, d. h., die Muster werden entweder durch einen Experten (mit Kosten) oder transduktiv (kostenfrei) klassifiziert. Andererseits ermöglicht er es, die Datenmodellierung iterativ auf Basis der bekannten Klassenlabels zu verfeinern. In [RC14] wird ein modifizierter (kNN: k-Nearest-Neighbors)-Klassifikator namens Resp-kNN für spärlich gelabelte Daten vorgestellt, der auch innerhalb des transduktiven Prozesses verwendet wird, um die Datenmodellierung durch lokale Änderungen zu modifizieren. Der RWM-Kernel wird in [RS15] präsentiert und für das halbüberwachte Training von SVM untersucht. Für eine detailliertere Beschreibung der vorgestellten Lösungsansätze und weitere Literaturhinweise sei auf die Dissertation [Re15] verwiesen.

3 Experimente und Ergebnisse

Dieser Abschnitt erläutert den Aufbau der experimentellen Untersuchungen, die darin erzielten Ergebnisse und schließt mit einem anschaulichen Beispiel.

Experimenteller Aufbau: Zur Untersuchungen der vorgestellten Lösungsansätze wurden 20 öffentlich zugängliche Benchmark-Datensätze verwendet, wobei 16 dieser Datensätzen

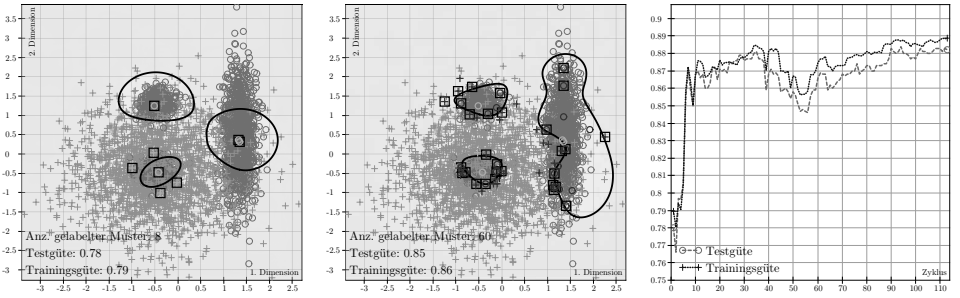
realen Anwendungen entstammen. Um repräsentative und aussagekräftige Ergebnisse zu erhalten, wurde zudem darauf geachtet, dass viele Datensätze (1) unterschiedliche Muster- und Klassenanzahlen, (2) unausgeglichene Klassenverteilungen und (3) kontinuierliche und kategorische Merkmale besitzen. Zudem wurden alle Datensätze *z-normalisiert* und mittels einer 5-fachen Kreuzvalidierung in *Trainings-* und *Testmengen* aufgeteilt, wobei die Testmengen zu *keinen* Parametrierungszwecken verwendet wurden. Des Weiteren wurden allen Ansätzen die gleichen Teildatenmengen zur Verfügung gestellt und der AL-Prozess stets nach der Anfrage von maximal 500 Mustern abgebrochen.

Aufgrund der umfassenden experimentellen Untersuchungen und um alle Lösungsansätze *numerisch* miteinander vergleichen zu können, wurden vier Bewertungsmaße *Ranked Performance* (RP), *Data Utilization Ratio* (DUR), *Area under the Learning Curve* (AULC) und *Class Distribution Match* (CDM) verwendet. Das Maß RP vergleicht die untersuchten Ansätze anhand ihrer erzielten Testgüten mit Hilfe eines nicht-parametrischen, statistischen *Friedman-Tests* gefolgt von einem *Nemenyi-Test*. Somit kann mit Hilfe des RP-Maßes Aussagen über die statistische Signifikanz der Ergebnisse getroffen werden. Das Maß DUR vergleicht die Ansätze anhand der Anzahl gelabelter Muster, die benötigt werden um eine vorgegebene Klassifikationsgüte zu erreichen. Das Maß AULC bewertet die Ansätzen auf Basis der Fläche unter ihrer Lernkurve. Das letzte Maß CDM wurde in der Dissertation neu definiert und verwendet als Vergleichskriterium die Abweichung der Klassenverteilung der Menge L von der Klassenverteilung der Gesamtdaten. Folglich bewertet das Maß RP die Effektivität der untersuchten Ansätze, während die Maße DUR, AULC und CDM die Effizienz bewerten.

Ergebnisse: Die Selektionsstrategie 4DS wurde für das AL-Training des Klassifikators CMM_{sha} anhand der 20 Datensätzen mit sieben anderen Selektionsstrategien (3DS, ITDS, PBAC, DUAL, US, DWUS und Random Sampling) numerisch verglichen. Aus diesem Vergleich resultierte, dass der CMM_{sha}-Klassifikator mittels 4DS effizienter und effektiver aktiv zu trainieren ist als mit den zuvor genannten Selektionsstrategien. 4DS benötigte im Mittel mindestens 24% weniger Anfragen, um eine vergleichbare Güte zu erreichen. Zudem erreichte der mit 4DS trainierte CMM_{sha} auf 12 Datensätzen die höchste Testgüte.

Hinsichtlich des zweiten Lösungsansatzes ist festzuhalten, dass anhand des neuen, erweiterten PAL-Prozesses ein mit 4DS aktiv trainierter CMM_{sep}-Klassifikator auf 15 der 20 Datensätze signifikant bessere Klassifikationsergebnisse erzielte als ein mit 4DS trainierter CMM_{sha}-Klassifikator. Des Weiteren wurde hierfür eine ähnliche Anzahl an Expertenanfragen benötigt. Aufgrund der vorherigen Ergebnisse wurde 4DS lediglich mit den Strategien US und Random Sampling numerisch verglichen.

Die Untersuchung des dritten Ansatzes ergab, dass eine SVM mit RWM-Kernel effektiver (auf 11 der 20 Datensätze signifikant höhere Gütewerte) und effizienter (im Mittel mindestens 15% weniger Anfragen) aktiv trainiert werden kann als eine SVM mit RBF-Kernel oder den datenabhängigen Kernels GMM (Gaussian Mixture Models) und LAP (Laplacian). Außerdem konnte gezeigt werden, dass eine SVM mit RWM-, GMM- und RBF-Kernel anhand 4DS effizienter aktiv trainiert werden kann als mit US oder Random Sampling. Zudem erzielte dieser Ansatz im Vergleich zu den beiden vorhergehenden signifikant höhere Klassifikationsgüten unter ähnlich hohen Kosten (Expertenanfragen).



(a) RBF-Kernel ($i=1$) (b) RBF-Kernel ($i=53$) (c) Güte vs. Anfragerunde

(d) RWM-Kernel ($i=1$) (e) RWM-Kernel ($i=53$) (f) Güte vs. Anfragerunde

Abb. 1: SVM mit RBF- bzw. RWM-Kernel aktiv trainiert mit US bzw. 4DS auf dem Datensatz Clouds. Die initial (Iteration $i=1$) gewählten Muster sind orange eingefärbt. Die gelabelten Muster, die in der aktuellen Anfragerunde aktiv ausgewählt wurden, sind rot eingefärbt sonst violett.

Exemplarischer Vergleich zweier aktiv trainierter SVM: Nachfolgend wird anhand des Datensatzes Clouds [Re15] exemplarisch das Lernverhalten zweier SVM verglichen. Hierzu wird eine SVM mit RBF-Kernel mittels Uncertainty Sampling[3](US) und eine SVM mit RWM-Kernel mittels 4DS aktiv trainiert. Abb. 1 zeigt den AL-Prozess beider SVM in zwei ausgewählten Anfragerunden. In jeder Anfragerunde $i > 0$ wurde mit Hilfe beider Strategien jeweils ein Muster aktiv gewählt, durch $\mathcal{O}$ gelabelt und die SVM daraufhin neu trainiert. Die Abb. 1(a) und 1(d) zeigen die SVM nach Ausführung der Initialisierungsrunde, in der acht Muster mit einem dichtebasierten Ansatz gewählt wurden (orange eingefärbt), und die Abb. 1(b) und 1(e) nach der aktiven Anfrage von weiteren 52 Mustern (violett oder rot eingefärbt). Die schwarz durchgezogene Linie stellt die Entscheidungsgrenze der SVM zwischen den Klassen „blaues Kreuz“ und „grüner Kreis“ dar und die mit einem Rechteck markierten Muster entsprechen ihren Stützvektoren. Im Fall des RWM-Kernels werden zusätzlich mit grau eingefärbten Ellipsen die Komponenten des Mischmodells $\mathbf{M}$ dargestellt, deren Zentren sich an den Stellen befinden, die durch ein großes $\times$ markiert sind. In den linken Abbildungen ist zu erkennen, dass die SVM mit RWM-Kernel, der Strukturinformation berücksichtigt, bereits initial eine deutlich höhere Testgüte (88%) erreicht als die SVM mit RBF-Kernel (78%), der diese Information nicht nutzt. Die mittleren Abbildungen zeigen, dass Muster mit 4DS besser entlang der „wahren“ Position der Entscheidungsgrenze verteilt gewählt werden als mittels US, wodurch die SVM mit RWM-Kernel

[3] Die Strategie US wählt jeweils das Muster bzgl. dessen Klassenzugehörigkeit die SVM am unsichersten ist.

bereits nach wenigen Anfragen eine Testgüte von 89% erzielt. In den rechten Abbildungen sind die Lernverläufe der beiden SVM dargestellt. Hier ist zu sehen, dass die SVM mit RBF-Kernel erst nach der Anfrage von fast 100 Mustern eine Testgüte von 88% erreicht. Diese Testgüte wird von der SVM mit RWM-Kernel bereits nach acht gelabelten Mustern ($i = 1$) erzielt.

4 Zusammenfassung und Ausblick

Ziel der Arbeit war die Entwicklung eines effektiveren, effizienteren und praxistaglicheren AL-Ansatz, wobei drei aufeinander aufbauende Lösungsansätze präsentiert wurden:

1. Die Struktur der Daten wurde mit probabilistischen, generativen Modellen einmalig geschätzt und davon ausgehend eine neue selbstadaptive, (fast) parameterfreie Selektionsstrategie namens 4DS entwickelt. 4DS nutzt Strukturinformation für ihre Musterauswahl und löst ein Schlüsselproblem des AL: In jedem Anfragezyklus eine gelabelte Mustermenge zu „erfragen“, deren Klassenverteilung die „wahre“, unbekannte Klassenverteilung der Gesamtdaten bestmöglich approximiert.
2. Zur Feinabstimmung der Datenmodellierung wurde der konventionelle PAL-Zyklus um einen transduktiven Prozess erweitert. Dieser adaptiert während des AL-Prozesses anhand der bekanntwerdenden Klassenlabels das Dichtemodell derart, dass jede seine Komponenten, möglichst nur Muster derselben Klasse modellieren.
3. Zur Kombination der Vorteile generativer und diskriminativer Ansätze für das AL-Training einer SVM wurde der neue, datenabhängige Kernel RWM entwickelt, der im Gegensatz zu verwandten Kernels keine zusätzlichen Parameter besitzt und ohne Anpassungen mit jeder SVM-Standardimplementierung verwendet werden kann.

In umfangreichen Untersuchungen wurde gezeigt, dass mit den vorgestellten Lösungen generative und diskriminative Klassifikatoren effizient (d. h., mit möglichst wenig Expertenanfragen) und effektiv aktiv trainiert werden können, da diese auf Basis statistischer Maße signifikant höhere Klassifikationsgüten erreichen als verwandte AL-Techniken.

Die vorliegende Arbeit stellt einen wichtigen Schritt dar, um AL-Techniken für reale Anwendungen effizient und effektiv nutzen zu können. Für dies wurden einige Restriktionen angenommen, auf denen auch „bisherige“ AL-Techniken beruhen, wie es existiert ein allgegenwärtiges und allwissendes Orakel oder die Klassen der Problemstellung sind vor Beginn des Lernprozesses bekannt. Das Ziel zukünftiger Forschungsarbeiten besteht somit darin, diese restriktiven Annahmen weiter aufzuheben, wodurch viele anwendungsnahe Probleme effizient lösbar werden. Prinzipiell lassen sich hierbei zwei mögliche Szenarien unterscheiden [Ca16]: Erstens Szenarien, in die eine kleinere Anzahl an Spezialisten längerfristig in einem kollaborativen AL-Prozess eingebunden sind. Anwendungsgebiete sind typische industrielle Problemstellungen, z. B. im Bereich der Qualitätskontrolle oder Produktentwicklung. Zweitens Szenarien, in die sehr viele „Nicht-Experten“ involviert sind, wobei diese meist nur kurzzeitig zur Verfügung stehen. Typisch hierfür sind

Crowdsourcing-Anwendungen zur Beantwortung von Anfragen basierend auf digitalen Medien, um z. B. Empfehlungsdienste kostengünstig zu realisieren. Darüber hinaus müssen sich zukünftige Informationssysteme zur Laufzeit weiterentwickeln, d. h. hochgradig autonom lernen, ihre eigenen Fähigkeiten bewerten und sich an ihre Umgebung anpassen können. Unterschiedliche Gebiete der Informatik leisten hierfür substantielle Beiträge, wobei im Bereich des maschinellen Lernens vor allem neue Verfahren entwickelt werden, die halbüberwachte, verstärkende und aktive Lerntechniken kombinieren.

Literaturverzeichnis

[Ca16] Calma, A.; Leimeister, J. M.; Lukowicz, P.; Oeste-Reiß, S.; Reitmaier, T.; Schmidt, A.; Sick, B.; Zweig, K. A.: From Active Learning to Dedicated Collaborative Interactive Learning. In: Proceedings of the 4th International Workshop on Self-optimisation in Autonomic and Organic Systems. Augsburg, Germany, 2016. (accepted).

[RC14] Reitmaier, T.; Calma, A.: Resp-*k*NN: A Semi-Supervised Classifier for Sparsely Labeled Data in the Field of Organic Computing. In: Organic Computing: Doctoral Dissertation Colloquium 2014. Kassel, Germany, S. 85–99, 2014.

[RCS14] Reitmaier, T.; Calma, A.; Sick, B.: Transductive active learning – A new semi-supervised learning approach based on iteratively refined generative models to capture structure in data. Inf. Sci., 293:275–298, 2014.

[Re15] Reitmaier, T.: Aktives Lernen für Klassifikationsprobleme unter der Nutzung von Strukturinformationen. Intelligent Embedded Systems. Kassel University Press, 2015.

[RS11] Reitmaier, T.; Sick, B.: Active Classifier Training with the 3DS Strategy. In: Symposium on Computational Intelligence and Data Mining. Paris, France, S. 88–95, 2011.

[RS13] Reitmaier, T.; Sick, B.: Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS. Inf. Sci., 230:106–131, 2013.

[RS15] Reitmaier, T.; Sick, B.: The Responsibility Weighted Mahalanobis Kernel for Semi-Supervised Training of Support Vector Machines for Classification. Inf. Sci., 323:179–198, 2015.

[Se12] Settles, B.: Active Learning. Morgan & Claypool Publishers, 2012.

Tobias Reitmaier studierte Informatik an der Universität Passau, an der er 2009 das Diplom erwarb. Anschließend war er an der Universität Passau für eineinhalb Jahre als wissenschaftlicher Mitarbeiter am Lehrstuhl für Rechnerstrukturen und auch am Institut für Softwaresysteme in technischen Anwendungen der Informatik (FORWISS) in den Bereichen des maschinellen Lernens und der Deflektometrie tätig. Daraufhin wechselte er an das Fachgebiet Intelligente Eingebettete Systeme der Universität Kassel, an dem er innerhalb von vier Jahren promovierte (2015). Zuletzt wirkte er als Post-Doktorand im DFG-geförderten Projekt „Techniken des Organic Computing für die Laufzeit-Selbst-Adaption ubiquitärer, multi-modaler Kontext- und Aktivitätserkennungssysteme“ mit. Seine Forschungsinteressen umfassen insbesondere aktives Lernen, Pattern Recognition und Methoden des Data Mining und des maschinellen Lernens.