Active Learning mit dem GUIDE-Entscheidungsbaum

1st Justus Kösters
Center for Applied Science Gütersloh
Bielefeld University of Applied Science
Gütersloh, Germany
justus.koesters@hsbi.de

2st Marvin Schöne
Center for Applied Data Science Gütersloh
Bielefeld University of Applied Science
Gütersloh, Germany
marvin.schoene@hsbi.de

***Zusammenfassung*—Industrielle Auslegungsversuche sind mit hohem Ressourcenaufwand verbunden, weshalb eine auf maschinellem Lernen (ML) basierende Unterstützung für die Auslegung von Maschinen oder Prozessen anzustreben ist. Diese gilt es auf Basis von wenigen historischen Daten zu entwickeln. Entscheidungsbäume mit dem GUIDE-Algorithmus sind für das Training mit geringen Datenmengen geeignet. Durch den Query-By-Commitee Ansatz des Active Learnings können zudem informative Datenpunkte identifiziert werden, mit denen eine bessere Modellperformance erreicht wird. Des Weiteren sind Bereiche, in denen das Modell unsichere Prognosen liefert, mithilfe des Query-By-Commitee Ansatzes quantifizierbar. Im Vergleich zu einer zufälligen Auswahl von zusätzlichen Trainingsdatenpunkten, weist der Active Learning Ansatz ein höheres Genauigkeitspotential auf. Die Prädiktionen von einem Ensemble aus Entscheidungsbäumen und einem einzelnen Entscheidungsbaum weichen nur gering voneinander ab, was relevant ist, da die Datenpunkte im Active Learning durch Ensembles ausgewählt werden, aber ein Einzelmodell aufgrund der Interpretierbarkeit als späteres Unterstützungstool verwendet wird.**

***Index Terms*—Scarce Data, Active Learning, GUIDE-Algorithmus, Modellunsicherheiten**

I. Einleitung

Für die industrielle Auslegung von Prozessen oder Maschinen führen Unternehmen Versuche durch, bei denen Daten generiert werden. Da diese Daten mit domänenspezifischen Wissen angereichert sind, weisen sie eine hohe Datenqualität auf, was die Nutzung dieser Daten zur Unterstützung in zukünftigen Produkt- oder Prozessauslegungen rechtfertigt [1]. Um diese Daten für zukünftige Auslegungen zu nutzen, kann ML verwendet werden, womit komplexe Zusammenhänge in den existierenden Daten approximiert werden können [2]. Vorteile von der ML-gestützen Auslegung ist die Ressourceneinsparung von Versuchen und die Sicherung des Domänenwissens in einem ML Modell. Während viele ML Algorithmen große Datenmengen für das Training benötigen, gibt es einige Anwendungen, wie die benannten Versuche für industrielle Auslegungen, wo die Datenerfassung mit hohen Kosten verbunden oder zeitintensiv ist [3]. Aufgrund dessen liegt bei industriellen Auslegungen häufig nur eine spärliche Datenbasis vor, worauf die verwendeten ML Modelle und Methoden abgestimmt sein müssen. In vorherigen Untersuchungen konnten Entscheidungsbäume mit dem GUIDE-Algorithmus als etablierte ML Verfahren für den Umgang mit spärlichen Datenmengen identifiziert werden. Neben der Auswahl eines geeigneten ML Modells für die geringen Datenmengen, gibt es weitere Methoden, die den Umgang mit diesen vereinfachen. Dazu zählt unter anderem das Active Learning [4]. Unter Active Learning wird eine Lernmethode verstanden, bei der die Lernalgorithmen die Datenpunkte auswählen, mit denen diese trainiert werden, um eine maximale Modellperformance mit so wenig Datenpunkten wie möglich zu erreichen [5]. Bei einer vorliegenden, geringen Datenbasis kann der Active Learning Ansatz die informativsten oder repräsentativsten Datenpunkte auswählen und zusätzliche benötigte Datenpunkte für die Verbesserung der Modellqualität auswählen bzw. erzeugen. Die Anzahl dieser ausgewählten, ungelabelten Datenpunkte wird durch den Active Learning Ansatz minimiert, sodass ein minimaler Aufwand für das Labeln durch Domänenexperten o.ä. anfällt [6].

Im Rahmen dieser Arbeit werden zunächst aus der Literatur bekannte Anwendungsfälle von Active Learning und Entscheidungsbäumen im industriellen Umfeld vorgestellt. Daraufhin werden die theoretischen Grundlagen für die Klassifikation mit Entscheidungsbäumen und dem GUIDE-Algorithmus sowie für das Active Learning erläutert, bevor die erarbeiteten Ansätze an einem gängigen Datensatz untersucht werden. Mit einer Beschreibung der Ergebnisse und einem Ausblick für weitere Untersuchungen wird der Forschungsbericht beendet.

II. Stand der Technik

Active Learning wird in der Literatur beim Umgang mit Bilddaten und tabellarischen Daten genutzt [6]. Die Nutzung von Active Learning bei einer tabellarischen Datenbasis kann zum Finden einer optimalen Lösung, beispielsweise einer Zusammensetzung von Arzneimitteln zur Bildung einer Zielzusammensetzung [7], verwendet werden. Ein weiterer Anwendungsfall ist die Auswahl von hochqualitativen Datenpunkten, mit denen Redundanzen im Datensatz vermieden und effiziente ML Modelle mit dieser reduzierten Datenbasis trainiert werden [8]. Außerdem kann das Active Learning bei einer unzureichenden Modellperformance aufgrund einer geringen Datenbasis durch die Auswahl der informativsten Datenpunkte den Aufwand für ein weiteres Labeling minimie-

ren. Durch das Training mit den aktualisierten Daten wird die Modellperformance gesteigert [9].

Die Verbindung von Entscheidungsbäumen mit Active Learning wird in [10] beispielsweise für Empfehlungssysteme verwendet. Die Entscheidungsbäume werden dabei neben ihrer intrinsischen Interpretierbarkeit aufgrund ihrer Möglichkeit zur Ensemblebildung genutzt. Dadurch bieten sie zum einen eine hohe Modellperformance und können zum anderen Sequenzen von informativen Datenpunkten für das Active Learning erzeugen. Die kontinuierliche Datenstromanalyse in [11] ist ein weiterer Anwendungsfall. Dabei ist es relevant den Informationszuwachs von Analysemodellen zu bestimmen, damit das Training bei fehlendem Zuwachs abgebrochen werden kann. Hierbei werden ebenfalls Entscheidungsbäume verwendet, da bei diesen Konfidenzintervalle für den Informationszuwachs erstellt werden können, wodurch das Training von Datenströmen effektiver wird. Neben der Nutzung von Konfidenzintervallen konnte die Performance der Entscheidungsbäume auch bei einer, durch das Active Learning selektierten, deutlich geringeren Datenbasis erhalten werden [11].

Des Weiteren werden in [12] Entscheidungsbäume für ein ML-gestütztes Tool zur Einstellung der Privatsphäre in Sozialen Netzen verwendet. Der Grund für die Nutzung von Entscheidungsbäumen liegt in dem Anwendungsfall bei der Interpretierbarkeit. Durch den Active Learning Ansatz werden dem Nutzer gezielte Fragen zum Teilen seiner Information mit Kontakten gestellt, sodass das ML-basierte Tools auf Basis dieser Angaben die Privatsphäreeinstellungen festlegt.

III. Theoretische Grundlagen

In diesem Kapitel werden nach der Erläuterung des GUIDE-Algorithmus zur Erstellung von Entscheidungsbäumen, die einzelnen Teilbereiche des Active Learnings aufgezählt und die für die Untersuchung relevanten Methoden beschrieben.

A. GUIDE-Algorithmus zur Entwicklung von Entscheidungsbäumen für die Klassifikation

Entscheidungsbäume werden durch Methoden des überwachten Lernens erzeugt, welches ein Teilgebiet des ML ist. Bei dem überwachten Lernen liegt ein Datensatz $\mathcal{D}$

$$\mathcal{D} = \{(\mathbf{x}_1, y_1),\ \dots\ , (\mathbf{x}_N, y_N)\} \tag{1}$$

mit N Datenpunkten $(\mathbf{x}_i, y_i)$ und $i \in \mathbb{N} \mid 1 \leq i \leq N$, wobei $\mathbf{x}_i$ Eingangs- und y_i Zielwerte sind, vor. Die Eingänge $\mathbf{x}_i$ enthalten M Merkmale $x_{i,m}$ mit $m \in \mathbb{N} \mid 1 \leq m \leq M$, welche entweder numerisch $x_{i,m} \in \mathbb{R}$ oder kategorisch $x_{i,m} \in \mathbb{G}_m$ sein können, wobei $\mathbb{G}_m$ die Menge der möglichen Ausprägungen des Merkmals m beschreibt. Die Zielgrößen $y_i \in \mathbb{N} \mid 1 \leq y_i \leq J$ stellen bei der Klassifikation J Klassen dar [13].

Die Entscheidungsbaum-Algorithmen partitionieren den Datenraum rekursiv über einen *divide and conquer* Ansatz. Aufgrund ihrer Struktur ist die Entscheidungsfindung der Entscheidungsbäume für die Anwender nachvollziehbar [14]. Sie bestehen aus Knoten t, die mit ihren Nachfolgern verbunden sind und einen gerichteten Graphen aufstellen. Auf den Initialknoten, welcher als Wurzel bezeichnet wird, folgen interne Knoten und Blattknoten, wobei letztere keine Nachfolger haben. In jedem internen Knoten werden die Datenmengen an einem oder einer Kombination von mehreren Merkmalen partitioniert und auf eine kleinere Datenmenge aufgeteilt. Die Blattknoten enthalten lokale Modelle, die die jeweiligen Zusammenhänge in ihrer Partition approximieren.

Die Performance eines Entscheidungsbaums wird maßgeblich durch die Partitionierungsqualität in den einzelnen Knoten bestimmt. Die Aufteilung wird durch die Kombination von Teilungsmerkmal k und dem dazugehörigen Schwellenwert c_k definiert [15]. Neben dem in dieser Arbeit vorgestellten GUIDE-Algorithmus, gibt es weitere Algorithmen zur Erstellung von Entscheidungsbäumen, wie CART [16], CHAID [17], ID3 [18] oder C4.5 [19]. Ein Problem, welches in vielen Algorithmen zur Baumerstellung auftritt, ist die verzerrte Auswahl der Teilungsmerkmale, welche durch eine unterschiedliche Anzahl der Merkmalsausprägungen auftritt. Dabei haben Merkmale mit mehr Ausprägungen eine höhere Wahrscheinlichkeit als Teilungsmerkmal ausgewählt zu werden. Neben der Minimierung der Verzerrung in der Merkmalsauswahl durch einen speziellen statistischen Test, erhöht der GUIDE-Algorithmus die Modellperformance durch eine Interaktionsanalyse zwischen den Merkmalen und ermöglicht die simultane Verwendung von numerischen und kategorischen Features [20].

Im Folgenden wird der Teilungsschritt im GUIDE-Algorithmus für die Klassifikation mit einer Datenmenge $\mathcal{D}$ beschrieben, wobei zunächst das geeignetste Teilungsmerkmal m^* und daraufhin der zugehörige Schwellenwert c_{m^*}, welcher bei numerischen Features ein realer Wert $c_{m^*} \in \mathbb{R}$ und bei kategorischen Merkmalen $c_{m^*} \subset \mathbb{G}_m^*$ eine Teilmenge der möglichen Ausprägungen des Merkmals ist, bestimmt.

1) Auswahl Teilungsmerkmal: Bei der Auswahl des Teilungsfeatures wird zunächst ein *main effect* Test durchgeführt, bei dem der Einfluss der einzelnen Merkmale m auf die Zielgröße $\mathbf{y}$ über einen Chi-Quadrat Test $\chi^2_\nu(\mathbf{x}_m, \mathbf{y})$ bestimmt wird. Zur Durchführung des Chi-Quadrat Testes werden bei kategorischen Features die jeweiligen Ausprägungen $G \in \mathbb{G}_m$ mit den J Klassen des Zielfeatures in einer Kontingenztabelle zusammengefasst. Numerische Features werden je nach Verhältnis von Anzahl Datenpunkte zu Ausprägung der Zielgröße in drei oder vier Intervalle diskretisiert und daraufhin in einer Kontingenztabelle mit den Klassen der Zielgröße zusammengefasst. Nachdem leere Zeilen und Spalten in der Kontingenztabelle entfernt sind, wird der Chi-Quadrat Test durchgeführt und das Ergebnis dessen, welches aus dem Chi-Quadrat Wert χ^2_ν und dem Freiheitsgrad des Testes ν besteht, mittels der Wilson-Hilferty Approximation $W(m)$

$$W(m) = \max\Bigg(0, \left[\frac{7}{9} + \sqrt{\nu}\left\{\left(\frac{\chi^2_\nu(\mathbf{x}_m, \mathbf{y})}{\nu}\right)^{1/3} - 1 + \frac{2}{9\nu}\right\}\right]^3\Bigg) \tag{2}$$

korrigiert, um eine Verschiebung der Chi-Quadrat Verteilung zu realisieren. Mit den Ergebnissen aus dem main effect Test wird überprüft, ob eine univariate Teilung mit einem Eingangsfeature einen ausreichenden Einfluss auf die Zielgröße hat. Dafür wird die Variable $\alpha = 0.05/K$ eingeführt, bei der K die Anzahl der numerischen Eingangsgrößen des Datensatzes ist. Wird die Bedingung

$$\max_m W(m) > \chi^2_{1,\alpha}, \tag{3}$$

bei der der Wert $\chi^2_{1,\alpha}$ den Chi-Quadrat Wert bei dem Freiheitsgrad 1 und einem Signifikanzniveau α beschreibt, erfüllt, wird die univariate Partitionierung anhand des Merkmals m mit dem maximalen Wilson-Hilferty Wert aus dem main effect Test durchgeführt.

Ist die Bedingung in (3) nicht erfüllt, wird die *interaction chi-squared statistic* zwischen zwei Merkmalen $m', m'' \in \mathbb{N} \mid 1 \leq m', m'' \leq M$, wobei $m' \neq m''$ berechnet. Dafür wird erneut eine Tabelle aufgestellt, bei der die Kontingenztabellen der einzelnen Featurepaare spaltenweise zusammengeführt werden. Bei den numerischen Merkmalen wird wie im main effect Test eine Diskretisierung der Werte in Intervalle durchgeführt. Das Ergebnis der interaction chi-squared statistic wird erneut mit der Wilson-Hilferty Approximation überführt und für alle möglichen Merkmalspaare berechnet. Der maximale Wilson-Hilferty Wert wird mit der Ungleichung

$$\max_{m',m''} W(m', m'') > \chi^2_{1,\beta} \tag{4}$$

überprüft, wofür der Parameter $\beta = 0.05/K(K-1)$ eingeführt wird. Wenn diese Bedingung erfüllt ist, wird das Merkmalspaar mit dem maximalen Wilson-Hilferty Wert für den weiteren Teilungsprozess verwendet. Wird die Ungleichung nicht erfüllt, wird auf das Merkmal mit dem maximalen Wilson-Hilferty Wert aus dem main effect Test zurückgegriffen.

2) Bestimmung Schwellenwert: Aus der Featureauswahl gehen entweder ein oder zwei Feature hervor. Je nach Konstellation der vorherigen Featureauswahl können die in Figur 1 abgebildeten Fälle eintreten.

Bei der Auswahl eines Features (Fall 1)), wird der vorliegende Datensatz $\mathcal{D}_t$ des Entscheidungsknotens t mit dem Schwellenwert c_m an dem jeweiligen Feature m in $\mathcal{D}_{t,\mathrm{L}}$ und $\mathcal{D}_{t,\mathrm{R}}$ aufgeteilt. Bei den numerischen Merkmalen (Fall 1) a)) entspricht $\mathcal{D}_{t,\mathrm{L}} = \mathcal{D}_t \leq c_m$ und $\mathcal{D}_{t,\mathrm{R}} = \mathcal{D}_t > c_m$, bei den kategorischen (Fall 1) b)) $\mathcal{D}_{t,\mathrm{L}} = \{\mathcal{D}_t \in c_m\}$ und $\mathcal{D}_{t,\mathrm{R}} = \{\mathcal{D}_t \notin c_m\}$. Der Schwellenwert wird dabei so gewählt, dass die gewichtete Summe der Gini Unreinheiten

$$\left|\frac{\mathcal{D}_{t,\mathrm{L}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{L}}) + \left|\frac{\mathcal{D}_{t,\mathrm{R}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{R}}) \tag{5}$$

minimiert wird. Die Werte $\left|\frac{\mathcal{D}_{t,\mathrm{L}}}{\mathcal{D}_t}\right|$ und $\left|\frac{\mathcal{D}_{t,\mathrm{R}}}{\mathcal{D}_t}\right|$ geben dabei die Proportionalität der Teilmengen am Gesamtdatensatz an. Die Gini Verunreinigung g eines Datensatzes $\mathcal{D}_{t,(\cdot)}$ wird durch

$$g(\mathcal{D}_{t,(\cdot)}) = 1 - \sum_{j=1}^{J} P(j \mid \mathcal{D}_{t,(\cdot)})^2 \tag{6}$$

beschrieben, wobei $P(j \mid \mathcal{D}_{t,(\cdot)})$ den Anteil an Datenpunkten, die zur jeweiligen Klasse j gehören, definiert. Um den Suchaufwand nach einem optimalen Schwellenwert c^*_m, besonders bei kategorischen Features mit vielen Ausprägungen, zu minimieren, wird eine eingeschränkte Suche über einige der möglich Teilmengen durchgeführt. In der Implementierung des GUIDE-Algorithmus für die spätere Anwendung wird lediglich die vollständige Suche durchgeführt.

Sobald zwei Features für eine Partitionierung relevant sind (Fall 2)), wird eine zweistufige Partitionierung für verschiedene Schwellenwerte $c_{m'}$, $c_{m''}$ der Features m' und m'' überprüft. Da in dieser Arbeit nur univariate Splits und keine Linearkombinationen aus verschiedenen Features betrachtet werden, ist das Resultat der Schwellenwertbestimmung ein Paar aus optimalem Teilungsfeature m^* und zugehörigem, optimalem Schwellenwert c^*_m. Für die Auswahl der optimalen Teilung wird erneut die gewichtete Summe der Gini Unreinheiten

$$\left|\frac{\mathcal{D}_{t,\mathrm{LL}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{LL}}) + \left|\frac{\mathcal{D}_{t,\mathrm{LR}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{LR}}) + \left|\frac{\mathcal{D}_{t,\mathrm{RL}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{RL}}) + \left|\frac{\mathcal{D}_{t,\mathrm{RR}}}{\mathcal{D}_t}\right| g(\mathcal{D}_{t,\mathrm{RR}}) \tag{7}$$

minimiert. Aufgrund der zwei möglichen Teilungsmerkmale wird hier eine zweistufige Betrachtung durchgeführt, wobei einmal Merkmal m' als erstes und Merkmal m'' als zweites Teilungsfeature betrachtet wird, bevor die umgekehrte Teilung (Merkmal m'' als erstes, m' als zweites) durchgeführt wird. Die Aufteilungen der unteren Teilmengen finden an einem Merkmal statt, sodass die eingeführten Verfahren nach Fall 1) a) oder Fall 1) b), je nach Art des Teilungsmerkmals angewandt werden können [21].

In den jeweiligen Knoten werden die Partitionierungen nach dem beschriebenen Verfahren durchgeführt und rekursiv fortgesetzt. Abbruchkriterien für weitere Splits sind das Erreichen einer definierten maximalen Tiefe, eine minimale Datensatzgröße oder eine einzige verbleibende Klasse im Datensatz.

B. Etablierte Active Learning Methoden

In diesem Kapitel werden die Grundlagen des Active Learnings beschrieben. Dazu gehören die Einteilungen der verschiedenen Abfrageszenarien und -strategien sowie eine Erläuterung der relevanten Methoden für den erläuterten Anwendungsfall.

1) Abfrageszenarien: Im Active Learning gibt es mehrere Abfrageszenarien, welche aus verschiedenen Problemstellungen resultieren. Die drei bekanntesten Szenarien sind die *membership query synthesis*, das *stream-based selective sampling* und das *pool-based sampling* [6]. Im Rahmen dieser Arbeit werden nur das pool-based sampling und die membership query synthesis beschrieben, da diese relevant für den erläuterten industriellen Anwendungsfall sind. Beim pool-based sampling wird angenommen, dass ein kleiner gelabelter und ein großer ungelabelter Datensatz vorliegen. Ein ML Modell wird zunächst mit den gelabelten Daten trainiert, bevor das Active Learning aus dem ungelabelten Pool die

Konstellationen für Featureauswahl				
1) ein Feature main effect		2) zwei Feature interaction chi-squared statistic		
a) numerisches Feature	b) kategorisches Feature	a) 2 numerische Features	b) 2 kategorische Features	c) 1 numerisches/ 1 kategorisches Feature

Abbildung 1: Konstellationen Featureauswahl

Datenpunkte auswählt, welche die Modellperformance nach einem bestimmten Kriterium verbessern können. Diese werden von einem Orakel, welches ein Domänenexperte sein kann, gelabelt und das ML Modell mit dem aktualisierten gelabelten Datenpool trainiert. Dieses Vorgehen kann iterativ durchgeführt werden, beispielsweise bis das Modell eine ausreichende Performance aufweist, oder die Ressourcen für das Labeling aufgebraucht sind. Bei der membership query synthesis fragt der Lernalgorithmus Labels für die informativsten Datenpunkte, die synthetisch im Eingangsdatenraum erzeugt werden, beim Orakel an. Problematisch bei der Abfrage dieser künstlich erzeugten Datenpunkte ist, dass es für das Orakel teilweise unmöglich ist, diese zu labeln [4]. Ein Beispiel für diese Problematik liegt in der Erkennung von handgeschriebenen Zeichen, da solche synthetisch erzeugten Zeichen in Grenzbereichen zwischen Zahlen nicht eindeutig für den Menschen identifizierbar sind, beziehungsweise auf verschiedene Zeichen deduzieren [22].

2) Abfragestrategien: Weitere Unterschiede bei den Active Learning Methoden liegen in der Strategie zur Auswahl der abzufragenden Datenpunkte. Die unterschiedlichen Abfragestrategien wählen die Instanzen zur Abfrage durch den vorher bestimmten Nutzen der ungelabelten Instanzen aus [6]. Bei den Active Learning Strategien kann eine Unterteilung in informationsbasierte und repräsentative Abfragestrategien durchgeführt werden. Der Nutzen bei den informationsbasierten Abfragestrategien ist der Informationsgehalt der einzelnen Instanzen. Der Informationsgehalt einzelner Instanzen ist in den Eingangsdatenbereichen am größten, in denen das Modell die unsichersten Prognosen durchführt, was im Bereich der Entscheidungsgrenzen der Fall ist [23]. Bekannte informationsbezogene Methoden sind *Uncertainty Sampling*, *Query-By-Committee*, *Expected Model Change*, *Expected Error Reduction*, *Variance Reduction* und *Density-Weighted Methods* [4]. Die repräsentationsbasierten Abfragestrategien gestalten die Abfragen so, dass der gesamte Datenbereich durch die Datenpunkte repräsentiert wird. Im Gegensatz zu den informationsbasierten Ansätzen, wird hierbei nicht explizit der Bereich nahe der Entscheidungsgrenzen abgefragt [6]. Ansätze, die repräsentationsbasiert agieren, sind *Density-Based*, *Diversity-Based* und *Cluster-Based* [23].

Im Folgenden wird die informationsbezogene Abfragestrategie des Uncertainty Samplings, welche als eine der meist genutzten und simpelsten Methoden gilt, sowie der Query-By-Commitee Ansatz beschrieben [4]. Beim ML werden grundsätzlich Unsicherheiten aufgrund von Widersprüchen in Daten und aufgrund von dem Mangel an Daten unterschieden. Bei ersterem gibt es widersprüchliche Datenpunkte im Bereich der Entscheidungsgrenzen und bei letzterem tritt die Unsicherheit aufgrund einer fehlenden Datengrundlage im Bereich der zugrunde liegenden Entscheidungsgrenzen auf [24]. Das Uncertainty Sampling kann durch das Beispiel eines probabilistischen Modells mit einem binären Klassifikator einfach erklärt werden. Das Uncertainty Sampling stuft den Datenpunkt $\mathbf{x}_i$, bei dem die posteriori Wahrscheinlichkeit am nächsten an 0.5 liegt, als unsichersten Datenpunkt ein [25]. Eine generalisiertere Variante, welche für nicht-binäre Klassifikation verwendet werden kann, ist *least-confident* LC

$$\mathbf{x}^*_{\mathrm{LC}} = \underset{\mathbf{x}}{\operatorname{argmax}} \big(1 - P_\theta(\hat{y}_1|\mathbf{x})\big), \tag{8}$$

bei der der unsicherste Datenpunkt $\mathbf{x}^*_{\mathrm{LC}}$ aus dem Datenpunkt $\mathbf{x}$ resultiert, bei dem die prädizierte Klasse $\hat{y}_1$ das Label für den Datenpunkt ist, dass bei dem verwendeten Modell θ die höchste Wahrscheinlichkeit aufweist. Eine Erweiterung ist die Variante *margin sampling* M

$$\mathbf{x}^*_{\mathrm{M}} = \underset{\mathbf{x}}{\operatorname{argmin}} \big(P_\theta(\hat{y}_1|\mathbf{x}) - P_\theta(\hat{y}_2|\mathbf{x})\big), \tag{9}$$

welche den least-confident Ansatz, bei der nur die wahrscheinlichste Prädiktionsklasse $\hat{y}_1$ verwendet wird, um die zweit wahrscheinlichste Prädiktion $\hat{y}_2$ des Modells erweitert. Beim margin sampling wird also der unsicherste Datenpunkt über die minimale posteriori Wahrscheinlichkeitsdifferenz zwischen den beiden wahrscheinlichsten Prädiktionen $\hat{y}_1$ und $\hat{y}_2$ bestimmt. Der Ansatz zur Berechnung über die *Entropie* H

$$\mathbf{x}^*_{\mathrm{H}} = \underset{\mathbf{x}}{\operatorname{argmax}} \left(-\sum_{j=1}^{J} P_\theta(y_j|\mathbf{x}) \log P_\theta(y_j|\mathbf{x}) \right) \tag{10}$$

beinhaltet die Betrachtung aller J Klassen. Die Entropie ist ein üblicher Wert für die Bestimmung von Unsicherheiten oder Ungleichheiten im ML [26].

Bei dem Query-By-Commitee Ansatz werden die informativsten Datenpunkte über den größten Widerspruch des Komitees $\mathcal{C} = \{\theta^{(1)} \ \ldots \ \theta^{(C)}\}$ mit C Einzelmodellen $\theta^{(r)}$ mit $r \in \mathbb{N} \mid 1 \leq r \leq C$ detektiert. Die einzelnen Modelle $\theta^{(r)}$ des Komitees werden auf dem gelabelten Datensatz trainiert und stellen verschiedene Hypothesen dar. Um ein

Set von verschiedenen Hypothesen, die unterschiedliche Eingangsdatenräume abbilden, aufzustellen, gibt es verschiedene Möglichkeiten [4]. Für diskriminative, nicht-probabilistische Modelle werden die Ensemble Learning Ansätze *bagging* und *boosting* für die Erstellung von verschiedenen Datensätzen für das Training der Modelle verwendet [27]. Beim bagging (bootstrap aggregating) werden Modelle auf einer Teilmenge der gelabelten Datensätze, die zufällig mit mehrmaliger Auswahl zusammengesetzt werden, trainiert. Die Prädiktion des bagging Modells wird bei der Klassifikation beispielsweise durch die Auswahl der dominanten Klasse aus den einzelnen Modellen gewählt. Durch bagging werden Instabilitäten der Modelle ausgeglichen und Varianzen sowie die Möglichkeit von Overfitting verringert [28]. Die Unsicherheit des Ensembles wird über die verschiedenen Vorhersagen der einzelnen Modelle bestimmt. Dabei kann die in Formel (10) eingeführte Entropie oder die Erweiterung *vote entropy* VE

$$\mathbf{x}^*_{\mathrm{VE}} = \underset{\mathbf{x}}{\operatorname{argmax}} \left(-\sum_{j=1}^{J} \frac{V(y_j)}{C} \log \frac{V(y_j)}{C} \right) \tag{11}$$

zur Bestimmung des unsichersten Datenpunkts $\mathbf{x}^*$ genutzt werden. Der Wert $V(y_j)$ stellt dabei die Anzahl der Prädiktionen des Ensembles für eine Klasse y_j dar.

IV. Untersuchung des GUIDE-Algorithmus

In diesem Kapitel werden die durchgeführten Versuche zum GUIDE-Algorithmus erläutert. Dabei gilt es zunächst den allgemeinen Versuchsaufbau zu beschreiben, bevor auf die spezifischen qualitativen und quantitativen Versuche sowie deren Ergebnisse eingegangen wird.

A. Versuchsaufbau

Für die Untersuchungen im Rahmen dieser Arbeit wird der durch eine Principal Component Analyse auf zwei Hauptkomponenten reduzierte Iris-Datensatz herangezogen. Der Iris-Datensatz [29] besteht initial aus 150 Datenpunkten und 4 Merkmalen, welche 3 verschiedene Schwertlilienarten darstellen, die balanciert in den Daten vorkommenden. Der reduzierte Datensatz ist der Abbildung 2 zu entnehmen.

In den folgenden Unterkapiteln werden die Durchführung eines quantitativen Proof of Concepts für das Active Learning mit Entscheidungsbäumen und einer qualitativen Untersuchung von Abweichungen zwischen Vorhersagen von Ensembles und einzelnen Entscheidungsbäumen behandelt.

1) Quantitative Untersuchung: Aufgrund der im Kapitel III vorgestellten Grundlagen wird ein Proof of Concept für das Active Learning mit einem Query-By-Commitee Ansatz, bei dem das Ensemble durch bagging erstellt wird, verfolgt. Ein Grund für die Nutzung von Query-By-Commitee ist, dass einzelne Entscheidungsbäume aufgrund ihrer Partitionierungen nur grobe räumliche Bezüge zu den Unsicherheiten herstellen können und eine Abfrageauswahl daher aus mehreren unsicheren Datenpunkten durchgeführt werden muss. Der Fokus der Betrachtung liegt auf dem Einfluss der Datenpunktauswahl durch Ensembles auf das finale Einzelmodell. Dafür wird

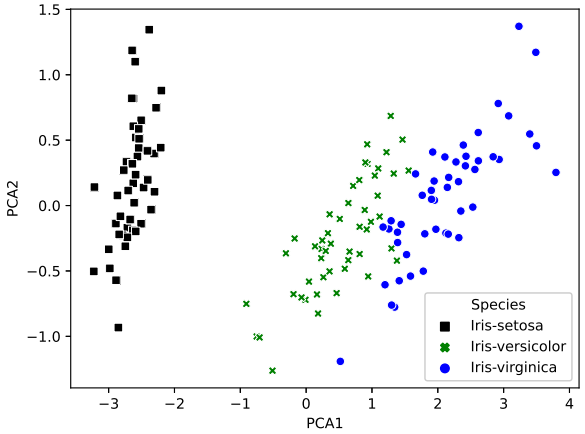


Abbildung 2: Iris-Datensatz mit PCA auf 2 Dimensionen reduziert

der Güteverlauf der Prädiktionen anhand verschiedener Auswahlmethoden betrachtet. Die Auswahlmethoden enthalten 2 Query-By-Commitee Ansätze mit unterschiedlichen maximalen Baumtiefen, wodurch die Komplexität variiert wird, und einer zufälligen Auswahl. Durch die Variation der Komplexität weisen die Ensembles unterschiedliche Neigungen zum Under- oder Overfitting auf.

Für diesen Vergleich verschiedener Datenpunktauswahlmethoden wird ein Benchmarking mit 20 unabhängigen Durchläufen durchgeführt. Die initiale Datenbasis besteht aus 10 zufällig ausgewählten Datenpunkten, welche im Verlauf mit 25 weiteren Datenpunkten mittels Active Learning sowie einer zufälligen Auswahl angereichert wird. Ein Einzelmodell mit einer maximalen Baumtiefe von 3 wird mit jedem verfügbaren Datensatz trainiert, während die Klassifikationsgenauigkeit über den Anteil der richtig klassifizierten Datenpunkte über den gesamten Iris-Datensatz bestimmt wird. Unterschiede in den Query-By-Commitee Auswahlmethoden bestehen in der maximalen Tiefe der einzelnen Bäume im Ensemble. Es werden die Ensembles mit einer maximalen Tiefe von 3 und 8 verglichen. Erstere haben also das selbe Tiefenabbruchkriterium wie die Einzelmodelle, auf Basis derer die Genauigkeit bestimmt wird. Die Ensembles bei den Query-By-Commitee Ansätzen bestehen aus 100 mit bagging erstellten Entscheidungsbäumen, wodurch eine ausreichende Varianz in den Prädiktionen vorhanden ist.

2) Qualitative Untersuchung: Des Weiteren werden die Unsicherheiten im Dateneingangsraum qualitativ betrachtet. Die Betrachtung der Unsicherheiten wird auf Basis eines Ensembles aus 100 Einzelmodellen mit einer maximalen Baumtiefe von 8 auf dem gesamten Iris-Datensatz durchgeführt. Ein Vergleich der Unsicherheiten eines Ensembles vor und nach der Anreicherung von informativen Datenpunkten wird exemplarisch anhand der Ergebnisse einer der 20 vorherigen Benchmarking-Durchläufe durchgeführt. Die Unsicherheiten werden mit der vote entropy aus den Klassifikationen

der einzelnen, im Ensemble enthaltenen Modellen bestimmt. Dafür wird für ein Meshgrid aus 62500 Datenpunkten im ursprünglichen Dateneingangsraum die Klassifikationen des Ensembles betrachtet.

Des Weiteren wird das Prädiktionsverhalten eines Ensembles und eines Einzelmodells untersucht, um zu betrachten inwiefern das Modellverhalten übereinstimmen muss, damit in der Anwendung Einzelmodelle aufgrund ihres interpretierbaren Aufbaus verwendet werden können. Für diese Untersuchung werden Unsicherheitsbereiche hinzugezogen und die Prädiktionen beider Modellarten klassenweise miteinander verglichen. Grund für die Untersuchung ist die Anreicherung der Datenbasis mittels Ensembles, welche später für das Training eines Einzelmodells verwendet wird. Der Vergleich zwischen den Modellausgaben vom Einzelmodell und dem Ensemble wird auf Basis eines 2500 Datenpunkten enthaltenden Meshgrids im Eingangsdatenraum durchgeführt. Bei der Untersuchung wird die Datensatz- sowie die Ensemblegröße variiert. Für den Datensatz wird die volle sowie halbe Datensatzmenge verwendet, sodass die Ausprägungen des Zielfeatures weiterhin balanciert sind. Das Ensemble enthält entweder 100 oder 150 mit bagging erzeugte GUIDE-Entscheidungsbäume.

B. Quantitative Ergebnisse

Die Ergebnisse aus der Durchführung des Benchmarkings für das Active Learning mit verschiedenen Abfragestrategien ist in der Darstellung 3 zu sehen. Es ist zu erkennen, dass die Genauigkeiten bei allen 3 Abfrageszenarien mit zunehmender Anzahl an vorhandenen Datenpunkten steigt. Die Genauigkeiten bei dem initialen Datensatz variieren aufgrund einer zufälligen Auswahl der Datenpunkte. Bei der zufälligen Auswahl der Datenpunkte tritt eine Stagnation der Klassifikationsgüte nach ca 10 hinzugefügten Datenpunkten ein. Die maximale Genauigkeit liegt hier bei ca. 0.92. Bei den Abfragen nach dem Query-By-Commitee Ansatz ist keine Stagnation der Genauigkeit nach 25 hinzugefügten Datenpunkte zu erkennen, was auf ein höheres Genauigkeitspotential als das Training mit einer zufälligen Auswahl der Datenpunkte hinweist. Die Prädiktionsgenauigkeiten liegen bei der Query-By-Commitee Anreicherung bei ca. 0.95. Die Betrachtung der Standardabweichung der Klassifikationsgüten (Fig. 3 b)), die aus der Durchführung des 20-fachen Benchmarkings resultieren, zeigt, dass die Resultate bei dem Query-By-Commitee Ansatz mit einer höheren Einzelmodellkomplexität am meisten variieren. Bei einer geringeren Komplexität der Einzelmodelle liegt die Standardabweichung bereits nach dem Hinzufügen von 9 Datenpunkten auf einem vergleichsweise geringen Niveau. Beim Random Sampling ist die Standardabweichung nach dem Eintreten der Stagnation bei der Klassifikationsgüte auf dem niedrigsten Niveau und stagniert ebenfalls.

Eine Übersicht über die ausgewählten Datenpunkte der einzelnen Verfahren ist beispielhaft für einen Durchlauf im Benchmarking in Darstellung 4 abgebildet. Dort ist zu erkennen, dass das Random Sampling die Datenpunkte eher im Datenraum verteilt, während die Query-By-Commitee Ansätze Datenpunkte im Bereich der Entscheidungsgrenzen fokussiert.

C. Qualitative Ergebnisse

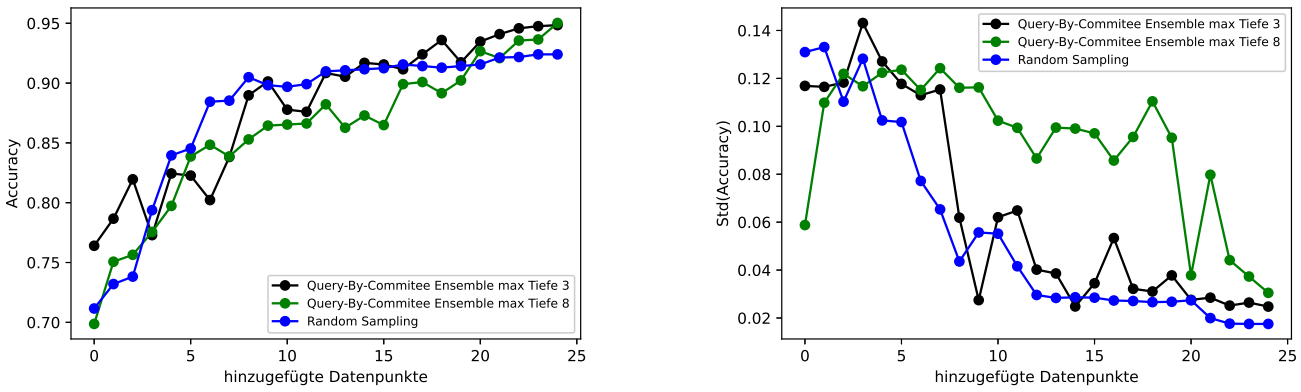
Ein Vergleich der Unsicherheiten im Eingangsdatenraum, basierend auf der vote entropy der Ergebnisse eines Ensembles, ist der Darstellung 5 abgebildet. Das Ensemble enthält 100 Einzelmodelle, die auf die vor und nach der beschriebenen Anreicherung reduzierten Iris-Datensätze trainiert sind. Es ist zu erkennen, dass die Unsicherheiten im Eingangsdatenraum drastisch reduziert werden. Im Folgenden wird diese Aussage quantifiziert. Mit dem initialen Datensatz weisen alle Datenpunkte eine Unsicherheit größer 0 auf. Das arithmetische Mittel[1] der Unsicherheit beträgt 0.68. Nach der gezielten Anreicherung des Datensatzes sind noch 79.8 Prozent der Datenpunkte mit Unsicherheiten behaftet. Die mittlere Unsicherheit aller Datenpunkte beträgt 0.23, während die unsicheren Datenpunkte einen Mittelwert von 0.29 aufweisen. Des Weiteren liegen die unsicheren Bereiche im Extrapolationsbereich und nahe der Entscheidungsgrenzen. Sie verlaufen ebenfalls achsen orthogonal.

Der Anteil gleicher Modellvorhersagen von einem Einzelmodelle und einem Ensemble auf Basis von 2500 Datenpunkten im Eingangsdatenraum liegen je nach Trainingsdatensatz- und Ensemblegröße zwischen 94 und 96.6 Prozent. Die Punkte, in denen Unterschiede zwischen den Prädiktionen von Einzelmodell und Ensemble auftreten, sind in Figur 6 abgebildet. Unterschiede in der Klassifikation treten bis auf eine Ausnahme nur unter den Klassen Iris-versicolor und Iris-virginica auf.

V. Diskussion

Im Folgenden gilt es die vorher beschriebenen Ergebnisse zu diskutieren. Eine Möglichkeit für die ähnlichen Prädiktionsgüten im Active Learning Proof of Concept liegen in der Einfachheit des Datensatzes. Bereits mit wenigen Datenpunkten und Partitionierungen ist es möglich, die grobe Struktur des Datensatzes abzubilden und die Klassen zu separieren. Diese These wird durch die Stagnation der Genauigkeit im Random Sampling bestätigt, da mit wenigen Datenpunkten ein gewisses Genauigkeitsniveau erreicht wird. Würde eine kleinere initiale Datenmenge verwendet werden, wäre der Genauigkeitszuwachs bei der Auswahl von informativen Datenpunkten vermutlich größer als der vom Random Sampling. Des Weiteren sollte eine gleiche Ausgangsdatenbasis für alle Modelle verwendet werden, um Verzerrungen, die auf der Initiierung der Datensätze basieren, zu entfernen. Unterschiede zwischen den Ensembles mit verschiedenen Komplexitäten sind primär im Verlauf der Standardabweichung und nicht in der Klassifikationsgüte zu erkennen. Dabei weist das Ensemble mit der höheren maximalen Baumtiefe eine höhere Standardabweichung auf, was auf die größeren Unterschiede der einzelnen Ensembles in der Entscheidungsfindung durch Overfitting zurückzuführen ist. Bei der Betrachtung der Datenpunktanreicherung ist zu erwähnen, dass der Query-By-

[1] im folgenden standardmäßig ohne genaue Benennung verwendet



(a) Klassifikationsgenauigkeit

(b) Standardabweichung Klassifikationsgenauigkeit

Abbildung 3: Klassifikationsgenauigkeitsverlauf sowie Standardabweichung aus des Active Learnings mit verschiedenen Abfragestrategien

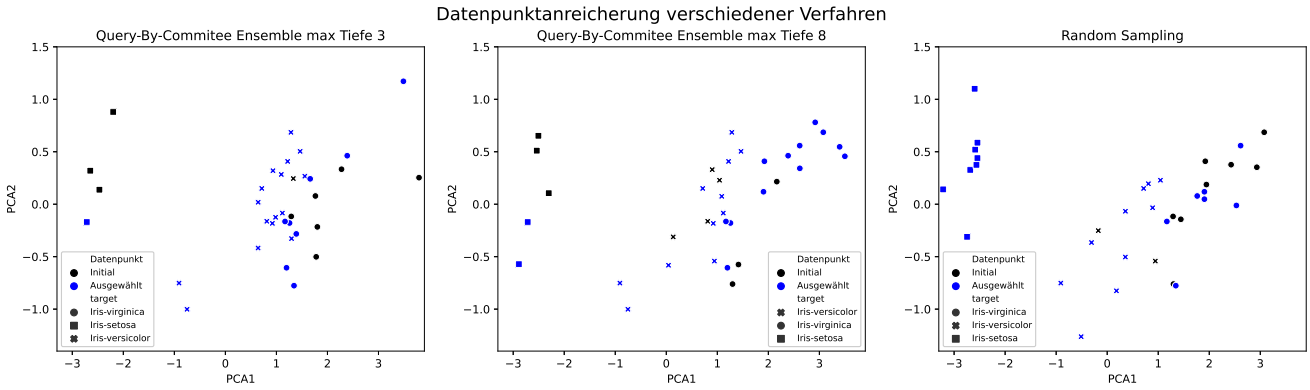


Abbildung 4: Auswahl der Datenpunkte nach Auswahlverfahren

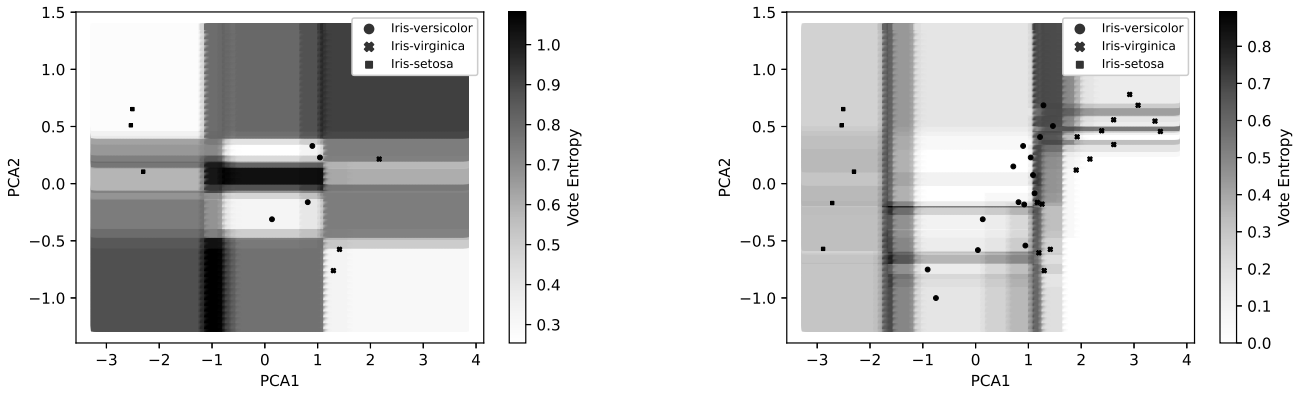
Commitee Ansatz Datenpunkte im Bereich der Entscheidungsgrenzen favorisiert auswählt. Dies stimmt mit der Aussage über informationsbasierte Abfragestrategien aus dem Kapitel III-B überein.

Die Unsicherheiten bilden die Entscheidungsgrenzen aufgrund der variierenden Datenauswahl der Ensembles ab. Dementsprechend verlaufen diese, wie die Entscheidungsgrenzen auch, achsen orthogonal. Die Unsicherheiten im Eingangsdatenraum werden aufgrund der aussagekräftigeren Datenbasis minimiert.

Zur Betrachtung der Prädiktionsunterschiede zwischen Ensemble- und Einzellmodellklassifikation kann gesagt werden, dass die Unterschiede hauptsächlich zwischen den Klassen Iris-versicolor und Iris-virginica auftreten. Grund dafür ist die teilweise Überschneidung der Klassen im Eingangsdatenraum. Die Abweichungen zwischen Ensemble und Einzelmodell werden zum einen durch die Auswahl von verschiedenen Datenpunkten beim Training des Ensembles hervorgerufen und zum anderen durch die Möglichkeit der Ensembles zur Bildung komplexerer Entscheidungsgrenzen aufgrund der höheren Komplexität. Diese abweichenden Prädiktionen von Einzelmodellen und den Ensembles treten selten auf und könnten durch eine generelle Verwendung von ausschließlich relevanten Datenpunkten für das Einzelmodelltraining verhindert werden.

VI. Fazit und Ausblick

Abschließend kann gesagt werden, dass Active Learning in ML Anwendungen mit Entscheidungsbäumen eingesetzt werden kann. Für die Erstellung der Entscheidungsbäume im Trainingsprozess ist der GUIDE-Algorithmus besonders geeignet, da dieser neben der Minimierung von Verzerrungen bei der Trennungsmerkmalsauswahl auch mit fehlenden Werten sowie kategorischen Merkmalen umgehen kann. Eine geeignete Abfragestrategie für das Active Learning mit Entscheidungsbäumen ist der Query-By-Commitee Ansatz, welcher informative Datenpunkte für das Modell auf Basis der Abweichungen eines Ensembles aus Entscheidungsbäumen auswählt. Im Rahmen der Untersuchung weisen die Active Learning Ansätze ein höheres Genauigkeitspotential als eine



(a) Unsicherheiten im initialen Datensatz

(b) Unsicherheiten im angereicherten Datensatz

Abbildung 5: Unsicherheiten im Dateneingangsraum vor und nach aktiver Datenreicherung durch GUIDE-Ensemble

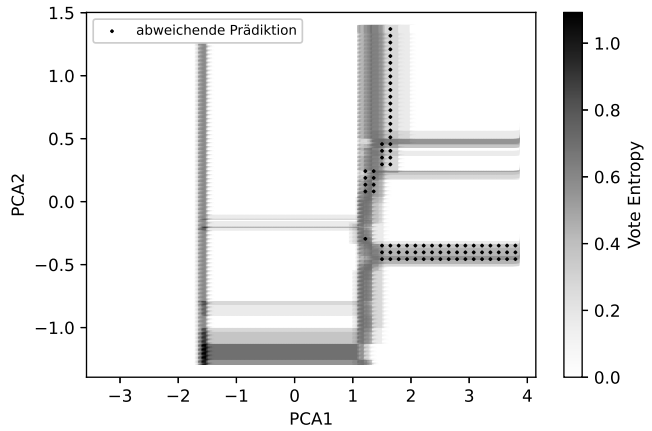


Abbildung 6: Abweichende Prädiktionen von Einzelmodell und Ensemble

zufällige Auswahl von Datenpunkten auf, obwohl ein Datensatz mit simplen Entscheidungsgrenzen verwendet wird. Die Unsicherheiten der Vorhersagemodelle im Eingangsdatenraum können mithilfe des Active Learnings präzise bestimmt werden, wodurch informative Abfragen aus einem Pool von Datenpunkten (pool-based sampling) oder von synthetisch im Datenraum gewählten Datenpunkten (membership query synthesis) erstellt werden können.

In weiteren Betrachtungen sollte dieses Vorgehen auf komplexere Datensätze aus industriellen Anwendungen überführt und getestet werden. Interessant ist ebenfalls die Betrachtung der Modellperformance, wenn die Trainingsdatenpunkte aktiv ausgewählt werden, im Vergleich zur Verwendung der gesamten Datenbasis zum Training. In der Industrie ist das Nachtrainieren des Modells mit einzelnen Datenpunkten aufgrund einer aktualisierten Datenbasis durch später durchführte Versuche üblich. Durch das instabile Verhalten von Entscheidungsbäumen beim Training mit leicht unterschiedlichen Datenpunkten gilt es, inkrementell aufgebaute Entscheidungsbäume zu nutzen und die Unterschiede zu quantifizieren.

LITERATUR

[1] K. Guo, Z. Yang, C.-H. Yu, and M. J. Buehler, "Artificial intelligence and machine learning in design of mechanical materials," vol. 8, no. 4, pp. 1153–1172. [Online]. Available: http://xlink.rsc.org/?DOI=D0MH01451F

[2] P. Larrañaga, D. Atienza Alonso, J. Diaz-Rozo, A. Ogbechie, C. Puerto-Santana, and C. Bielza, *Industrial Applications of Machine Learning*, first issued in paperback ed., ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, Taylor & Francis Group.

[3] A. Adadi, "A survey on data-efficient algorithms in big data era," vol. 8, no. 1, p. 24. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00419-9

[4] B. Settles, "Active Learning Literature Survey."

[5] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," vol. 15, no. 2, pp. 201–221. [Online]. Available: http://link.springer.com/10.1007/BF00993277

[6] A. Tharwat and W. Schenck, "A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions," vol. 11, no. 4, p. 820. [Online]. Available: https://www.mdpi.com/2227-7390/11/4/820

[7] D. Reker and G. Schneider, "Active-learning strategies in computer-assisted drug discovery," vol. 20, no. 4, pp. 458–465. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1359644614004735

[8] J. Yang, G. Lan, Y. Li, Y. Gong, Z. Zhang, and S. Ercisli, "Data quality assessment and analysis for pest identification in smart agriculture," vol. 103, p. 108322. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0045790622005444

[9] P. Peng, W. Zhang, Y. Zhang, Y. Xu, H. Wang, and H. Zhang, "Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis," vol. 407, pp. 232–245. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231220306603

[10] R. Karimi, A. Nanopoulos, and L. Schmidt-Thieme, "A supervised active learning framework for recommender systems based on decision trees," vol. 25, no. 1, pp. 39–64. [Online]. Available: http://link.springer.com/10.1007/s11257-014-9153-z

[11] R. De Rosa and N. Cesa-Bianchi, "Confidence Decision Trees via Online and Active Learning for Streaming Data," vol. 60, pp. 1031–1055. [Online]. Available: https://jair.org/index.php/jair/article/view/11102

[12] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, pp. 351–360. [Online]. Available: https://dl.acm.org/doi/10.1145/1772690.1772727

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. Springer New York. [Online]. Available: http://link.springer.com/10.1007/978-0-387-84858-7

[14] W. Loh, "Classification and regression trees," vol. 1, no. 1, pp. 14–23. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/widm.8

[15] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, second edition ed. O'Reilly Media, Inc.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Routledge. [Online]. Available: https://www.taylorfrancis.com/books/9781351460491

[17] J. Sonquist and J. Morgan, *The Detection of Interaction Effects: A Report on a Computer Program for the Selection of Optimal Combinations of Explanatory Variables*, ser. (Monograph. Survey Research Center, Institute for Social Research, the University of Michigan). Survey Research Center, Institute for Social Research, University of Michigan. [Online]. Available: https://books.google.de/books?id=0bxOvuPoQ3wC

[18] J. R. Quinlan, "Induction of decision trees," vol. 1, no. 1, pp. 81–106. [Online]. Available: http://link.springer.com/10.1007/BF00116251

[19] ——, "C4. 5: Programs for machine learning," Morgan Kaufmann Publishers Inc.

[20] W.-Y. Loh, "Regression Trees with Unbiased Variable Selection and Interaction Detection."

[21] ——, "Improving the precision of classification trees," vol. 3, no. 4. [Online]. Available: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-3/issue-4/Improving-the-precision-of-classification-trees/10.1214/09-AOAS260.full

[22] E. B. Baum and K. Lang, "Query learning can work poorly when a human oracle is used," in *International Joint Conference on Neural Networks*, vol. 8. Beijing China, p. 8.

[23] P. Kumar and A. Gupta, "Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey," vol. 35, no. 4, pp. 913–945. [Online]. Available: https://link.springer.com/10.1007/s11390-020-9487-4

[24] M. Sharma and M. Bilgic, "Evidence-based uncertainty sampling for active learning," vol. 31, no. 1, pp. 164–202. [Online]. Available: http://link.springer.com/10.1007/s10618-016-0460-3

[25] D. D. Lewis, T. B. Laboratories, and M. Hill, "A Sequential Algorithm for Training Text Classi ers."

[26] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing. [Online]. Available: https://link.springer.com/10.1007/978-3-031-01560-1

[27] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98. Morgan Kaufmann Publishers Inc., pp. 1–9.

[28] L. Breiman, "Bagging predictors," vol. 24, no. 2, pp. 123–140. [Online]. Available: http://link.springer.com/10.1007/BF00058655

[29] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," vol. 7, no. 2, pp. 179–188. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x