Einsatzmöglichkeiten der topologischen Datenanalyse zur Verbesserung von Modellen und des Modellverständnisses

Martin Alexander Wilk

Hochschule Darmstadt, Fachbereich Mathematik & Naturwissenschaften und Fachbereich Informatik

Problemstellung

Die entstandene Masterthesis entstand in Kooperation mit Dr. Jochen Papenbrock des Fintechs Firamis. Der Schwerpunkt dieses in Frankfurt ansässigen Unternehmens besteht in der Anwendung künstlicher Intelligenz im Finanzbereich. Die Ayasdi Inc., ein kalifornisches Unternehmen, dessen Tätigkeitsbereich die Analyse von Daten mithilfe von Machine Learning umfasst, verwendet als Kerntechnologie die topologische Datenanalyse als komprimierte Darstellung der Form der Datensätze ihrer Kunden. In zahlreichen, auf der Webseite des Unternehmens Ayasdi einsehbaren Whitepapers, wird die topologische Datenanalyse mit bekannten Machine Learning Verfahren kombiniert, welches vielversprechende Ergebnisse zeigt.

In der Masterthesis soll nun herausgefunden werden, ob die Kombination der topologischen Datenanalyse mit Machine Learning Verfahren nach dem Vorbild Ayasdi im Bereich Finance erfolgreich anwendbar ist. Hierzu soll die Performance herkömmlicher Klassifikationsverfahren durch die topologische Datenanalyse verbessert werden. Zur Zeit erfreuen sich Modelle aus dem Deep Learning wegen ihrer oftmals hohen Qualität großer Beliebtheit, allerdings sind diese so komplex, dass eine Interpretierbarkeit oftmals nicht gegeben ist. Nach einer Vorgabe der BaFin sind alle auf Basis von Modellen getroffenen Entscheidungen zu begründen. Dadurch entsteht ein Bedarf nach erklärbaren Modellen. Leicht erklärbare Modelle weisen jedoch oftmals eine geringere Performance auf, weswegen eine Anhebung der Modellqualität gewünscht ist.

Zur Lösung des beschriebenen Problems wurde im Rahmen der Masterthesis ein Prozess basierend auf der topologischen Datenanalyse erarbeitet, welcher exemplarisch auf drei Datensätzen getestet wurden. Die Themenbereiche dieser Datensätze sind die Erkennung von Kreditkartenbetrug und der Vorhersage, ob ein Kredit ausfällt.

Vorgehen

Als Lösung der Problemstellung wurde der Prozess aus Abbildung 1 entworfen, welcher die topologische Datenanalyse nutzt, um eine Verbesserung der Modellqualität herbeizuführen. Die topologische Datenanalyse ist ein Mittel zur komprimierten Darstellung der Form (engl. *shape*) eines Datensatzes. Daher scheint eine Zunahme der Performance eines Entscheidungsbaumklassifikators oder einer logistischen Regression durch die in Abbildung 2 dargestellten Schritte plausibel.

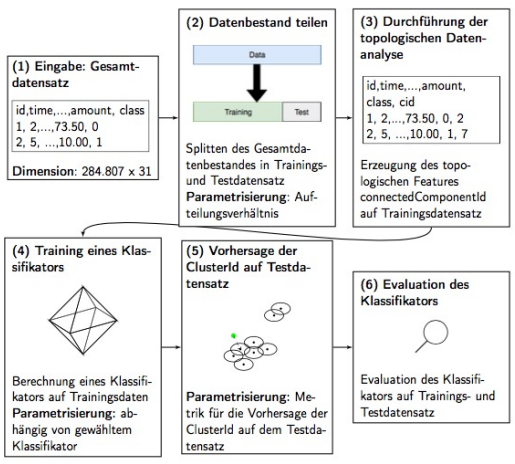

Abbildung 1:Schematische Darstellung des Gesamtprozesses der Masterthesis

Die Erzeugung des neuen topologischen Features, welche dem 6. Schritt des in Abbildung 2 dargestellten Prozesses entspricht, erfolgt durch eine Analyse des erstellten Mapper Graphen hinsichtlich der Zusammenhangskomponenten. Als neues Feature wird jeweils die ID der zusammenhängenden Komponente des Nodes des Mapper Graphens verwendet, in dem der jeweilige Datensatz liegt. Die Linse, welche in Schritt 2 des Prozesses der topologischen Datenanalyse bestimmt wird, soll möglichst eine Trennung der Objekte unterschiedlicher Klassen (wie zum Beispiel normale und betrügerische Transaktionen) vornehmen, sodass die Nodes des Graphen möglichst aus Beobachtungen mit gleichem Targetwert bestehen.

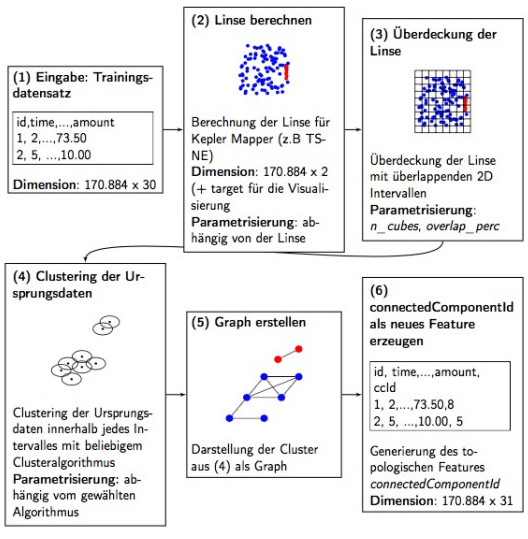

Abbildung 2:Schematische Darstellung des Prozesses zur Bestimmung des neuen topologischen Features

Die Bestimmung der Ergebnisse erfolgt in Abhängigkeit der Linsen lineare Diskriminanzanalyse, Kombination aus Isolation Forest und $\boldsymbol{L_2}$-Norm, und die beiden Verfahren zur Dimensionsreduktion TSNE und UMAP.

Ergebnisse

Die Anwendung des erarbeiteten Prozesses erfolgt auf drei Datensätzen. Aus Platzgründen sind in diesem Poster die Veränderungen der Performance durch die Hinzunahme des topologischen Features beim Kreditkartentransaktionsdatensatz und dem Lending Club Datensatz dargestellt.

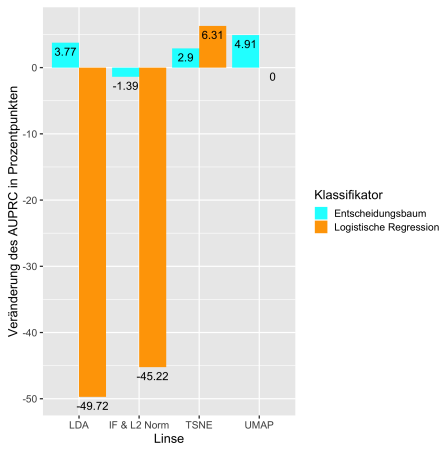

Abbildung 3:Kreditkartendatensatz: Veränderungen des Test-AUPRC durch topologisches Feature. Baseline: 78,06% (Entscheidungsbaum), 50,22% logistische Regression

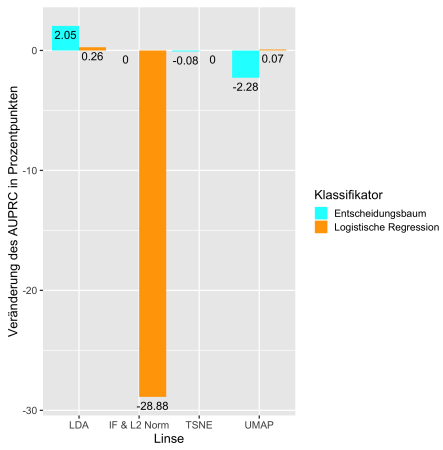

Abbildung 4:Lending Club Datensatz: Veränderungen des Test-AUPRC durch topologisches Feature. Baseline: 76,44% (Entscheidungsbaum), 75,44% logistische Regression

Alle durchgeführten Untersuchungen zeigen, dass der vorgestellte Prozess in einigen Fällen zu einer Verbesserung des Entscheidungsbaumklassifikators führen kann. Eine ausgeprägte Zunahme der Performance der logistischen Regression zeigt sich nur beim Kreditkartentransaktionsdatensatz mit der TSNE Linse. Beim Lending Club Datensatz führt der vorgestellte Prozess nicht zu einer Zunahme der Performance der Klassifikatoren. Die topologische Datenanalyse besitzt viele Parameter, für die keine Optimierungsstrategien zur Verfügung stehen. Dies erfordert die Suche passender Parameterwerte auf manuellem Weg durch Experimente.

Implementierung

Die Masterthesis schließt eine Implementierung des vorgestellten Gesamtprozesses mit ein, da sich die topologische Datenanalyse als parameterintensiv herausstellt. Ein interaktiver Prototyp, welcher auf die Eingaben des Benutzers durch Visualisierungen der Linse beziehungsweise des Mapper Graphen reagiert, dient zur experimentellen Bestimmung optimaler Einstellungen. Das Design des Prototyps begründet sich auf die Flussdiagramme 1 und 2. Abbildung 5 zeigt den Ablauf der Bestimmung des topologischen Features in der interaktiven Software, da dieser Schritt des Prozesses zentral ist.

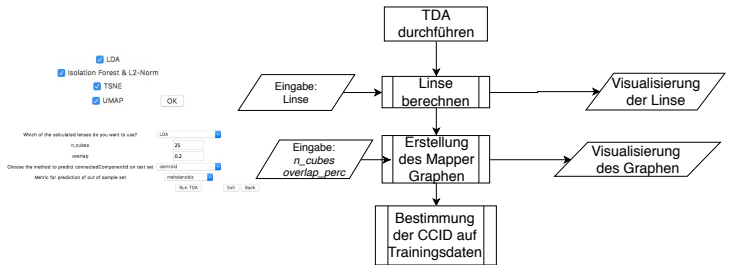

Abbildung 5:Prozessablauf der Erzeugung des topologischen Features auf dem Trainingsdatensatz

Ausblick

Zur Weiterentwicklung des vorgestellten Ansatzes ist die Klärung der folgenden Fragen sinnvoll:

- Existiert eine Strategie zur Findung sinnvoller Parameter?
- Gibt es Eigenschaften des Datensatzes, welche auf eine passende Linse hindeuten?
- Kann man anhand des Datensatzes erkennen, ob die topologische Datenanalyse angewendet werden sollte?
- Welche anderen Methoden zur Extraktion von Informationen aus dem Mapper Graph gibt es?