Hauptkomponentenanalyse

Grundlagen und Anwendung in Machine Learning bei Versicherungen

Zusammenfassung der Forschungsarbeit an der Universität Ulm

Tim Remiger

Motivation und Gegenstand der Arbeit

Die Motivation für diese Arbeit liegt im Wesentlichen begründet in der zunehmenden Relevanz für Versicherer sehr große Datenmengen effizient zu verarbeiten. Dies führt zur Generierung von umfangreichen Datensätzen, welche für einen Menschen schwer zu verstehen sind und deren rechenintensive Weiterverarbeitung mit Kosten verbunden ist. Eine Möglichkeit diesen Herausforderungen zu begegnen, ist eine Reduktion der verfügbaren Daten. Gegenstand dieser Arbeit ist die Hauptkomponentenanalyse, durch welche sich die zu verarbeitende Datenmenge verringern lässt, indem die Dimension einzelner Datensätze reduziert wird.

Theoretische Grundlagen und „klassische“ Anwendung der Hauptkomponentenanalyse

Zunächst wird die mathematische Theorie der Hauptkomponentenanalyse erläutert. Dabei werden die Hauptkomponenten anhand von Zufallsvektoren hergeleitet und relevante Eigenschaften für reale Daten angeführt. Die Unkorreliertheit der Hauptkomponenten wird gezeigt und geometrische Eigenschaften beschrieben, welche die Anwendung der Hauptkomponentenanalyse für reale Daten begründen. Anhand eines Datensatzes mit Verträgen aus der privaten Krankenversicherung wird die erarbeitete Theorie praktisch angewendet. Es wird erklärt, wie Hauptkomponenten allgemein interpretiert werden können und wie ein Datensatz für eine Hauptkomponentenanalyse vorbereitet werden muss. Darauf aufbauend wird versucht durch eine Dimensionsreduktion ein besseres Verständnis für den betrachteten Datensatz aus Sicht eines Versicherers zu schaffen. Dabei wird deutlich, dass eine Hauptkomponentenanalyse ein einfaches und schnell anzuwendendes Werkzeug zur Datenanalyse darstellt. Gleichzeitig zeigt sich auch, dass der Wert einer

Analyse, welche ausschließlich auf den Hauptkomponenten basiert, abhängt von den speziellen Strukturen in den Daten. So führt eine Hauptkomponentenanalyse für den betrachteten Datensatz nur bedingt zu aussagekräftigen Erkenntnissen, da eine anschauliche Betrachtung von wenigen Hauptkomponenten hier nicht ausreicht, um die Daten umfassend zu erklären. Dies ist unter anderem darauf zurückzuführen, dass der betrachtete Datensatz bereits aus einer von Experten getroffenen Auswahl von Merkmalen besteht. Die Ergebnisse der Hauptkomponentenanalyse können somit als Bestätigung dieser Auswahl angesehen werden.

Hauptkomponenten in Verbindung mit Entscheidungsbäumen

Über den „konventionellen“ Einsatz der Hauptkomponentenanalyse hinaus geht die Betrachtung von Hauptkomponenten in Verbindung mit einem modernen statistischen Verfahren.
Im Speziellen werden hier sogenannte „Entscheidungsbäume“ für ein binäres Klassifizierungsproblem untersucht. Für den betrachteten Datensatz werden solche Modelle erstellt, um Stornierungen von Verträgen zu prognostizieren. Dabei werden vor allem die Auswirkungen einer Dimensionsreduktion mittels der Hauptkomponentenanalyse auf die Güte der Vorhersagen sowie die benötigte Zeit zur Erstellung der Modelle untersucht. Insbesondere soll damit geklärt werden, inwieweit es möglich ist durch eine Reduktion der Daten in kürzerer Zeit Entscheidungsbäume zu erstellen, deren Vorhersagequalität vergleichbar ist mit der Güte von Entscheidungsbäumen für nicht reduzierte Daten.
Es stellt sich heraus, dass Entscheidungsbäume, welche auf Hauptkomponenten trainiert werden, tendenziell etwas schlechtere Vorhersagen liefern als Entscheidungsbäume mit den entsprechenden untransformierten Daten. Eine Hauptkomponentenanalyse scheint für den vorliegenden Datensatz also zu Strukturen zu führen, welche sich negativ auf die Qualität von Entscheidungsbäumen auswirken. Insbesondere ergibt sich daraus, dass die Güte der Vorhersagen von Entscheidungsbäumen im Allgemeinen nicht invariant unter orthogonalen Transformationen der Daten ist.
Andererseits ergibt sich für das betrachtete Problem, dass es durch eine Hauptkomponentenanalyse möglich ist, die Daten soweit zu reduzieren, dass Entscheidungsbäume doppelt so schnell erstellt werden können

und dabei nur geringfügig schlechtere Vorhersagen liefern. Konkret lässt sich der betrachtete Datensatz auf ein Viertel seiner ursprünglichen Dimension reduzieren ohne erhebliche Einbußen in der Güte der erstellten Entscheidungsbäume in Kauf nehmen zu müssen.
Dies kann zum Beispiel für sehr umfassende Datensätze genutzt werden, um in kürzerer Zeit Entscheidungsbäume zu trainieren, welche als Vergleichsmaßstab für die Erstellung weiterer Modelle dienen können. Besonders hervorzuheben ist dabei, dass durch die Verwendung der Hauptkomponenten eine zeitintensive Merkmalsauswahl durch Experten nicht zwingend erfolgen muss. So werden etwa Multikollinearitäten durch die Hauptkomponentenanalyse erkannt und durch Hauptkomponenten repräsentiert, welche nicht für die Erstellung der Modelle verwendet werden.
Das präsentierte Vorgehen lässt sich damit als „off-the-shelf“-Methode zur Erstellung einfacher Modelle einsetzen, für welche ein Datensatz nur in geringem Maße vorbereitet werden muss.

Illustration der Trainingsmethodik und Vergleich der Hauptkomponenten von zufälligen Teildatensätzen

Das Vorgehen zur Generierung der beschriebenen Ergebnisse wird in der Arbeit ausführlich dokumentiert. Insbesondere wird die verwendete Trainingsmethodik zur Erstellung von Modellen mit Hauptkomponenten allgemein veranschaulicht und illustriert. Die Untersuchungen können somit für beliebige Klassifizierungsverfahren analog durchgeführt werden.
Zudem werden die Hauptkomponenten verschiedener zufällig gewählter Teildatensätze verglichen. Diese Betrachtungen sind relevant für die Aussagekraft der präsentierten Ergebnisse, da die erstellten Entscheidungsbäume mit unterschiedlichen Teildatensätzen trainiert und bewertet werden. Damit soll sichergestellt werden, dass Resultate nicht auf die zufälligen Aufteilungen des Datensatzes zurückgeführt werden können. Es stellt sich für die betrachteten Daten heraus, dass die ersten sieben Hauptkomponenten von zufällig gewählten Trainingsdatensätzen einem einheitlichen Muster folgen, während es mitunter große Unterschiede bezüglich der weiteren Hauptkomponenten gibt. Demnach lassen sich anhand von Modellen, welche mit den ersten sieben Hauptkomponenten erstellt werden, besonders gute Aussagen zu den Auswirkungen einer

Dimensionsreduktion mittels Hauptkomponenten auf Entscheidungsbäume treffen.

Die Arbeit schließt mit der Formulierung einer Idee zur Weiterführung des Einsatzes der Hauptkomponentenanalyse in Verbindung mit Entscheidungsbäumen. Es kann untersucht werden, ob durch eine Vereinfachung der tatsächlichen Hauptkomponenten Strukturen entstehen, welche sich günstig auf die Vorhersagequalität von Entscheidungsbäumen auswirken. Als mögliche Verfahren zur Vereinfachung der Hauptkomponenten werden die „Varimax“- und „Quartimax“-Rotation genannt.