

> Genauere Ergebnisse durch verbesserte Datenaufbereitung

Vor der Analyse steht die Datenaufbereitung. PASW Statistics Base* enthält Werkzeuge für die Datenaufbereitung, doch benötigen Sie in manchen Fällen speziellere Verfahren, um Ihre Daten für die Analyse vorzubereiten. Mit dem Zusatzmodul PASW Data Preparation* können Sie verdächtige und ungültige Fälle, Variablen und Datenwerte ermitteln, die Muster fehlender Daten anzeigen und Variablenverteilungen auswerten. Außerdem können Sie mit besseren Algorithmen arbeiten, die auf nominal skalierte Variablen ausgerichtet sind. Auf diese Weise wird die Datenaufbereitung optimiert, sodass Sie Analysen schneller vorbereiten und daraus genauere Schlussfolgerungen erzielen können.

PASW Data Preparation ist als Client-Software zur Installation verfügbar. Zur Steigerung der Leistung und Skalierbarkeit kann es jedoch auch in einer Client-/Server-Installation mit PASW Statistics Server* verwendet werden.

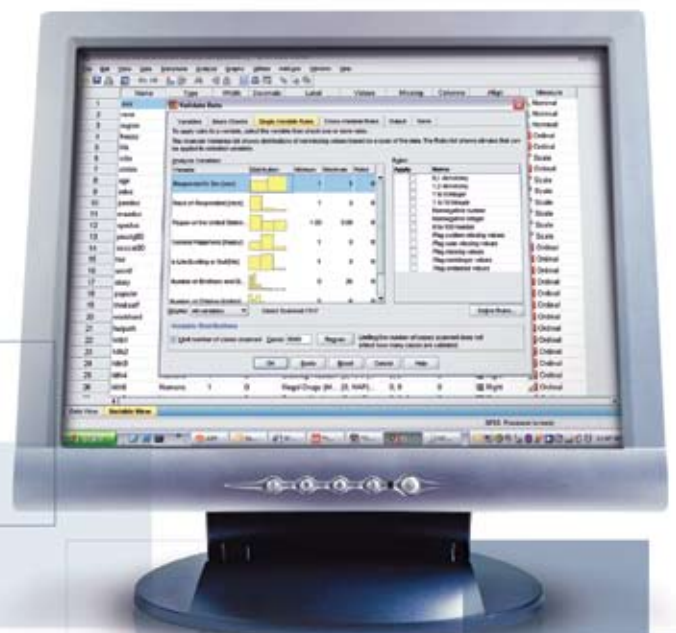
Prüfen der Daten

Die Datenvalidierung ist häufig ein rein manueller Vorgang. So haben Sie beispielsweise Häufigkeiten für die Daten berechnet und anzeigen lassen, verdächtige Fälle angestrichen und über IDs geprüft. Dies ist natürlich sehr zeitraubend. Und da die Analysten in Ihrer Organisation möglicherweise unterschiedliche Methoden verwenden, kann es eine Herausforderung darstellen, die Einheitlichkeit zwischen einzelnen Projekten sicherzustellen.

* PASW Data Preparation, PASW Statistics Base, und PASW Statistics Base Server, früher bekannt als SPSS Data Preparation™, SPSS Statistics Base, und SPSS Statistics Base Server, sind Teil des SPSS Inc.'s Predictive Analytics Software Portfolios.

Sie können sich das manuelle Prüfen ersparen, indem Sie die Prozedur „Daten validieren“ verwenden. Diese Prozedur ermöglicht es, anhand des Messniveaus der einzelnen Variablen (kategorial oder stetig) Validierungsregeln auf die Daten anzuwenden. Wenn Sie beispielsweise Umfragedaten analysieren, die Variablen auf einer fünfstufigen Likert-Skala enthalten, wenden Sie mit der Prozedur „Daten validieren“ eine Regel für Fünf-Punkt-Skalen an und markieren alle Fälle mit Werten außerhalb des Bereichs 1–5. Sie können sich die ungültigen Fälle sowie Zusammenfassungen der Regelverletzungen und die Anzahl der betroffenen Fälle anzeigen lassen. Sie können Validierungsregeln für einzelne Variablen (z. B. Bereichsprüfungen) und für mehrere Variablen (z. B. „schwangere Männer“) angeben.

Anhand dieser Informationen können Sie die Validität der Daten feststellen und fragwürdige Fälle vor der Analyse ggf. entfernen oder korrigieren.



Schnelles Auffinden multivariater Ausreißer

Mit der Prozedur „Anomalie-Erkennung“ können Sie verhindern, dass Ausreißer eine Analyse verzerren. Diese Prozedur sucht die Fälle, die verhältnismäßig stark von den anderen Fällen abweichen, und es werden Gründe für solche Abweichungen angegeben. Sie können Ausreißer markieren, indem Sie eine neue Variable erstellen. Wenn Sie ungewöhnliche Fälle festgestellt haben, können Sie diese weiter untersuchen und bestimmen, ob sie weiterhin in die Analysen einbezogen werden sollen.

Vorverarbeiten der Daten vor der Modellerstellung

Um Algorithmen verwenden zu können, die für nominal skalierte Variablen (z. B. Naïve Bayes- und Logit-Modelle) konzipiert sind, müssen Sie metrische Variablen vor der Modellerstellung in Klassen einteilen. Wenn metrische Variablen nicht in Klassen eingeteilt sind, nimmt die Verarbeitung von Algorithmen wie der multinominalen logistischen Regression eine extrem lange Zeit in Anspruch, oder der Algorithmus konvergiert möglicherweise gar nicht. Dies gilt insbesondere bei großen Datensätzen. Außerdem sind die Ergebnisse möglicherweise schwierig zu deuten oder zu interpretieren.

Mit Optimal Binning können Sie jedoch die Trennwerte bestimmen, mit denen in Algorithmen für nominal skalierte Variablen das bestmögliche Ergebnis erzielt wird.

Bei dieser Prozedur können Sie aus drei Möglichkeiten zur Einteilung zum Aufbereiten der Daten vor dem Erstellen eines Modells auswählen:

- **Unüberwacht** – Es werden Klassen mit gleicher Häufigkeit erstellt.
- **Überwacht** – Die Zielvariable wird zum Bestimmen der Trennwerte hinzugezogen. Diese Methode ist genauer als die unüberwachte Methode, jedoch auch rechenintensiver.
- **Hybrider Ansatz** – Die unüberwachte und die überwachte Methode werden kombiniert. Diese Methode eignet sich besonders, wenn viele einzelne Werte vorliegen.



Mit Optimal Binning werden bei Algorithmen für nominale Attribute exaktere Ergebnisse erzielt.



Funktionen

Daten validieren

Mit der Prozedur „Daten validieren“ werden die Daten in der Arbeitsdatei validiert.

- Grundlegende Prüfungen: Sie können grundlegende Prüfungen angeben, die auf Variablen und Fälle in der Datei angewendet werden sollen. Beispielsweise können Berichte zu Variablen mit einem hohen Prozentsatz fehlender Werte oder leerer Fälle abgerufen werden.
 - Maximaler Prozentsatz fehlender Werte
 - Maximaler Prozentsatz der Fälle in einer einzelnen Kategorie
 - Maximaler Prozentsatz der Kategorien mit Anzahl 1
 - Minimaler Variationskoeffizient
 - Minimale Standardabweichung
 - Unvollständige IDs markieren
 - Doppelte IDs markieren
 - Leere Fälle markieren
- Standardregeln: Deskriptiv analysieren, Regeln für eine Variable anzeigen und diese auf Analysevariablen anwenden.
 - Deskriptive Analyse:
 - Verteilung: Anzeige eines Balkendiagramms in Miniaturgröße für kategoriale Variablen oder eines Histogramms für metrische Variablen
 - Minimale und maximale Datenwerte werden gezeigt
 - Regeln für eine Variable:
 - Anwenden von Regeln auf einzelne Variablen, um fehlende oder ungültige Werte zu ermitteln, z. B. Werte außerhalb eines zulässigen Bereichs
 - Benutzerdefinierte Regeln für eine Variable sind ebenfalls möglich.
- Benutzerdefinierte Regeln: Sie können Ausdrücke zu Regeln für mehrere Variablen für Fälle definieren, in denen die Antworten der Befragten die Logik verletzen (z. B. „schwängere Männer“).
- Ausgabe: Berichte mit Beschreibungen ungültiger Daten
 - Fallweiser Bericht, in dem die Verletzungen von Validierungsregeln nach Fällen aufgezeichnet sind
 - Sie können die Mindestzahl von Verletzungen angeben, die vorliegen müssen, damit ein Fall in den Bericht aufgenommen wird.
 - Sie können die Höchstzahl von Fällen im Bericht festlegen.

- Standardberichte zu Validierungsregeln
 - Zusammenfassung der Verletzungen nach Analysevariable
 - Zusammenfassung der Verletzungen nach Regel
 - Anzeigen deskriptiver Statistiken für Analysevariablen
- Speichern: Hiermit können Sie Variablen speichern, die Regelverletzungen aufzeichnen und zur Datenbereinigung verwenden.
 - Auswertungsvariablen:
 - Indikator für leere Fälle
 - Indikator für doppelte IDs
 - Indikator für unvollständige IDs
 - Verletzungen von Validierungsregeln (Gesamtzahl)
 - Indikatorvariablen, die alle Verletzungen von Validierungsregeln aufzeichnen

Ungewöhnliche Fälle identifizieren

Mit der Prozedur „Anomalie-Erkennung“ wird anhand von Abweichungen in der jeweiligen Gruppe nach ungewöhnlichen Fällen gesucht, und es werden Gründe für solche Abweichungen angegeben.

- Mit dem Unterbefehl VARIABLES geben Sie die von der Prozedur zu verwendenden Variablen an. Sie können kategoriale, stetige und ID-Variablen (zum Identifizieren von Fällen) verwenden und die Variablen festlegen, die aus der Analyse ausgeschlossen werden sollen.
- Mit dem Unterbefehl HANDLEMISSING geben Sie die Methoden für die Behandlung fehlender Werte in dieser Prozedur an.
 - Behandeln fehlender Werte. Wenn diese Option ausgewählt ist, werden fehlende Werte von stetigen Variablen durch deren Gesamtmittelwert ersetzt, und fehlende Kategorien von kategorialen Variablen werden gruppiert und als gültige Kategorie behandelt. Die verarbeiteten Variablen werden anschließend in der Analyse verwendet. Wenn diese Option nicht ausgewählt ist, werden Fälle mit fehlenden Werten aus der Analyse ausgeschlossen.
 - Erstellen einer Variablen für den Anteil fehlender Werte. Wenn diese Option ausgewählt ist, wird eine zusätzliche Variable für den Anteil fehlender Werte erstellt, die den Anteil der fehlenden Variablen in jedem Datensatz darstellt. Diese Variable wird dann in der Analyse verwendet. Wenn die Option nicht ausgewählt ist, wird die Variable für den fehlenden Anteil nicht erstellt.
- Mit dem Unterbefehl CRITERIA können Sie die folgenden Einstellungen angeben:
 - Die minimale und maximale Anzahl von Gruppen
 - Gewichtung in Abhängigkeit des Messniveaus
 - Anzahl von Gründen in der Anomalie-Liste
 - Prozentsatz der Fälle, die als Anomalien behandelt und in die Anomalie-Liste aufgenommen werden
 - Anzahl der Fälle, die als Anomalien behandelt und in die Anomalie-Liste aufgenommen werden
 - Trennwert des Anomalie-Index, anhand dessen bestimmt wird, ob ein Fall als Anomalie behandelt wird
- Mit dem Unterbefehl SAVE können Sie zusätzliche Variablen in der Arbeitsdatei speichern.
 - Anomalie-Index
 - Gruppen-ID
 - Gruppengröße
 - Gruppengröße in Prozent
 - Die Variable, der ein Grund zugeordnet ist
 - Das Variablen-Einflussmaß, dem ein Grund zugeordnet ist
 - Der Variablenwert, dem ein Grund zugeordnet ist
 - Der Normwert, dem ein Grund zugeordnet ist
- Mit dem Unterbefehl OUTFILE lässt sich das Modell im XML-Format unter einem bestimmten Dateinamen speichern.
- Mit dem Unterbefehl PRINT steuern Sie die Anzeige der ausgegebenen Ergebnisse. Folgende Ausgaben sind möglich:
 - Zusammenfassung der Fallverarbeitung
 - Liste der Anomalie-Indizes, Liste der Anomalie-Gruppen-IDs und Liste der Anomalie-Gründe
 - Die Tabelle „Normwerte der stetigen Variablen“, wenn die Analyse stetige Variablen umfasst, und die Tabelle „Normwerte der kategorialen Variablen“, wenn die Analyse kategoriale Variable umfasst
 - Auswertung des Anomalie-Index
 - Tabelle „Auswertung der Gründe“ für jeden Grund
 - Die gesamte Ausgabe mit Ausnahme der Anmerkungstabelle und Warnungen kann unterdrückt werden.

Optimal Binning

Sie können die Daten mit der Prozedur „Optimal Binning“ aufbereiten. Hiermit werden eine oder mehrere stetige Variablen durch Einteilen der Werte jeder Variablen in Klassen kategorisiert. Diese Prozedur ist nützlich, um die Anzahl der in den angegebenen Eingabevariablen vorhandenen Werte zu reduzieren. Das kann die Performanz bei den Algorithmen beträchtlich steigern. Bei der Verwendung bestimmter Optimal Binning-Methoden wird zum Bestimmen der Trennwerte eine Zielvariable verwendet. So wird die Beziehung zwischen der Zielvariablen und der einzuteilenden Variablen deutlich herausgestellt.

- Wählen Sie aus folgenden Statistiken:
 - Unüberwachte Einteilung in Klassen durch den Algorithmus für gleiche Häufigkeiten. Bei dieser Methode werden die einzuteilenden Eingabevariablen mit dem Algorithmus für gleiche Häufigkeiten klassifiziert. Eine Zielvariable ist nicht erforderlich.
 - Überwachte Einteilung in Klassen durch den MDLP-Algorithmus (Minimal Description Length Principle). Bei dieser Methode werden die einzuteilenden Eingabevariablen mit dem MDLP-Algorithmus ohne Vorverarbeitung klassifiziert. Diese Methode eignet sich für Datensätze mit einer kleinen Anzahl von Fällen. Eine Zielvariable ist erforderlich.

- Hybrider MDLP-Ansatz. Hierbei erfolgt eine Vorverarbeitung mit dem Algorithmus für gleiche Häufigkeiten und dann die Einteilung mit dem MDLP-Algorithmus. Diese Methode eignet sich für Datensätze mit einer großen Anzahl von Fällen. Eine Zielvariable ist erforderlich.
- Geben Sie die folgenden Kriterien an:
 - Wie wird der minimale Trennwert für jede Eingabevariable definiert?
 - Wie wird der maximale Trennwert für jede Eingabevariable definiert?
 - Wie wird die Untergrenze eines Intervalls definiert?
 - Soll das Zusammenführen von gering besetzten Klassen erzwungen werden?
 - Sollen fehlende Werte listenweise oder paarweise ausgeschlossen werden?
- Folgende Elemente können gespeichert werden:
 - Neue Variablen mit in Klassen eingeteilten Werten
 - Syntax in einer PASW Statistics Base-Syntaxdatei

- Mit dem Unterbefehl PRINT steuern Sie die Anzeige der ausgegebenen Ergebnisse. Folgende Ausgaben sind möglich:
 - Endpunkte der Klassen
 - Deskriptive Statistiken für alle Eingabevariablen
 - Modellentropie für in Klassen eingeteilte Variablen

Systemanforderungen

- Software: PASW Statistics Base* 17.0
- Andere Systemanforderungen können je nach Plattform abweichen.

Die endgültige Version kann geänderte Funktionen enthalten.



Weitere Informationen erhalten Sie auf unserer Website www.spss.de.
SPSS GmbH Software – Theresienhöhe 13 –
80339 München – Tel. +49.89 48 90 74-0, Fax +49.89.448 31 15.

SPSS ist eine eingetragene Marke, und alle weiteren genannten SPSS Inc.-Produkte sind Marken von SPSS Inc. Alle anderen Namen sind Marken ihrer jeweiligen Eigentümer.
© 2009 SPSS Inc. Alle Rechte vorbehalten.
SDP1702SPC-0209-DE