Data Mining und Predictive Process Control

Michael Brugger
0523161

Seminararbeit
erstellt im Rahmen der LV
Seminar aus Informatik, SS 2010
LV-Leiter: Prof. Wolfgang Pree

Universität Salzburg, Masterstudim Angewandte Informatik

July 4, 2010

Inhaltsverzeichnis

1 Einführung

"The Handbook of Statistical Analysis and Data Mining Applications" als Grundlage sowohl quantitativer als auch qualitativer Aspekte für das Ableiten und die Generierung von Information und Wissen aus konsolidierten und transformierten Datenbeständen. Dieses ehrgeizige Ziel wird diesem Buch zugrunde gelegt. Dabei wird unter anderem ein Überblick bezüglich analytischer Aspekete der Datenakquisition, Datenaufbereitung und statistischen Auswertung von Betriebsdaten in Produktionsanlagen, als Basis einer kontinuierlichen Prozessoptimierung gewährt und damit ein Einblick in die Komplexität und Beeinflussbarkeit von produktionsrelevanten Parametern geleistet. Diese Thematik wird unter dem Begriff *Predictive Process Control* angesprochen. Die Autoren liefern mit der folgenden Kurzbeschreibung eine Anregung, welchen Fokus dieses Buch betrachten will und damit als Hilfestellung für interdisziplinäre Betrachtungen im Bereich der Wissensgenerierung aus konsolidierten Datenbeständen dienen kann:

... it is a comprehensive professional reference book that guides business analysts, scientists, engineers and researchers (both academic and industrial) through all stages of data analysis, model building and implementation. The Handbook helps one discern the technical and business problem, understand the strengths and weaknesses of modern data mining algorithms, and employ the right statistical methods for practical application. Use this book to address massive and complex datasets with novel statistical approaches and be able to objectively evaluate analyses and solutions. It has clear, intuitive explanations of the principles and tools for solving problems using modern analytic techniques, and discusses their application to real problems, in ways accessible and beneficial to practitioners across industries - from science and engineering, to medicine, academia and commerce ... [RN10]

Data Mining ist ein relativ junges Forschungsgebiet, die ersten bedeutenden Ansätze wurden in den 1990er Jahren getroffen. Es repräsentiert einen Zusammenschluss von verschiedenen, etablierten Forschungsgebieten wie:

- Statistische Analyse
- Künstliche Intelligenz
- Maschinelles Lernen
- Entwicklung von großen Datenbanken

Traditionelle statistische Analysen folgen dabei einer deduktiven Methode um Beziehungen in Datensätzen zu finden und zu erklären. Künstliche Intelligenz und Techniken des maschinellen Lernens (z.B. Expertensysteme, künstliche Neuronale Netze, Entscheidungsbäume, etc.) folgen der induktiven Methode um Patterns oder Abhängigkeiten in Datenbeständen zu fixieren.

2 Definition

Der Begriff Data Mining stammt ursprünglich aus dem Bereich der Statistik und kennzeichnet dort die selektive Methodenanwendung zur Bestätigung vorformulierter Hypothesen. Mit der Weiterentwicklung von Data Mining zur eigenen Disziplin ist es erforderlich, eine Unterscheidung mit unterschiedlicher Ausprägung für den Begriff *Data Mining* zu definieren. Dies erfolgt jedoch unter dem gemeinsamen Aspekt der Erforschung und Analyse großer Datenmengen mit automatischen oder halbautomatischen Werkzeugen, um bedeutungsvolle Muster und Regeln in Datensätzen aufzufinden:

- **Statistisches Modellieren:** die Verwendung von statistischen Algorithmen um ein Ereignis zu definieren oder voraus zu sagen, basierend auf der Verwendung von Prädiktor- Variablen
- **Data Mining:** die Verwendung von Algorithmen für maschinelles Lernen um Muster von Abhängigkeiten zwischen Datenelementen in großen, potentiell unvollständigen Datensätzen zu erkennen. Als Auslöser für Aktionen um Vorteile zu generieren (Diagnose, Erlös, Detektion, Produktionsqualität, etc.)
- **KDD (Knowledge Discovery in Databases):** beschreibt den umfassenden Prozess der Daten- Akquisition, Daten- Exploration, Daten-Aufbereitung, Modellierung, Anwenden und Auswerten der Modellierung

In nachfolgender Darstellung ist die Positionierung von Data Mining im Vergleich zu Knowledge Discovery in Databases als übergeordnete Instanz verdeutlicht:

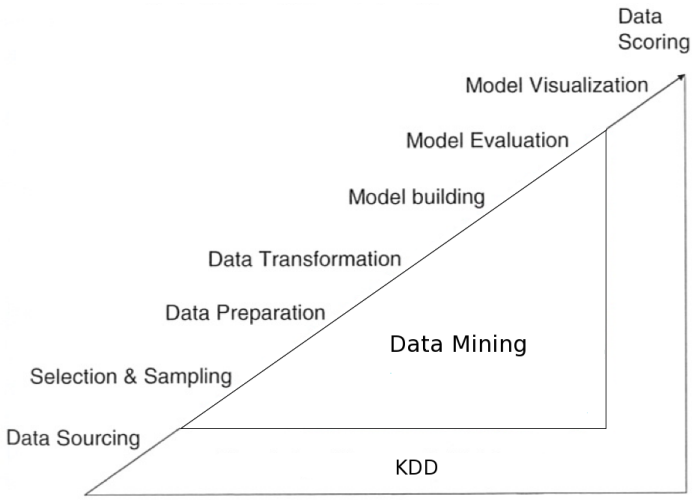

Figure 1: [RN10, Data Mining und Knowledge Discovery in Databases]

3 Predictive Process Control

... is an approach to identify variations of controllable parameters to stabilize within a manufacturing unit to maintain/ enhance quality ... [RN10]

Die Erhöhung des wirtschaftlichen Drucks verstärkt in produzierenden Industriesektoren zunehmend die Erfordernis an die Unternehmensleitung, Rationalisierungsreserven und Innovationsmöglichkeiten in allen Bereichen auszuschöpfen. Der Ansatz des Erfahrungskurveneffekts (jede Verdoppelung der im Zeitablauf kumulierten Erfahrungen, also der Produktionsmengen, senkt die auf die Wertschöpfung bezogenen und in konstanten Einheiten ausgedrückten Stückkosten auf 70%, 80% oder 90% des Ausgangswertes) liefert unter Betrachtung klassischer Parameter mit zunehmendem Reifegrad des jeweiligen Industriesektors jedoch nur mehr unzureichende Verbesserungsmaßnahmen wie:

- Produkt- und Verfahrensinnovation
- Mechanisierung und Automatisierung
- Erhöhung der Produktivität durch Verbesserung der Arbeitsorganisation

Es wird durch den bereits erreichten, sehr hohen Automatisierungs- und Optimierungsgrad vorausgesetzt in der Lage zu sein, zusätzlich und verstärkt innovative Instrumente einzusetzen um entscheidende Wettbewerbsvorteile erzielen zu können. Die Akquisition und Aufbereitung von Prozessdaten kann dabei von der einfachen Datenanalyse bis zur Verbesserung von vorhandenen Prozesswissen und Aufzeigen von versteckten Prozessabhängigkeiten reichen. Um entsprechendes Optimierungspotential generieren zu können, ist eine Verbindung von Datenerfassung auf technischer Ebene, vorhandener verfahrenstechnischer Expertise sowie der Organisation der gesammelten Daten mit Hilfe hochperformanter Datenstrukturen und der Auswertung mit Hilfe mathematischer Modelle zu realisieren. Die gewählte Methode soll dementsprechend die Möglichkeit bieten, aus gesammelten Daten eine umfassende Repräsentationsbasis für die Akquisition von Wissen zu ermöglichen. Predictive Process Control (PPC) ist ein Ansatz um Varianten von kontrollierbaren Parametern zu identifizieren um Prozesse innerhalb einer Produktionseinheit oder Fertigungsstufen zu stabilisieren um Qualitätsparameter halten - und erweitern zu können. PPC erklärt dabei fortgeschrittene Prädiktionsmodelle um Disfunktionalitäten innerhalb von Prozessen vorhersagen zu können. Je früher Funktionsstörungen erkannt werden, desto weniger Ausschussware fällt innerhalb eines Produktionscharge an. Dies wiederum führt zu einer Steigerung der Auslastung von Produktionsanlagen. Es wird oft beobachtet, dass die Qualität durch eine Verringerung der Variabilität von Prozessen und Rohmaterialien erhöht werden kann. Variabilität kann nur in statistischen Maßstäben beschrieben werden, daraus leitet sich die grundlegende Bedeutung von statistischen Methoden im Rahmen der Verbesserung von Qualitätsparametern ab.

3.1 Verstehen von Produktionsprozessen

Die Entwicklung von komplexen Produktionsprozessen und die Beherrschbarkeit von physikalischen Einflüssen und Seiteneffekten auf die Produktion erweist sich in der Praxis als langwierige Aufgabenstellung mit zumeist explorativem Charakter. Maschinenparameter wie Temperaturvorgaben, Drücke, Geschwindigkeiten, etc. werden oft in einem kostenintensiven "trial-and-error" Prozess ermittelt, um die Ergebnisse in Prototypen einfließen lassen zu können. Den Status einer qualitativ hochwertigen Produktion mit geringem Ausschuss und hoher Produktgüte in dieser suboptimalen Umgebung zu erreichen, erweist sich als sehr schwierig. Zu betrachtende Aspekte bei der Evaluierung von Produktionsprozessen im Fokus von Predictive Process Control:

- sehr hohe Komplexität erfordert Expertenwissen in Bezug auf verfahrenstechnische und prozessbedingte Zusammenhänge
- Entwicklung und Verbesserung von Produktionsanlagen als langwieriger und experimenteller Prozess
- Umfassendes Einfließen von Prozessparametern
- Betrachtung aller relevanten Dimensionen eines optimalen Prozessablaufes
- Datenhoheit über verknüpfte Datenstrukturen im Fokus der Produktionsprozesse, zum Beispiel eine realisierte Auftragsverwaltung in SAP, verwaltet auf administrativer Ebene im Gegensatz zu resultierenden, auftragsbezogenen Parameter auf Produktionsebene

3.2 Aufteilung der Variablen in drei Hauptgruppen

Die Aufzählung kann beispielhaft für eine große Anzahl weiterer möglicher Abhängigkeiten in der Datenakquisition gesehen werden.

- **Unabhängige (invariante) Variablen:** z.B. Laborwerte über Rohmaterial für den nachfolgenden Produktionsprozess, ebenso Rezeptwerte einer Produktionsanlage (für eine Charge) sowie Rezeptur/ Mischverhältnisse, etc.
- **Abhängige (variante) Variablen:** z.B. resultierende Delta- Werte aus Differenz der Soll-/Istwerte aus Rezeptvorgaben von Produktionsstufen und den erfassten Istwerten (z.B. Sollwert von Temperaturzonen im Abgleich mit dem erfassten Istwert - mögliche Auswirkung der Temperaturabweichungen von bestimmten Zonen auf die weitere Produktionsqualität - mögliche Beeinträchtigung der weiteren Fertigungsstufen, etc.), sowie Prozessvariablen (Sensorwerte, etc.)
- **Erzeugte abhängige Variablen:** z.B. das Ergebnis einer Regressionsanalyse (Feststellung von Beziehungen zwischen abhängigen und einer oder mehreren unabhängigen Variablen)

3.3 Ziel

Der Verbund der nachfolgend aufgelisteten Stichworte ergibt einen Mix aus Ökonomie, Produktivität und Qualität. Jedes industrielle Unternehmen sollte sich vor Augen halten, dass diese drei Aspekte durch die eindimensionale Betrachtung von Qualitätskriterien bereits erreicht werden können. Erhöhte Qualität ist ein unmittelbares Kriterium für die Erhöhung der Produktivität und Kostenersparnis.

- Vorhersagen von Disfunktionalitäten innerhalb eines Prozesses, einer Fertigungsstufe oder einer Produktionsniederlassung
- Zeit- und Geldersparnis durch Verbessern der Produktqualität
- Verringerung der Standzeiten von Produktionseinheiten - bei gleichzeitig hoher Produktgüte - und dadurch höhere Maschinenauslastung

3.4 Methoden

- **Statistische Analysen:** z.B. *Regressionsanalyse* mit dem Ziel, Beziehungen zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen festzustellen

  $y = f(x_1, x_2, ..., x_n) + e$

- **Statische Analysen:** z.B. *Chi-square Automatic Interaction Detector*, eingesetzt bei der Konstruktion von Entscheidungsbäumen. CHAIDs kommen zur Anwendung, wenn eine Aussage über die Abhängigkeit von Variablen gemacht werden muss, dazu wird der Chi-Quadrat-Abstand berechnet

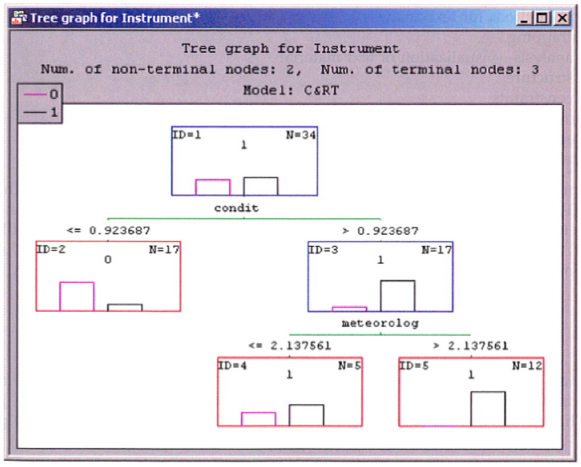

Figure 2: [RN10, Entscheidungsbaum]

- **Dynamische Analysen:** z.B. *Support Vector Machine* als mathematisches Verfahren der Mustererkennung, eine Klassifzierungsmethode zur Unterteilung von Datenpunkten in 2 Klassen. Dabei wird versucht, Datenpunkte durch eine Hyperebene zu trennen

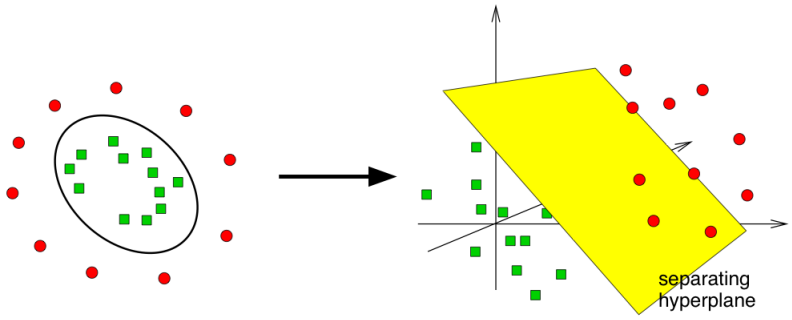

Figure 3: [Ula10, Prinzip Support Vector Machine]

4 Data Mining Software-Tools

Statistik-Software im Allgemeinen, versetzt leistungsfähige Computer in die Lage, mit teilweise rechenintensiven Methoden sehr große Datenmengen zu analysieren. In dem Buch werden folgende kommerzielle Plattformen als *"... the three most common Data Mining Software Tools ..."* angeführt:

- **SPSS Clementine:** modular aufgebautes Programmpaket zur statistischen Analyse von Daten
- **SAS-Enterprise Miner:** Software für statistische Analysen, Data-Mining, Data-Warehouse-Lösungen und Business-Intelligence einschließlich OLAP-Servern und Komponenten
- **STATISTICA Data Miner:** Software für statistische und grafische Datenanalysen, ist modular aufgebaut und bietet ein breites Spektrum an Methoden

Die Autoren geben in Ihren Ausführungen explizit keine Präferenz für eines dieser Software-Tools an. In diesem Kapitel wird nachfolgend noch auf die Frage *kommerzielle vs OpenSource - Plattformen* näher eingegangen, des Weiteren wird vor allem ein Augenmerk auf STATISTICA gelegt und die modulare Verwendbarkeit dieser Plattform umrissen.

4.1 Kommerzielle vs OpenSource Systeme

Der größte Vorteil von OpenSource CAS gegenüber kommerziellen Computeralgebrasystemen ist natürlich, dass keine Anschaffungs- und Lizenzkosten an den Hersteller zu entrichten sind. Dieser Vorteil kommt vor allem an Universitäten mit knappen Bildungsbudget zum Tragen, ebenso können OpenSource Derivate von StudentInnen ohne weitere Kosten verwendet werden. Aus Forschungsprojekten namhafter Universitäten entstehen innovative Projekte, welche im Anschluss durch eine aktive Community weitergetragen werden. Als Beispiel werden zwei wichtige Systeme aus der OpenSource Community angeführt:

- **RapidMiner:** zuvor YALE (Yet Another Learning Environment) genannt, ist eine Umgebung für maschinelles Lernen und Data Mining. Experimente können aus einer großen Zahl von nahezu beliebig schachtelbaren Operatoren erzeugt werden
- **GGobi:** Statistik Software um hochdimensionale multivariate Daten mit dynamischer Grafik zu visualisieren und zu analysieren

Entsprechend finden die meisten OpenSource Systeme im wissenschaftlichen Bereich für Forschung und Lehre ihr Einsatzgebiet. Dem gegenüber stehen kommerzielle Produkte. Die Lizenzierungskosten einer Predictive Process Control Applikation im industriellen Kontext, sind im Vergleich zum Nutzen welchen derartige Systeme erbringen können, sehr gering. Dazu ist eine differenzierte Betrachtungsweise zwischen dem meist akademisch angesiedelten Community-Charakter und der ökonomisch orientierten Entwicklung von proprietärer Software, festzuhalten. Anwender einer Plattform wie zum Beispiel STATISTICA gehen davon aus, eine hochleistungsfähige Datenanalyse- und Statistik-Software zur Verfügung zu haben und sich nicht um Produktverbesserung, fehlenden Support, etc. zu kümmern. Kommerzielle Systeme beinhalten folgende Features als "Softfacts" bei dem Kauf einer Lizenz, die meisten dieser Servicemaßnahmen sucht man im OpenSource Bereich vergeblich:

- Schulungen, Auswahl aus einem breiten Programm an produktunabhängigen Methodenkursen sowie Trainings zur Bedienung der Plattformen
- Consulting in Form von Mitarbeiter-Workshops bis hin zu individuellen Beratungsleistungen zu Fragen der statistischen Datenanalyse
- Umfassendes Einfließen von Prozessparametern
- Automatisierung, integrieren von Plattformen als "Knopfdrucklösung" mit einfacher Benutzerführung in die Systemumgebung
- Validierung, für Anwender in regulierten Branchen mit hohen Sicherheitsstandards wie Pharma und Life Science
- Technischer Support

4.2 STATISTICA

STATISTICA ist eine universelle Software für statistische und grafische Datenanalyse. Sie ist modular aufgebaut und bietet ein breites Spektrum an unterschiedlichen Methoden. Im *Basismodul* finden sich grundlegende statistische Auswerteverfahren wie statistische Kennziffern, Korrelationen und Varianzanalyse. Das Zusatzmodul *Professionell* bietet weitere Auswertemethoden. Im Zusatzmodul *Industrie* sind Prozeduren zusammengefasst, die man vornehmlich im industriellen Kontext anwendet. Es verbindet auf die Industrie zugeschnittene statistische Verfahren und Grafiken mit der Leistungsfähigkeit und der leichten Bedienbarkeit von STATISTICA. Hierzu gehören u.a.

- Versuchsplanung
- Prozessanalyse
- Regelkarten
- Poweranalyse
- Qualitätsregelkarten

Weitere Spezialprodukte zum Beispiel für Data Mining und Multivariate SPC werden ebenfalls angeboten. *STATISTICA Enterprise* bietet zusätzliche Funktionen für einen unternehmensweiten Einsatz: Datenzugriffe auf externe Datenbanken und Standardanalysen lassen sich automatisieren und Ergebnisse an definierte Personen verteilen. Die Ansicht auf das System und die Bedienungsoptionen lassen sich benutzerspezifisch einstellen. Das Modul *Data Miner* erkennt entscheidende Muster, Trends und Zusammenhänge. Der Anwender kann zwischen verschiedenen Modi wählen – vom Expertenmodus bis zum Assistenten, der Schritt für Schritt durch den Data-Mining-Prozess führt. Auch hier ist eine weitere Modularisierung möglich:

- Data Miner als Universallösung
- Process Optimization als Industrielösung einschließlich Verfahren zur Ursachenanalyse und Optimierung
- Text Miner zur Extraktion und Analyse unstrukturierter Textdaten

Durch verschiedene Anwendungsmodi ist es möglich, Bedienerfreiheit und Automatisierungsgrad zu variieren:

- Expertenmodus, alle Methoden und Optionen
- Assistent unterstützt eine klare Benutzerführung durch wichtigste Analyseschritte
- Projektoberfläche visualisiert und automatisiert den gesamten Data-Mining-Prozess, von der Datenaufnahme bis zur Ergebnisausgabe

Data Mining bewegt oft sehr große Datenbestände und führt aufwendige Modellberechnungen durch – eine gute Performance der Software ist daher Voraussetzung für den Einsatz im Kontex von Predictive Process Control:

- Optimierte Algorithmen, 64-Bit-Version und Verteilung von Rechenlast auf zwei Prozessoren
- Optimierter Lese- und Schreibzugriff auf große Datenbanken: Die IDP-Technologie (In-Place Database Processing) liest Daten asynchron direkt von Datenbankservern
- Eine Serverversion ermöglicht zusätzlich die Verlagerung aller Berechnungen auf leistungsstarke Multiprozessorsysteme und kann außerdem von einem PC in einem Intra- oder Internet über eine Web-Oberfläche – also ohne installierte STATISTICA-Komponenten gesteuert werden

Die "Beste aller Methoden" für Data Mining gibt es ohne Betrachtung des erforderlichen Analyse- Kontext nicht. Ein gutes Data-Mining-Werkzeug muss daher ein breites Methodenspektrum anbieten. Nachfolgend ein Überblick der wichtigsten Werkzeuge, welche STATISTICA zur Verfügung stellt:

- zahlreiche Verfahren der Datenvorbereitung wie Filtern, Selektieren und Transformieren
- Feature Selection zur Identifikation relevanter Prädiktoren
- Klassifikations- und Regressionsbäume (CART, CHAID), Interaktive Entscheidungsbäume, Boosted Trees und Random Forests
- Verallgemeinerte Additive Modelle, MAR Splines
- EM- und k-Means-Clusterverfahren, Independent-Component-Analysen, einfache und sequenzielle Assoziationsregeln
- Neuronale Netzverfahren einschließlich Self-Organizing Maps (SOM)
- K-Nearest Neighbors, Support Vector Machines (SVM), Bayessche Verfahren
- Techniken wie Voting, Bagging, Boosting, Meta-Lernen
- einfache und höhere statistische Methoden (wie Regressions-, Diskriminanz- und Zeitreihenanalysen)
- interaktive Visualisierungswerkzeuge

4.2.1 Beispiele für Module in STATISTICA

- **Allgemeine nichtlineare Regression:** Das Modul Nichtlineare Regression ermöglicht dem Benutzer die Anpassung beliebiger Typen nichtlinearer Modelle. Eine der speziellen Eigenschaften dieses Moduls besteht darin, dass – im Unterschied zu traditionellen Programmen der nichtlinearen Regression – die Größe der Datendatei keine Rolle spielt.

  Schätzmethoden: Die Modelle können unter Verwendung von KQ- oder Maximum-Likelihood-Schätzverfahren bzw. basierend auf benutzerdefinierten Verlustfunktionen angepasst werden. Auf der Basis des Kleinste-Quadrate-Kriteriums lassen sich der hocheffiziente Levenberg-Marquardt- und der Gauss-Newton-Algorithmus zur Parameterschätzung für beliebige lineare und nichtlineare Regressionsprobleme einsetzen. Für große Datensätze oder schwierige nichtlineare Regressionsprobleme auf der Basis Kleinster Quadrate ist dies die empfohlene Methode zur Berechnung präziser Parameterschätzwerte. Der Benutzer kann aus vier leistungsfähigen Optimierungsverfahren zur konkreten Parameterschätzung auswählen: Quasi-Newton, Simplex, Koordinatensuche nach Hooke-Jeeves sowie Rosenbrock-Suchverfahren der rotierenden Koordinaten. Damit erhält man stabile Parameterschätzungen in nahezu allen Fällen, selbst bei numerisch anspruchsvollen Problemen.

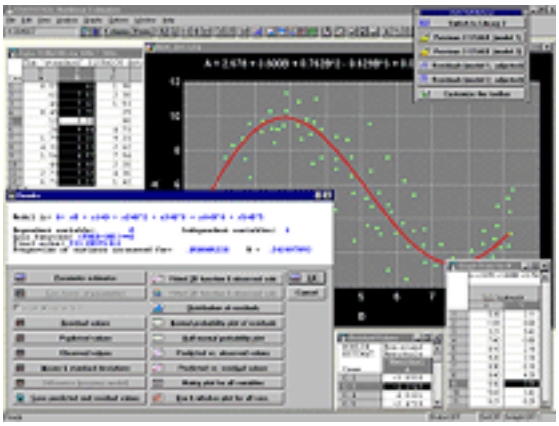

Figure 4: [STA10, Allgemeine nichtlineare Regression]

Modelle: Der Benutzer kann die Modellgleichung einfach dadurch spezifizieren, dass er die Gleichung in einem speziellen Editor eingibt. Die Gleichungen können logische Operatoren enthalten, wodurch es möglich wird, unstetige, d.h. stückweise definierte, Regressionsmodelle und Modelle mit Indikatorvariablen zu spezifizieren. In den Gleichungen kann auf eine breite Auswahl von Verteilungen Bezug genommen werden. Dazu gehören Beta-, Binomial-, Cauchy-, Chi-Quadrat-, Exponential-, Extremwert-, F-

, Gamma-, Geometrische, Laplace-, Logistische, Normal-, Lognormal-, Pareto-, Poisson-, Rayleigh-, Students t- sowie Weibull-Verteilung (Dichte- bzw. Wahrscheinlichkeitsfunktion, Verteilungsfunktion und deren Inverse). Der Benutzer kann alle Aspekte des Schätzverfahrens, wie z.B. Startwerte, Schrittweiten, Konvergenzkriterien, in vollem Umfang kontrollieren. Die am häufigsten benötigten Regressionsmodelle sind im Modul Nichtlineare Regression vordefiniert und können einfach als Menüoptionen abgerufen werden. Diese Modelle schließen schrittweise Probit- und Logit-Regression, das exponentielle Regressionsmodell und stückweise lineare Regression (mit Strukturbruch) ein. Zu beachten ist, dass STATISTICA auch Implementierungen mächtiger Algorithmen zur Anpassung von verallgemeinerten linearen Modellen enthält, einschließlich Probit und multinomialer Logit-Modelle, sowie verallgemeinerte additive Modelle.

Ergebnisse: Zusätzlich zu verschiedenen deskriptiven Statistiken enthält die Standardausgabe die Parameterschätzungen, deren Standardfehler, die unabhängig von den Schätzungen selbst berechnet werden, die Kovarianzmatrix der Parameterschätzungen, die Prognosewerte, Residuen und Maße für die Anpassungsgüte (z.B. die Log-Likelihood der geschätzten/Nullmodelle und den Chi-Quadrat-Test der Differenz an erklärter Varianz, Klassifikation der Fälle und Odds-Ratios für Logit- und Probit-Modelle). Die Prognosewerte und die Residuen können der Datendatei für weitere Analysen hinzugefügt werden. Für Probit- und Logit-Modelle wird der Gewinn bzw. Verlust an Anpassung automatisch berechnet, wenn Parameter dem Modell hinzugefügt oder aus diesem entfernt werden, d.h. der Benutzer kann die Modelle anhand schrittweiser nichtlinearer Verfahren an die Daten erkunden. Optionen zur automatischen schrittweisen Regression (vorwärts und rückwärts) sowie Beste-Subset-Auswahl von Prädiktoren in Logit- und Probit-Modellen werden im Modul Verallgemeinerte Lineare/Nichtlineare Modelle angeboten.

Grafiken: In die Ausgabe der Ergebnisse ist eine umfassende Auswahl von Grafiken integriert. Dazu gehören 2D- und 3D-Flächenplots, die dem Benutzer die Güte der Anpassung verdeutlichen und die Identifikation von Ausreißern ermöglichen. Der Benutzer kann interaktiv die Gleichung der angepassten Funktion korrigieren, ohne die Daten neu verarbeiten zu müssen, und nahezu alle Aspekte des Schätzprozesses visualisieren. Viele weitere spezielle Grafiken dienen der Bewertung der Güte der Anpassung und der Visualisierung der Ergebnisse, wie z.B. Histogramme aller ausgewählten Variablen und der Residuen, Scatterplots der Beobachtungswerte gegen die Prognosewerte sowie Prognosewerte gegen Residuen, einfache und einseitige Normalverteilungsplots der Residuen und weitere.

- **Klassifikations- und Regressionsbäume:** Das Modul Klassifikations- und Regressionsbäume bietet eine umfassende Implementierung der aktuellsten Algorithmen für die effektive Erstellung und für das Testen der Robustheit von Klassifikationsbäumen. Ein Klassifikationsbaum ist eine Regel für die Prognose der Klassenzugehörigkeit eines Objektes aus den Werten seiner Prädiktor-Variablen. Höhere Methoden für Klassifikationsbäume, einschließlich flexibler Optionen zur Modellentwicklung und interaktive Werkzeuge zur Exploration von Bäumen sind im STATISTICA Data Miner mit den General Classification and Regression Tree Models (GTrees) und General CHAID (Chi-square Automatic Interaction Detection) enthalten. Klassifikationsbäume können auf der Basis von kategorialen oder ordinalen Prädiktor-Variablen erstellt werden. Dabei können sowohl univariate als auch multivariate Splits oder Linearkombinationen von Splits eingesetzt werden. Die Optionen der Analyse enthalten die Durchführung von umfassenden Splits oder auf Diskrimination basierende Splits; unverzerrte (unbiased) Variablenauswahl (wie in QUEST); direkte Stopregeln (direct stopping rules, wie in FACT) oder "Aufwärtsabschneiden" (bottom-up pruning, wie in CART); Abschneiden basierend auf Fehlklassifikationsraten oder der "Deviance"-Funktion; verallgemeinerte Chi-Quadrat-, G-Quadrat- oder Gini-Index-Maße für die Güte der Anpassung. Priors und Fehlklassifikationskosten können als identisch spezifiziert, aus den Daten geschätzt oder benutzerspezifiziert werden. Der Benutzer kann außerdem den v-Wert für v-fache Kreuzvalidierung während der Baumerstellung, den v-Wert für v-fache Kreuzvalidierung für die Fehlerschätzung, die Größe der SE-Regel, die minimale Knotengröße vor dem Abschneiden, Startwerte für die Zufallszahlengenerierung und Alpha-Werte für die Variablenselektion spezifizieren. Für die Unterstützung der Analysen stehen integrierte Grafikoptionen zur Verfügung.

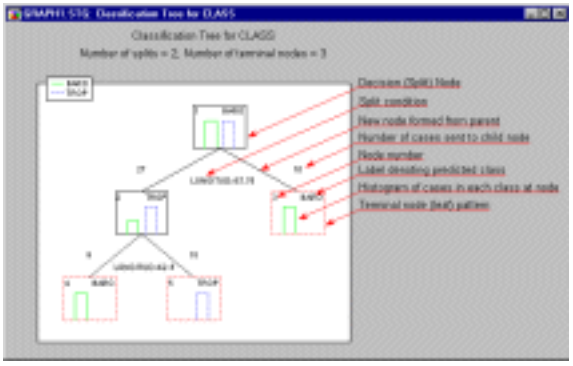

Figure 5: [STA10, Klassifikations- und Regressionsbäume]

- **Clusteranalyse:** Dieses Modul enthält eine umfassende Implementierung von Methoden zur Clusterung (k-Means, hierarchisch, 2-fach agglomerativ). Die Software kann sowohl Einzeldaten als auch Distanzmatrizen wie z.B. Korrelationsmatrizen verarbeiten. Der Benutzer kann Fälle, Variablen oder beides basierend auf einer Vielzahl von Distanzmaßen [Euklidisch, quadriert Euklidisch, City-block (Manhattan), Chebychev, Power-Distanzen, Prozent Nichtübereinstimmung und 1-Pearsons r] clustern. Als Fusionregeln stehen Single Linkage, Complete Linkage, Weighted und Unweighted Group Average oder Centroid, Ward-Methode und weitere Verfahren zur Verfügung. Die Distanzmatrizen können für weitere Analysen gespeichert werden. Beim k-Means-Verfahren hat der Benutzer die vollständige Kontrolle über die anfänglichen Cluster-Zentren. Dabei können Designs von extremer Größe verarbeitet werden: Die hierarchischen Verfahren können Matrizen von 1000 Variablen oder einer Million Distanzen behandeln. Zusätzlich zu den üblichen Ergebnissen einer Clusteranalyse ist ein breiter Satz deskriptiver Statistiken und Diagnose-Kenngrößen verfügbar. So wird z.B. das vollständige Fusionsprotokoll bei hierarchischen Verfahren oder die ANOVA-Tabelle bei k-Means ausgegeben. Die Information über die Clusterzugehörigkeit kann der Datendatei zur weiteren Bearbeitung angefügt werden. Die Grafikoptionen des Moduls beinhalten Baumdiagramme, diskrete Matrixplots, grafische Darstellungen des Fusionsprotokolls, Plots der Mittelwerte bei k-Means-Verfahren und viele weitere.

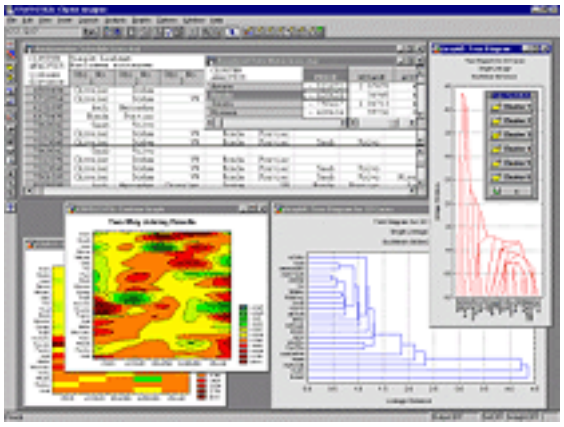

Figure 6: [STA10, Clusteranalyse]

5 Zusammenfassung

Durch Data Mining Verfahren am Allgemeinen und Predictive Process Control im Speziellen, ist es möglich, selbst komplexe Produktionsprozesse mit Hilfe von routinemäßig gesammelten Betriebsdaten zu analysieren und zu verbessern. Das Potential und der Informationsgehalt von strukturierten Daten, nach verfahrens- und prozesstechnisch sinnvoll ausgelegten Maßstäben aufbereitet, wird in der heutigen Zeit noch immer unterschätzt. Mit zunehmender Komplexität von produktionstechnischen Aspekten und damit verbundenem erhöhtem Output, gelingt es immer weniger, Produktionsverfahren weiter zu optimieren. Der Herausforderung, verschiedene Produktionsstufen, unterschiedliche Technologien und daraus resultierend divergierende Anforderungen an/von das/dem Produktionspersonal durch eine übergreifende Datenbasis analysieren und verbessern zu können, kann mit Hilfe einer PPC Applikation begegnet werden. Die Methoden: Analyse, Vorhersage und Optimierung von Produktionsfaktoren lassen einen immanenten Vorteil gegenüber Verfahren, die auf eine solche Auswertung von Produktionsdaten verzichten, erkennen. Das "Handbook of Statistical Analysis and Data Mining Applications" dient für diese Betrachtung in hervorragender Weise, Untersuchungsgegenstände sowohl quantitativ als auch qualitativ zu bearbeiten. Wer in der Praxis statistische Analysen sowie Data- u Textmining verwendet, findet in diesem Werk zahlreiche Hilfestellungen. Besonders hervorzuheben ist, dass zu jedem Kapitel Fallbeispiele inklusive Berechungen und Anleitungen zur Vorgehensweise mit entsprechenden Plattformen (SPSS, STATISTICA, MaxQDA, SAS...) existieren. Der Anspruch, eine Überleitung aus der Theorie in die Praxis zu leisten, wird durch zahlreiche Fallstudien bestätigt. Das Buch wird mit einer beigelegten DVD ausgeliefert, welche noch mehr Fallstudien inklusive einer detaillierten Problembeschreibung und dem resultierenden Researchdesign, enthält. Zusätzlich liegt eine Testversion von STATISTICA (4 Monate gültig) bei, sowie Code zur Elsevier-Seite, welche noch mehr Case-Studies vereint.

Literaturnachweis

[RN10] Robert Nisbet, John Elder, Gary Miner: *Handbook of Statistical Analysis and Data Mining Applications.* Academic Press, July 2010.

[STA10] STATSOFT: *STATISTICA*, July 2010. [Online; accessed 1-July-2010].

[Ula10] Ulamec, Norbert: *Advanced Database Systems*, 2010. [Online; accessed 2-April-2010].