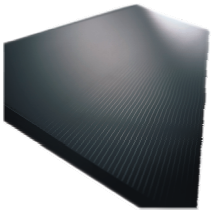
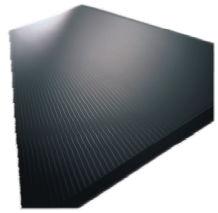
Fachhochschule Brandenburg
Fachbereich Informatik und Medien



Anwendung von Methoden des Data Mining bei der Produktion von Dünnschicht-Solarmodulen

Bachelorarbeit, vorgelegt von David Saro



Aufgabenstellung

Ziel des Themas ist die Untersuchung der Anwendbarkeit von Methoden des Data Mining zum Finden und Modellieren von Abhängigkeiten aus den vorhandenen Stell- und Messgrößen, die während der Produktion von Solarmodulen erfasst wurden.

Durch das Aufstellen einfacher Thesen von den Technologen soll versucht werden, bekannte zu erkennen und neue Wirkzusammenhänge zu bestätigen und darzustellen.

Unternehmen

Abb. 1 Fabrik

Die Johanna Solar Technology GmbH ist im Jahr 2006 gegründet worden und beschäftigt sich mit der Produktion von Dünnschicht-Solarmodulen. Die dabei eingesetzte Technologie wird in der Form nur von diesem Unternehmen eingesetzt.

Das relativ junge Unternehmen befindet sich im Aufbau. Der Produktionsprozess wird ebenfalls noch stabilisiert.

Während der Prozessierung der Solarmodule sollen alle Stell- und Messgrößen erfasst werden. Zum Zeitpunkt der Ausarbeitung der Bachelorarbeit ist die Erfassung nicht komplett. Erfassungslücken und Messfehler sind zu erwarten, aufzuzeigen und wenn möglich zu beheben.

Data Mining

Das Data Mining erfolgt in 7 Phasen.

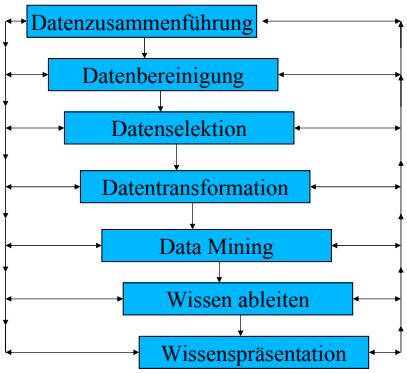


Abb. 2 Die 7 Phasen des Data Mining

Die Pfeile in Abb. 2 deuten an, dass es keine strikte einzuhaltende Abfolge der Phasen gibt. In der Datenzusammenführungsphase werden die unterschiedlichen Quellen in einer Ziel-Datenquelle zusammengeführt. Oft handelt es sich hierbei um eine Datenbank. In der Datenbereinigungsphase werden Fehler gesucht, die die Datenmenge für den Einsatz von Data Mining Methoden vorbereiten. Die Datenselektionsphase soll die Dimension des zu nutzendem Datenbestandes verkleinern, indem unwichtige und redundante Attribute gefiltert werden.

Die übrigen Attribute werden wenn nötig in der Datentransformationsphase bearbeitet. Beispielhaft sei hier die Diskretisierung und die Aggregation erwähnt. In der Data Mining Phase werden verschiedene Lernverfahren mit der Datenmenge getestet. Um die Leistungsfähigkeit auf eine unbekannte Testmenge beurteilen zu können, werden hier mittels n-facher Kreuzvalidierung die zur Verfügung stehende Datenmenge in *n* gleich große Teile gesplittet. Jedes dieser Teile wird als Validierungsmenge genutzt, während alle anderen dem Training dienen. Nachdem jeder Teil einmal als Validierungsmenge genutzt wurde, stehen *n* Performanzwerte zur Verfügung. Deren Mittelwert ist der Schätzwert des Generalisierungsfehlers, also den zu erwartenden Fehler bei der Anwendung auf unbekannten Datenmengen. Aus den Ergebnissen kann Wissen abgeleitet werden, das weiteren Experimenten als Grundlage dienen, oder zur Ansicht in die Wissenpräsentationsphase gebracht werden kann.

Rapidminer

Für die Anwendung der Data Mining Methoden, wurde die Software Rapidminer verwendet. Sie beinhaltet Operatoren für die Aufbereitung der Daten, Anwendung von Lernalgorithmen, sowie die Erzeugung von grafischen Darstellungen deskriptiver Statistik.

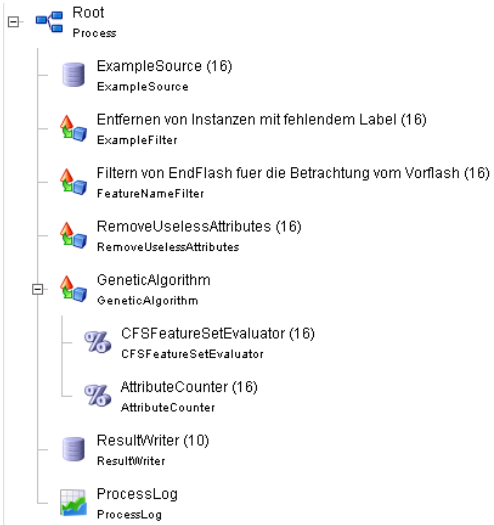


Abb. 3 Aufbau eines Experiments im Rapidminer für die Datenselektion mit genetischen Algorithmen

Das in Abbildung 3 dargestellte Experiment liest aus dem Dateisystem eine Datenmenge ein, filtert alle Datensätze mit fehlender Klassifizierung aus und entfernt einige Attribute deren Name einem regulärem Ausdruck entsprechen. Anschließend werden alle Attribute verworfen, die durchgängig den selben oder keinen Wert enthalten. Der GeneticAlgorithm-Operator bildet eine erste Generation mit mehreren Individuen, die aus unterschiedlichen Attributskombinationen bestehen. Die inneren zwei Operatoren sind zwei Zielfunktionen, die jeweils einen Performanzvektor zurückgeben. Schlecht bewertete Individuen werden verworfen und gut bewertete gehen in veränderter Form in die nächste Generation über. Die in der Abbildung 4 dargestellte Pareto-Front stellt beide Zielfunktionsergebnisse einer Generation gegenüber. Der GeneticAlgorithm-Operator gibt als Ergebnis Attributsgewichtungen zurück. Die letzten beiden Operatoren speichern die Ergebnisse persistent im Dateisystem.

Probleme, Ergebnisse und Ausblick

Die Software Rapidminer in der Version 4.2 enthält einige Fehler, die nach Aussage der Entwickler in der nächsten Version korrigiert sein werden. Hierzu zählt der Operator für die Umwandlung von nominal behandelten Zeitstempeln in numerische Unixepochen. Dieser verweigert bisher bei leeren Attributwerten seine Arbeit.

Desweiteren fielen nicht vorhandene Operatoren für die Ersetzung fehlender Werte ins Gewicht. Sollte hier von den Rapidminerentwicklern eine Implementierung nachgeliefert werden, die die Klassifizierungsspalte bei der Ersetzung einbezieht, ist dieser wahrscheinlich vielversprechender als die bestehenden Implementierungen.

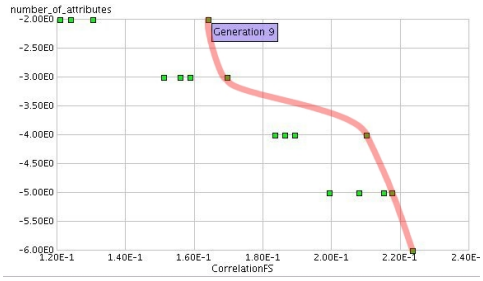


Abb.4 Pareto-Front

Einige vermutete Zusammenhänge konnten durch die Anwendung von Lernalgorithmen oder durch deskriptive Statistik bestätigt werden. Neue Erkenntnisse konnten aus Gründen der bisher noch unvollständigen Datenerfassung nicht gewonnen werden. Trotz umfangreicher Datenbereinigung, die aus zeitlichen Gründen abgebrochen werden musste, wurden durch die Algorithmen weitere Schwächen im Datenbestand ersichtlich. Wege zur Beseitigung dieser Schwächen sind bereits in Planung.

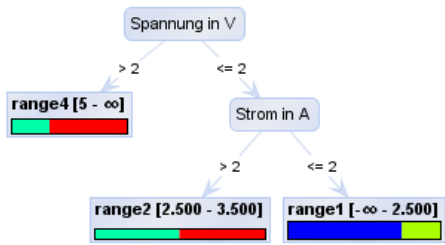


Abb. 5 Univariater Entscheidungsbaum für die Vorhersage der elektrischen Leistung aus den Attributen Strom und Spannung

Ebenfalls wurden Korrelationen zwischen Attributen entdeckt, die bisher nicht erklärt werden konnten. Diesen gilt es nachzugehen. Durch die Datenauswahlphase ging hervor, welche Attribute im bestehenden Datenbestand relevant für die Vorhersage der Klassifikationsspalte sind. Mit entsprechenden Attributsgenerierungsalgorithmen konnten teilweise bekannte Redundanzen von Attributen bestätigt werden.

Zu Beginn der Datenselektionsphase wurde ein Schnappschuss des Datenbestandes erstellt, der als Grundlage für weitere Untersuchungen diente. Im Laufe der Bachelorarbeit wurden einige bekannte Schwächen der Datenerfassung beseitigt, so dass bereits jetzt ein weiteres Data Mining zu anderen, vielleicht besseren Ergebnissen führen würde.

Betreuer: Dipl.-Inform. I. Boersch
Prof. Dr. J. Heinsohn
C. Kneisel

Abgabe: 18.09.2008