ERKENNUNG NICHT RELEVANTER UNTERNEHMEN IN DEN HANDWERKSSTATISTIKEN

Einsatz von Support Vector Machines zur maschinellen Klassifikation

Jörg Feuerhake, Florian Dumpert

↘ **Schlüsselwörter:** Support Vector Machines – Handwerksstatistik – nicht-parametrische Methoden – statistisches Unternehmensregister

ZUSAMMENFASSUNG

Die Handwerksstatistiken werden seit dem Berichtsjahr 2008 vollständig aus Verwaltungsdaten gewonnen. Dabei treten teilweise Klassifizierungsprobleme auf, die derzeit aufwendig in den Statistischen Ämtern der Länder manuell gelöst werden. Der vorliegende Aufsatz zeigt, wie sich maschinelle Lernalgorithmen einsetzen lassen, um mit relativ geringem personellen Aufwand Beobachtungen in Gruppen einzuteilen. Es wird der Einsatz von Support Vector Machines in Verbindung mit dem Klassifikationsverfahren Random Forests für die Klassifizierung der Handwerkseigenschaft im Unternehmensregister dargestellt.

↘ ***Keywords:*** *support vector machines – crafts statistics – non-parametric methods – statistical business register*

ABSTRACT

Since the reference year of 2008, all crafts statistics have been compiled from administrative data. Sometimes, this causes time-consuming classification problems which have to be fixed manually by staff at the statistical offices of the Länder. This article shows how machine learning algorithms can be used to group observations with relatively little staff effort. It describes the use of support vector machines combined with the Random Forests classification method to identify the businesses' affiliation to the crafts sector in the business register.

Jörg Feuerhake

ist Diplom-Volkswirt und Referent im Bereich Handwerksstatistik des Statistischen Bundesamtes. Er betreut die Durchführung und die methodische Weiterentwicklung der Handwerksstatistiken.

Florian Dumpert

ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Stochastik der Universität Bayreuth. Der Diplom-Mathematiker forscht im Bereich maschineller statistischer Lernverfahren, insbesondere Support Vector Machines.

1

Hintergrund und Problem

Die Handwerksstatistiken werden derzeit vollständig aus Verwaltungsdaten erstellt (Neuhäuser, 2008; Feuerhake, 2012). Für die Aufbereitung der Statistiken über das Handwerk ist das statistische Unternehmensregister (URS) von zentraler Bedeutung. Dort werden einmal jährlich Lieferungen der Handwerkskammern sowie Verwaltungsdaten aus anderen Quellen, insbesondere der Bundesagentur für Arbeit und der Finanzverwaltung, verarbeitet. Die Handwerkskammern liefern für diesen Zweck den handwerklichen Gewerbezweig und Hilfsmerkmale, wie Gewerbesteuernummer und Adressinformationen, anhand derer die Unternehmen identifiziert werden können. Nach der Verarbeitung stehen die Handwerksinformationen im URS für statistische Auswertungen zur Verfügung.

Allerdings sind nicht alle in den Datenlieferungen der Handwerkskammern enthaltenen Unternehmen relevant für die Handwerksstatistiken. Nach § 2 Handwerkstatistikgesetz sind in den Handwerksstatistiken nur selbstständige Handwerksunternehmen zu erfassen. Daneben gibt es aber eine Gruppe von Unternehmen, die selbst nicht Handwerksunternehmen sind, aber handwerkliche innerbetriebliche Abteilungen oder handwerkliche Nebenbetriebe unterhalten. Dies sind zum Beispiel Speditionen, die eigene Kfz-Werkstätten haben, Supermärkte mit Fleischer- oder Bäckertheken oder Energieversorgungsunternehmen, die Lehrwerkstätten für bestimmte Handwerksberufe betreiben. Diese Unternehmen müssen identifiziert werden, weil sie bei der Aufbereitung der Handwerksstatistiken nicht einbezogen werden sollen.

Die Handwerkskammern sind anhand der ihnen vorliegenden Informationen hierzu nicht in der Lage. Diese Aufgabe wird deswegen jährlich von den Fachbereichen Handwerk der Statistischen Ämter der Länder übernommen. Hierbei werden für die einzelnen Fälle Informationen aus verfügbaren Quellen, wie dem Internet, gegebenenfalls vorliegenden Geschäftsberichten oder dem Handelsregister, recherchiert. Da die Arbeiten zur manuellen Klassifizierung in erheblichem Maße personelle Ressourcen binden, sollte im Rahmen eines internen Projekts geprüft werden, ob die fraglichen Unternehmen hinsichtlich ihrer Relevanz für die Handwerksstatistiken mit ausreichender Genauigkeit maschinell klassifiziert werden können. Das Projekt startete Mitte 2014. Erste Ergebnisse wurden Mitte 2015 im Rahmen des Arbeitskreises für mathematisch-statistische Methoden und einer internen Kurzveranstaltung jeweils im Statistischen Bundesamt zur Diskussion gestellt.

Der vorliegende Aufsatz beschreibt einen Ansatz, mit dem dieses Problem angegangen werden kann. Hierzu wird im folgenden Kapitel 2 kurz auf statistische Eigenschaften des Problems eingegangen. Kapitel 3 stellt die maschinellen Lernverfahren Support Vector Machines und Random Forests vor, die im Projekt verwendet wurden. In den weiteren Kapiteln wird dargestellt, wie die maschinellen Lernverfahren im Projekt praktisch eingesetzt wurden, welche Ergebnisse erzielt wurden und welche weiteren Schritte sich daraus ableiten lassen.

2

Statistische Eckdaten und Komplexität des Problems

Aus dem URS mit dem Bezugsjahr 2012[1] und den Ergebnissen der Handwerkszählung desselben Jahres lassen sich folgende Angaben zu Umfang und Komplexität des vorliegenden Klassifikationsproblems ermitteln:

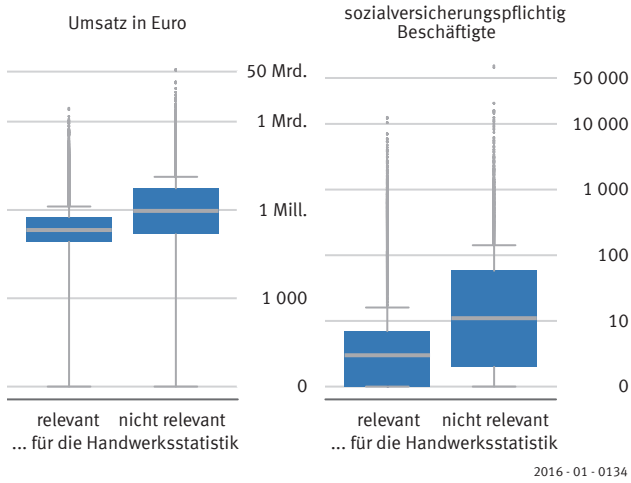
Von den rund 600 000 Unternehmen, die von den Handwerkskammern als Handwerksunternehmen gemeldet wurden, waren 1,3 % nicht relevant für die Handwerksstatistik. Sie repräsentierten 19 % der sozialversicherungspflichtig Beschäftigten und erzielten 35 % der Umsätze. Die entsprechenden Verteilungen zeigt ↘ Grafik 1.

Die oben genannten Anteile an Umsatz und sozialversicherungspflichtig Beschäftigten sowie die Verteilung der Umsätze in beiden Gruppen legen nahe, dass vorwiegend große Unternehmen nicht relevant für die Handwerksstatistiken sind.

Die im ersten Kapitel „Hintergrund und Problem“ genannten Beispiele lassen weiter vermuten, dass

1 Das Bezugsjahr gibt an, auf welches Jahr sich die Angaben des URS beziehen, insbesondere die verarbeiteten Verwaltungsdaten der Bundesagentur für Arbeit und der Finanzbehörden.

Grafik 1
Verteilung von Umsätzen und sozialversicherungspflichtig Beschäftigten der gemeldeten Handwerksunternehmen 2012



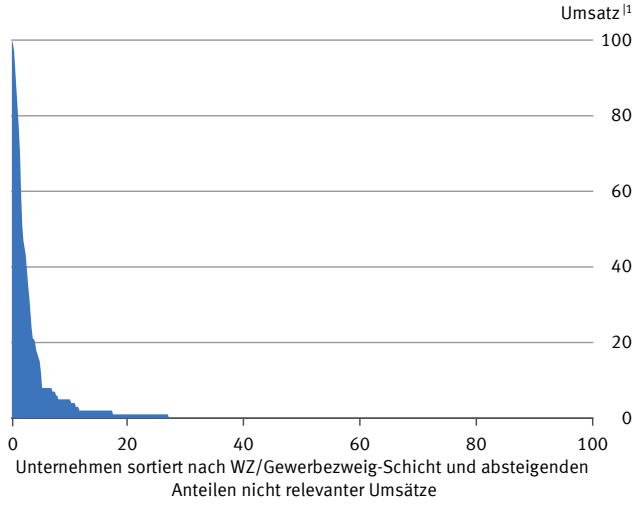
2016 - 01 - 0134

bei bestimmten Kombinationen aus Wirtschaftszweig (WZ)[2] und handwerklichem Gewerbezweig besonders viele Unternehmen als nicht relevant für die Handwerksstatistiken klassifiziert werden. Die bisher vorgenommenen Zuordnungen haben gezeigt, dass 7 % der etwa 7 700 möglichen WZ/Gewerbezweig-Kombinationen nur nicht relevante Handwerksunternehmen enthalten, während in 72 % der WZ/Gewerbezweig-Kombinationen keine nicht relevanten Handwerksunternehmen vorkommen. ↘ Grafik 2

Man kann festhalten, dass ein Großteil der vorkommenden WZ/Gewerbezweig-Kombinationen relativ leicht klassifizierbar ist, weil dort entweder für die Handwerksstatistiken nur relevante oder nur nicht relevante Unternehmen vorkommen. Es gibt allerdings eine Gruppe von Unternehmen, bei der zurzeit eine manuelle Prüfung erforderlich ist, um sie zu klassifizieren. Diese Arbeiten müssen jährlich für neu gelieferte Unternehmen durchgeführt werden. Die einmal festgelegte Klassifizierung wird dann für die Zukunft beibehalten. Im Berichtsjahr 2012 wurden bundesweit 6 100 von den Handwerkskammern erstmals gemeldete Unternehmen von den Fachbereichen Handwerk der Statistischen Ämter der Länder geprüft. Wir sind der Frage nachgegangen, ob maschinelle Klassifikationsverfahren die manuellen Prüfungen ersetzen können.

2 Klassen der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008).

Grafik 2
Nicht relevante Umsätze in WZ/Gewerbezweig-Kombinationen 2012
in %



1 Anteil des nicht relevanten Umsatzes in WZ/Gewerbezweig-Schicht.

2016 - 01 - 0135

3 Eingesetzte maschinelle Klassifikationsverfahren

Bevor das Vorgehen bei der maschinellen Klassifizierung erläutert wird, sollen die zwei eingesetzten maschinellen Lernverfahren – Support Vector Machines und Random Forests – vorgestellt werden.

3.1 Support Vector Machines

Zur Darstellung dieser Methode wird in diesem Aufsatz der anschauliche, geometrische Ansatz gewählt. Hinsichtlich des dazu äquivalenten analytischen Ansatzes verweisen wir auf Dumpert/von Eschwege/Beck (2016). Mathematische Details zu den folgenden Ausführungen finden sich in den ursprünglichen Arbeiten von Boser und andere (1992) und Cortes/Vapnik (1995) sowie beispielsweise in den Übersichtsdarstellungen von Campbell/Ying (2011, hier: Seite 1 ff.), Steinwart/Christmann (2008, hier: Seite 13 ff.) sowie James und andere (2013, hier: Seite 337 ff.). Für einen

allgemeinen, leicht verständlichen Überblick über Support Vector Machines (SVM) sei darüber hinaus auf Hamel (2009) verwiesen.

Die Klassifikationsmethode soll diejenigen Unternehmen, die für die Handwerksstatistiken relevant sind, von solchen trennen, die diese Eigenschaft nicht besitzen. Die Menge der Unternehmen soll also in zwei überschneidungsfreie Teilmengen geteilt werden.

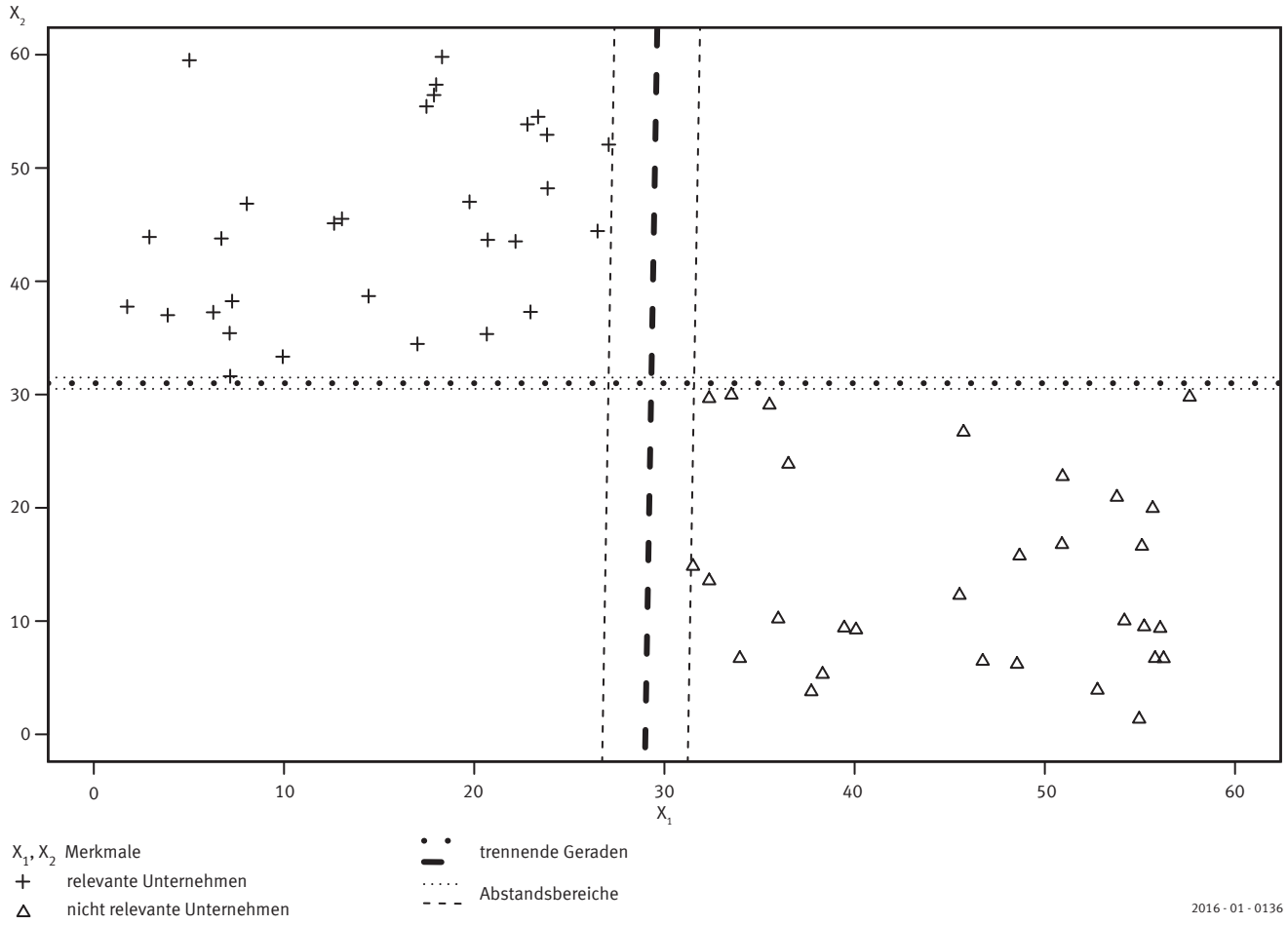
Beispiel

Da die geometrische Veranschaulichung von SVM in diesem Aufsatz auf zwei Dimensionen beschränkt ist, soll hier für die Darstellung vereinfachend davon ausgegangen werden, dass die Zugehörigkeit zum Handwerk vom Zusammenspiel von lediglich zwei Merkmalen abhängt.[3] Diese könnten zum Beispiel der Umsatz und die Anzahl der sozialversicherungspflichtig Beschäftigten sein. Wir bezeichnen die beiden Merkmale mit X_1 und X_2. Für das Handwerk relevante Unternehmen sind in den Grafiken mit „+“ markiert, die nicht relevanten mit „△“.

Grundidee einer SVM ist es, zwei Gruppen mit bekannter Klassifizierung linear zu trennen. In ↘ Grafik 3 erscheint die lineare Trennung leicht, da die Punktemengen vollständig durch eine Gerade zu trennen sind. Unklar ist zunächst, welche Gerade man als Trennung wählen soll.

3 Somit ist es möglich, trotz der Einschränkungen in der Darstellung einen Eindruck von der Wirkungsweise von SVM zu gewinnen. Welche Merkmale tatsächlich für die Klassifizierung zur Verfügung stehen, ist in Übersicht 1 dargestellt.

Grafik 3
Punktmengen und willkürliche Trennungsgeraden



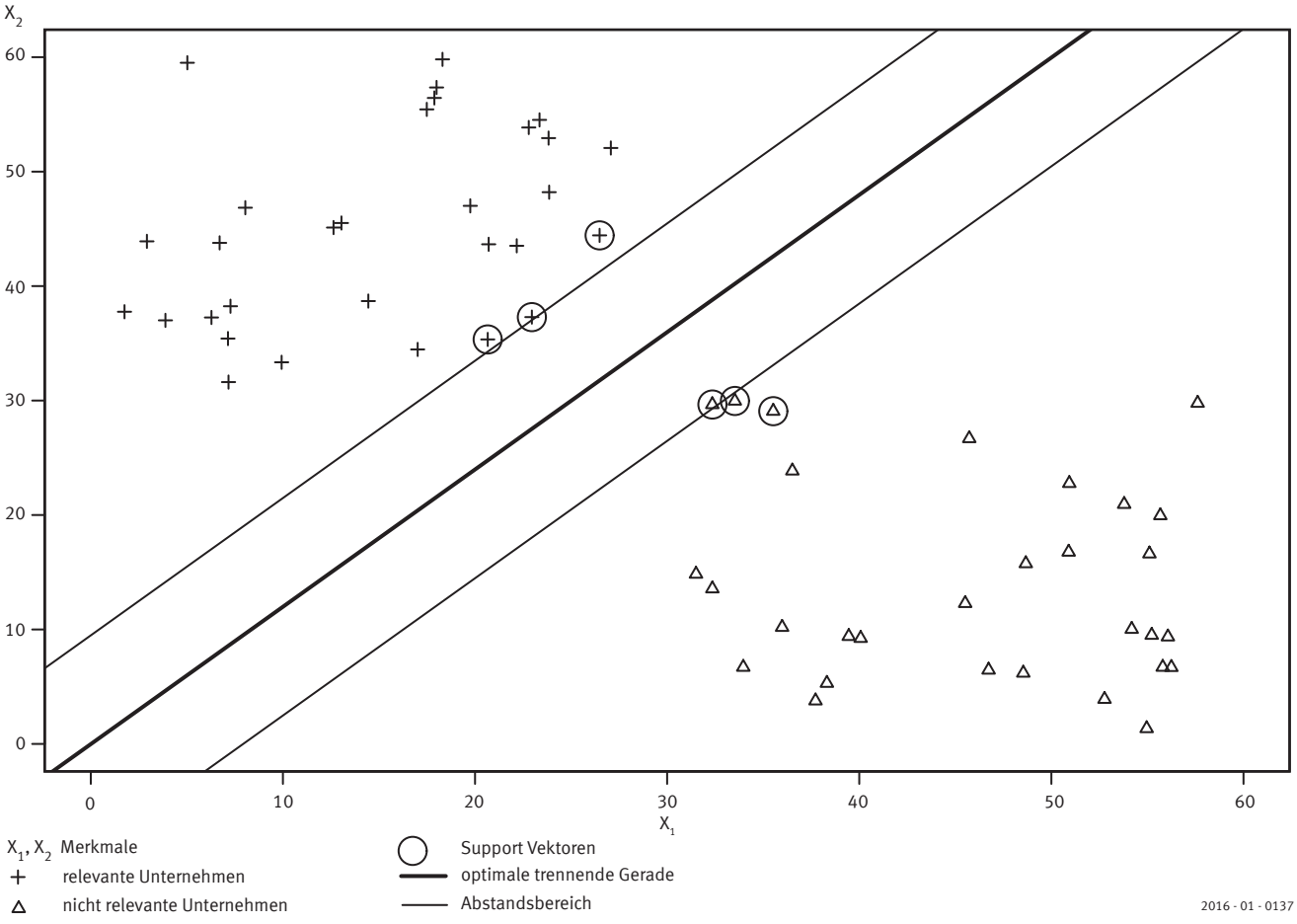
Support Vector Machines wählen, als Ergebnis eines Optimierungsproblems, diejenige Gerade, welche den größten Abstand zu den Punktemengen aufweist. Dies ist zweckmäßig, da dadurch das Risiko von Fehlzuordnungen neu zu klassifizierender Unternehmen minimiert wird. Die dicke, gepunktete, horizontale Gerade in der Grafik 3 eignet sich hierfür offenbar nicht, denn schon der zweite, dick gestrichelte, willkürliche Vorschlag liefert einen breiteren Abstandsbereich (gekennzeichnet durch die dünnen, parallelen Geraden). Die Support Vector Machine basiert auf den diesen Abstand am meisten beeinflussenden Datenpunkten, den sogenannten Support Vektoren. In ↘ Grafik 4 wurden letztere eingekreist und die optimale trennende Gerade eingezeichnet.

Die optimale trennende Gerade wurde anhand von bereits klassifizierten Unternehmen ermittelt, das heißt anhand von Unternehmen, für welche sowohl die Ausprägungen von X_1 und X_2 als auch die Ausprägung von „Zugehörigkeit zum Handwerk“ bereits bekannt sind. Einen solchen sogenannten Trainingsdatensatz muss man bei maschinellen Lernverfahren wie SVM voraussetzen.[4] Wie bei jeder statistischen Methode bedarf es also auch hier einer Stichprobe mit gesicherten Informationen über die zu betrachtende Grundgesamtheit.

Im hier vorgestellten Projekt entstammen die Trainingsdaten dem URS der Bezugsjahre 2011 und 2012.

4 Für Details zu statistischen maschinellen Lernverfahren siehe Dumpert/von Eschwege/Beck (2016).

Grafik 4
Support Vektoren und die optimale trennende Gerade



Nach dem Lernen der Support Vector Machine ist die Klassifizierung eines neuen Unternehmens ein einfacher Vorgang: Je nachdem, ob das Unternehmen mit seinen Ausprägungen von X_1 und X_2 ober- beziehungsweise unterhalb der trennenden Gerade liegt, wird es von der SVM als dem Handwerk zugehörig (+) beziehungsweise als dem Handwerk nicht zugehörig (△) klassifiziert. Bei mehr als zwei erklärenden Merkmalen würde keine trennende Gerade, sondern eine trennende Ebene (im Fall von drei erklärenden Merkmalen) beziehungsweise eine trennende Hyperebene (im Fall von mehr als drei erklärenden Merkmalen) ermittelt werden.

Vorzüge von SVM

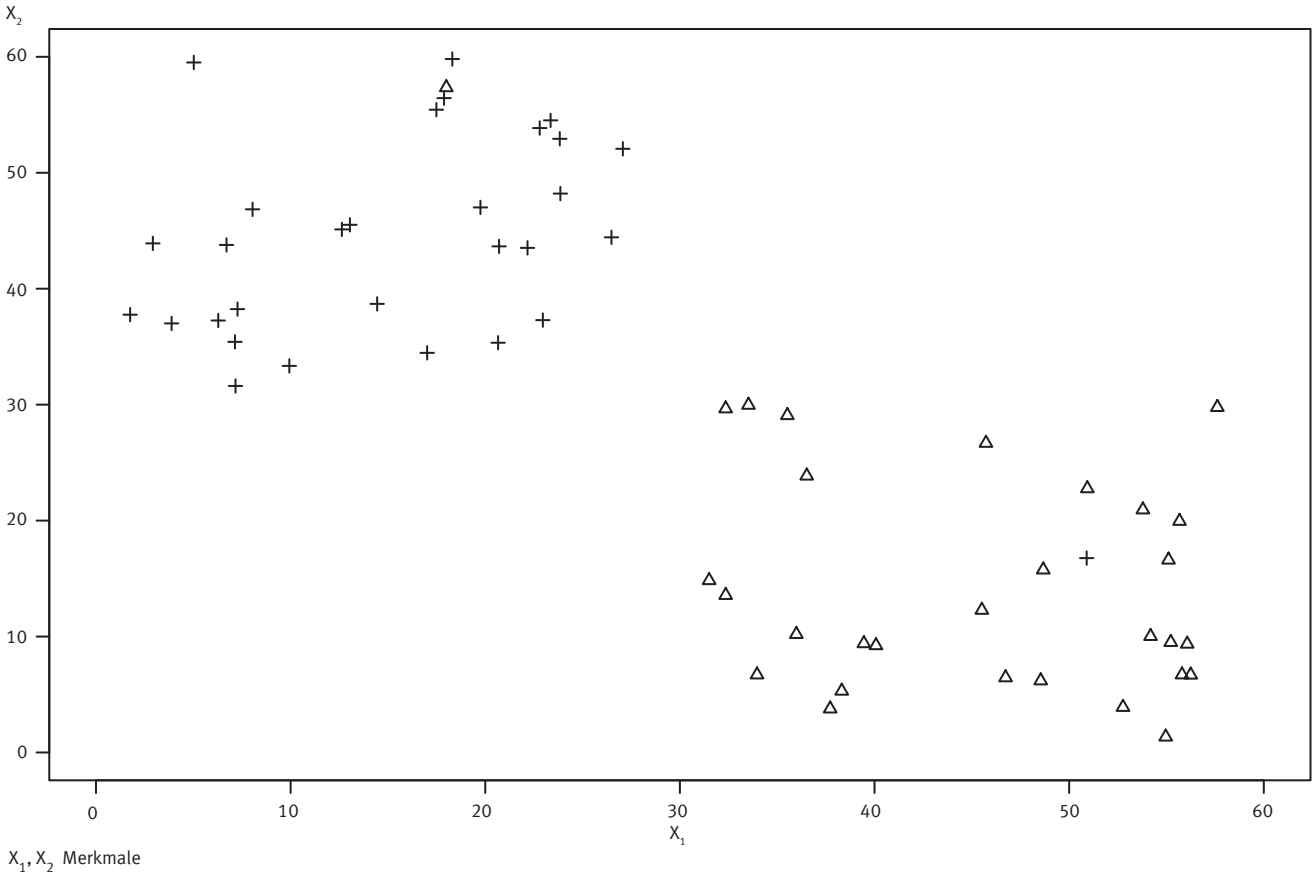
SVM sind robust, das heißt die SVM verändert sich nicht oder nur minimal, wenn sich die Daten im Trainingsdatensatz geringfügig ändern.[5] Dies erscheint wünschenswert, da Merkmalsausprägungen einzelner Unternehmen gegebenenfalls nur geschätzt oder ungenau erfasst wurden. Solche geringfügigen Ungenauigkeiten sollen sich nicht erheblich auf das Ergebnis auswirken. Weiterhin stellen Korrelationen zwischen den erklärenden Merkmalen kein Problem für das Lernen einer SVM dar. Eine weitere Stärke von SVM liegt darin, dass es völlig unerheblich ist, welcher Verteilung die Merkmale folgen.[6] Insbesondere wird keine (approximative) Normalverteilung benötigt.[7]

5 Anders als die logistische Regression im Fall von linear trennbaren Punktemengen.

6 Beispielsweise im Unterschied zur linearen oder quadratischen Diskriminanzanalyse.

7 Für Details zu nichtparametrischen Lernverfahren siehe ebenfalls Dumpert/von Eschwege/Beck (2016).

Grafik 5
Datensatz, bei dem eine lineare Trennung nicht möglich ist



2016 - 01 - 0138

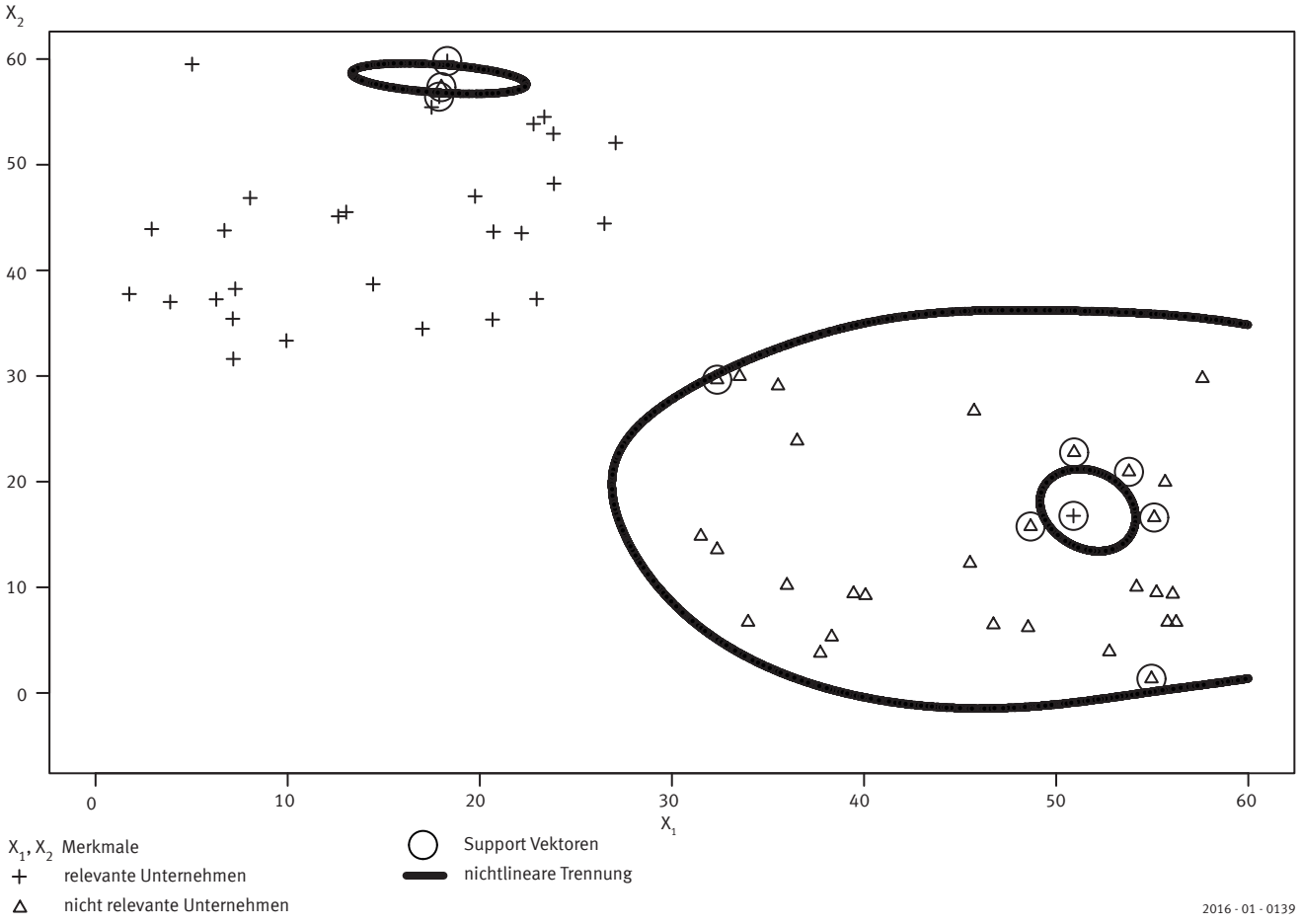
Überlappende Punktemengen und die Gefahr der Überanpassung

Das Beispiel aus den Grafiken 3 und 4 ist ein Spezialfall: Die Punktemengen lassen sich linear trennen. Dies ist bei dem in ↘ Grafik 5 gezeigten Beispiel anders. Jeweils ein Unternehmen aus dem Handwerks- beziehungsweise Nicht-Handwerks-Bereich liegt in der Punktemenge der anderen Gruppe.

Dies ist kein Problem für SVM. Es können auch nichtlineare Trennungsstrukturen ermittelt werden. Ein Beispiel hierfür ist in ↘ Grafik 6 dargestellt, in der die Support Vektoren wieder eingekreist sind. Im Vergleich zu Grafik 4 sind sehr viel mehr Support Vektoren notwendig, um die nichtlineare Trennung ohne Fehlzuordnung zu ermöglichen.

Dem Statistiker steht mit dem Trainingsdatensatz in der Regel nur eine Stichprobe oder ein irgendwie fehlerbehafteter Datenbestand zur Verfügung. Passt man die Support Vector Machine wie in Grafik 6 zu gut an den Trainingsdatensatz an, so läuft man Gefahr, dass der so gewonnene Klassifikator nicht mehr zu einem neuen Datensatz passt. Unternehmen, die hinsichtlich der Handwerkszugehörigkeit zur gleichen Klasse gehören, sich in den Merkmalsausprägungen von X_1 und X_2 aber geringfügig von Unternehmen im Trainingsdatensatz unterscheiden, können somit leicht der falschen Klasse hinsichtlich der Handwerksklassifikation zugeordnet werden. Um dies zu konkretisieren: Das dem Handwerk zugehörige Unternehmen mit $X_1 = 52$ und $X_2 = 18$ liegt inmitten von Nicht-Handwerks-Unternehmen. Davon ausgehend, dass der Trainingsdatensatz korrekt ist, handelt es sich vermutlich um ein atypisches Unter-

Grafik 6
Nichtlineare Trennung des Datensatzes



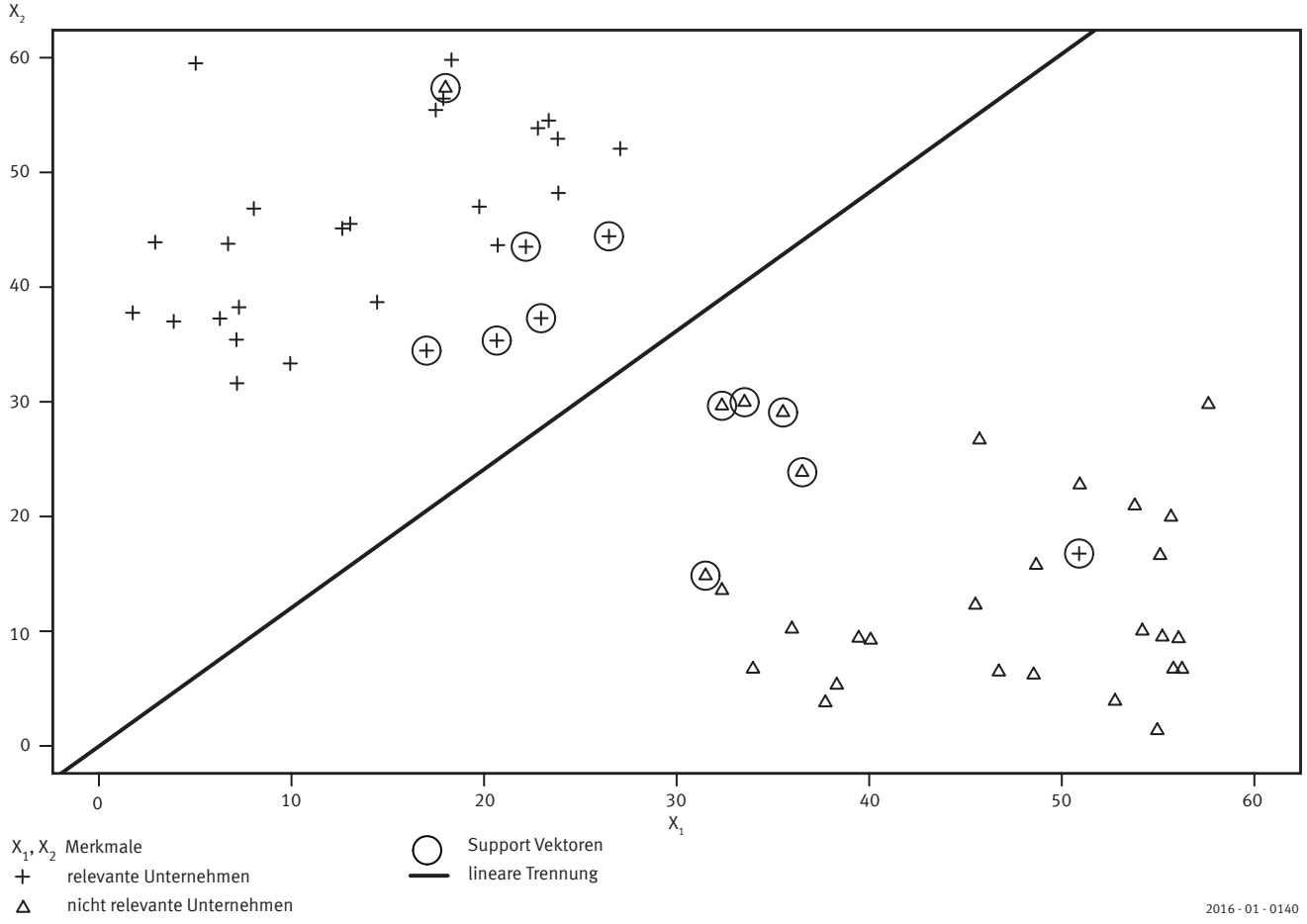
nehmen. In dieser Situation ist es zweckmäßig, auf die komplexere Trennung zu verzichten und stattdessen eine einfachere, zum Beispiel lineare Trennung zu verwenden, die jedoch Fehlzuordnungen zulässt. Letztere gehen dann mit einem negativen Wert in die Zielfunktion des Optimierungsproblems[8] der SVM ein. Sie werden dort mit einer Kostenkomponente (C) bewertet.[9] Wählt man diese sehr groß, um Fehlzuordnungen zu vermeiden, wird im Extremfall eine Art Insel im Bereich der Nicht-Handwerks-Unternehmen erzeugt: Neu zu klassifizierende Unternehmen, die mit ihrer Merkmalsausprägung im Bereich dieser Insel liegen, zum Beispiel mit $X_1 = 53, X_2 = 19$, würden (vermutlich irrtümlich) als dem Handwerk zugehörig klassifiziert. Entsprechendes gilt für die Insel im Bereich der Handwerks-Unternehmen. Eine derartige Überanpassung (overfitting) der SVM an die Trainingsdaten sollte daher vermieden werden. Die Trennung des Datensatzes durch eine Gerade, bei der zwei Fehlzuordnungen auftreten, zeigt ↘ Grafik 7.

Mit der Zahl der Unternehmen, die zum Lernen zur Verfügung stehen, steigt in der Regel die Qualität des Trainingsdatensatzes, da immer mehr Abhängigkeiten zwischen den Merkmalen und der Klassenzuordnung von Unternehmen durch die SVM erkannt werden können. In diesem Fall nehmen aber, bedingt durch die steigende Zahl an Trainingsunternehmen, auch die

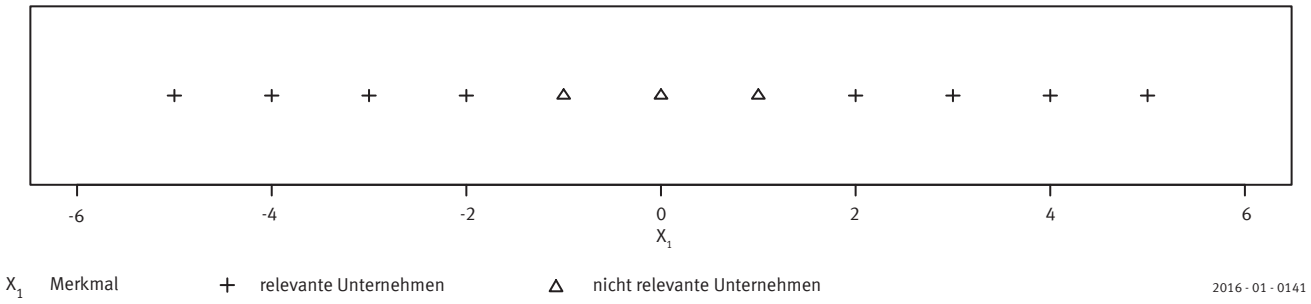
8 Für Details zur Überführung des hier beschriebenen geometrischen Optimierungsproblems in das von Dumpert/von Eschwege/Beck (2016) dargestellte analytische siehe Steinwart/Christmann (2008, hier: Seite 16 ff.).

9 Die Höhe der Kosten von Fehlzuordnungen kann frei bestimmt werden. Sie beeinflusst das Ergebnis des Optimierungsproblems erheblich. Deswegen ist es nötig, sie in dem im Abschnitt 4.3 beschriebenen Tuningprozess optimal zu wählen.

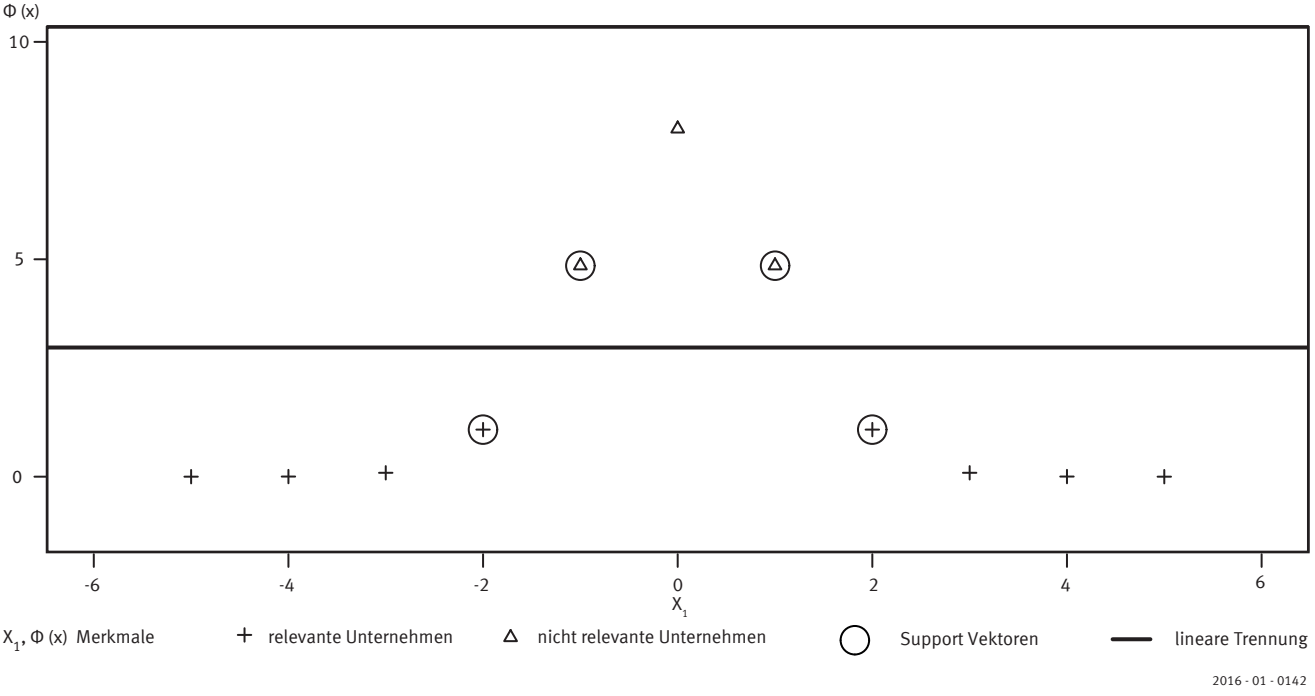
Grafik 7
Lineare Trennung mit Fehlzuordnungen



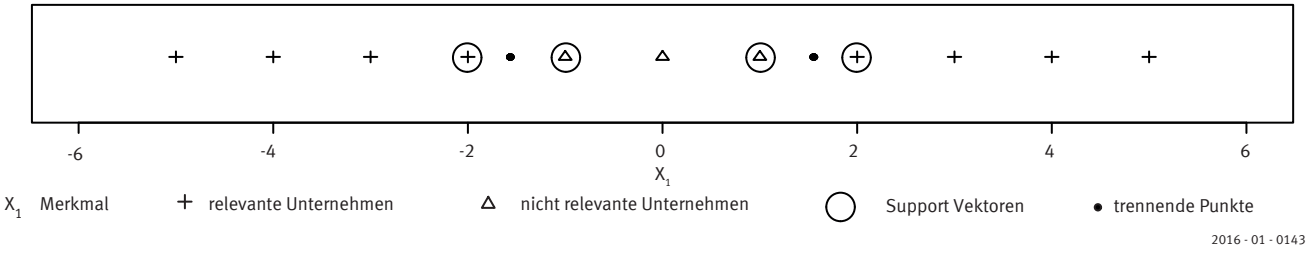
Grafik 8
Linear nicht zu trennender Datensatz im eindimensionalen Raum



Grafik 9
Lineare Trennung im höherdimensionalen Raum



Grafik 10
Nichtlineare Trennung im ursprünglichen eindimensionalen Raum (zwei trennende Punkte statt nur einem)



Fehlzuordnungen zu und gegebenenfalls ist eine lineare Trennung bei einer zu hohen Fehlerzahl nicht mehr sinnvoll. Die Aufgabe der SVM besteht nun darin, ein optimales Maß zwischen dem völligen Verkennen der in den Trainingsdaten enthaltenen Strukturen und einer Überanpassung an diese Strukturen zu finden. Dies bedeutet meist eine Abkehr von einer Trennung durch eine Gerade, Ebene oder Hyperebene: Die Trennungslinie muss geschwungen sein, um die zugrunde liegenden Strukturen abbilden zu können. Die Abkehr von einer linearen Trennung ist möglich, ohne das oben eingeführte Grundprinzip der Abstandsmaximierung aufgeben zu müssen.

Finden einer geschwungenen Trennlinie

Das Auffinden einer linearen Trennung (gegebenenfalls mit in Kauf genommenen Fehlzuordnungen) ist verhältnismäßig einfach und wird in den eingangs angegebenen Literaturstellen eingehend beschrieben. Erscheint eine lineare Trennung aber nicht mehr sinnvoll, muss auf geschwungene Trennlinien zurückgegriffen werden. Der im Folgenden anhand eines neuen, einfachen Beispiels beschriebene Ansatz, der sogenannte Kern-Trick, ist ein charakteristisches Merkmal von SVM und verwandten weiteren kernbasierten statistischen Methoden.

↘ Grafik 8 zeigt einen eindimensionalen Datensatz, der nicht durch einen einzigen Punkt (zu verstehen als „nulldimensionale Gerade“) zu trennen ist. Der Kern-Trick besteht darin, die Punkte derart in einer höheren Dimension anzuordnen, dass dort eine lineare Trennung möglich ist. Dazu wird jeder Datenpunkt in eine Funktion Φ eingesetzt.[10] Im vorliegenden Beispiel ist $\Phi(x) = 2^3 \cdot e^{-x^2/2}$.

Durch die weitere Dimension wird die lineare Trennung möglich, wie ↘ Grafik 9 zeigt. SVM transformieren das Problem also in einen Raum höherer Dimension, lösen es dort unter schwachen, leicht zu prüfenden Voraussetzungen linear und transformieren anschließend die Lösung wieder zurück in den ursprünglichen Raum. Die Rücktransformation für das Beispiel zeigt ↘ Grafik 10.

Wie eine Rücktransformation in einem Raum mit zwei die Klasse bestimmenden Merkmalen aussehen kann, wurde in Grafik 6 gezeigt.

10 Die Tupel $(x,\Phi(x))$ sind dann Elemente eines höherdimensionalen (meist sogar unendlich-dimensionalen) Raumes. Im konkret vorliegenden Beispiel ist $(x,\Phi(x))$ zweidimensional.

Die Art der dazu notwendigen Funktion Φ (zum Beispiel Polynom- oder Exponentialfunktion) wird durch die Wahl des sogenannten Kerns festgelegt, woher die Bezeichnung Kern-Trick für dieses Vorgehen stammt. Der Anwender kann zur Vorfestlegung eine Menge von Kernen gleichen Typs vorgeben und damit den Anpassungsspielraum der SVM festlegen. Die Menge der auf einer Exponentialfunktion basierenden sogenannten Gauß-Kerne liefert Funktionen Φ, die gewünschte statistische Eigenschaften der SVM garantieren. Hierzu gehören insbesondere die Existenz und Eindeutigkeit der punktetrennenden und damit der die Klassenzugehörigkeit vorhersagenden Linie im Klassifikationsfall sowie, dass mit zunehmender Größe des Trainingsdatensatzes die Trennung immer näher an der für die Grundgesamtheit aller Unternehmen unbekannten besten Trennlinie verläuft, die Klassenzuordnung also immer besser wird.

Schwierigkeiten beim Einsatz von SVM

Die Transformation des Trainingsdatensatzes sowie die Optimierung im höherdimensionalen Raum stellen in der praktischen Anwendung hohe Anforderungen an Speicher- und Rechenkapazität. Zwar ist eine SVM theoretisch in der Lage, mit sehr vielen erklärenden Merkmalen umzugehen, allerdings steigt mit der Zahl der Merkmale der Speicher- und Rechenaufwand in den meisten Fällen stärker als linear (Steinwart/Christmann, 2008, hier: Seite 420 ff.). Als Statistiker ist man aber natürlich daran interessiert, alle die Klassifikation beeinflussenden Abhängigkeiten zwischen den Merkmalen zu erfassen. Eine Vorauswahl der „wichtigsten“ erklärenden Variablen durch Random Forests ist gegebenenfalls geeignet, dieses Problem zu entschärfen.

3.2 Random Forests

Übersicht und wesentliche Idee

Grundlegend für die Entwicklung der Idee von Random Forests ist die Veröffentlichung von Breiman und anderen (1984). Sie stellen dabei sogenannte Bäume (tree-based methods) als Klassifikations- und Regressionsmethode vor. Ebenso wie Support Vector Machines stellen Bäume eine nicht-parametrische, das heißt ver-

teilungsunabhängige, statistische Lernmethode dar.[11] Mathematische Details und weitergehende Informationen zu dieser Methode finden sich beispielsweise in James und andere (2013, hier: Seite 303 ff.).

Die wesentliche Idee eines Baumes besteht darin, die Grundgesamtheit anhand der Informationen so in nicht überlappende Regionen aufzuteilen, dass die Unternehmen innerhalb derselben Region sehr große Übereinstimmung hinsichtlich ihrer Merkmalsausprägungen (insbesondere der Handwerkszugehörigkeit) haben (also sehr homogen sind), sich aber möglichst stark von den Unternehmen außerhalb dieser Region unterscheiden. In jedem Schritt ist an dieser Stelle daher für jede an diesem Punkt existierende Region ein Optimierungsproblem zu lösen, welches davon abhängt, wie der Anwender den Begriff der Homogenität definiert. Im Falle der Klassifikation ist dies regelmäßig die Reinheit einer Region, das heißt die Eigenschaft, hauptsächlich Unternehmen von nur einer Klasse, also entweder relevante oder nicht relevante Handwerksunternehmen, in der Region zu enthalten.[12] Vereinfacht formuliert geht der Algorithmus für jede der im jeweiligen Schritt betrachteten Regionen wie folgt vor: Die Region wird probeweise in alle denkbaren, beziehungsweise im Hinblick auf die erklärenden Merkmale sinnvollen, Paare von Teilregionen aufgeteilt. Anschließend werden die dadurch entstandenen neuen Reinheitsgrade berechnet, die in der Summe größer sind als der Reinheitsgrad der ursprünglichen Region. Die Aufteilung, welche die stärkste Zunahme an Reinheit erreicht, wird gewählt (Breiman und andere, 1984, hier: Seite 26 ff.).

Ein später neu hinzukommendes, also zu klassifizierendes Unternehmen wird anhand der erklärenden Merkmale einer Region zugewiesen. Die Frage, ob es die statistische Handwerkseigenschaft besitzt oder nicht, wird anhand der in dieser Region vorherrschenden Unternehmen (Handwerks-Unternehmen oder Nicht-Handwerks-Unternehmen) beantwortet.

11 Die Frage, ob ein Unternehmen dem Handwerk zugehörig ist oder nicht, könnte man theoretisch also auch mit der Methode der Random Forests alleine lösen. Im Allgemeinen ist die Güte von Bäumen jedoch schlechter als die anderer Klassifikationsmethoden (James und andere, 2013, hier: Seite 316). Konkret zeigte sich, dass die Klassifikation mittels Random Forests weniger zufriedenstellende Ergebnisse lieferte als die hier vorgestellte kombinierte Methode (siehe Tabelle 1).

12 Hinsichtlich verschiedener Maße für die Homogenität im Klassifikationsfall siehe beispielsweise James und andere (2013, hier: Seite 312).

Ermittlung der Regionen

Um die Regionen zu ermitteln, wird der Trainingsdatensatz, also die Menge der Unternehmen, anhand derer der Baum gelernt werden soll, schrittweise aufgegliedert. Im ersten Schritt wird der Trainingsdatensatz in zwei Untermengen aufgeteilt, im nächsten Schritt jede der Untermengen in zwei weitere Untermengen und so weiter. Theoretisch kann dieses Vorgehen so lange fortgesetzt werden, bis jedes Unternehmen im Trainingsdatensatz seine eigene Region bildet. Aus Anwendersicht ist dies aber nicht sinnvoll, denn es impliziert eine Überanpassung des Baumes an den Trainingsdatensatz und birgt die Gefahr schlechter Resultate bei späteren Klassifizierungen neuer Unternehmen. Erneut besteht also ein Trade-Off zwischen sehr guter Anpassung (und somit Erfassung der in den Trainingsdaten enthaltenen Zusammenhänge in der Grundgesamtheit) und späterer Einsetzbarkeit der Methode. Während sich der Baum also zunächst sehr tief verzweigen darf, muss er anschließend wieder zurückgeschnitten werden (pruning), um dem Trade-Off zu begegnen. Dabei wird anschaulich gesprochen für jeden möglichen Teilbaum, der durch Wiedervereinigen von Regionen entsteht, die Rate der Fehlzuordnungen geschätzt; der Teilbaum mit der kleinsten solchen Rate wird schließlich ausgewählt.[13]

Random Forests und Nutzung des Ergebnisses

Random Forests berechnen nun mehrere Bäume, basierend auf zufälligen Stichproben[14] aus dem Trainingsdatensatz und mitteln in geeigneter Weise die Ergebnisse.

Ein wesentlicher Vorteil von Random Forests ist nun die Tatsache, dass die vorgenommene Einteilung der Grundgesamtheit in Regionen Informationen über hierfür relevante erklärende Merkmale liefert. So kann man in jedem Schritt nachvollziehen, welches Merkmal für die Aufteilung verantwortlich ist. Die Beiträge der einzelnen Merkmale im schließlich entstandenen Random Forest ermöglichen einen Überblick darüber,

13 In der Praxis werden fortgeschrittene Verfahren (insbesondere cost complexity pruning) eingesetzt, um den Rechenaufwand zu reduzieren.

14 Die Stichproben werden durch Ziehen mit Zurücklegen aus dem Trainingsdatensatz gewonnen, es handelt sich um ein sogenanntes Bootstrap-Verfahren (James und andere, 2013, hier: Seite 187 ff.).

welche Merkmale entscheidend für die Klassifikation sind.[15]

Die so ermittelten Merkmale werden anschließend für das Lernen der Support Vector Machine genutzt.

4

Das Verfahren zur maschinellen Klassifizierung nicht relevanter Unternehmen

Nachdem das grundlegende Problem und die theoretischen Grundlagen zur Lösung eingeführt wurden, wird nun der Lösungsansatz dargestellt. Ziel ist es, ein Modell für die Klassifizierung der Relevanz von Unternehmen für die Handwerksstatistiken zu ermitteln. Dafür werden die oben beschriebenen nicht-parametrischen maschinellen Lerner und ein Trainingsdatensatz aus dem URS, in dem die Relevanz der Unternehmen für die Handwerksstatistik zugeordnet ist, verwendet.

4.1 Der Trainingsdatensatz

Als Trainingsmaterial steht ein URS-Auszug zur Verfügung, der verschiedene Informationen zu Größe, Handwerkseigenschaft und Struktur der zu klassifizierenden Unternehmen enthält. ↘ Übersicht 1 Er umfasst rund eine Million Beobachtungen handwerksrelevanter Unternehmen aus dem URS der Bezugsjahre 2011 und 2012. Zwei Bezugsjahre wurden gewählt, um mehr Beobachtungen seltener Gewerbezweige auswerten zu können. Die Fälle folgen in den Verteilungseigenschaften weitgehend den Angaben aus dem Kapitel 2 „Statistische Eckdaten und Komplexität des Problems“. Es konnten also in erheblichem Maße trivial zu klassifizierende Einheiten entfernt werden, um die Handhabung des Problems zu verbessern. Nach Ausschluss dieser Unternehmen blieben rund 68 000 Einheiten, die maschinell zu klassifizieren waren. Der Datensatz enthielt nun 6,5 % für die Handwerksstatistik nicht relevante Unternehmen, die 34,6 % der Umsätze repräsentieren.

Betrachtet man die vorliegenden Merkmale, stellt man zusätzlich fest, dass viele dieser Merkmale untereinander korreliert und teils hoch korreliert sind. Verfahren, die unabhängig verteilte erklärende Variablen als Voraussetzung haben, könnten nicht direkt angewandt werden. Sowohl SVM- als auch Random-Forest-Algorithmen haben diese Voraussetzung nicht.

4.2 Ermitteln besonders aussagekräftiger Merkmale mit Random Forest

Auf die rund 68 000 Unternehmen mit jeweils 330 relevanten Merkmalen[16] aus dem Trainingsmaterial wurde im ersten Schritt der oben beschriebene Random-Forest-Klassifikator angewandt. Mit ihm wurden die für das Klassifizierungsproblem relevantesten Variablen ermittelt. Als Nebenprodukt liefert er zusätzlich ein Modell, mit dem sich die Unternehmen ebenfalls klassifizieren lassen.

Da nominal skalierte Variablen ihren Ausprägungen entsprechend in Dummys zerlegt werden, gelingt es, nicht nur besonders relevante Merkmale, sondern auch besonders relevante Ausprägungen einzelner Merkmale zu identifizieren. Das heißt Wirtschafts- beziehungsweise Gewerbezweige, in denen nicht relevante Unternehmen keine beziehungsweise eine sehr kleine Rolle spielen, werden nicht in das Klassifizierungsproblem für die SVM übernommen.

4.3 Ansatz der SVM auf den reduzierten „Merkmalskranz“

Aus dem Trainingsmaterial mit rund 68 000 Unternehmen wurden mit der Information aus dem ersten Schritt die 30 relevantesten Variablen ausgewählt.[17] Auf diesen Datensatz wurde im zweiten Schritt der SVM-Algorithmus mit einem Gauß-Kern angesetzt.

15 Diese Übersicht erhält man somit gegebenenfalls schneller als über den Weg der Faktoren- oder der Hauptkomponentenanalyse, insbesondere bei sehr vielen infrage kommenden Merkmalen wie im vorliegenden Fall.

16 Alle nominal skalierten Variablen gehen mit je einer Dummy-Variable pro Ausprägung in den Trainingsdatensatz ein. Übersicht 1 legt nahe, dass es dann weit mehr als 330 Ausprägungen geben sollte. Die geringere Zahl ergibt sich, weil die Ausprägungen, die ausschließlich bei trivial klassifizierbaren Unternehmen auftreten, nicht berücksichtigt werden müssen.

17 Es wurden 30 ausgewählt, weil diese Anzahl mit der zur Verfügung stehenden Hardware in vertretbarer Zeit zu bewältigen war.

Übersicht 1

Merkmale des Trainingsdatensatzes

Merkmal	Beschreibung	Bemerkung
Land des Unternehmenssitzes	01 = Schleswig-Holstein 02 = Hamburg 03 = Niedersachsen 04 = Bremen 05 = Nordrhein-Westfalen 06 = Hessen 07 = Rheinland-Pfalz 08 = Baden-Württemberg 09 = Bayern 10 = Saarland 11 = Berlin 12 = Brandenburg 13 = Mecklenburg-Vorpommern 14 = Sachsen 15 = Sachsen-Anhalt 16 = Thüringen	Das Merkmal ist als erklärendes Merkmal relevant, weil es zwischen den Ländern teils erhebliche Unterschiede in der Verteilung von Umsatz und Beschäftigten in den Unternehmen gibt, die bei der Klassifizierung berücksichtigt werden sollen.
Änderungsart	1 = Neuaufnahme 2 = Änderung mit Erhalt des ursprünglichen Satzes, neues Änderungsdatum 3 = Einheit ruht 4 = Einheit erloschen 5 = Einheit in ein anderes Land gewechselt 6 = Einheit aus einem anderen Land zugezogen 8 = Einheit wieder aktiv A = Änderung mit Erhalt der Änderungsart und altem Änderungsdatum	In diesem Merkmal ist der Grund kodiert, warum ein Unternehmen im URS zuletzt geändert wurde. Man findet hier auch Informationen, ob Unternehmen wirtschaftlich inaktiv sind. Inaktive Unternehmen sind oft auf nicht relevant für die Statistik gesetzt, wenn von den Handwerkskammern noch kein Löschdatum gemeldet wurde.
Art der Einheit	1 = Einbetriebsunternehmen 2 = Mehrbetriebsunternehmen 3 = Mehrländerunternehmen	Dieses Merkmal ist mit der Größe der Unternehmen korreliert. Zusätzlich ist die Wahrscheinlichkeit bei Mehrländer- und Mehrbetriebsunternehmen höher, handwerkliche Nebenbetriebe zu haben.
Verbundenes Unternehmen	0 = nicht verbunden 1 = verbundenes Unternehmen	Für dieses Merkmal gilt die gleiche Vermutung wie bei „Art der Einheit“.
Wirtschaftszweig	Wirtschaftliche Aktivität des Unternehmens nach WZ 2008 klassifiziert	Das Merkmal hat im Trainingsdatensatz 556 nominale Ausprägungen und ist in Verbindung mit dem Gewerbezweig relevant für die Klassifikation. Es ist stark mit dem Merkmal Gewerbezweig korreliert.
Gewerbezweig	Hauptgewerbezweig nach Handwerksordnung klassifiziert	Das Merkmal hat 94 Ausprägungen und ist stark mit dem Merkmal Wirtschaftszweig korreliert.
Organschaftszugehörigkeit	0 = nicht in Organschaft 1 = in Organschaft	Das Merkmal Organschaftszugehörigkeit eines Unternehmens ist, wie das zu klassifizierende Relevanz-Merkmal, mit der Unternehmensgröße korreliert.
Umsatz	Jahresumsatz in 1 000 Euro	–
Tätige Personen	Tätige Personen am 31.12. des Bezugsjahres	–
Sozialversicherungspflichtig Beschäftigte	Sozialversicherungspflichtig Beschäftigte am 31.12. des Bezugsjahres	–

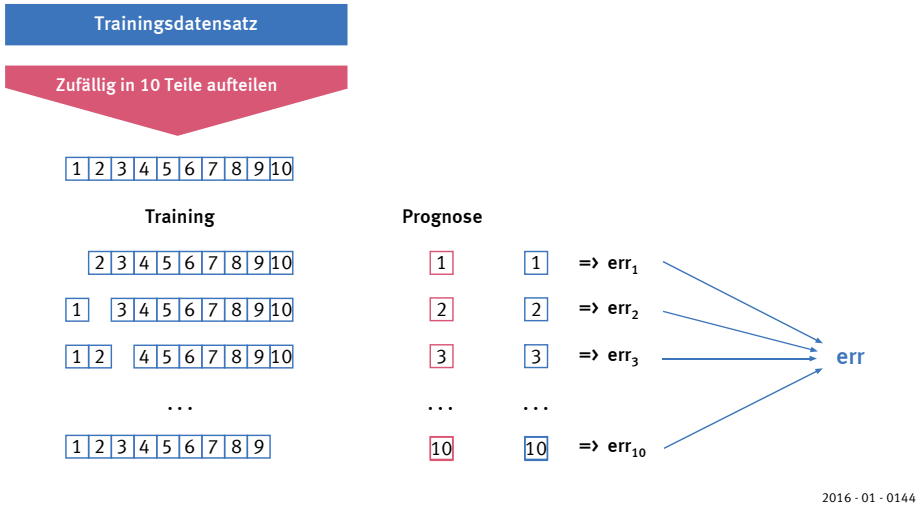
Wie oben erwähnt ist das Tuning einer SVM erfolgskritisch. Bei Gauß-Kernen müssen für zwei Parameter möglichst optimale Werte ermittelt werden. Erstens für die Kosten (C), die ein Ausreißer bei der Anpassung verursachen soll. Ist dieser Parameter hoch, werden in überlappend verteilten Punktemengen auch um sehr kleine Gruppen von Fällen oder um Einzelfälle, die weit entfernt vom Zentrum der Verteilung liegen, noch Klassifikationsgrenzen ermittelt.

Zweitens muss ein Wert für die Form des Gauß-Kerns gewählt werden (γ). Vereinfacht gesagt gibt er die Kurveneigenschaften der Kernfunktion an. Ist der Parameter hoch, können im Grenzbereich überlappender Verteilungen sehr fein geschwungene Grenzen um die Grenzfälle gezogen werden.[18]

Die Parameter sollen so gewählt werden, dass die Fehlklassifikationsrate möglichst gering ist und gleichzeitig keine Überanpassung auftritt. Als Qualitätsmerkmal wurde die Fehlklassifikationsrate bei zehnfacher Kreuzvalidierung verwendet. Man teilt dazu den Trainingsdatensatz zufällig in zehn gleich große Teile. Dann

18 Betrachtet man analog die analytische Herangehensweise aus Dumpert/von Eschwege/Beck (2016), so entsprechen sich die Parameter γ exakt. Auch für die Kosten C gibt es dort einen korrespondierenden Parameter: λ. C und λ verhalten sich jedoch indirekt proportional zueinander.

Grafik 11
Vorgehen bei zehnfacher Kreuzvalidierung



ermittelt man für jeden der zehn Teile, jeweils mit den verbliebenen neun Teilen als Trainingsmaterial, ein Modell. Mit dem Modell klassifiziert man den Teil, der nicht zum Trainieren verwendet wurde, und vergleicht die geschätzte Klassifikation mit der bekannten. So erhält man eine Fehlklassifikationsrate für den Trainingsdatensatz. Gleichzeitig vermeidet man, dass Unternehmen mit einem Modell klassifiziert werden, das zuvor auf ihrer Grundlage ermittelt wurde. Eine detaillierte Beschreibung findet sich in Witten und andere (2011). Das Tuning der SVM wurde für die oben genannten Parameter mit einer Rastersuche im Bereich $10^{-5} \leq \gamma \leq 10^{5}$ und $10^{-5} \leq C \leq 10^{5}$ durchgeführt. Die beste Anpassung des Modells wurde bei einem Kostenparameter C von 1 000 und einem γ-Wert von 1 erreicht. ↘ Grafik 11

5 Ergebnis

Als Parameter zur Bewertung der Qualität des beschriebenen Verfahrens bietet sich die Fehlklassifikationsrate an. Sie gibt den Anteil der Einheiten des Trainingsdatensatzes an, die mit dem ermittelten Modell und bezogen auf die bekannte Klassenzugehörigkeit falsch klassifiziert werden. Mit dem oben beschriebenen Verfahren konnte eine Fehlklassifikationsrate von 3,7 % erreicht werden. Da die Klassifikation stark von der Unternehmensgröße abhängt, muss zusätzlich geprüft werden, ob unter den falsch klassifizierten Einheiten überproportional viele relativ große Einheiten sind. Nimmt man den Umsatz als Gewichtungsmaßstab, so ergibt sich, dass nur 1,8 % des Umsatzes falsch zugeordnet wurden. Zum Vergleich: Der Random-Forest-Algorithmus lieferte eine Fehlklassifikationsrate von 5,6 %. Dies wäre ein akzeptabler Wert, besonders, wenn man berücksichtigt, dass dieser Algorithmus viel weniger Ressourcen benötigt. Er ordnete jedoch 16,7 % des Umsatzes falsch zu. Bedenkt man, dass knapp 35 % der Umsätze von den 6,5 % nicht relevanten Unternehmen erwirtschaftet werden, ist dieser Wert bedenklich hoch. Die SVM lieferte also im Test bessere Ergebnisse.[19] ↘ Tabelle 1

Tabelle 1
Klassifikationsraten

Algorithmus	Vorhersage	Originalwert	Fallzahl	Umsatz
			%	
SVM	relevant	relevant	93,1	65,2
	nicht relevant	nicht relevant	3,2	33,1
	korrekt zugeordnet:		**96,3**	**98,2**
	nicht relevant	relevant	0,3	0,3
	relevant	nicht relevant	3,4	1,5
	falsch zugeordnet:		**3,7**	**1,8**
Random Forest	relevant	relevant	93,0	62,8
	nicht relevant	nicht relevant	1,4	20,5
	korrekt zugeordnet:		**94,4**	**83,3**
	nicht relevant	relevant	0,4	2,7
	relevant	nicht relevant	5,2	14,1
	falsch zugeordnet:		**5,6**	**16,7**

19 Häufige Wiederholungen mit kleinen Stichproben scheinen die Beobachtung zu stützen. Die SVM erreicht gewichtet mit dem Umsatz ähnliche Fehlklassifikationsraten wie im ungewichteten Fall. Der Random-Forest-Algorithmus klassifiziert die Umsätze in der Regel in größerem Maße falsch.

6

Fazit und Ausblick

Das vorgestellte Verfahren ist geeignet, im dargestellten Beispiel verwendbare Modelle zur Klassifizierung zu ermitteln. Ein mindestens unterstützender Einsatz bei der Klassifizierung nicht relevanter Unternehmen im Handwerk ist im laufenden Betrieb der Handwerksstatistiken möglich. Nach Klärung einiger noch offener Fragen soll im Rahmen der Gremien des Statistischen Verbundes über das Ob und Wie des produktiven Einsatzes maschineller Klassifikationsmethoden in den Handwerksstatistiken entschieden werden.

Wichtig wird jedoch auch weiterhin eine stichprobenartige manuelle Prüfung der Handwerkseigenschaft sein. Damit wird vermieden, dass auf längere Sicht künftige maschinelle Klassifizierungsverfahren nur noch auf Basis der Ergebnisse früherer maschineller Klassifizierungsläufe lernen. Wie groß und in welcher Weise strukturiert die Stichproben sein müssen, wird Gegenstand weiterer Untersuchungen im Rahmen des vorgestellten Projekts sein.

Abschließend lässt sich festhalten, dass maschinelle Lernverfahren das Potenzial haben, Klassifizierungsprobleme auch in anderen Bereichen erfolgreich anzugehen. Dies zeigt auch ein Projekt zur Sektorzuordnung von Unternehmen (Dumpert/von Eschwege/Beck, 2016). Weiterhin soll untersucht werden, ob die Schätzmodelle zum sogenannten bereinigten Verdienstunterschied von Männern und Frauen mittels SVM verbessert werden können.

LITERATURVERZEICHNIS

Boser, Bernhard E./Guyon, Isabelle M./Vapnik, Vladimir N. *A Training Algorithm for Optimal Margin Classifiers*. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. Pittsburgh 1992, Seite 144 ff.

Breiman, Leo/Friedman, Jerome H./Olshen, Richard A./Stone, Charles J. *CART: Classification and Regression Trees*. Boca Raton 1984.

Campbell, Colin/Ying, Yiming. *Learning with Support Vector Machines*. San Rafael 2011.

Cortes, Corinna/Vapnik, Vladimir N. *Support-Vector Networks*. In: Machine learning. Jahrgang 20 (1995). Band 3, Seite 273 ff.

Dumpert, Florian/von Eschwege, Katja/Beck, Martin. *Einsatz von Support Vector Machines bei der Sektorzuordnung von Unternehmen*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2016, Seite 87 ff.

Feuerhake, Jörg. *Handwerkszählung 2008*. In: Wirtschaft und Statistik. Ausgabe 1/2012, Seite 51 ff.

Hamel, Lutz. *Knowledge Discovery with Support Vector Machines*. Hoboken 2009.

James, Gareth/Witten, Daniela/Hastie, Trevor/Tibshirani, Robert. *An Introduction to Statistical Learning*. New York 2013.

Neuhäuser, Jenny. *Verwaltungsdaten ersetzen Konjunkturerhebungen im Handwerk*. In: Wirtschaft und Statistik. Ausgabe 5/2008, Seite 398 ff.

Statistisches Bundesamt. *Klassifikation der Wirtschaftszweige, Ausgabe 2008*. Wiesbaden 2009.

Steinwart, Ingo/Christmann, Andreas. *Support Vector Machines*. New York 2008.

Witten, Ian H./Frank, Eibe/Hall, Mark A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. Auflage. Burlington 2011.

Herausgeber
Statistisches Bundesamt, Wiesbaden
www.destatis.de

Schriftleitung
Dieter Sarreither, Präsident des Statistischen Bundesamtes
Redaktionsleitung: Kerstin Hänsel
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im April 2016
Das Archiv aller Ausgaben ab Januar 2001 finden Sie unter www.destatis.de/publikationen

Print
Einzelpreis: EUR 18,– (zzgl. Versand)
Jahresbezugspreis: EUR 108,– (zzgl. Versand)
Bestellnummer: 1010200-16002-1
ISSN 0043-6143
ISBN 978-3-8246-1044-0

Download (PDF)
Artikelnummer: 1010200-16002-4, ISSN 1619-2907

Vertriebspartner
IBRo Versandservice GmbH
Bereich Statistisches Bundesamt
Kastanienweg 1
D-18184 Roggentin
Telefon: +49 (0) 382 04 / 6 65 43
Telefax: +49 (0) 382 04 / 6 69 19
destatis@ibro.de

Papier: Metapaper Smooth, FSC-zertifiziert, klimaneutral, zu 61% aus regenerativen Energien