EINSATZ VON SUPPORT VECTOR MACHINES BEI DER SEKTOR-ZUORDNUNG VON UNTERNEHMEN

Florian Dumpert, Katja von Eschwege, Martin Beck

↘ **Schlüsselwörter:** Support Vector Machines – Dritter Sektor – Nichtparametrische Methoden – Statistisches Unternehmensregister

ZUSAMMENFASSUNG

Der vorliegende Artikel beschreibt die Motivation, die Herangehensweise und die Ergebnisse der probeweisen Anwendung von Support Vector Machines im Bereich des statistischen Unternehmensregisters. Eine Support Vector Machine ist ein universell einsetzbares maschinelles Lern- und Klassifikationsverfahren. Auf der Basis eines mathematischen Optimierungsansatzes können Objekte nach bestimmten Merkmalen klassifiziert und den entsprechenden Klassen zugeordnet werden. Diese nichtparametrische statistische Methode klassifiziert Unternehmen erfolgreich hinsichtlich ihrer Zugehörigkeit zum sogenannten Dritten Sektor und ist daher geeignet, das bislang dafür eingesetzte Verfahren zu verbessern und zu ergänzen.

↘ ***Keywords:*** *Support vector machines – third sector – non-parametric methods – business register*

ABSTRACT

This article shows the motivation for, approach to, and results of applying support vector machines in official statistics concerning the business register for test purposes. A support vector machine is a universally applicable machine learning and classification method. Based on a mathematical optimisation approach, objects can be classified by specific variables and be allocated to corresponding classes. The non-parametric statistical method succeeded in classifying enterprises with respect to the so-called third sector and is therefore suitable to improve and complement the method used up to now.

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.

Katja von Eschwege

ist Diplom-Kauffrau und Referentin im Bereich „Unternehmensregister“ des Statistischen Bundesamtes. Sie ist mit verschiedenen Aufgaben zur Weiterentwicklung und Methodik des statistischen Unternehmensregisters befasst. Hierzu gehört auch die Sektorkennzeichnung von Unternehmen, unter anderem für Zwecke der Volkswirtschaftlichen Gesamtrechnungen.

Martin Beck

ist Diplom-Ökonom und leitet seit 2007 im Statistischen Bundesamt die Gruppe „Unternehmensregister, Klassifikationen, Verdienste, übergreifende Unternehmensstatistiken“. Er befasst sich derzeit unter anderem damit, die Prozesse der Datengewinnung durch Einführung neuer statistischer Methoden effizienter zu gestalten.

1 Einleitung

Um Arbeitsprozesse effizienter zu gestalten, ist die amtliche Statistik bestrebt, den Einsatz neuer statistischer Verfahren voranzubringen. In der Strategie 2016 des Statistischen Bundesamtes heißt es dementsprechend unter dem Punkt Wirtschaftlichkeit: „Wir schaffen Handlungsspielräume durch Veränderung unserer Methoden, Verfahren und Strukturen." Support Vector Machines (SVM) wurden in den letzten Jahren erfolgreich in ganz unterschiedlichen Anwendungsbereichen, wie zum Beispiel Handschriften- und Bilderkennung, Genomanalyse und Astrophysik, eingesetzt. In der amtlichen Statistik spielten sie hingegen bislang keine Rolle. Mit dem Einsatz von Support Vector Machines bei der Zuordnung von Unternehmen zu den institutionellen Sektoren nach dem Europäischen System Volkswirtschaftlicher Gesamtrechnungen (ESVG) 2010 betritt die amtliche Statistik Neuland und schafft die Möglichkeit, unter bestimmten Bedingungen aufwendige Recherchen zur Klassifizierung vieler Einzelfälle künftig durch diese Methode des maschinellen Lernens zu ersetzen und sich dabei der Erkenntnisse einmal(ig) durchgeführter Recherchen zu bedienen. Entsprechend der Zielsetzung „Wir nutzen die innovative Kraft der Wissenschaft." entwickelte das Statistische Bundesamt die konkrete Methodik in Zusammenarbeit mit einem an der Universität Bayreuth tätigen Mathematiker.

Der vorliegende Beitrag beschreibt zunächst den fachlichen Hintergrund und Bedarf für die durchgeführten Tests des maschinellen Lern- und Klassifikationsverfahrens SVM, gibt einen Überblick über die Theorie von Support Vector Machines und stellt die Ergebnisse der Testrechnungen vor.

2 Hintergrund

Die institutionellen Sektoren nach dem ESVG 2010 sind nach einer Vorgabe der Europäischen Union (EU) als Merkmal im Unternehmensregister zu führen und werden unterschieden in Nichtfinanzielle Kapitalgesellschaften, Finanzielle Kapitalgesellschaften, Staat, Private Haushalte, Private Organisationen ohne Erwerbszweck (im Folgenden mit POOE abgekürzt) und den Sektor „Übrige Welt". Als wichtige Vorbedingung zur Bestimmung der Unternehmen des institutionellen Sektors „POOE" nimmt das Statistische Bundesamt jährlich auf der Basis des Unternehmensregisters eine maschinelle Zuordnung der Unternehmen zum sogenannten Dritten Sektor vor.

2.1 Dritter Sektor

Der Dritte Sektor, auch Non-Profit-Sektor genannt, bezeichnet einen eigenständigen Bereich jenseits des Staates und der privaten Unternehmen.[1]

Dritter Sektor

Entsprechend dem Handbuch der Vereinten Nationen umfasst der Dritte Sektor alle Unternehmen, die die folgenden fünf Kriterien erfüllen: (1) Sie agieren als Organisationen, (2) treten privat auf, (3) sind nicht gewinnorientiert, (4) sind selbstverwaltet und (5) zeichnen sich durch Freiwilligkeit aus (Vereinte Nationen, 2003).

Im April 2011 veröffentlichte das Statistische Bundesamt erstmals Daten zur wirtschaftlichen Bedeutung des Dritten Sektors. Diese wurden im Gemeinschaftsprojekt „Zivilgesellschaft in Zahlen" – initiiert und finanziert vom Stifterverband für die Deutsche Wissenschaft, der Bertelsmann Stiftung und der Fritz Thyssen Stiftung – gewonnen (Statistisches Bundesamt/Centrum für soziale Investitionen und Innovationen, 2011, sowie Rosenski, 2012).

2.2 Datengrundlage Unternehmensregister

Als Datenbasis für das oben genannte Projekt diente das statistische Unternehmensregister (im Folgenden Unternehmensregister genannt), dessen Unternehmen bezüglich ihrer Zugehörigkeit zum Dritten Sektor gekennzeichnet wurden.

1 Klar abzugrenzen vom Dritten Sektor ist der Tertiärsektor beziehungsweise Dienstleistungssektor.

Statistisches Unternehmensregister

Das statistische Unternehmensregister ist eine regelmäßig aktualisierte Datenbank mit Angaben zu Unternehmen und Betrieben aus nahezu allen Wirtschaftsbereichen mit steuerbarem Umsatz aus Lieferungen und Leistungen und/oder sozialversicherungspflichtig Beschäftigten. Quellen zur Pflege des Unternehmensregisters sind zum einen Dateien aus Verwaltungsbereichen, wie die Bundesagentur für Arbeit oder die Finanzbehörden, und zum anderen Angaben aus einzelnen Bereichsstatistiken, beispielsweise aus Erhebungen des Produzierenden Gewerbes, des Handels oder des Dienstleistungsbereichs. Das Unternehmensregister wird von den statistischen Ämtern der einzelnen Bundesländer sowie dem Statistischen Bundesamt gemeinsam geführt. Es dient als wichtiges Instrument zur rationellen Unterstützung statistischer Erhebungen und ermöglicht eigenständige Auswertungen (Nahm/Stock, 2004).

2.3 Maschineller Algorithmus zur Dritt-Sektor-Klassifizierung

Für einen Großteil der Unternehmen konnte die Zugehörigkeit zum Dritten Sektor anhand eines maschinellen Algorithmus ermittelt werden. Dieser griff auf Merkmale zurück, die im Unternehmensregister vorliegen, wie Wirtschaftszweig, Rechtsform und Name, sowie auf abgeleitete Informationen des Unternehmensregisters und weitere, in der amtlichen Statistik vorliegende Angaben, die sich mit dem Unternehmensregister verknüpfen lassen.

Für eine Restmenge, die nicht automatisiert zugeordnet werden konnte, wurden im Rahmen des Projektes einmalig in großem Umfang zeit- und arbeitsintensive Einzelfallrecherchen zur Dritt-Sektor-Eigenschaft von Unternehmen vorgenommen. Als Ergebnis lagen für eine große Zahl von Unternehmen belastbare Informationen über ihre Zugehörigkeit oder Nichtzugehörigkeit zum Dritten Sektor vor.

Seit Abschluss des Projektes nimmt das Statistische Bundesamt jährlich auf der Basis des Unternehmensregisters eine maschinelle Zuordnung der Unternehmen zum Dritten Sektor vor. Die Dritt-Sektor-Zuordnung dient in einem zweiten Schritt – neben weiteren Kriterien – als wichtige Vorbedingung zur Bestimmung der Unternehmen des institutionellen Sektors „POOE“, einer echten Teilmenge des Dritten Sektors.

Während der maschinelle Algorithmus zur Kennzeichnung des Dritten Sektors und der institutionellen Sektoren von Jahr zu Jahr weiterentwickelt wird, stehen dem Statistischen Bundesamt für manuelle Einzelfallrecherchen, wie sie im Rahmen des genannten Projektes durchgeführt wurden, keine ausreichenden Ressourcen zur Verfügung. So erhält eine Restmenge von nicht zuzuordnenden Einheiten kein eindeutiges Dritt-Sektor-Kennzeichen. Für diese Einheiten steht das Kriterium „Dritter Sektor“ als Hilfsmerkmal zur Abgrenzung der institutionellen Sektoren und insbesondere der POOE nicht zur Verfügung. Die Zahl der Einheiten des Dritten Sektors und in der Folge der POOE wird daher möglicherweise unterschätzt. Zwar werden die Ergebnisse der Einzelfallrecherchen zu Unternehmen, die noch aktiv sind, unter der Annahme der Kontinuität dieser Unternehmen weiterhin im Rahmen der jährlichen Sektorkennzeichnung verarbeitet; die Anzahl der Einzelfallrecherchen, die sich mit dem Unternehmensregister verknüpfen lassen, sinkt jedoch aufgrund von demografischen Veränderungen (zum Beispiel Unternehmensschließungen, Zusammenschlüssen von Unternehmen) von Jahr zu Jahr. Damit stellt sich die Frage nach einer Qualitätssicherung der exakten Abgrenzung des Dritten Sektors.

2.4 Aktuelle Datenlage

Anforderungen an die Qualität der Daten stellt zum einen der Hauptnutzer der Sektorkennzeichnung des Unternehmensregisters, also die Volkswirtschaftlichen Gesamtrechnungen. Zum anderen gewinnen Qualitätsanforderungen im Hinblick auf eine neue Rahmenverordnung über die gesamte Unternehmensstatistik[2] an Bedeutung: Künftig sollen im Rahmen der strukturellen Unternehmensstatistiken und Konjunkturstatistiken, die ihre Stichproben aus dem Unternehmensregister ziehen, ausschließlich Marktproduzenten erfasst werden, also keine Unternehmen des Sektors „Staat“ oder des Sektors „POOE“. Auch die Verdienststatistiken nutzen Sektorinformationen, um die relevante Grundgesamtheit im Rahmen der Stichprobenziehungen zu bestimmen. So wird auch im Zusammenhang mit der Erhebungsunterstützung eine trennscharfe Abgrenzung des Dritten Sektors als Vorbedingung zur Abgrenzung des Sektors der POOE bedeutsam.

2 Es handelt sich hierbei um FRIBS (Framework Regulation Integrating Business Statistics), ein von der EU geplantes Gesetzesvorhaben (Waldmüller/Weisbrod, 2015).

2.5 Ziel der Tests

Die Datenlage für den Bereich der nicht eindeutig dem Dritten Sektor zuzuordnenden Unternehmen soll künftig verbessert werden, ohne personalaufwendige Einzelfallrecherchen durchführen zu müssen. Daher wurde geprüft, ob eine ressourcenschonende Methode des maschinellen Lernens zum Einsatz kommen kann. Für diesen Zweck wurde ein Verfahren gewählt, das auf einer Support Vector Machine basiert.

2.6 Vorbereiten der Datenbasis, Inputvariablen

Die Grundlage für Testrechnungen zum maschinellen Lernen bildeten etwa 45 000 Unternehmen, die im Rahmen des Projektes „Zivilgesellschaft in Zahlen“ nicht maschinell zugeordnet werden konnten, für die aber aus den weiter oben beschriebenen Einzelfallrecherchen präzise und gesicherte Angaben zum Dritten Sektor ermittelt worden waren. Die Datengrundlage enthielt für jedes Unternehmen die folgenden Merkmale:

- Identifikationsnummer aus dem Unternehmensregister
- Wirtschaftszweig nach der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008)
- Rechtsform (qualitativ verbessert im Rahmen der maschinellen Sektorkennzeichnung)
- Wirtschaftszweig-Kategorie
- Rechtsform-Kategorie
- Beschaffenheit als öffentliches Unternehmen[3]
- Sitz des Unternehmens (Bundesland) auf Basis des amtlichen Gemeindeschlüssels
- Sitz des Unternehmens in einem Ballungsgebiet, abgeleitet aus dem amtlichen Gemeindeschlüssel[4]
- Dritter Sektor auf der Basis von Einzelfallrecherchen
- Teilnahme am Intrahandel (auf Basis der Identitätsnummer aus dem Intrahandelsregister)
- Anzahl der sozialversicherungspflichtig Beschäftigten, auch untergliedert nach Vollzeit- und Teilzeitbeschäftigten
- Anzahl der geringfügig entlohnt Beschäftigten aus dem Verwaltungsdatenspeicher der amtlichen Statistik
- Höhe des steuerbaren Umsatzes.

Kategorienbildung

Die Wirtschaftszweig-Kategorien werden auf Basis der Wirtschaftszweige gebildet und besitzen die Ausprägungen „Wirtschaftszweig ist untypisch für den Dritten Sektor“, „Wirtschaftszweig ist potenziell Dritter Sektor“ und „Wirtschaftszweig ist typisch für den Dritten Sektor“. Analog dazu werden die drei Rechtsform-Kategorien „Rechtsform ist untypisch für den Dritten Sektor“, „Rechtsform ist potenziell Dritter Sektor“ und „Rechtsform ist typisch für den Dritten Sektor“ unterschieden (Rosenski, 2012).

3 Darstellung der SVM-Methodik

Support Vector Machines (SVM) stellen eine Methode nichtparametrisch-statistischen maschinellen Lernens dar. Eingeführt durch die Arbeiten von Boser/Guyon/Vapnik (1992) und Cortes/Vapnik (1995) haben sich SVM zu einer verbreiteten statistischen Methode entwickelt. Eine allgemeinverständliche Einführung ist Hamel (2009). Die nachfolgenden Ausführungen beziehen sich in der Regel auf die Klassifikation. SVM eignen sich darüber hinaus auch für Regressionen und die sogenannte Outlier oder Novelty Detection (Hamel, 2009, hier: Kapitel 12/13).

Maschinelles Lernen

Simon (1983) definiert den maschinellen Lernvorgang als adaptive Änderungen des Systems, in dem Sinne, dass sie das System in die Lage versetzen, die gleiche(n) Aufgabe(n) auf Basis der gleichen Population bei Wiederholung effizienter oder effektiver zu erfüllen (Simon, 1983, hier: Seite 28). Das zu erstellende Programm (genauer: der resultierende Prädiktor zur Klassifikation) ist ein solches System. Hat es einmal gelernt, so soll die

3 Zu den öffentlichen Unternehmen zählen nach § 2 Absatz 3 Finanz- und Personalstatistikgesetz alle Unternehmen, an deren Nennkapital die öffentliche Hand (Bund, Länder, Gemeinden) mit mehr als 50 % beteiligt ist.

4 Bei der Ermittlung des Ballungsgebiets wurden Informationen des Gemeindeverzeichnisses genutzt.

zu erfüllende Aufgabe im Anschluss effizienter (hier: ressourcenschonender) und/oder effektiver gelöst werden. Zum einen wird dadurch klar, dass maschinelles Lernen in diesem Sinne nur funktionieren kann, wenn hinreichend gutes Material vorhanden ist, anhand dessen gelernt werden kann. Zum anderen kann das Ziel des maschinellen Lernens ins Leere laufen, wenn sich die Aufgabe nach dem Lernen nicht mehr auf die gleiche Population bezieht: Hat das System anhand von Population A gelernt, so ist seine Anwendung auf Population B zwar gegebenenfalls noch möglich, aber sicherlich nicht mehr sinnvoll, wenn sich die Populationen A und B deutlich unterscheiden.

↘ Nichtparametrisch-statistische Methode

Beim Lernen von SVM handelt es sich um einen statistischen Vorgang in dem Sinne, dass (im Wesentlichen unbekannte) Eigenschaften und Zusammenhänge einer Population anhand einer Stichprobe geschätzt werden sollen. Kommen neue Daten hinzu, ist es nicht mehr notwendig, die gesamte Analyse durchzuführen. Das Einsetzen in den ermittelten Zusammenhang genügt. Im vorliegenden Fall soll die Zuordnung (ja/nein) zu einer Klasse (im Fließtext auch allgemein als Output oder als zu erklärende Variable Y bezeichnet) anhand der beobachtbaren Merkmale (im Fließtext allgemein als Input oder erklärende Variable X bezeichnet) geschätzt werden (binäre Klassifikation). Im Unterschied zu parametrischen Verfahren unterstellt die Methode der Support Vector Machines nicht bereits von vornherein ein bestimmtes Modell von Verteilungen, welchem die Population vermeintlich folgt. Die in der Population vorhandenen Zusammenhänge zwischen Input- und Outputvariablen werden durch das nichtparametrische Verfahren erst entdeckt. Somit lösen nichtparametrische Verfahren das Grundproblem der Statistik, die zugrunde liegende Verteilung anhand der gegebenen Daten zu schätzen, wesentlich allgemeiner als parametrische Verfahren.

3.1 Das Lernen der SVM

Das Lernmaterial wird im Folgenden als Trainingsdatensatz bezeichnet. Anhand dieses Trainingsdatensatzes lernt die SVM, das heißt sie betrachtet die Zusammenhänge zwischen Input- und Outputvariablen und adaptiert diese. Würde man an dieser Stelle das Verfahren beenden, so müsste man hoffen, dass die SVM gut gelernt hat; dass also ein guter Prädiktor entstanden ist, der neue Inputdaten richtig verarbeiten und den hoffentlich korrekten zugehörigen Output liefern wird. Eine Überprüfung des Ergebnisses ist nicht möglich, da die wahren Werte der neuen Inputdaten nicht bekannt sind (wären sie es, bräuchte man keine Statistik jenseits des Deskriptiven zu betreiben). Um die Güte der SVM zu messen, wird daher neben dem Trainings- ein Testdatensatz benötigt, von welchem ebenfalls Input- und Outputwerte verlässlich bekannt sind. Anhand dieses Testdatensatzes kann nach dem Lernvorgang geprüft werden, wie gut die Vorhersagen der SVM sind.

Aus Anwendersicht hätte man gerne, dass alle Vorhersagen zutreffen, die Missklassifikationsrate also 0 ist. Ob dies aus statistischer Sicht erreichbar ist, ist nicht klar, da nur eine Stichprobe und nicht die gesamte Population zur Verfügung steht. Support Vector Machines versuchen daher, das „Bestmögliche“ aus den Daten herauszuholen und minimieren das empirische Risiko einer Missklassifikation zuzüglich eines Strafterms, das heißt die SVM $f_{L,\lambda,\gamma}$ minimiert

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} L\left(x_i, y_i, f(x_i)\right)}_{\text{Missklassifikationsrisiko}} + \underbrace{\lambda \|f\|_H^2}_{\text{Strafterm}}$$

über alle zulässigen Funktionen f.[5] Welche dies sind, legt der Anwender über einen Hyperparameter $\gamma > 0$ fest (zum Einfluss von γ siehe die funktionale Form der SVM weiter unten). Die x_i sind dabei die Vektoren der Inputwerte (also die Werte der erklärenden Variablen) der Unternehmen 1 bis n aus dem Trainingsdatensatz. Die y_i repräsentieren die Klassenzugehörigkeit dieser Unternehmen und sind entweder 1, falls das Unternehmen dem Dritten Sektor angehört, oder – 1, falls nicht. L stellt allgemein die Verlustfunktion (Loss function) dar, die betrachtet werden soll. Im Kontext des Projektes sollen Missklassifikationen bestraft werden, hierfür bietet sich konkret die sogenannte Hinge-Verlustfunktion (Steinwart/Christmann, 2008, hier: Seite 8 f. und Seite 310 ff.) an:

$$L\left(x_i, y_i, f(x_i)\right) = \max\left\{0, 1 - y_i f(x_i)\right\}.$$

Ziel ist es – wie weiter oben bereits beschrieben –, durch die SVM den mittleren Verlust, das heißt die mittlere Missklassifikation, zu minimieren, allerdings nicht

5 Die Herleitung der SVM wird hier aus statistisch-analytischer Sicht beschrieben. Eine weitere, geometrische Erläuterung der Funktionsweise einer SVM wird in Feuerhake/Dumpert (2016) erscheinen.

um jeden Preis. Der Anwender könnte nämlich so viele Prädiktoren zulassen, dass man das Missklassifikationsrisiko im Trainingsdatensatz auf 0 drücken könnte. Dies wäre ein typischer Fall von Overfitting mit der Folge, dass – bezogen auf die Trainingsdaten – die Klassifikation perfekt, bezogen auf neue Daten, für die eine überwiegend zutreffende Klassifikation angestrebt wird, die Qualität hingegen nicht gesichert wäre. Um das zu vermeiden, bestraft man stark interpolierende Funktionen, die sich zu sehr den Trainingsdaten anpassen, durch den Strafterm $\|f\|_H^2$.[6] Wie stark dieser die Lösung des Minimierungsproblems beeinflussen soll, regelt ein zweiter Hyperparameter $\lambda > 0$, den der Anwender ebenfalls wählen kann.

Das Lernen der SVM selbst besteht in der Lösung des Optimierungsproblems

$$\text{minimiere} \quad \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1-y_i f(x_i)\} + \lambda\|f\|_H^2$$

über alle zulässigen Funktionen f. In Abhängigkeit von λ und γ kann man einen funktionalen Zusammenhang für die optimale Lösung, also für die SVM $f_{L,\lambda,\gamma}$ angeben. Im Rahmen der Testrechnungen ergab sich der folgende Zusammenhang:

$$f_{L,\lambda,\gamma}(x_{neu}) = \sum_{i=1}^{n} \alpha_i \, e^{-\gamma\|x_{neu}-x_i\|^2}$$

Die Koeffizienten α_i entstehen technisch durch die Optimierung. Inhaltlich geben sie an, welchen Einfluss ein Unternehmen aus dem Trainingsdatensatz auf die SVM hat. Die Klassifikation eines neuen Unternehmens erfolgt nun durch Einsetzen der Werte seiner erklärenden Variablen x_{neu} in $f_{L,\lambda,\gamma}$ und Prüfung, welches Vorzeichen das Ergebnis hat. Negative Ergebnisse werden der einen, die übrigen Ergebnisse werden der anderen Klasse zugeordnet.

3.2 Das Testen der SVM

Das Testen der SVM stellt, anders als das Lernen der SVM, keinen zwingend notwendigen Teil der Arbeit mit SVM dar. Um allerdings eine Schätzung dafür zu erhalten, wie gut die auf oben beschriebene Art und Weise zustande gekommene Support Vector Machine arbeitet, das heißt mit welchem Missklassifikationsrisiko der Anwender bei künftiger Verwendung der SVM leben muss, werden die Inputdatenpunkte des vom Trainingsdatensatz unabhängigen Testdatensatzes in die SVM eingesetzt und das Vorzeichen des Ergebnisses betrachtet, das die Klassenzuordnung liefert. Die resultierende Klassifikation wird mit der wahren Klasse verglichen. Der Anteil der falsch klassifizierten Testdatenpunkte ist dann eine Schätzung für die Missklassifikationsrate der SVM. Der Vorteil, diesen Wert zu kennen, wird durch den Nachteil erkauft, etwas weniger Daten für den Trainingsdatensatz zur Verfügung zu haben. Bei dem für die Testrechnungen vorliegenden Datensatz spielt dieser Aspekt aufgrund des großen Umfangs der zur Verfügung stehenden Daten allerdings keine Rolle.

3.3 Eigenschaften von SVM

Bennett/Campbell diskutieren die Vorzüge von Support Vector Machines in ihrem Übersichtsartikel (Bennett/Campbell, 2000, hier: Seite 9 f.), insbesondere besitzen SVM statistisch wünschenswerte Eigenschaften. Sie sind beispielsweise konsistent in dem Sinne, dass das Missklassifikationsrisiko für größer werdende Trainingsdatensätze gegen das (für die gewählte Verlustfunktion) bestmögliche (also kleinstmögliche) Missklassifikationsrisiko konvergiert. Weiterhin sind Support Vector Machines robust in dem Sinne, dass kleine Veränderungen in den gegebenen Daten das Ergebnis der SVM, das heißt die erlernten Regeln, nur wenig verändern. Insbesondere kommt es nicht zu einem vollkommen anderen Verhalten der SVM, wenn man den vorliegenden Datensatz geringfügig ändert oder wenn eine neue Stichprobe aus der unbekannten Verteilung der Klassen gezogen wird. Auch können Existenz und Eindeutigkeit von Support Vector Machines unter schwachen Voraussetzungen, welche im Rahmen der Testrechnungen sämtlich erfüllt sind, gezeigt werden.

Die Frage, ob Support Vector Machines eine bessere Methode darstellen als andere parametrische (zum Beispiel logistische Regression) oder nichtparametrische Ansätze (zum Beispiel Classification Trees), ist im Allgemeinen nicht zu beantworten. Bennett/Campbell halten jedoch fest, dass SVM im Mittel anderen Methoden überlegen sind. Gegebenenfalls sind auch Kombi-

6 Hierbei handelt es sich technisch um eine Norm der Funktion f in einem unendlich-dimensionalen Hilbertraum (H) (Steinwart/Christmann, 2008, hier: Seite 121).

nationen aus mehreren Verfahren denkbar (Feuerhake/Dumpert, 2016).

4

Implementierung einer SVM für die Sektorzuordnung von Unternehmen

Eines der Ziele der Implementierung war es, die Handhabung des Programms für den Anwender möglichst einfach zu halten. Dies wird durch zwei Umsetzungsaspekte erreicht:

1. Einfacher Umgang mit den Daten: Der Anwender stellt die notwendigen Daten als csv-Datei zur Verfügung und erhält am Ende wieder csv-Dateien mit den Resultaten. Eine besondere Behandlung der Daten ist nicht erforderlich.
2. Einfacher Funktionsaufruf: Sowohl das Lernen als auch das Anwenden der SVM ist durch je einen Funktionsaufruf in R zu bewerkstelligen. Der Anwender bedarf keiner weiteren, vertieften Kenntnisse zur Implementierung.

Die Programmierung erfolgte in R. R ist eine Open Source Software zur Datenanalyse, die aus grundlegenden Funktionen (base) und ergänzenden Packages (hier zum Beispiel kernlab) besteht. Letztere werden vom Nutzer nach Bedarf installiert. R rechnet in der Regel lokal, also auf dem Rechner des Nutzers. In einer grafischen Oberfläche werden eventuelle Kontrollausgaben und, sofern vorhanden, Grafiken angezeigt.

5

Vorgehen bei den Testrechnungen und Ergebnisse

5.1 Auswahl der erklärenden Variablen

Im Rahmen der Testrechnungen sollte untersucht werden, ob Support Vector Machines geeignet sind, die bisher notwendigen, sehr personal- und zeitaufwendigen Einzelfallrecherchen bezüglich der Zugehörigkeit zum Dritten Sektor für durch den maschinellen Algorithmus nicht eindeutig zuzuordnende Unternehmen zu ersetzen. Geprüft werden sollte, ob die SVM in der Lage wäre, weitere Strukturen in den Daten zu entdecken, die über die im maschinellen Algorithmus bereits abgebildeten Regeln hinausgehen. Hierzu musste insbesondere ermittelt werden, welche erklärenden Variablen beim Lernen der SVM genutzt werden sollten. Kriterien für eine solche Auswahl sind:

1. Optimalität: Es ist diejenige Variablenkombination zu wählen, welche im Mittel (über verschiedene zufällige Aufteilungen des Datensatzes in Trainings- und Testdatensatz) das geringste Missklassifikationsrisiko liefert.
2. Stabilität: Es ist diejenige Variablenkombination zu wählen, welche im Mittel die geringste Streuung des Missklassifikationsrisikos aufweist.
3. Recheneffizienz beziehungsweise Variableneffizienz: Es ist diejenige Variablenkombination zu wählen, welche den geringsten Rechenaufwand verursacht beziehungsweise welche die geringste Anzahl an erklärenden Variablen benötigt.

Der zur Verfügung stehende Datensatz enthält einige Variablen (Merkmale), die für alle Einheiten besetzt sind. Variablen, für die das nicht zutrifft, wurden nicht untersucht, sofern die Datenlücke nicht kodierungstechnisch, sondern inhaltlich begründet war (zum Beispiel fehlende Erhebung oder Ähnliches). Die betrachteten Variablenkombinationen[7] enthält ↘ Tabelle 1.

Anmerkungen zu den betrachteten Variablenkombinationen:

(K1) Es ist naheliegend, alle zur Verfügung stehenden und fachstatistisch als sinnvoll eingeschätzten Variablen in die Untersuchung einzubeziehen. Jede verfügbare Information über die Daten wird damit in das Lernen der SVM eingebunden. Wesentlicher Nachteil dieser Kombination ist die mit zunehmender Dimension [Anzahl der betrachteten (Dummy-)Variablen] des Problems

7 Es ist hierbei zu beachten, dass alle nominal codierten Variablen, das heißt Wirtschaftszweig-Kategorie, Rechtsform-Kategorie, Beschaffenheit als öffentliches Unternehmen, Wirtschaftszweig, Rechtsform und Sitz des Unternehmens, zunächst in entsprechend der Anzahl ihrer auftretenden Ausprägungen viele Dummy-Variablen (0/1-Variablen) zerlegt werden.

Tabelle 1
Betrachtete Variablenkombinationen

	Kombination (K1)	Kombination (K2)	Kombination (K3)	Kombination (K4)	Kombination (K5)	Kombination (K6)
Steuerbarer Umsatz	X				X	
Anzahl der sozialversicherungspflichtig Beschäftigten	X				X	
Anzahl der Vollzeitbeschäftigten	X				X	
Anzahl der Teilzeitbeschäftigten	X				X	
Anzahl der geringfügig entlohnt Beschäftigten	X				X	
Wirtschaftszweig	X		X	X	X	X
Wirtschaftszweig-Kategorie[1]	X	X		X		
Beschaffenheit als öffentliches Unternehmen	X	X	X	X	X	X
Rechtsform	X		X	X	X	X
Rechtsform-Kategorie[1]	X	X		X		X
Sitz des Unternehmens	X	X	X	X	X	X

1 Siehe Erläuterung in Abschnitt 2.6.

überproportional ansteigende benötigte Lern- und Auswertungszeit. Auch die Variableneffizienz ist hier nicht gegeben.

(K2) Aus inhaltlichen Überlegungen bei der Entwicklung des bisher schon eingesetzten maschinellen Algorithmus ist bekannt, dass die Variable „Beschaffenheit als öffentliches Unternehmen“ sowie die präklassifizierenden Variablen „Rechtsform-Kategorie“ und „Wirtschaftszweig-Kategorie“ die vermeintlich wichtigsten Variablen darstellen, um die Zugehörigkeit zum Dritten Sektor zu überprüfen (Rosenski, 2012). Sowohl Variablen- als auch Recheneffizienz sind hier gegeben.

(K3) Diese Kombination ähnelt (K2), enthält aber nicht die zusammenfassenden Variablen „Rechtsform-Kategorie“ und „Wirtschaftszweig-Kategorie“, sondern deren Grundlagen (also alle Ausprägungen der Rechtsformen und der Wirtschaftszweige). Die Idee, diese Kombination zu untersuchen, bestand darin, dass die Präklassifizierung nicht alle möglichen Fälle berücksichtigen kann, durch diese also Informationen verloren gehen. Der SVM sollte ermöglicht werden, alle im Wirtschaftszweig und in der Rechtsform enthaltenen Informationen zu nutzen.

(K4) Hier liegt eine Synthese aus (K2) und (K3) vor. Zwar korrigiert (K3) im oben beschriebenen Sinne (K2); gleichzeitig fehlt (K3) nun das fachstatistische Wissen, das in die Präklassifizierung von Wirtschaftszweig und Rechtsform eingeflossen ist. Variablenkombination (K4) nutzt beide Informationen.

(K5) Diese Kombination ist an (K1) angelehnt, verzichtet aber auf die präklassifizierenden Variablen „Rechtsform-Kategorie“ und „Wirtschaftszweig-Kategorie“. Zur Begründung siehe (K3).

(K6) Wie (K4), allerdings wird auf die Variable „Wirtschaftszweig-Kategorie“ verzichtet, die Rechtsform also weiterhin präklassifiziert. Diese Variablenkombination wurde erst im Laufe der Untersuchungen interessant, als festgestellt wurde, dass viele falsch klassifizierte Unternehmen eine übereinstimmende Wirtschaftszweig-Kategorie (und zwar „Wirtschaftszweig typisch für den Dritten Sektor“) aufwiesen. Um sicherzustellen, dass die Variable „Wirtschaftszweig-Kategorie“ keine schädliche Wirkung auf die Missklassifikationsrate hat, wurde diese Kombination zusätzlich betrachtet.

Hinzu kam bei allen sechs Kombinationen jeweils noch das Bundesland, in welchem der Sitz der jeweiligen Einheit liegt („Sitz des Unternehmens“).

Erste Untersuchungen zeigten folgende Ergebnisse:

1. Die Einbeziehung des Bundeslandes, in dem der Sitz des jeweiligen Unternehmens liegt, liefert für alle Variablenkombinationen geringere Missklassifikationsraten im Vergleich zu Durchläufen ohne Einbeziehung dieser Variable. Daher wurden alle Kombinatio-

Tabelle 2
Geschätzte Missklassifikationsraten

	Fälschlich nicht dem Dritten Sektor zugeordnet	Fälschlich dem Dritten Sektor zugeordnet	Summe	Saldo bezüglich der Zuordnung zum Dritten Sektor
	(1)	(2)	(1 + 2)	(2 – 1)
	%			
Anzahl Unternehmen	4,95	8,90	13,85	+ 3,95
Steuerbarer Umsatz	11,57	7,91	19,48	– 3,66
Sozialversicherungspflichtig Beschäftigte	6,10	8,87	14,97	+ 2,77

nen stets um das Merkmal „Sitz des Unternehmens“ ergänzt.

2. Variablenkombination (K2) liefert stets Missklassifikationsraten von etwa 30 % und liegt damit über den Raten aller anderen Kombinationen. Auf eine genauere Untersuchung dieser Kombination wurde daher verzichtet.

Die geringste mittlere Missklassifikationsrate weist Variablenkombination (K1) auf. Es wird allerdings deutlich, dass keine der übrigen Variablenkombinationen deutlich schlechter ist. Allein in Bezug auf das Entscheidungskriterium der Optimalität kann somit keine Auswahl der besten Variablenkombination getroffen werden. Betrachtet man die Stabilität, so liefert Variablenkombination (K4) das beste Ergebnis, allerdings ist auch hier festzuhalten, dass die anderen Kombinationen keine deutlich größeren Standardabweichungen aufweisen. Auch die Stabilität alleine ist also nicht ausreichend, um eine Entscheidung zu treffen. Betrachtet man den numerischen Aufwand, so wären die Variablenkombinationen (K3) und (K4) auszuwählen. Als Synthese dieser Überlegungen fiel die Entscheidung schließlich zugunsten von Variablenkombination (K4), da die Rechen- und Variableneffizienz am schwerwiegendsten für die Praxis erschienen, Optimalität und Stabilität hier keine Trennschärfe liefern, Variablenkombination (K4) aber die fachstatistische Präklassifizierung noch miterfassen kann. Alle weiteren Berechnungen wurden daher auf Basis von Variablenkombination (K4) durchgeführt.

Weitere mögliche erklärende Variablen, die während des Projektverlaufs in Verbindung mit Variablenkombination (K4) untersucht wurden, liefern keinen verbessernden (aber auch keinen das Ergebnis deutlich verschlechternden) Beitrag. Zusätzlich untersucht wurden die Merkmale: „Teilnahme am Intrahandel“ und „Sitz des Unternehmens in einem Ballungsgebiet“ sowie deren Kombinationen. Es zeigte sich darüber hinaus, dass das Merkmal „Sitz des Unternehmens ist in einem Ballungsgebiet“ das Merkmal „Sitz des Unternehmens“ nicht ohne Verschlechterung des Ergebnisses ersetzen kann.

5.2 Ergebnisse der Berechnungen

Insgesamt lagen 45 662 Datensätze vor, davon wurden 36 531 zum Trainieren genutzt. Die verbleibenden 9 131 Datensätze bildeten den Testdatensatz. Tabelle 2 stellt die für diese Testdaten ermittelten Resultate zusammengefasst dar. ↘ Tabelle 2

Tabelle 3 zeigt die geschätzten Sensitivitäten und Spezifitäten bezüglich des Testdatensatzes. ↘ Tabelle 3

Die erzielten Missklassifikationsraten reihen sich in eine große Zahl von Resultaten ein. Bennett/Campbell (2000) nennen exemplarisch für Klassifikationsstudien mit realen, nicht naturwissenschaftlichen Daten Raten zwischen 3 % und 36 %. Zu kleine Missklassifikationsraten ließen darauf schließen, dass sich (a) die Daten in

Tabelle 3
Geschätzte Sensitivitäten und Spezifitäten

	Anteil der durch die SVM dem Dritten Sektor zugeordneten Unternehmen an den tatsächlichen Dritt-Sektor-Unternehmen (Sensitivität)	Anteil der durch die SVM dem Dritten Sektor nicht zugeordneten Unternehmen an den tatsächlichen Nicht-Dritt-Sektor-Unternehmen (Spezifität)
	%	
Anzahl Unternehmen	90,47	81,47
Steuerbarer Umsatz	78,62	82,76
Sozialversicherungspflichtig Beschäftigte	87,65	82,47

offensichtlicher Weise trennen lassen (was den Einsatz fortgeschrittener statistischer Methoden erübrigt), (b) die Abhängigkeit zwischen Input- und Outputvariablen vollständig deterministisch ist oder (c) die Ermittlung der geschätzten Missklassifikationsrate nicht anhand von vom Trainingsdatensatz unabhängigen Daten (zum Beispiel aus dem Testdatensatz) vorgenommen wurde[8].

6

Fazit und Ausblick

Manuelle Prüfungen (durch geschultes Personal), die über Recherchen zusätzliche Informationen zu den betreffenden Einheiten liefern, stellen auf den ersten Blick die ideale Lösung dar, um in besonders komplexen Fällen die Zuordnung von Unternehmen zum Dritten Sektor vorzunehmen. Sie sind jedoch sehr zeit- und personalaufwendig und daher nicht in jedem Jahr im benötigten Umfang durchführbar.

Support Vector Machines sind eine praktikable und kostengünstige Alternative. Wann und in welchem Umfang sie für die Zuordnung von Unternehmen zum Dritten Sektor im Unternehmensregister eingesetzt werden sollen, ist noch offen und hängt auch von den Wünschen und Qualitätsanforderungen der Datennutzer ab. Denkbar ist zudem, SVM zu nutzen, um eine Vorselektion zu prüfender, gewichtiger Einheiten zu treffen.

Mit den Testrechnungen wurden die SVM bezüglich ihrer Anwendbarkeit für die Klassifizierung von Unternehmen hinsichtlich ihrer Zugehörigkeit zum Dritten Sektor erprobt. Dieser Test kann in Bezug auf die vorliegende Datenbasis als erfolgreich gewertet werden und bietet Raum für weitere Einsatzfelder. In der Handwerksstatistik ist eine weitere Erprobung weitgehend abgeschlossen. Hierüber wird voraussichtlich in der nächsten Ausgabe dieser Zeitschrift berichtet werden (Feuerhake/Dumpert, 2016). Darüber hinaus soll in einem weiteren gemeinsamen Projekt untersucht werden, ob die Schätzmodelle zum sogenannten bereinigten Verdienstunterschied von Männern und Frauen mittels SVM verbessert werden können.

8 Es erscheint auf den ersten Blick unklar, weshalb man die Güte der SVM nicht anhand des (sowieso zur Verfügung stehenden) Trainingsdatensatzes überprüft. Technisch steht dem nichts im Wege. In der statistischen Praxis sähe man sich aber mit der Frage konfrontiert, ob das dann kommunizierte Resultat nicht zu günstig für die SVM ausfiele, da der Prädiktor speziell für diesen Datensatz optimiert wurde (theoretisch sind so Missklassifikationsraten von 0 % erreichbar). Interessiert ist man aber daran, wie sich die SVM bei neuen, unbekannten, nicht zum Lernen herangezogenen Daten verhält. Wählt man dagegen einen vom Trainingsdatensatz unabhängigen Testdatensatz, ist die SVM über diesen Zweifel erhaben.

LITERATURVERZEICHNIS

Bennett, Kristin P./Campbell, Colin. *Support Vector Machines: Hype or Hallelujah?* In: SIGKDD Explorations Newsletter. Band 2, Heft 2 (2000), Seite 1 ff.

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

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

Feuerhake, Jörg/Dumpert, Florian. *Erkennung nicht relevanter Unternehmen in den Handwerksstatistiken – Einsatz von Support Vector Machines zur maschinellen Klassifikation*. Veröffentlichung vorgesehen in: WISTA Wirtschaft und Statistik, Ausgabe 2/2016.

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

Hothorn, Torsten. *CRAN task view: Machine learning & statistical learning*. 2014. [Zugriff am 15. Januar 2016]. Verfügbar unter: https://cran.r-project.org/

Nahm, Matthias/Stock, Gerhard. *Erstmalige Veröffentlichung von Strukturdaten aus dem Unternehmensregister*. In: Wirtschaft und Statistik. Ausgabe 7/2004, Seite 723 ff.

Rosenski, Natalie. *Die wirtschaftliche Bedeutung des Dritten Sektors*. In: Wirtschaft und Statistik. Ausgabe 3/2012, Seite 209 ff.

Simon, Herbert A. *WHY SHOULD MACHINES LEARN?* In: Michalski, Ryszard S./Carbonell, Jaime G./Mitchell, Tom M. (Herausgeber). Machine Learning – An Artificial Intelligence Approach. Palo Alto 1983, Seite 25 ff.

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

Statistisches Bundesamt/Centrum für soziale Investitionen und Innovationen (Herausgeber). *Zivilgesellschaft in Zahlen – Modul 1 – Endbericht*. [Zugriff am 15. Januar 2016]. Verfügbar unter: http://stifterverband.info

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

Vereinte Nationen. *Handbook on Non-Profit Institutions in the System of National Accounts*. New York 2003, Seite 16. [Zugriff am 15. Januar 2016]. Verfügbar unter: http://unstats.un.org

Waldmüller, Bernd/Weisbrod, Joachim. *Neuere Entwicklungen in den Unternehmensstatistiken*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2015, Seite 33 ff.

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 Februar 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: 1010200-16001-1
ISSN 0043-6143
ISBN 978-3-8246-1043-3

Download (PDF)
Artikelnummer: 1010200-16001-4, 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