

Social Lending: Credit Scoring für Normalbürger (mit Open Source-Software)

Version 1.0

Dr. Rainer Typke, Theano GmbH (rainer.typke@theano.de)

Copyright 2009 by Rainer Typke. CC-BY-NC-ND

You are free to share this work under the following conditions: attribution, noncommercial, no derivative works. For details, see <http://creativecommons.org/licenses/by-nc-nd/3.0>

Banken verwenden ausgefeilte Methoden, um potentielle Kreditnehmer zu bewerten, bevor sie über eine Kreditvergabe entscheiden. Seit es Social Lending-Plattformen wie beispielsweise smava.de gibt, stellt sich das Problem der Bewertung potentieller Kreditnehmer plötzlich auch Normalbürgern, die bisher höchstens Gegenstand solcher Bewertungen waren. In diesem Vortrag wird gezeigt, wie man durch das Kombinieren diverser Open Source-Produkte (Weka als Support Vector Machine, CRM114 zur Bewertung von Kreditbeschreibungen, Image_Graph zur Visualisierung, MySQL, Linux, Apache uvm.) auch für Normalbürger gute Entscheidungen über das Gewähren oder Verweigern von Krediten ermöglichen kann. Alle diese Produkte sind in die webbasierte Kreditbeurteilungs-Software "smava.beobach.de" eingeflossen, die der Smava-Gemeinde (Investoren und Kreditnehmern) zur Verfügung steht und bei der Entscheidungsfindung benutzt wird.

Social Lending in Deutschland

In Deutschland gibt es mehrere Internetplattformen, auf der Privatleute in Ratenkredite investieren können, die wiederum von anderen Privatleuten aufgenommen werden. Die wahrscheinlich bedeutendste Plattform in Deutschland ist smava.de.

Funktionsweise von smava.de

Kreditprojekte, Gebotsphase

Menschen mit Geldbedarf erstellen auf Smavas Website „Kreditprojekte“, für die in einer bis zu 14 Tage dauernden Phase durch Kreditgeber Gebote abgegeben werden können. Der Zinssatz und Gesamtbetrag werden dabei vom Kreditnehmer bestimmt; für den Fall, dass nicht genug Gebote durch Kreditgeber eingegeben werden, kann der Kreditnehmer den Zinssatz während der Gebotsphase erhöhen.

Wenn entweder die gewünschte Summe erreicht wurde oder aber die 14 Tage

vorbei sind und mindestens ein Viertel der gewünschten Summe erreicht ist, erhält der Kreditnehmer die Option, die Gebote anzunehmen und damit wirklich zum Kreditnehmer zu werden.

Mögliche Szenarien während der Kreditlaufzeit

Während der Laufzeit kann es verschiedene Szenarien geben: planmäßige Abwicklung, Ausfall, Verzug und vorzeitige Rückzahlung. Je nachdem, in welchem dieser Zustände sich ein Kredit befindet, nehmen seine Investoren auf verschiedene Weise an einem Pool teil, mit dem Investoren das Ausfallrisiko teilen. Die Grundidee dieses Pools ist, dass jeden Monat die glücklichen Investoren, deren Kredit nicht ausgefallen ist, den weniger glücklichen mit Ausgleichszahlungen helfen. Dabei wird nie Geld im Pool gespeichert, sondern jeden Monat sind die Einzahlungen in den Pool genauso hoch wie die Auszahlungen.

Planmäßige Abwicklung: Anleger tragen die Ausfälle anderer Anleger mit („Pools“)

Solange der Kreditnehmer seine Raten pünktlich und vollständig zahlt, erhalten die Anleger, die in seinen Kredit investiert haben, monatlich ihre planmäßigen Tilgungs- und Zinszahlungen. Von den Tilgungszahlungen erhalten die Investoren allerdings nur den Prozentsatz, der der Zahlungsquote entspricht, die im jeweiligen Monat bei allen Smava-Krediten im selben „Pool“ zu beobachten ist. In jedem Pool sind Kreditprojekte von Kreditnehmern mit derselben Schufa-Bonität und mit derselben Laufzeit zusammengefasst. Beträgt die Zahlungsquote für 36-Monats-Projekte der Bonität F beispielsweise in einem bestimmten Monat 90 %, so erhalten die Anleger, deren 36er-F-Kreditprojekte im Plan sind, 90 % der ihnen zustehenden Tilgungsanteile (und 100 % der ihnen zustehenden Zinsen). Die anderen 10 % der Tilgungsanteile wandern in diesem Monat stattdessen an die Anleger, die im selben Pool in ausgefallene Kreditprojekte investiert haben.

Ausfall: Anleger erhalten Großteil der Tilgung von Inkassounternehmen und anderen Anlegern

Ausgefallene Kredite werden an ein Inkassounternehmen verkauft, das für den Rest der Laufzeit nur einen Bruchteil der Tilgungen und keine Zinsen bezahlt. Ausgefallene Kredite verringern also in ihrem Pool die Zahlungsquote. Kreditgeber, die in ein ausgefallenes Projekt investiert haben, erhalten neben den Zahlungen des Inkassounternehmens Ausgleichszahlungen von Investoren, die im selben Pool in nicht ausgefallene Kredite investiert haben. Dadurch erhalten alle Investoren, die im selben Pool investiert haben, genau den der Zahlungsquote entsprechenden Tilgungsanteil, unabhängig davon, ob ihr eigenes Projekt ausgefallen ist oder nicht. Die Investoren, deren Projekt nicht ausgefallen ist, erhalten darüber hinaus die ihnen zustehende Zinszahlung.

Verzug: meist nur temporäre Auswirkungen

Kredite können bis zu 6 Wochen lang in Verzug sein, ohne dass die Restforderung an das Inkassounternehmen verkauft wird. Sollten sie noch rechtzeitig komplett nachbezahlt werden, hat ein Verzug außer einer Schwankung in den monatlichen Einkünften der Investoren keine Auswirkungen.

Merkmale eines Kreditprojektes

Ein Kreditgeber erhält folgende Informationen, die er seiner Investitionsentscheidung zugrundelegen kann:

- eine Beschreibung der beabsichtigten Verwendung des Geldes durch den Kreditnehmer (Freitext, wird in keiner Weise auf Plausibilität geprüft),
- der Wunschbetrag (Euro-Betrag, fix),
- der gebotene Zinssatz (Zahl zwischen 0 und 18, zunächst variabel, kann ab dem Vorliegen eines Investor-Gebots nur noch erhöht oder gleich gelassen werden),
- die Laufzeit (36 oder 60 Monate); außerdem werden
- die Schufa-Bonität des Kreditnehmers (langfristig variabel, aber meist während der Gebotsphase fix; durch Smava geprüft),
- „Kapitaldienstfähigkeit“ („KDF“), d. h. der Prozentsatz des monatlich verfügbaren Einkommens, formuliert als Zugehörigkeit zu einer von vier Klassen (Plausibilitätsprüfung durch Smava, aber grundsätzlich nur auf Angaben des Kreditnehmers beruhend; langfristig variabel, aber meist während der Gebotsphase fix),
- Alter des Kreditnehmers (durch Smava/PostIdent geprüft),
- Geschlecht (durch Smava/PostIdent geprüft, fast immer während der Gebotsphase fix),
- berufliche Situation (Arbeiter oder Angestellter/Beamter/Freiberufler/Gewerbetreibender/Rentner/sonstiges; durch Smava geprüft)
- und Bundesland des Kreditnehmers (durch Smava/PostIdent geprüft) veröffentlicht.

Rationale Investitions-Entscheidungen

Aus der oben skizzierten Funktionsweise von Smava-Krediten ergibt sich für rationale Investoren:

- Es ist trotz des Pool-Mechanismus nicht wünschenswert, in früh ausfallende Projekte zu investieren. Ab einem Ausfall wird die laufende Rendite eines Projektes zwangsläufig negativ, weil die Zahlungsquote kleiner ist als 100 % und keine Zinsen mehr bezahlt werden. Allerdings kann die Rendite des ausgefallenen Projektes über die Gesamtlaufzeit betrachtet trotzdem positiv sein, wenn es nämlich so spät ausfällt, dass die bereits erfolgten Zinszahlungen den Forderungsausfall übersteigen.
- Es ist wünschenswert, in Projekte mit hohen Zinssätzen zu investieren. Schließlich gehen die Zinseinnahmen, solange ein Projekt nicht ausfällt, ungeschmälert an die individuellen Investoren, ohne in den Poolmechanismus einbezogen zu werden.
- Aus den beiden vorhergehenden Punkten ergibt sich manchmal ein „Moral Hazard“: falls man ein Kreditprojekt so einschätzt, dass es mit hoher Wahrscheinlichkeit irgendwann ausfällt, aber dank seines hohen Zinssatzes trotzdem schnell profitabel sein wird, kann das Vorhandensein des Pool-Mechanismus Anlegerentscheidungen so beeinflussen, dass sie zwar für individuelle Anleger gut, aber für die Gesamtheit aller Anleger trotzdem schlecht sind. Zinseinnahmen können privatisiert und spätere Forderungsverluste auf andere Anleger mit umgelegt werden.
- Zur Beurteilung der Attraktivität eines gebotenen Zinssatzes muss außer dem individuellen Projekt-Ausfallrisiko auch der Poolzustand, d. h. der Anteil, den bereits ausgefallene Kredite am Pool ausmachen sowie deren Restlaufzeit, berücksichtigt werden. Schließlich teilen Investoren nicht nur Risiken, sondern auch bereits bekannte und eingetretene Forderungsausfälle miteinander.

Bei den Investitionsentscheidungen können Investoren auf die von beobach.de bereitgestellten Werkzeuge zurückgreifen, die im folgenden Abschnitt dargestellt werden.

Werkzeuge zur Entscheidungs-Unterstützung: beobach.de

Auf beobach.de stehen einige Werkzeuge bereit, für die Open Source-Komponenten verwendet werden. Sie können dabei helfen, Ausfall-Kandidaten zu identifizieren und das Pool-Konzept bei der Entscheidungsfindung angemessen zu berücksichtigen. Zu den verwendeten Open Source-Komponenten gehören: Debian, MySQL, Typo3, CRM114 (<http://crm114.sourceforge.net>, Spamfilter von William S. Yerazunis), Image_Graph (http://pear.veggerby.dk/wiki/image_graph, Visualisierung unter PHP, Autoren: Neufeind und Veggerby), Weka (<http://www.cs.waikato.ac.nz/ml/weka/>, Support Vector Machine). Zur Illustration dessen, was dank Image_Graph möglich ist, siehe beispielsweise die Darstellung des Poolzustandes auf <http://beobach.de/pools.html> oder auch die Darstellung eines einzelnen Kreditprojekts, beispielsweise <http://www.beobach.de/k/80251839-schaffung-eines-neuen-arbeitsplatzes>. Der Rest dieses Artikels ist dem Vorhersagen von

Ausfällen mittels einer Support Vector Machine (SVM) gewidmet.

Ausfallkandidaten mittels SVM identifizieren: Credit Scoring durch Machine Learning

Wer eine Weile als Investor auf smava.de aktiv ist, wird wahrscheinlich früher oder später die Erfahrung machen, dass es einem normalen Menschen fast unmöglich ist, anhand der oben aufgelisteten Merkmale eines Kreditprojektes zuverlässig vorauszusagen, ob und wann der Kredit ausfallen wird. Das ist aber die zentrale Voraussage, die jeder rationalen Entscheidung über Investitionen zugrundeliegen muss. Glücklicherweise ist es mit Machine Learning-Methoden möglich, aus den Eigenschaften aller bisherigen Smava-Kreditprojekte eine Formel für die Bewertung neuer Projekte abzuleiten, die deutlich bessere Ergebnisse liefert als ein triviales, uninformiertes Orakel.

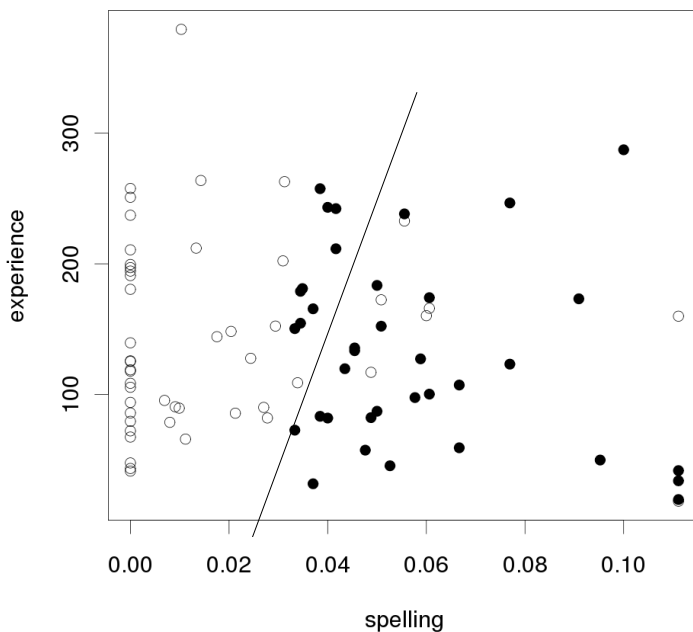
Was ist eine SVM?

„SVM“ steht für „Support Vector Machine“. Eine SVM wird zur Klassifizierung von Daten benutzt. Sie wird zunächst mit Trainingsdaten versorgt, von denen bekannt ist, zu welchen Klassen sie gehören. In unserem Fall sind die Trainingsdaten Merkmale von Krediten, die entweder zur Klasse der Ausfälle oder zur Klasse der Nicht-Ausfälle gehören. Die Gesamtheit aller Merkmale eines Kredits kann als ein Vektor aufgefasst werden (letztlich schlicht eine Liste von Zahlen). Jeder Vektor entspricht einem Punkt in einem mehrdimensionalen Raum, wobei jedes Merkmal eine Dimension ist. Irgendwo in diesem Raum verläuft eine Grenze, die die beiden Klassen optimal trennt. Es ist möglich, dass die Klassen nicht perfekt getrennt werden können, aber man kann immerhin eine Grenze finden, die sie so trennt, dass möglichst wenige Datenpunkte auf der falschen Seite liegen. Zusätzlich ist auch gewünscht, dass der Abstand der nächsten Datenpunkte zur Trennungsgrenze so groß wie möglich ist. Die Vektoren, die am dichtesten an dieser Grenze liegen, werden „Supportvektoren“ genannt, weil sie sozusagen die Grenze „unterstützen“.

Beim Trainieren der SVM konstruiert die SVM die beste Trennungsgrenze, die sie aufgrund der Trainingsdaten finden kann. Man hofft dann, dass diese Grenze auch später anfallende Daten gut klassifiziert, dass also spätere Ausfälle auf der Ausfall-Seite zu liegen kommen und Nicht-Ausfälle auf der anderen Seite der Grenze. Das sollte tendenziell so sein, wenn die als Punkt-Koordinaten verwendeten Merkmale wirklich etwas mit der Ausfallwahrscheinlichkeit zu tun haben.

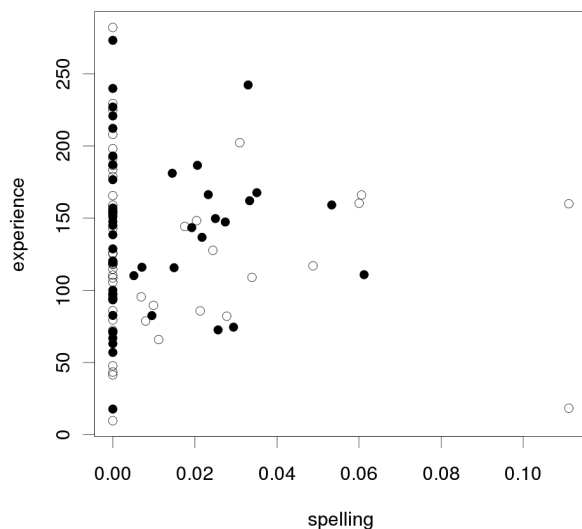
Es wurden verschiedene Varianten einer SVM ausprobiert, eine lineare sowie drei nichtlineare („PolyKernel“). Eine lineare SVM darf die Grenze nicht beliebig durch den Raum ziehen, sondern die Grenze muss eine Hyperebene sein, die eine Dimension weniger hat als die Datenpunkte. Wenn also beispielsweise alle Kredite zwei Merkmale hätten (wobei die Klassenzugehörigkeit nicht mitgezählt wird), wäre die Grenze eine eindimensionale Hyperebene, mit anderen Worten, eine Gerade.

Es ist durchaus denkbar, dass die Daten nicht so im Raum verteilt sind, dass sie sich durch eine solche Hyperebene sauber trennen lassen. In einem zweidimensionalen Raum wäre z. B. denkbar, dass beide Klassen eine bananenförmige Verteilung im Raum haben, so dass die Bananen sich zwar nicht berühren, aber doch so ineinanderliegen, dass sie von keiner Geraden getrennt werden können. Für solche Fälle gibt es den „PolyKernel“, eine SVM-Variante, die kompliziertere Grenzen ziehen darf. Mit einer gekrümmten Grenze kann man auch ineinanderliegende, sich nicht berührende Bananen sauber trennen. Je höher der Grad des Polynoms, desto komplizierter können die Grenzen aussehen.



Ein hypothetisches Beispiel mit den zwei Dimensionen Investorerfahrung (in Tagen, gewichtet nach investierter Summe, vertikale Achse) und Rechtschreibfähigkeiten (Anteil nicht im Wörterbuch stehender Wörter an der Beschreibung, horizontale Achse; hier wurden die 50 ältesten Ausfälle sowie 50 Nicht-Ausfälle mit besonders vielen Rechtschreibfehlern ausgewählt und eingezeichnet). Ausfälle sind als gefüllte Kreise, Nicht-Ausfälle als leere Kreise dargestellt. Eine SVM könnte hier als Grenze eine schräge Gerade einziehen, die die

Punkte aber nicht perfekt trennen könnte. Für neue Projekte, die rechts unter der Geraden zu liegen kommen, würde dann ein Ausfall vorhergesagt, während die SVM für Projekte auf der anderen Seite der Grenze optimistisch wäre.



So sieht dieses Schaubild mit den realen Rechtschreibfehlern aus – man sieht, dass diese beiden Merkmale höchstwahrscheinlich nicht für das Training einer SVM geeignet sind, weil kein Zusammenhang zwischen ihnen und der Ausfallwahrscheinlichkeit erkennbar ist.

Wie kann man die Qualität eines Klassifizierers beurteilen?

Confusion Matrix, durchschnittlicher Ertrag pro Klasse

Eine „Confusion Matrix“ sagt recht viel aus über die Qualität eines Klassifizierers. Sie besteht aus vier Angaben: True positives, false negatives, false positives und true negatives, mit anderen Worten: korrekt vorhergesagte Ausfälle, inkorrekt vorhergesagte Nicht-Ausfälle, inkorrekt vorhergesagte Ausfälle und korrekt vorhergesagte Nicht-Ausfälle. Anhand dieser Angaben kann man beurteilen, ob der Klassifizierer besser ist als ein Zufalls-Orakel; außerdem kann man sehen, wie oft überhaupt ein Ausfall vorhergesagt wird, wie sehr also die Auswahl an Projekten eingeschränkt würde bzw. wieviele der tatsächlichen Ausfälle überhaupt erkannt werden können. Ein Klassifizierer, der beispielsweise nur einen einzigen Ausfall vorhersagt, diesen aber korrekt, hätte zwar einen perfekten Recall in der Klasse der vorhergesagten Ausfälle, wäre aber in der Praxis nicht besonders nützlich, weil er fast keine verwertbaren Empfehlungen liefert.

Zusätzlich zur Confusion Matrix berechnen wir den durchschnittlichen Ertrag pro 250 investierten EUR in den Klassen „Ausfall vorhergesagt“, „kein Ausfall vorhergesagt“, „Ausfall eingetreten“ und „kein Ausfall eingetreten“. Letztlich kommt es Investoren ja darauf an, was hinten herauskommt.

Ausfall vorhergesagt	Kein Ausfall vorhergesagt		
12 Projekte	43,75 Projekte	Ausfall eingetreten	-14,74 EUR Ertrag pro 250 inv. EUR
18,5 Projekte	209 Projekte	Kein Ausfall eingetreten	29,38 EUR Ertrag pro 250 inv. EUR
10,39 EUR Ertrag pro 250 inv. EUR	21,54 EUR Ertrag pro 250 inv. EUR		

Confusion Matrix und durchschnittlicher Ertrag pro Klasse für lineare SVM, Track-Record-Grenze 0,8 (siehe unten), Durchschnitt für in 4 Zeiträumen von je 6 Monaten Länge hinzugekommene Projekte. Man sieht: wenn man die Zukunft perfekt vorhersehen kann und daher in alle Nicht-Ausfälle, aber in keine Ausfälle investiert, kann man im Schnitt 29,38 EUR Ertrag pro 250 investierten EUR erwarten. Richtet man sich nach dieser SVM, sind es immerhin 21,54 EUR – und man vermeidet die wesentlich niedrigeren Erträge von durchschnittlich nur 10,39 EUR in den Projekten, deren Ausfall vorhergesagt wird. Insgesamt wird vor ca. 12 Prozent der Projekte gewarnt. Der durchschnittliche Ertrag über alle Ausfälle und Nicht-Ausfälle zusammengenommen beträgt für diesen Zeitraum 20,69 EUR.

Ein brauchbarer Klassifizierer erreicht einen deutlich besseren Durchschnittsertrag in der Klasse „kein Ausfall vorhergesagt“ als in der Klasse „Ausfall vorhergesagt“; außerdem sollte er vor einem vernünftigen Prozentsatz

der Projekte warnen, so dass weder die Auswahl allzu sehr eingeschränkt wird noch allzu wenige der tatsächlichen Ausfälle vermieden werden.

Im folgenden wird beschrieben, wie man das erreichen kann: durch die Auswahl und Konstruktion geeigneter Merkmale, das Trainieren einer Support Vector Machine (SVM) anhand der Merkmale historischer Projekte, und die Anwendung der SVM auf zukünftige Projekte. Die Effektivität dieser Methode kann dadurch beurteilt werden, wie gut die Vorhersagen für Daten sind, die zeitlich später entstanden sind als die fürs Training verwendeten Daten. Die Qualität der Vorhersagen kann anhand der durchschnittlichen Kapitalwerte in den beiden Klassen „Ausfall vorhergesagt“ und „kein Ausfall vorhergesagt“ beurteilt werden.

Als SVM wurde das Open Source-Produkt „Weka“ verwendet (<http://www.cs.waikato.ac.nz/ml/weka/>), aber andere Implementierungen wären auch verwendbar (http://www.support-vector-machines.org/SVM_soft.html).

Die vorhergesagte Eigenschaft: „Ausfall in den ersten 10 Monaten“

Da es Smava noch nicht lange gibt, gibt es noch keine Daten von Kreditprojekten, die bereits ihre gesamte Laufzeit hinter sich haben. Man kann also eine SVM kaum darauf trainieren, vorherzusagen, ob ein Projekt irgendwann innerhalb seiner Laufzeit ausfällt. Es gibt aber schon hunderte von Projekten, die älter sind als 10 Monate. Außerdem ist wegen der Poolkonstruktion auch die Vorhersage eines frühen Ausfalls besonders interessant – späte Ausfälle belasten schließlich den individuellen Investor kaum, führen oft sogar nicht einmal zu individuellen Verlusten. Daher wurde die SVM darauf trainiert, zwischen Projekten zu unterscheiden, die innerhalb der ersten 10 Monate ausfallen (im folgenden „Ausfälle“ genannt), und solchen, die das nicht tun („Nicht-Ausfälle“).

Auswahl und Konstruktion geeigneter Merkmale

Die SVM wurde nicht anhand aller verfügbaren Kreditmerkmale trainiert. Im Rahmen einer Vorauswahl wurden solche Merkmale aussortiert, bei denen entweder keine Korrelation zur Ausfallwahrscheinlichkeit feststellbar war oder diese abwegig und zufallsbedingt erscheint. Siehe hierzu beispielsweise <http://www.beobach.de/krediteigenschaften.html> – wenn man Ausfälle gruppiert nach Geschlecht auswählt, sieht man, dass sowohl bei den Ausfällen als auch bei den Nicht-Ausfällen etwa 70 % der Kreditnehmer männlich und 30 % weiblich sind. Das Geschlecht scheint also für die Beurteilung der Ausfallwahrscheinlichkeit ungeeignet zu sein.

Andererseits wurden einige Merkmale neu konstruiert:

- Bietgeschwindigkeit: es ist nicht völlig ausgeschlossen, dass die Geschwindigkeit (gemessen in Euro pro Sekunde), mit der für ein Projekt geboten wird, etwas über seine Ausfallwahrscheinlichkeit aussagt. Also

wurde die Bietgeschwindigkeit für alle Projekte berechnet.

- Track Record der Bieter: wenn jemand, dem es in der Vergangenheit überdurchschnittlich gut gelungen ist, auf nicht ausfallende Projekte zu bieten, ein neues Gebot abgibt, könnte das ein gutes Zeichen sein.
- Erfahrung der Bieter (in Tagen, gewichtet mit den investierten Summen)
- Textähnlichkeit zur Beschreibung von ausgefallenen bzw. nicht ausgefallenen Projekten: Spammer haben viele Programmierer dazu angeregt, extrem leistungsfähige Software zur Klassifizierung von Texten zu entwickeln. Ein besonders leistungsfähiges derartiges Programm, das unter der GPL lizenziert ist, ist CRM114 (crm114.sourceforge.net). Statt es mit Spam und Ham zu trainieren, kann man es natürlich auch mit Beschreibungen von Ausfällen und gut laufenden Projekten trainieren und dann eine Kennzahl erhalten, die aussagt, wie sehr eine Projektbeschreibung nach Ausfall oder Nicht-Ausfall aussieht. Ob es wirklich einen Zusammenhang zwischen den Beschreibungstexten und der Ausfallwahrscheinlichkeit gibt, kann man dann dadurch beurteilen, ob dieses Merkmal die SVM-Vorhersagen verbessert.
- Rechtschreibfehler: mit verfügbaren Open Source-Spellcheckern wie beispielsweise ispell kann man abschätzen, wie gut ein Kreditnehmer die deutsche Orthografie beherrscht – oder zumindest wieviel Mühe er darauf zu verwenden bereit ist, einen orthografisch korrekten Text abzuliefern.

Schließlich wurden verschiedene Kombinationen von Merkmalen dazu verwendet, eine lineare, quadratische oder kubische SVM mit dem „Binary SMO“-Kernel zu trainieren. Dabei wurde von einer kleinen Anzahl von Merkmalen ausgegangen, bei denen die Korrelation mit der Ausfallwahrscheinlichkeit klar ist (Schufa-Bonität und berufliche Situation), und diese Gruppe wurde dann um Merkmale erweitert, die die SVM-Ergebnisse verbesserten.

Für die hier präsentierten Ergebnisse wurden letztlich die Merkmale Schufa-Bonität, KDF, berufliche Situation, Textähnlichkeit, Bietgeschwindigkeit und Track Record der Bieter ausgewählt. Diese Merkmale wurden jeweils so berechnet, wie sie dem Kenntnisstand in der Gebotsphase des jeweils betrachteten Projekts entsprachen.

Ergebnisse für verschiedene Strategien

Um zu beurteilen, ob die SVM nützlich ist, vergleichen wir sie mit den Ergebnissen, die man erhält, wenn man allein aufgrund der Schufa-Bonität Projekte auswählt.

Strategie 1: Auswahl bestimmter Bonitätsgruppen

Die Schufa-Bonitätsklassen sind explizit dafür gedacht, die Ausfallwahrscheinlichkeit der Kredite zu beurteilen. Daher liegt es nahe, zu prüfen, wie sehr in der Vergangenheit der durchschnittliche Ertrag von Krediten in bestimmten Bonitätsgruppen vom durchschnittlichen Ertrag in der Gesamtheit aller Kredite abwich. Man sollte insgesamt erwarten, dass das

höhere Risiko in den schlechten Bonitätsgruppen durch höhere Zinssätze in etwa ausgeglichen wird, so dass der durchschnittliche Ertrag in allen Bonitätsgruppen etwa gleich sein sollte. Für die größere Unsicherheit sollte in den schlechten Bonitäten allerdings noch ein Ertragsbonus zu verbuchen sein.

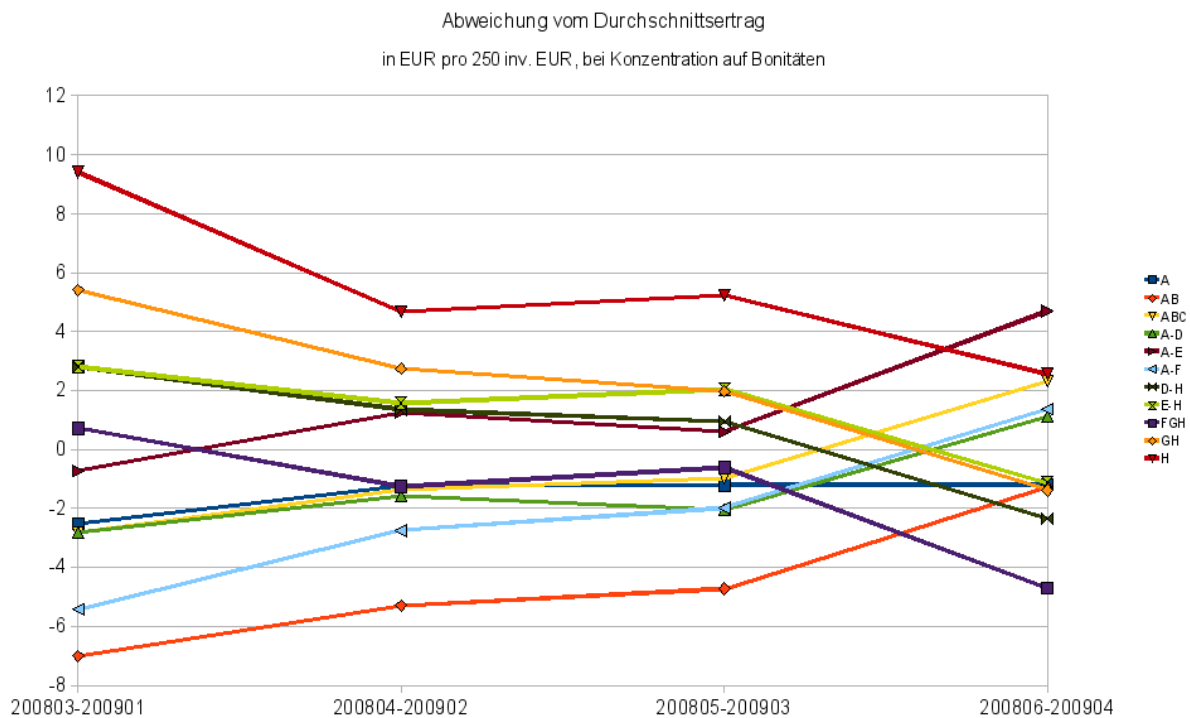


Schaubild 1: Abweichung vom Durchschnittsertrag, wenn Investitionen gleichmäßig auf bestimmte Bonitäten verteilt werden statt gleichmäßig auf alle Projekte, in EUR pro 250 investierten EUR. Es werden jeweils die in einem bestimmten Zeitraum hinzugekommenen Projekte gezählt, beispielsweise ganz links die zwischen März 2008 und Januar 2009 hinzugekommenen. Für den erwarteten Ertrag der Projekte wurde der Kenntnisstand von Juni 2009 verwendet (Ausfälle und Zahlungsquoten wurden als für den Rest der Projektlaufzeit ab Juni 2009 konstant angenommen; für die Zeit vorher wurden die realen Zahlungsquoten verwendet).

Schaubild 1 legt den Schluss nah, dass man in der Tat für die Übernahme höherer Risiken mit einem höheren Ertrag belohnt wird (denn im oberen Bereich sind die schlechten Bonitäten zu finden), und dass die Schufa-Bonitäten etwas über die Ausfallwahrscheinlichkeit aussagen (denn alle Kurven verlaufen einigermaßen nah bei Null – das entspricht dem ausgleichenden Effekt der höheren Zinssätze für schlechtere Bonitäten).

Wenn man statt einer SVM die Schufa-Bonität A als Vorhersage für Nicht-Ausfall nimmt (also annimmt, dass alle Projekte mit einer schlechteren Bonität ausfallen, aber keines der A-Projekte), ergibt sich im Durchschnitt über die in Schaubild 1 gezeigten Zeiträume folgende Confusion Matrix:

Ausfall vorhergesagt (nicht A)	Kein Ausfall vorherges. (A)		
55,5 Projekte	0,25 Projekte	Ausfall eingetreten	-14,74 EUR Ertrag pro 250 inv. EUR
197,75 Projekte	29,75 Projekte	Kein Ausfall eingetreten	29,38 EUR Ertrag pro 250 inv. EUR
20,86 EUR Ertrag pro 250 inv. EUR	19,32 EUR Ertrag pro 250 inv. EUR		

Ein Investor, der seinen Vorsteuer-Ertrag maximieren möchte, war für den betrachteten Zeitraum also schlecht beraten, einfach nur in alle A-Projekte zu investieren. Er vermeidet dadurch zwar fast alle Ausfälle, realisiert aber nur einen Ertrag von 19,32 EUR pro 250 investierten EUR – das ist schlechter als der Durchschnitt aller Nicht-A-Projekte, der bei 20,86 EUR liegt, der Investor wäre also besser gefahren, wenn er sich genau andersherum entschieden hätte. Anders betrachtet: der Recall für Ausfälle liegt bei fantastischen 99,55 %, aber dafür ist die Precision mit 21 Prozent recht bescheiden. Der Recall für Nicht-Ausfälle liegt bei nur 13 % - ein reiner A-Investor verzichtet also für 87 % der Nicht-Ausfälle aufs Investieren.

Strategie 2: Investieren in Nicht-Ausfälle laut SVM

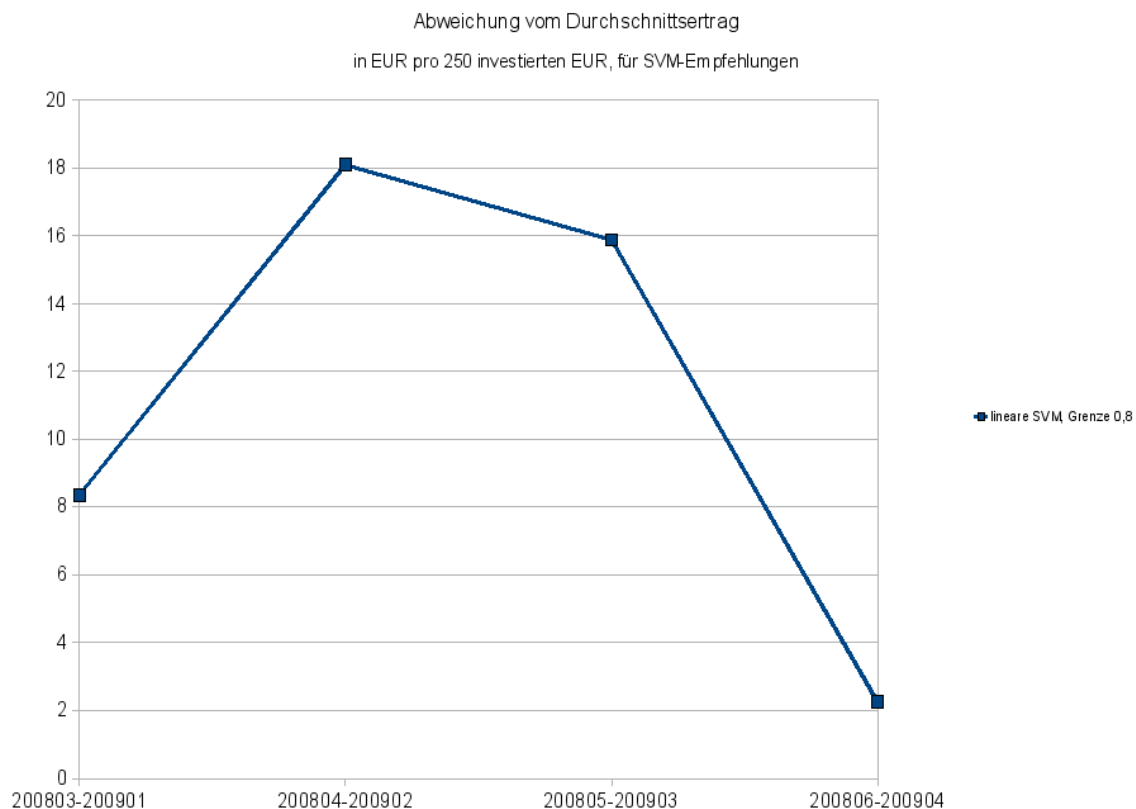


Schaubild 2: Das Gegenstück zu Schaubild 1, aber diesmal für die Strategie „Investiere in alle Projekte, die von der SVM empfohlen werden, und in keines, vor dem sie warnt“. Man sieht für alle betrachteten Monate einen Vorteil, in manchen Monaten ist die Differenz zwischen dem Durchschnittsertrag in den beiden Klassen beträchtlich.

Ausfall vorhergesagt	Kein Ausfall vorhergesagt		
12 Projekte	43,75 Projekte	Ausfall eingetreten	-14,74 EUR Ertrag pro 250 inv. EUR
18,5 Projekte	209 Projekte	Kein Ausfall eingetreten	29,38 EUR Ertrag pro 250 inv. EUR
10,39 EUR Ertrag pro 250 inv. EUR	21,54 EUR Ertrag pro 250 inv. EUR		

Confusion Matrix: Durchschnitt über die vier Zeiträume.

Wenn man eine lineare SVM verwendet, die die Merkmale Schufa-Bonität, KDF, berufliche Situation, Textähnlichkeit, Bietsgeschwindigkeit und Track Record der Bieter verwendet, erreicht man das in Schaubild 2 gezeigte Resultat. In manchen der vier Zeiträume liegen die Durchschnittserträge in der empfohlenen Klasse deutlich mehr als 10 EUR über denen der nicht empfohlenen Klasse. Die Confusion Matrix sieht besser aus als die für „Investiere nur in A“ - durch diese SVM wird die Auswahl wesentlich weniger eingeschränkt (auf 89 % aller Projekte).

Wenn man mit dem Wilcoxon Rank Sum Test die Strategien „Investiere nur in A“ und „Investiere laut SVM“ vergleicht, erhält man einen p-Value von 0,029. Wenn man also als Signifikanzlevel 5 % ansetzt, kommt man zu dem Schluss, dass der Unterschied zwischen der A-Folge „-2,51 EUR, -1,24 EUR, -1,21 EUR, -1,20 EUR“ und der SVM-Folge „8,35 EUR, 18,09 EUR, 15,87 EUR, 2,27 EUR“ (Vorteil für die vier betrachteten Zeiträume) statistisch signifikant ist.

Behandlung fehlender Track-Record-Daten

Ein praktisches Problem mit dieser erfolgreichen SVM ist, dass sie den Track Record der Bieter als Eingabe erwartet. Allerdings kommt es in der Praxis oft vor, dass eine überwiegende Mehrheit der Bieter eines Projektes, oder sogar alle, noch keinen verwertbaren Track Record haben. In so einem Fall kann aber keine SVM-Vorhersage berechnet werden.

Stattdessen wurde für Projekte, bei denen der Track Record der Bieter nicht definiert war, die Schwelle für den Track Record bestimmt, ab dem die SVM das Projekt ablehnen würde. Wenn diese Schwelle unter 0,8 liegt (d.h.: wenn die SVM das Projekt auch dann gutheißen würde, wenn die Bieter bisher mit weniger als 80 % ihrer Gebote erfolgreich waren), wurde das Projekt gutgeheißen und anderenfalls abgelehnt. Mit dieser Schwelle für die Schwelle kann man justieren, wie stark man seine Auswahl einschränken möchte.

SVM: Zusammenfassung und Ausblick

Selbst mit den begrenzten bisher vorliegenden Daten – Smava läuft erst seit wenigen Jahren, und noch kein Projekt ist über die gesamte Laufzeit von 36 Monaten gelaufen – kann man eine SVM so trainieren, dass ihre Vorhersagen besser sind als zufällige Vorhersagen, und ihre Empfehlungen erfolgreichere

Investmentstrategien erlauben als das gleichmäßige Streuen auf bestimmte Bonitäten.

Bei den Schaubildern und Confusion-Matrizen ist allerdings zu beachten, dass es sich teilweise noch um Prognosen handelt, da diese Projekte zwar schon bis zu 16 Monate alt sind, aber folglich auch noch mindestens 20 Monate Laufzeit vor sich haben, wenn sie nicht vorzeitig zurückgezahlt werden.

Die Merkmale für die SVM wurden von Hand ausgewählt (der automatische Merkmalsauswahl-Algorithmus, der in Weka eingebaut ist, war nicht erfolgreich). Daher kann es sein, dass es noch bessere Kombinationen von Merkmalen gibt.

Beobach.de wird von der Theano GmbH betrieben (www.theano.de)