Über statistische Probleme bei der Analyse von Daten aus dem Bereich der Kraftfahrzeugversicherung

Andreas Christmann
Universität Dortmund
Fachbereich Statistik
44221 Dortmund
christmann@statistik.uni-dortmund.de

Abstract:

Es wird die typische Struktur von Datensätzen aus der Kraftfahrzeugversicherung beschrieben und eine Strategie vorgestellt, die versucht, dieses Wissen zu nutzen, um versteckte Information in derartigen Daten zu erkennen und zu modellieren. Es werden sowohl klassische Methoden als auch neuere Methoden aus dem Bereich des Maschinellen Lernens genannt, die zur Analyse der Wahrscheinlichkeit für einen Schadenfall und zur Schätzung der Schadenhöhe verwendet werden können.

Keywords: Data Mining, Klassifikation, Kernel logistic regression, Machine Learning, Regression, Robustheit, Support Vector Machine; Support Vector Regression, Tarif, Versicherung.

1 Einführung

Unternehmen aus dem Bereich der Kraftfahrzeugversicherung benötigen Schätzer für die Wahrscheinlichkeit eines Versicherungsschadens und für die erwartete Summe der Schäden innerhalb eines Jahres, um einen Versicherungstarif zu erstellen. In diesem Artikel werden einige statistische Aspekte bei der Analyse von Datensätzen aus der Kraftfahrzeugversicherung diskutiert. Prinzipiell kann die vorgestellte Strategie auch für andere Fragestellungen interessant sein, wie z.B. für Credit Risk Scoring, Customer Relationship Management (CRM) oder CHURN Analysen.

2 Statistische Ziele

In diesem Abschnitt werden allgemeine statistische Ziele bei der Analyse von Daten aus der Kraftfahrzeugversicherung beschrieben. Versicherungsgesellschaften speichern pro Jahr viele Informationen über die einzelnen versicherten Personen auf. Einige dieser Informationen werden im folgenden genannt:
Persönliche Informationen: Name, Vorname, Art der Versicherung, Versicherungs-Nummer
Demographische Informationen: Geschlecht, Alter, Wohnort, Postleitzahl, Beruf, Bevölkerungsdichte des Wohngebiets
Information über den Fahrer: Hauptnutzer des Wagens, Fahrleistung in 1000 km pro Jahr, Garage
Information über die Familie: Alter und Geschlecht anderer Personen, die das Fahrzeug benutzen
Historische Informationen: Anzahl und Schadenhöhe von früheren Schadenfällen, Art des Schadens, Personenschäden, Kundenscore, Schadenfreiheitsklasse
Informationen über das Fahrzeug: Typ, Alter, Stärkeklasse bzw. PS
Response Informationen über das letzte Jahr: Schaden (ja/nein), Anzahl der Schäden, Schadenhöhen in EUR.

In der Praxis kennt man die exakten Schadenhöhen oft erst mit einiger zeitlicher Verzögerung, da die gesamte Schadenabwicklung mitunter Jahre in Anspruch nimmt. Aus diesem Grund sind manche der Schadenhöhen (insbesondere die extremen Schäden) bei der Tarifbildung nur mehr oder weniger gute Schätzungen für die tatsächlichen Schadenhöhen, so daß die empirische Verteilung der Schadenhöhen eine Mischung aus tatsächlichen Schadenhöhen und Schätzwerten ist.
Versicherungsgesellschaften berücksichtigen bei der Tarifberechnung, d.h. für die Festsetzung der tatsächlichen Prämie, unter anderem die sogenannte reine Prämie sowie Personalkosten, Verwaltungskosten, Sicherheitsrücklagen und natürlich Profite für die Unternehmen. In diesem Artikel wird nur die reine Prämie Y behandelt. Es wird hier stets vorausgesetzt, daß Y nichtnegativ ist. Für jeden Versicherungsnehmer wird die beobachtete reine Prämie y berechnet gemäß der folgenden Formel:

$$\text{reine Prämie} = \frac{\text{Summe der individuellen Schadenhöhen}}{\text{Anzahl der Tage unter Risiko / 360}}.$$

Der p-dimensionale Vektor der erklärenden Variablen (Risikofaktoren) sei mit x bezeichnet. Die *primäre Zielgröße* ist der bedingte Erwartungswert der reinen Prämie bei gegebenem x, d.h. $E(Y|X=x)$. Die *sekundäre Zielgröße* ist die bedingte Wahrscheinlichkeit für das Eintreten eines Versicherungsfalls, d.h. $P(Y>0|X=x)$.
Versicherungstarife sollten die folgenden vier Eigenschaften besitzen:
Fair: Die Schätzung für $E(Y|X=x)$ sollte zumindest approximativ unverzerrt sein, und dies idealerweise sowohl in der Gesamtpopulation als auch in hinreichend großen Teilpopulationen.
Präzise: Die Schätzungen sollten präzise sein. Ein denkbares Maß für die Präzision ist der mean squared error (MSE), aber andere Maße, etwa Statistiken vom Typ Pearson's Chi-Quadrat-Typ, sind ebenfalls denkbar.
Robust: Die Schätzungen sollten robust gegenüber moderaten Verletzungen der Modellannahmen sein, und der Einfluß von Ausreißern sollte beschränkt sein.
Einfach: Die Tarifstruktur sollte nur so komplex wie nötig sein, um in der Praxis tatsächlich einsetzbar zu sein. Insbesondere schränken hochdimensionale Wechselwirkungsterme die praktische Umsetzbarkeit teilweise ein, wenn diese Wechselwirkungsterme zwar statistisch signifikant aber dem Kunden nur schwer vermittelbar sind.

3 Charakteristische Eigenschaften von Daten aus der Kraftfahrzeugversicherung

Wie im vorigen Abschnitt bereits erwähnt wurde, sind die Dateien im Bereich der Kraftfahrzeugversicherung zum Teil sehr umfangreich, da pro Versicherungsnehmer viele Merkmale erhoben werden, wovon viele diskret sind. Zu den charakteristischen Eigenschaften von Datensätzen aus dem Bereich der Kraftfahrzeugversicherung zählen:

- Der Anteil der Personen, die einen Schaden innerhalb eines Jahres haben, ist relativ klein, z.B. in der Größenordnung von 5%. D.h. die empirische Verteilung von Y hat ein Atom in 0. Nur wenige Personen haben innerhalb eines Jahres mehr als einen Schadenfall.
- Die empirische Verteilung der reinen Prämie ist extrem rechtsschief. Oft haben nur ca. 0.1% der Versicherten einen extremen Schaden von mehr als 50000 EUR, jedoch beträgt die Gesamtsumme dieser Schäden oberhalb von 50000 EUR mitunter 30% bis 50% der Gesamtsumme aller Schadenhöhen.
- Es gibt eine hochdimensionale Abhängigkeitsstruktur zwischen den erklärenden Variablen und den beiden Zielgrößen.
- Es gibt sowohl monotone als auch nicht-monotone Zusammenhänge zwischen einzelnen Einflußgrößen und den beiden Zielgrößen.

4 Strategie

In diesem Abschnitt wird eine allgemeine Strategie beschrieben, um die primäre und die sekundäre Zielgröße zu schätzen. Die Strategie hat zwei Ziele. Einerseits wird versucht, das Wissen um die charakteristischen Eigenschaften von Datensätzen aus der Kraftfahrzeugversicherung, wie sie kurz im vorigen Abschnitt genannt wurden, zu nutzen. Andererseits soll möglichst viel Information, die in den Daten enthalten ist, in strukturierter Form transparent gemacht werden.
Definiere zunächst eine diskrete Hilfsvariable C, die die Schadenhöhen in k interpretierbare Klassen einteilt. Als Beispiel sei hier genannt:

C=0,	falls	Y = 0 EUR	(kein Schaden)
C=1,		$Y \in (0, 2000]$ EUR	(kleine Schäden)
C=2,		$Y \in (2000, 10000]$ EUR	(mittlere Schäden)
C=3,		$Y \in (10000, 50000]$ EUR	(große Schäden)
C=4,		Y > 50000 EUR	(extrem hohe Schäden).

Offenbar gilt E(Y|C=0, X=x) = 0. Aus dem Satz der totalen Wahrscheinlichkeit folgt daher für die primäre Zielgröße:

$$E(Y \mid X = x) =$$

$$P(C > 0 \mid X = x) \sum_{c=1}^{k} P(C = c \mid C > 0, X = x) E(Y \mid C = c, X = x).$$

Ein Vorteil dieser Strategie ist, daß man nicht nur Schätzungen für die reine Prämie E(Y|X=x) erhält, sondern darüber hinaus auch Schätzungen für die primäre und die sekundäre Zielgröße in interpretierbaren Subgruppen. Für C=1 erhält man zum Beispiel Schätzungen für die Wahrscheinlichkeit eines geringfügigen Schadens zwischen 0 und 2000 EUR und für die Schadenhöhe, jeweils bedingt bezüglich des Vektors x der erklärenden Variablen. Damit kann man vergleichen, ob die Risikostruktur in verschiedenen Subgruppen vergleichbar ist oder nicht.
Ein anderer Vorteil dieser Strategie ist ein rechentechnischer: nur die Wahrscheinlichkeit P(C>0|X=x) muß für den *gesamten* Datensatz geschätzt werden, welche in unserem Fall natürlich gleich der sekundären Zielgröße P(Y>0|X=x) ist. Alle anderen bedingten Wahrscheinlichkeiten und Erwartungswerte müssen nur aus Teildatensätzen geschätzt werden. Da oft nur wenige Prozent aller Versicherungsnehmer innerhalb eines Jahres überhaupt einen Schaden haben, kann dies eine große Einsparung an Rechenzeit bedeuten.
Als weiterer Vorteil ist zu nennen, daß mit diesem Ansatz die Datensätze mit y=0 kein Problem mehr für die Anpassung mit stetigen Modellen wie der Gamma-Regression darstellen. So eliminiert beispielsweise die SAS-Prozedur PROC GENMOD mit der OPTION DIST=GAMMA im MODEL-Statement zur Anpassung einer Gamma-Regression automatisch alle Beobachtungen aus dem Datensatz, für die die Zielgröße den Wert Null hat, was für den Fall der Tarifberechnung im Rahmen der Versicherungsproblematik sicherlich i.a. nicht gewünscht wird.
Schließlich sei darauf hingewiesen, daß die obige Strategie zu einer Reduktion der benötigen Wechselwirkungsterme höherer Ordnung beitragen kann.
Verwendet man konsistente Schätzer für die bedingten Wahrscheinlichkeiten und bedingten Erwartungswerte, so schätzt man nach Slutzky's Theorem auch die primäre Zielgröße E(Y|X=x) konsistent (in Wahrscheinlichkeit oder fast sicher), jedoch nicht notwendig unverzerrt. Beim Data Mining splittet man oft den gesamten Datensatz auf in 3 Teile: Trainingsdaten, Validationsdaten und Testdaten. Der Validationsdatensatz kann verwendet werden, um eine etwaige Biasreduktion durchzuführen.

Einige Methoden zur Modellierung der bedingten Wahrscheinlichkeiten und der bedingten Erwartungswerte werden im folgenden aufgeführt:

- Logistische Regression + Gamma-Regression
  (SAS/PROC GENMOD mit DIST=BINOMIAL bzw. DIST=MULTINOMIAL bzw. DIST=GAMMA, SAS/PROC LOGISTIC)
- Logistische Regression + semi-parametrische Regression
  (SAS/PROC GENMOD + SAS/PROC GAM)
- Entscheidungsbäume und Regressionsbäume
  (SAS/Enterprise Miner)
- Kernel logistic regression + Support Vector Regression
  (in SAS 8e nicht verfügbar, Programme: myKLR + SVMlight)
- Kombination der obigen Verfahren mit Methoden der Extremwertstatistik, z.B. Generalisierten Pareto-Modellen

In Christmann (2004) wird gezeigt, daß diese Strategie auch für große Datensätze einsetzbar ist. Zur Modellierung hochdimensionaler Abhängigkeitsstrukturen wird dort zur Modellierung auf Methoden des maschinellen Lernens zurückgegriffen.

5 Diskussion

Die in diesem Artikel dargestellte recht allgemeine Strategie erlaubt es einerseits, Vorwissen über die charakteristischen Eigenschaften von Datensätzen aus der Kraftfahrzeugversicherung zu nutzen, und andererseits versteckte Information, die in den Daten enthalten ist, in strukturierter Form sichtbar zu machen. Derartige Information kann aus Sicht des Autors für Unternehmen dann wichtig und interessant sein, wenn wie im Fall der Kraftfahrzeugversicherung gesichertes Vorwissen über die Daten vorliegt und Interesse an Subgruppen oder hochdimensionalen Abhängigkeitsstrukturen besteht. In diesem Zusammenhang erscheint ein Vergleich von Ergebnissen mit klassischen Techniken wie dem Marginalsummenmodell oder einer Gamma-Regression mit Methoden des maschinellen Lernens wie etwa einer Kombination aus Kernel Logistic Regression, Support Vector Regression und Methoden der Extremwertstatistik interessant und sinnvoll.

Acknowledgements

Diese Arbeit wurde vom Sonderforschungsbereich 475 „Komplexitätsreduktion in multivariaten Datenstrukturen“ der Deutschen Forschungsgesellschaft (DFG) unterstützt. Der Autor dankt Herrn A. Wolfstein and Herrn Dr. W. Terbeck vom Verband öffentlicher Versicherer in Düsseldorf für hilfreiche Diskussionen.

Literatur

[1] Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley, New York.

[2] Christmann, A. (2004). On a strategy to develop robust and simple tariffs from motor vehicle insurance data. Universität Dortmund, SFB-475, TR 16/04.

[3] Christmann, A., Steinwart, I. (2003). On robust properties of convex risk minimization methods for pattern recognition. Universität Dortmund, SFB-475, TR 15/03. Erscheint in: Journal of Machine Learning Research.

[4] Cox, D.R., Snell, E.J. (1989). Analysis of Binary Data. 2nd ed. Chapman & Hall, London.

[5] Hastie, T., Tibshirani, R., Friedman, J. (2001). The elements of statistical learning. Springer, New York.

[6] Joachims, T. (1999). Making large-Scale SVM Learning Practical. In: B. Schölkopf, C. Burges, A. Smola (ed.), Advances in Kernel Methods - Support Vector Learning, MIT-Press. http://svmlight.joachims.org/

[7] Keerthi, S.S., Duan, K., Shevade, S.K., Poo, A.N. (2002). A fast dual algorithm for kernel logistic regression. National University of Singapore. Preprint. http://guppy.mpe.nus.edu.sg/~mpessk

[8] Rüping, S. (2003). myKLR - kernel logistic regression. Universität Dortmund, FB Informatik, http://www-ai.cs.uni-dortmund.de/SOFTWARE

[9] SAS/STAT Users Guide I und II (1990). SAS Institute Inc.

[10] SAS/STAT Software: Changes and Enhancements, through Release 6.11 (1996). SAS Institute Inc.

[11] SAS/STAT Software: Changes and Enhancements, Release 8.2 (2001). SAS Institute Inc.

[12] Schölkopf, B, Smola, A.J. (2002). Learning with kernels. Support vector machines, regularization, optimization, and beyond. MIT-Press, Cambridge.

[13] Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York.