Thema:

Association Rule Mining: Tests auf Hoch-Korrelation zur Regelextraktion Analyse, Erweiterungen und prototypische Implementierung

Diplomhausarbeit

im Fachgebiet Wirtschaftsinformatik
am Lehrstuhl für Quantitative Methoden

Themensteller: Prof. Dr. Ulrich Müller-Funk

vorgelegt von: Michael Höing
Bismarckallee 49, C101
48151 Münster
0251/83851770
hoeingm@uni-muenster.de

Abgabetermin: 2005-09-16

Inhaltsverzeichnis

1 Einleitung

Beim Association Rule Mining wird mit Hilfe von Data Mining Tools nach Mustern gesucht, die sprachlich in Form von Wenn-dann-Regeln – d. h. Assoziazionsregeln – dargestellt werden können. Eine typische Anwendung ist die *Warenkorbanalyse*, die das Kaufverhalten von Konsumenten untersucht. Dazu wird eine Menge von Kaufakten – d. h. Transaktionen – analysiert, um sachliche Verbundeffekte unter den gekauften Artikeln aufzudecken und daraus Aussagen über das Käuferverhalten entwickeln zu können.

Ein klassisches Ergebnis der Warenkorbanalyse ist z. B.:

*„**Wenn** Kunden Brot und Butter kaufen, **dann** kaufen 70% von ihnen ebenfalls Käse.“*

Solche Wenn-dann-Regeln lassen sich zur Absatzförderung nutzen, da sie als Entscheidungsgrundlage für z. B. Maßnahmen zur Verkaufsraumgestaltung oder Strategien für ein selektives Marketing dienen.[1]

Das Problem, Abhängigkeiten zwischen – allgemein gefasst – Merkmalen zu analysieren, ist bereits lange Gegenstand der deskriptiven Statistik und begegnet dort in Form von Kontingenztafeln und Assoziationsmaßen.[2] Klassische Verfahren der deskriptiven Statistik berücksichtigen jedoch nicht den das Data Mining prägenden Aspekt einer sehr großen zugrunde liegenden Datenmenge. Neben der *theoretischen* Betrachtung ergibt sich so immer die Notwendigkeit, gleichermaßen *numerische* bzw. *algorithmische* Aspekte bei der Wahl oder Entwicklung eines Verfahrens zu berücksichtigen. Die *empirische Brauchbarkeit* von Verfahren nimmt für ein Data Mining also eine ebenso wichtige Rolle wie ihre theoretische Fundierung ein; umgekehrt sind *theoretisch* zu rechtfertigende Verfahren nicht automatisch *empirisch brauchbar*.

Bei der Betrachtung implementierter Verfahren zum Association Rule Mining fällt nun auf, dass Ineffizienz, die auf die Größe der Datenbasis zurückzuführen ist, vielfach mit steigender Rechenleistung moderner Systeme kompensiert wird – bzw. versucht wird, zu kompensieren. *Empirische Brauchbarkeit* ist somit keinesfalls mit *Effizienz* gleichzusetzen; sie definiert sich anscheinend vielfach darin, dass ein Verfahren auch auf eine sehr große Datenmenge angewendet werden kann. State-of-the-art rule mining „kränkelt“ nun daran, dass aktuelle Software keine *effizienten* Verfahren zur Ableitung von Regeln bereitstellt. In der Folge ist die Menge gefundener Muster in Form von Wenn-dann-Regeln extrem groß und ihre Bewertung gestaltet sich entsprechend aufwendig.[3]

---

1 Vgl. Grob, Bensberg (1999) und Han, Kamber (2001), S. 225 f.
2 Vgl. Ferschl (1985).
3 Vgl. Müller-Funk (2004).

Der Grund für diese Entwicklung ist zum großen Teil darauf zurückzuführen, dass die Menge der potentiellen Wenn-dann-Regeln mit wachsender Datenmenge exponentiell steigt. Dieses Komplexitätsproblem hat dazu geführt, dass beim Association Rule Mining die Datenbasis zunächst sehr stark eingeschränkt wird, um die Menge der potentiellen Regeln zu beherrschen. Eingeschränkt wird, indem lediglich Artikelkombinationen betrachtet werden, die *häufig* auftreten. Muster werden dann im Anschluss nicht zwischen allen, sondern lediglich zwischen den Artikeln gesucht, die *häufig* gekauft werden. Das Association Rule Mining ist also ein 2-stufiger Prozess:

1. Finde Artikelsets, die *häufig* zusammen gekauft werden.

2. Finde Wenn-dann-Regeln – allerdings nur zwischen *häufig* gekauften Artikelsets, damit die Menge der Regeln beherrschbar bleibt.

Diente das Aufsplitten in zwei Teilprobleme erst einmal nur der Beherrschung der Komplexität, so führte es auch dazu, dass die beiden Stufen des rule mining völlig separat voneinander betrachtet werden können. Diese mögliche separate Betrachtung zog folgende Beurteilung der beiden Teilprobleme von z. B. HAN UND KAMBER[4] nach sich:

"[…]. The second step is the easiest of the two. The overall performance of mining association rules is determined by the first step. […]"

Da diese Bewertung korrekt ist, konzentrierte sich die Forschung auf die erste Stufe des rule mining und dieses erste Teilproblem kann als „gelöst" betrachtet werden; Konsequenz ist allerdings auch, dass der zweiten Stufe weniger Beachtung geschenkt wurde. Leider darf aus der Komparation „easiest" in obigem Zitat nicht auf „easy" geschlossen werden bzw. mit anderen Worten: Die Beherrschung von Schritt 1 des rule mining impliziert keinesfalls die Effizienz des Gesamtprozesses. Derzeitige Ansätze zur Lösung des zweiten Problems spiegeln diese Tatsache allerdings in keiner Weise wieder, was zur genannten Ineffizienz führt.

Im Fokus der vorliegenden Arbeit ist daher die zweite Stufe des rule mining. Nach einer Einordnung des Association Rule Mining in die umschließende Domäne der *Knowledge Discovery in Databases* und anschließender Formalisierung in Kapitel 2 wird in Kapitel 3 zunächst nachgewiesen, dass die erste Stufe des rule minings als beherrscht betrachtet werden kann. Hierzu werden zwei sehr unterschiedliche Algorithmen erläutert, die den Bogen zwischen *naivem Ansatz* und *Implementierung unter Nutzung effizienter Datenstrukturen* aufspannen. In Kapitel 4 werden state-of-the-art Verfahren zur Regelextraktion – der zweiten Stufe des Rule Minings – betrachtet. Neben der kritischen Beleuchtung der-

[4] Vgl. Han, Kamber (2001), S. 228.

zeit akzeptierter Verfahren, die Assoziationsmaße verwenden, werden insbesondere statistische Tests betrachtet. Zentral ist in jedem Fall die Prüfung der *empirischen Brauchbarkeit*, d. h. die *Effizienz* neben der bloßen Durchführbarkeit auf großen Datensätzen. Mit den *Tests auf Hoch-Korrelation* in Kapitel 5 wird gezeigt, wie die Regelextraktion mit Hilfe statistischer Tests effizient durchgeführt werden kann. Kapitel 6 stellt Erweiterungen des Verfahrens vor, die sich insbesondere dem Problem der exponentiell wachsenden Menge potentieller Regeln widmen. In Kapitel 7 wird eine prototypische Implementierung aller behandelten Verfahren unter Nutzung des R-frameworks gezeigt, die durchweg auch zur Ermittlung der in dieser Arbeit dargestellten Ergebnisse genutzt wurde. Dies ist u. a. die Konsequenz der Tatsache, dass die hier vorgestellten statistischen *Tests auf Hoch-Korrelation* in der Praxis – d. h. Software – derzeit keine Berücksichtigung finden.

Literaturverzeichnis

Agrawal, R.; Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Databases. Hrsg.: M. Kaufmann, 1994, S. 487-499.

Agrawal, R.; Imielinski, T.; Swami, A.: Mining Association Rules between Sets auf Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on the Management of Data, Jg. 22, 1993, S. 207-216.

Bastide, Y. et al.: Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets. In: Proceedings of the First International Conference on Computational Logic. Hrsg.: Springer-Verlag, 2000, S. 972-986.

Bayardo, R. J.; Agrawal, R.; Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. In: Proceedings of the 15th International Conference on Data Engineering. Hrsg.: IEEE Computer Society, 1999, S. 188-197.

Brin, S.; Motwani, R.; Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations. In: Proceedings ACM SIGMOD International Conference on Management of Data, 26(2), 1997, S. 265-276.

Fahrmeir, L.; Hamerle, A.; Tutz, G.: Multivariate statistische Verfahren. 2. Auflage, Berlin, New York, 1996.

Fahrmeir, L. u. a.: Statistik. Der Weg zur Datenanalyse. 3. Auflage, Berlin u. a., 2001.

Ferschl, F.: Deskriptive Statistik. Würzburg, 1985.

Gamma, E. et al.: Design Patterns: Elements of reusable object-oriented software. 1. Aufl., Amsterdam et al., 1995.

Goethals, B: Survey on Frequent Pattern Mining. 2003. http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG445/literatur/goethals_2003a.pdf.

Grob, H. L.; Bensberg, F.: Das Data-Mining-Konzept. In: Computergestütztes Controlling. Hrsg.: H. L. Grob, 1999.

Han, J. et. al.: Mining Frequent Patters without Candidate Generation: A Frequent-Pattern Tree Approach. In: Data Mining and Knowledga Discovery 8(1), 2004, S. 53-87.

Han, J.; Kamber, M.: Data Mining – Concepts and Techniques. San Francisco, 2001.

Hartung, J.; Elpelt, B.; Klösner, K.-H.: Statistik. Lehr- und Handbuch der angewandten Statistik. 10. Auflage, Oldenburg, München, Wien, 1995.

Hahsler, M.; Grün, B.; Hornik, K: A Computational Environment for Mining Associations Rules and Frequent Item Sets. 2005. http://wwwai.wu-wien.ac.at/~hahsler/research/arules_workingpaper15_2005/arules.pdf.

Hofmann, H.; Wilhelm, A.: Visual Comparison of Association Rules. In: Computational Statistics 16(3), 2001, S. 399-415.

Müller-Funk, U.: Statistische Aspekte von Assoziationsregeln. In: Trendberichte zum Controlling – Festschrift für Heinz Lothar Grob. Hrsg.: F. Bensberg , J. v. Brocke, Martin Schultz, 2004, S. 201-212.

Pasquier, N. et al.: Discovering Frequent Closed Itemsets for Association Rules. In: Proceedings of the 7th International Conference on Database Theory. Hrsg.: Springer-Verlag, 1999a, S. 398-416.

Pasquier, N. et al.: Closed Set Based Discovery of Small Covers for Association Rules. In: Proceedings 15emes Journees Bases de Donnees Avancees. 1999b, S. 361-381.

Pasquier, N. et al.: Generating a Condensed Representation for Association Rules. In: Journal of Intelligent Information Systems 24(1). Hrsg.: Kluwer Academic Publishers, 2005, S. 29-60.

Piatetsky-Shapiro, G.: Discovery, Analysis ans Presentation of Strong Rules. In: Knowledge Discovery in Databases. Hrsg.: G. Piatetsky-Shapiro, W. Frawley, 1991, S. 229-248.

Silverstein, C. et al.: Scalable Techniques for Mining Causal Structures. In: Proceedings of the 24th International Conference on Very Large Databases. Hrsg.: M. Kaufmann, 1998, S. 594-605.

Srikant, R.; Agrawal, R.: Mining Generalizes Association Rules. In: In: Proceedings of the 21st International Conference on Very Large Databases. Hrsg.: M. Kaufmann, 1995, S. 407-419.

Srikant, R.; Vu, Q.; Agrawal, R.: Mining Association Rules with Item Constraints. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining KDD-97, 1997, S. 67-73.

Witting, H.: Mathematische Statistik. Stuttgart, 1966.

Witting, H.; Müller-Funk, U.: Mathematische Statistik. Stuttgart, 1995.