

**Anwenderspezifische Reduzierung
von Mengen interessanter
Assoziationsregeln
mittels Evolutionärer Algorithmen**

Birgit Wenke

Universität der Bundeswehr München
Fakultät für Informatik

Thema der Dissertation: Anwenderspezifische Reduzierung von Mengen
interessanter Assoziationsregeln mittels
Evolutionärer Algorithmen

Verfasser: Birgit Wenke

Promotionsausschuss:

Vorsitzender: Prof. Dr. Markus Siegle
1. Berichterstatter: Prof. Dr. Ulrike Lechner
2. Berichterstatter: Prof. Dr. Martin Wirsing
3. Berichterstatter: Prof. Dr. Mark Minas
4. Berichterstatter: Prof. Dr. Karl Morasch

Tag der Prüfung: 23. Juli 2008
Mit der Promotion erlangter akademischer Grad Doktor der Naturwissenschaften
(Dr. rer. nat.)

Neubiberg, 31. Januar 2008

Der Druck der Arbeit wurde aus Haushaltsmitteln der Universität der
Bundeswehr München gefördert.

Berichte aus der Wirtschaftsinformatik

Birgit Wenke

**Anwenderspezifische Reduzierung
von Mengen interessanter Assoziationsregeln
mittels Evolutionärer Algorithmen**

Shaker Verlag
Aachen 2008

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Zugl.: München, Univ. der Bundeswehr, Diss., 2008

Copyright Shaker Verlag 2008

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8322-7564-8

ISSN 1438-8081

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: www.shaker.de • E-Mail: info@shaker.de

VORWORT

Die vorliegende Arbeit wurde im Wintersemester 2008 an der Fakultät für Informatik der Universität der Bundeswehr München als Dissertation angenommen. Das Kolloquium fand am 23. Juli 2008 statt.

Mein besonderer Dank gilt meiner Doktormutter Frau Prof. Dr. Ulrike Lechner, welche mir sowohl den Freiraum für die Erstellung dieser Arbeit gewährt hat als auch über den gesamten Entstehungszeitraum hinweg im Rahmen von Diskussionen wichtige und zielführende inhaltliche Impulse gegeben hat. Ihre wohlwollende Betreuung und Kritik war stets hilfreich und motivierend.

Bedanken möchte ich mich auch bei Herrn Prof. Dr. Martin Wirsing für die Erstellung des Zweitgutachtens und seine freundliche und faire Unterstützung. Gedenken möchte ich an dieser Stelle auch Herrn Prof. Dr. Jürgen Janas, meinem ursprünglichen Zweitgutachter und Ideengeber dieser Arbeit, welcher im Februar 2006 sehr überraschend verstorben ist.

Darüber hinaus danke ich meinem Ehemann, welcher mir durch sein Vertrauen und seine Zuversicht den Rückhalt für diesen Weg gab.

Meinen Eltern möchte ich besonders dafür danken, dass sie mich in meiner universitären Ausbildung gefördert haben und damit die Entscheidung für eine Promotion ermöglicht haben.

Abschließend möchte ich meinen Dank all denjenigen Familienangehörigen und Freunden aussprechen, die mich bei meiner Promotion unterstützt haben.

München im August 2008

Birgit Wenke

INHALTSÜBERSICHT

| | | |
|---|---|-----|
| 1 | Einleitung | 1 |
| 2 | Begriffe und Notationen | 19 |
| 3 | Data Mining | 23 |
| 4 | Evolutionäre Algorithmen | 217 |
| 5 | Iterative Selektion von Assoziationsregeln nach anwenderspezifischen Interessensmerkmalen mit Hilfe Evolutionärer Algorithmen | 293 |
| 6 | Schlussbetrachtung | 549 |

INHALTSVERZEICHNIS

| | |
|--|-----|
| Vorwort | V |
| Inhaltsübersicht | VII |
| Inhaltsverzeichnis | IX |
| 1 Einleitung | 1 |
| 2 Begriffe und Notationen | 19 |
| 3 Data Mining | 23 |
| 3.1 Der Knowledge Discovery Prozess | 27 |
| 3.2 Assoziationsregeln | 34 |
| 3.2.1 Grundlagen von Assoziationsregeln | 37 |
| 3.2.2 Die Apriori Algorithmen | 41 |
| 3.2.2.1 Berechnung der häufigen Attributwertmengen | 43 |
| 3.2.2.2 Berechnung der Assoziationsregeln | 52 |
| 3.2.3 Andere Algorithmen zur Ermittlung von häufigen Attributwertmengen und Assoziationsregeln | 58 |
| 3.2.3.1 Algorithmen der Apriori Familie | 65 |
| 3.2.3.2 Auf Baumstrukturen operierende Algorithmen | 85 |
| 3.2.3.3 Interaktive Algorithmen | 97 |
| 3.2.3.4 Algorithmen für Spezialfälle | 114 |
| 3.3 Klassifikationsregeln | 131 |
| 3.3.1 Grundlagen der Klassifikation | 131 |
| 3.3.2 Bewertungsparameter von Klassifikationsregeln binärer Klassifikationen | 140 |
| 3.4 Interessantheitsmaße | 142 |
| 3.4.1 Objektive Interessantheitsmaße für Assoziationsregeln | 147 |
| 3.4.2 Subjektive Interessantheitsmaße für Assoziationsregeln | 168 |
| 3.4.3 Objektive Interessantheitsmaße für Klassifikationsregeln | 177 |
| 3.4.3.1 Objektive Interessantheitsmaße zur Beurteilung der Qualität von Klassifikationsregeln | 178 |
| 3.4.3.2 Objektive Interessantheitsmaße zur Beurteilung der Komplexität von Klassifikationsregeln | 183 |

| | | |
|---------|--|-----|
| 3.4.4 | Subjektive Interessantheitsmaße für Klassifikationsregeln | 185 |
| 3.5 | Definitionen redundanter Assoziationsregeln | 190 |
| 3.5.1 | Definitionen redundanter Assoziationsregeln anhand von Ableitungsregeln | 194 |
| 3.5.2 | Definition redundanter Assoziationsregeln anhand von Attributwerthierarchien | 202 |
| 3.5.3 | Definitionen redundanter Assoziationsregeln anhand der Aussagekraft der Assoziationsregeln | 207 |
| 3.5.4 | Vergleich der Definitionen von Redundanz | 215 |
| 4 | Evolutionäre Algorithmen | 217 |
| 4.1 | Grundlagen Evolutionärer Algorithmen | 220 |
| 4.2 | Verwandte Gebiete | 224 |
| 4.3 | Die verschiedenen Arten Evolutionärer Algorithmen | 226 |
| 4.4 | Strukturen und grundlegende Operatoren Evolutionärer Algorithmen | 232 |
| 4.4.1 | Codierung | 234 |
| 4.4.2 | Initialisierung | 237 |
| 4.4.3 | Fitnessfunktion | 239 |
| 4.4.4 | Elternselektion | 246 |
| 4.4.4.1 | RouletteSelektion | 248 |
| 4.4.4.2 | Stochastic Universal Sampling | 250 |
| 4.4.4.3 | Turnierselektion | 252 |
| 4.4.4.4 | Truncation Selection | 254 |
| 4.4.5 | Rekombination | 256 |
| 4.4.5.1 | Multi-Point Crossover | 257 |
| 4.4.5.2 | Single-Point Crossover | 258 |
| 4.4.5.3 | Uniform Crossover | 258 |
| 4.4.5.4 | Crossover mit reduzierter Rekombination | 259 |
| 4.4.5.5 | Diagonal Crossover | 259 |
| 4.4.6 | Mutation | 261 |
| 4.4.7 | Umweltselektion | 263 |

| | | |
|---------|--|-----|
| 4.4.8 | Abbruchkriterien | 267 |
| 4.5 | Stärken und Schwächen Evolutionärer Algorithmen | 273 |
| 4.6 | Evolutionäre Algorithmen im Zusammenhang mit Klassifikations- und Assoziationsregeln | 276 |
| 4.6.1 | Wissenschaftliche Ansätze | 280 |
| 4.6.2 | Discipulus | 290 |
| 5 | Iterative Selektion von Assoziationsregeln nach anwenderspezifischen Interessensmerkmalen mit Hilfe Evolutionärer Algorithmen | 293 |
| 5.1 | Grundlegende Merkmale der entwickelten Vorgehensweise | 312 |
| 5.1.1 | Ablauf der Vorgehensweise | 313 |
| 5.1.1.1 | Verbale Beschreibung der Vorgehensweise | 313 |
| 5.1.1.2 | Exemplarisches Anwendungsbeispiel | 316 |
| 5.1.2 | Einordnung bezüglich der verwendeten Interessantheitsmaße | 323 |
| 5.1.3 | Art des Anwendungsgebiets und Abgrenzung zu Recommendation Engines | 325 |
| 5.2 | Die Architektur der Vorgehensweise | 330 |
| 5.2.1 | Anwendungsfalldiagramm | 331 |
| 5.2.2 | Aktivitätsdiagramm | 332 |
| 5.3 | Komponenten der Vorgehensweise und ihre Funktionsweisen | 337 |
| 5.3.1 | Codierung | 341 |
| 5.3.1.1 | Codierungsformen der Literatur | 341 |
| 5.3.1.2 | Codierung von Klassifikationsregeln in der Vorgehensweise | 346 |
| 5.3.1.3 | Codierung von Assoziationsregeln in der Vorgehensweise | 350 |
| 5.3.2 | Auswahl der zu präsentierenden Assoziationsregeln | 354 |
| 5.3.2.1 | Der Suchraum | 355 |
| 5.3.2.2 | Distanzmaße der Literatur | 357 |
| 5.3.2.3 | Das verwendete Auswahlverfahren in der Vorgehensweise | 360 |

| | | |
|-----------|---|-----|
| 5.3.3 | Bewertung durch den Anwender | 363 |
| 5.3.4 | Erstellung der Anfangspopulation von Klassifikationsregeln | 367 |
| 5.3.5 | Die Anwendung genetischer Operatoren | 371 |
| 5.3.5.1 | Rekombination | 371 |
| 5.3.5.2 | Mutation | 374 |
| 5.3.6 | Bewertung der Klassifikationsregeln | 377 |
| 5.3.6.1 | Mehrkriterielle Zielfunktionen der Literatur | 379 |
| 5.3.6.2 | Die verwendet Zielfunktion in der Vorgehensweise | 381 |
| 5.3.7 | Umweltselektion | 384 |
| 5.3.8 | Abbruch | 388 |
| 5.3.9 | Konvergenz | 390 |
| 5.4 | Evaluationen und Expertengespräche | 391 |
| 5.4.1 | Vorevaluation | 394 |
| 5.4.2 | Die Evaluationen | 401 |
| 5.4.2.1 | Eigenschaften der Evaluationen | 403 |
| 5.4.2.2 | Das Bewertungskriterium | 414 |
| 5.4.3 | Evaluationsergebnisse des Heart Disease Datensatzes | 420 |
| 5.4.3.1 | Eigenschaften des Heart Disease Datensatzes | 422 |
| 5.4.3.2 | Evaluationsergebnisse bei 100 Klassifikationsregeln | 423 |
| 5.4.3.2.1 | Evaluationsergebnisse der acht Verfahrensvarianten | 424 |
| 5.4.3.2.2 | Evaluationsergebnisse der sechs Interessenszenarien | 432 |
| 5.4.3.3 | Evaluationsergebnisse bei 500 Klassifikationsregeln | 445 |
| 5.4.3.3.1 | Evaluationsergebnisse der acht Verfahrensvarianten | 445 |
| 5.4.3.3.2 | Evaluationsergebnisse der sechs Interessenszenarien | 454 |
| 5.4.3.4 | Zusammenfassung der Evaluationsergebnisse des Heart Disease Datensatzes | 467 |
| 5.4.4 | Evaluationsergebnisse des flickr Datensatzes | 477 |
| 5.4.4.1 | Eigenschaften des flickr Datensatzes | 479 |

| | | |
|------------|--|--------|
| 5.4.4.2 | Evaluationsergebnisse bei 100 Klassifikationsregeln | 482 |
| 5.4.4.2.1 | Evaluationsergebnisse der acht Verfahrensvarianten | 482 |
| 5.4.4.2.2 | Evaluationsergebnisse der sechs Interessenszenarien | 490 |
| 5.4.4.3 | Evaluationsergebnisse bei 500 Klassifikationsregeln | 502 |
| 5.4.4.3.1 | Evaluationsergebnisse der acht Verfahrensvarianten | 502 |
| 5.4.4.3.2 | Evaluationsergebnisse der sechs Interessenszenarien | 510 |
| 5.4.4.4 | Zusammenfassung der Evaluationsergebnisse des flickr Datensatzes | 522 |
| 5.4.5 | Vergleich und zusammenfassende Interpretation der Evaluationsergebnisse der beiden Datensätze | 531 |
| 5.4.6 | Expertengespräche | 544 |
| 6 | Schlussbetrachtung | 549 |
| Anhang 1a: | Vorgehensweise mit Heart Disease Datensatz | CD-ROM |
| Anhang 1b: | Vorgehensweise mit flickr Datensatz | CD-ROM |
| Anhang 2a: | Evaluationsergebnisse mit Heart Disease Datensatz | CD-ROM |
| Anhang 2b: | Evaluationsergebnisse mit flickr Datensatz | CD-ROM |
| Anhang 3: | Fragebogen | 557 |
| Anhang 4: | Vorevaluation | 561 |
| Anhang 5: | Prozeduren in Pseudocode mit anschließender Komplexitätsbetrachtung | 575 |
| | Literaturverzeichnis | 591 |
| | Abbildungsverzeichnis | 609 |
| | Tabellenverzeichnis | 621 |
| | Index | 627 |