DATA MINING:
A NEW LABEL FOR AN OLD PROBLEM?

Wolfgang Gaul
University of Karlsruhe
Germany

Abstract:

Based on an overview on recent contributions with respect to data mining applications terms like „data mining", „data warehousing" and „knowledge discovery in databases" are related to the well-known discussion where "data analysis" and "decision support" are combined with "expert knowledge research". Next, commercial data mining tools are compared with the help of positioning and segmentation procedures to get a feeling for the support provided by software tailored for performing tasks concerning knowledge discovery in databases. Finally, an own modification of an association rule algorithm is used to handle buying histories in the area of consumer behavior interpretation.

Introduction via Recent Data Mining Applications

Instead of what is normally done in the introductory part of scientific papers (where the existing theory-oriented literature is used to structure the area under consideration), here, the starting point is a table in which recently published data mining applications are depicted. The collection in Table 1 is, of course, only a sample from the set of applications in this field, not all boxes in this table could be filled, and some entries have still question marks (?) but this listing can already make pretty clear what kind of messages can be extracted for a discussion of data mining topics:

(1) Quite a number of application areas are from economics (e.g., finance, insurance) and many problem descriptions stress economic objectives (e.g., cost controlling, cross sales, reduction of downtimes of plants or equipments).

(2) Data sets are generally large.

(3) The data mining techniques used are not of central interest to authors from the application side. They often only give hints to some methodological aspects (e.g., application of decision trees, neural nets, association rules) and not in all cases mention the software used.

(4) Authors from the research side are to some extend reluctant to relate merits of the methodology they propose to the efforts known from corresponding areas in data analysis and statistics in which similar problems are tackled.

Together with other references from the data analysis area one can recognize that there is a difference between the terms KDD (knowledge discovery in databases) and data mining. KDD denotes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data whereas data mining is just a step in the KDD process consisting of the application of particular data mining algorithms that,

References	Application areas	Problem descriptions	Data sets	Data Mining techniques used	Accuracy measures/ assessments	Software used
Anand et al. (1997)	Financial sector	Cross-sales problem	More than 100.000 records, more than 80 attributes	Deviation Detection Association Rules: EAR Algorithm	Support, Interest	
Borok (1997)	Health Care	Resource utilization for prospective patient populations	75.805 claims for patients	Rule Induction		Vantage Point, Inc. Software
Donato et al. (1997)	Financial sector (Credit cards)	Prediction of personal bankruptcy	9.521 accounts	Decision Trees Neural Networks	Decision Trees: Classification error rate Neural Networks: MSE	SNNS (Stuttgart Neural Network Simulator)
Evans (1997)	Printing industry	Reduction of downtime because of „ink cylinder bands"	177 records (?)	Decision Trees: ID3 (?)	Reduction of „ink cylinder bands" (1989: 538, 1995: 21)	
Fayyad et al. (1996 a)	Astronomy	Cataloging sky surveys	3.000 records, 40 attributes	Decision Trees: generalized ID3	Accuracy 94.2 %	SKICAT (Sky Image Cataloging and Analysis Tool)
Fürnkranz et al. (1997)	Politics	Prevention and termination of conflicts and wars	(1) 547 records, 70 attributes (2) 921 records, 33 attributes	K Nearest Neighbour Decision Trees: C4.5	(2) Decision Trees: Accuracy 67%	
Hätönen et al. (1996)	Telecommunication sector	Analysis of alarm sequences	Sequence of 73.679 alarms	Sequential association rules	Confidence, Support, Significance	TASA (Telec. Network Alarm Seq. Analyzer)
Hoffman et al. (1997)	Human Genome Project	Classification of DNA sequences	(1) 200 records (training), 800 records (test) (2) 6000 records (training), 41.000 records (test)	Neural Networks: MLP, Kohonen Decision Trees: C4.5	(1) Neural Networks: Accuracy 79.5%, Decision Trees: Accuracy 71.5% (2) Neural Networks: Accuracy 83%	Clementine SNNS (?) Tooldiag Visualization procedures
Matheus et al. (1996)	Health Care	Cost controlling		Deviation Detection	Interestingness of a deviation	KEFIR (Key Findings Reporter)
Mertens et al. (1997)	Pharmaceutical industry	Deviation detection for controlling issues	10.000 records	Cluster Analysis Deviation Detection		BETREX II
Sasisekharan et al. (1996)	Telecommunication sector	Description and prediction of faults in telecommunication networks	Tens of thousands of circuits	Rule Induction		
Williams, Huang (1996)	Insurance sector	Risk analysis	75.000 records, 23 attributes	Decision Trees: CART	Entropy, Gini, and Error as tree selection measures	Darwin (StarTree)

Table 1: Sample of data mining applications (1996 – 1998)

under some acceptable computational efficiency limitations, produces a particular enumeration of patterns (Fayyad et al. (1996 b)). Data warehousing is just a trendy label for all the issues raised in connection with data storage.

Process Descriptions for Data Problem Solving Tasks

Of course, given a specific process definition for KDD one can expect that different descriptions exist in the literature and Table 2 shows four possibilities which reveal a similar structure but offer different activities and numbers of steps.

	Task Analysis	Preprocessing	Data Mining	Postprocessing	Deployment
Brachman, Anand (1996)	Task Discovery	Data Discovery; Data Cleaning	Model Development	Data Analysis	Output Generation
Fayyad et al. (1996 b)	Selection	Preprocessing; Transformation	Data Mining	Interpretation/ Evaluation	
Mannila (1997)	Understanding the Domain	Preparing the Dataset	Discovering Patterns	Post-processing	Putting Results into Use
Wirth, Reinartz (1996)	Requirement Analysis	Knowledge Acquisition; Preprocessing	Pattern Extraction	Post-processing	Deployment

Table 2: Main stages of the KDD process (Gaul, Säuberlich (1998))

Brachman, Anand (1996) start with task discovery as a step, in which requirements with respect to tasks and resulting applications must be engineered. Data discovery and data cleaning activities follow before in model development and data analysis steps certain data mining techniques have to be selected and applied to the data. Finally, an output generation step is mentioned. We skip the process models of Fayyad et al. (1996 b) and Mannila (1997) and end with Wirth, Reinartz (1996) who formulate a requirement analysis step in the beginning, in which characteristics, needs and goals of the application are considered. In a knowledge acquisition step, availability and relevance of different types of knowledge are determined before preprocessing, actual pattern extraction, and postprocessing are performed. The label "deployment" for their last step stresses the point that more than just output generation is needed to turn scientific activities to successful applications.

Whenever various descriptions of an underlying phenomenon have to be taken into consideration attempts to unify the different perspectives are a must. Such a straightforward unification is already depicted in Table 2 and consists of the following five main steps: Task analysis, preprocessing, data mining, postprocessing, and deployment.

However, the idea to use process descriptions for data problem solving tasks and explain single process steps that should be performed has already some tradition. For activities where areas as AI (artificial intelligence), especially expert knowledge research, data analysis, and decision support intersect corresponding process definitions are given in e.g., Gaul et al. (1995) and Gaul, Schader (1989) and are combined in Table 3.

Depending on the points of view one would like to stress, emphasis could be laid on discussions with respect to expert systems or decision support tools as well as data analysis techniques or even research concerning mathematical modeling of underlying situations or recent trends in the visualization of information.

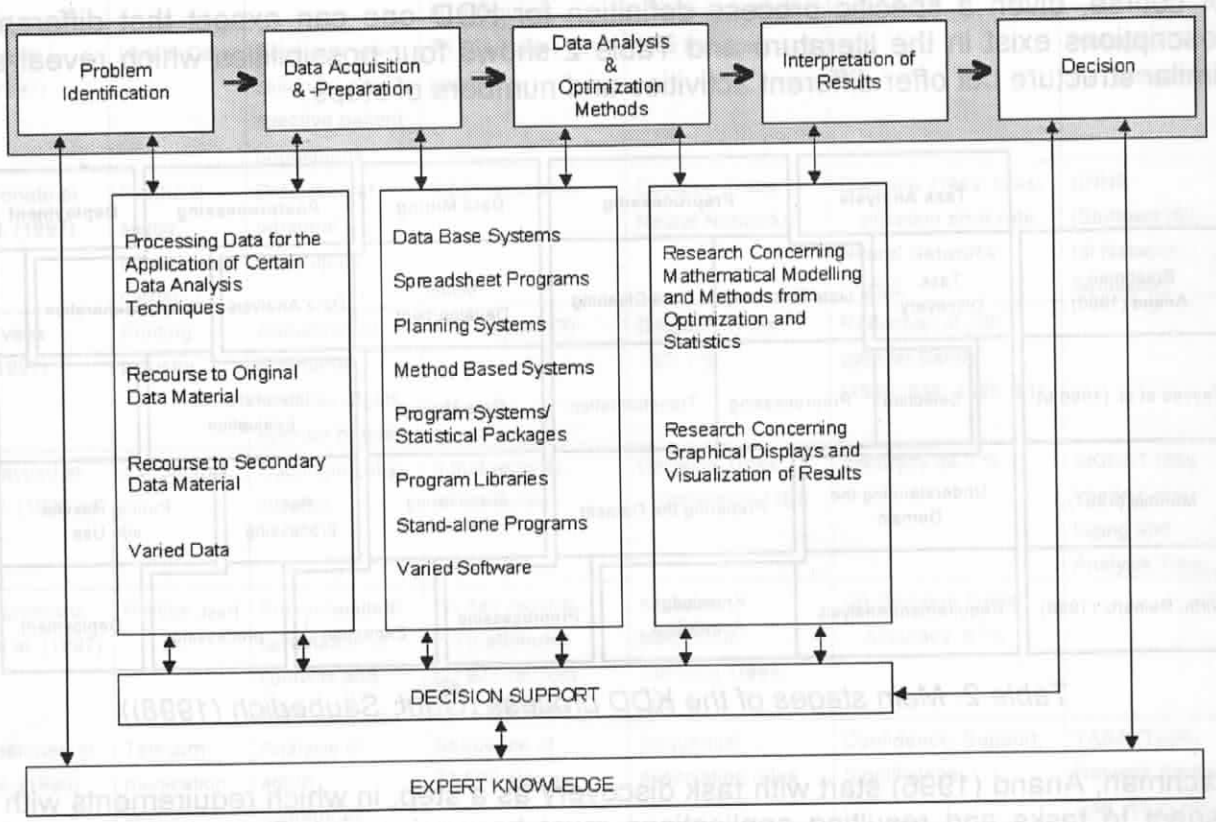

Table 3: Interdisciplinary research concerning data, expert knowledge, and decisions (Gaul, Schader (1989), Gaul et al. (1995))

Contributions concerning interdisciplinary research efforts from what could be labeled as DATA, EXPERT KNOWLEDGE, AND DECISIONS can be found in e.g. Gaul et al. (1995), Gaul, Schader (Eds.) (1988), Schader, Gaul (Eds.) (1990) or Gaul, Pfeifer (Eds.) (1996) and other issues of Studies in Classification, Data Analysis, and Knowledge Organization.

Thus, given the author's own experience with data handling activities a distinguished positioning of data mining that is well separated from already known data analysis methods needs further clarification. Of course, for large data sets there are computation time restrictions for some data analysis methods (e.g., with respect to the objects to be clustered by pyramidal clustering as described in Gaul, Schader (1994)) while for other techniques such a problem is of minor relevance (e.g., with respect to the number of consumers in the target segments for optimal product positioning as described in Baier, Gaul (1998)). Sometimes, problem sizes are restricted for demonstration purposes (as, e.g., in the MARK²MAN software of Gaul, Baier (1994)). Naturally, for "mountains" of data one would have to start with simple screening techniques. But is this reason enough to establish data mining as a new research direction?

Given the list of data mining applications of Table 1 and the background concerning interdisciplinary research efforts presented so far, KDD or data mining seems to be just an additional variant in the spectrum of data analysis possibilities.

An overview concerning commercially offered data mining software tools may help to further clarify the situation.

Comparison of Commercial Data Mining Software

In Table 4 a sample of data mining software tools is depicted. The name of the tool and the software company are mentioned together with the (data mining) techniques supported. Additional information is given concerning the platforms on which the software could be operated, price and year of the release of the first version (F.V.) as well as whether the software can be used on parallel environments (P.E.) and whether there are restrictions with respect to the size of the data sets.

Name	Company	Techniques	Platforms	Price	F. V.	P. E.	Restrictions
Clementine	Integral Solutions Ltd., GB	*Decision Trees*: ID3, C4.5 *Neural Networks*: MLP, Kohonen *Association Rules*: Apriori-Alg. *Regression*	UNIX: Sun SPARC, HP, Digital Alpha UNIX Windows NT	15.000 £	1994	No	n.a.
Darwin	Thinking Machines Corp., USA	*Decision Trees*: CART *Neural Networks*: MLP *k Nearest Neighbour*	UNIX: Solaris 2.5.1, IBM AIX 4.1.4 and others	from 30.000 US$	1996	SMP MPP	n.a.
Data Engine	MIT – Management Intelligenter Technologien GmbH, Aachen	*Decision Trees*: C4.5 (PlugIn) *Neural Networks*: MLP, Kohonen, Fuzzy Kohonen *Cluster Analysis*: Fuzzy C-Means *k Nearest Neighbour* (PlugIn) *Regression*	UNIX Windows 95, NT	Windows: 5.990 DM UNIX: 11.990 DM	1995	No	n.a.
Data Mining Tool	Syllogic, Netherlands	*Decision Trees*: C4.5 *Association Rules* *Cluster Analysis*: K-means *K Nearest Neighbour*	UNIX: Silic. Graph, IBM AIX Windows NT	UNIX/NT 30.000 US$	1996	SMP MPP	50.000 rows
Enterprise Miner	SAS Institute, USA	*Decision Trees*: CART, CHAID *Neural Networks*: MLP, RBF *Association Rules* *Cluster Analysis* *Regression*	Client: Windows 95, NT Server: Windows NT, all important UNIX platforms	from 45.000 US$ (unconfirmed)	1998	No	n.a.
Inspect	H. Lohninger, Vienna University of Technology	*Neural Networks*: RBF *Cluster Analysis*: K-means *K Nearest Neighbour* *Principal Component Analysis*	PC (DOS)	598 DM	1994	No	#variables × #rows < 8100
Intelligent Miner	IBM, USA	*Decision Trees*: Based on ID3 *Neural Networks*: MLP, RBF, Kohonen *Association Rules* *Cluster Analysis*: Propr. algorithm based on distance measure	Server: IBM AIX, OS/400, OS/390, MVS/ESA Clients: IBM AIX, Windows 95, NT, OS/2	from 42.000 US$	1995	SMP MPP	n.a.
KDD Explorer	SRA International Inc., USA	*Decision Trees*: C4.5 *Association Rules* *Cluster Analysis*: K-means	UNIX PC	from 39.500 US$	1998	SMP	n.a.
Knowl. Seeker	Angoss Softw. Corp., Canada	*Decision Trees*: CART, CHAID	UNIX Windows 95, NT	PC: 4.625 US$	1991	No	n.a.
MineSet	Silicon Graphics, USA	*Decision Trees*: C4.5 *Association Rules* *Simple Bayes Classifier*	UNIX: Silic. Graph. Challenge & Origin	from 20.000 US$	1996	No	n.a.
Neov. Decision Series	NeoVista, USA	*Neural Networks*: MLP *Association Rules* *Cluster Analysis*: Propr. algorithm based on distance measure	UNIX: HP, SUN, DEC, Oracle, Informix, Sybase	from 45.000 US$	1996	SMP	n.a.
Orchestrate	Torrent Systems Inc., USA	*Neural Networks*: MLP, RBF, Kohonen *Association Rules*: Apriori-Alg.	UNIX: Sun, IBM	from 12.500 US$	1996	SMP MPP	n.a.
Partek	Partek Inc., USA	*Neural Networks*: MLP *Cluster Analysis*: K-means *Regression* *Correspondence Analysis* *Principal Component Analysis*	UNIX: HP 900, IBM RS/6000, Silicon Graphics, Sun Microsystems	11.955 US$	1994	SMP	n.a.
Pattern Recogn. Work-	Unica Technologies, Inc., USA	*Neural Networks*: MLP, RBF *Cluster Analysis*: K-means *K Nearest Neighbour*	Windows 95, NT	from 995 US$	1993	No	n.a.

bench		*Regression*					
SIPINA	Lab. ERIC, Uni Lyon, France	*Decision Trees*: CART, Elisee, ID3, C4.5, CHAID, SIPINA	Windows 95	1.000 US$	1997	No	16.384 attr. 2^{32} - 1 rows
Xpert Rule Profiler	Attar Software, GB	*Decision Trees*: C4.5 *Association Rules* *Cluster Analysis*	Windows 95, NT	from 995 to 9.995 £	1996	No	Option 1 (995,- £): 2.000 rows

(F. V. = First Version; P. E. = Parallel Environment; Apriori-Alg. = Apriori-Algorithm; Propr. = Proprietary; MPP = massively parallel processing; SMP = symmetric multi processing; n.a = no answer)

Table 4: Sample of data mining software tools (Gaul, Säuberlich (1998))

The techniques supported most often are (ranked according to importance) decision trees, neural networks, cluster analysis, and association rules. Bayes classifier, correspondence analysis, k-nearest neighbour, principal component analysis and regression (in alphabetical order) have also been mentioned.

Information as provided by Table 4 can be aggregated in such a way that appropriate (dis)similarities between the listed data mining tools can be used as starting point for the application of positioning and segmentation techniques to try to dig out structures and discover knowledge that is hidden in the data.

For a subset of 12 data mining tools (selection criterion was comprehensiveness of the information provided) characteristics concerning the three main steps of the KDD process description of Table 2, i.e., preprocessing, data mining and postprocessing, as well as additional features concerning visual programming, parallel environments and platforms were combined to get the display shown in Figure 1 (see Gaul, Baier (1994) for details with respect to the application of standard positioning and segmentation procedures).

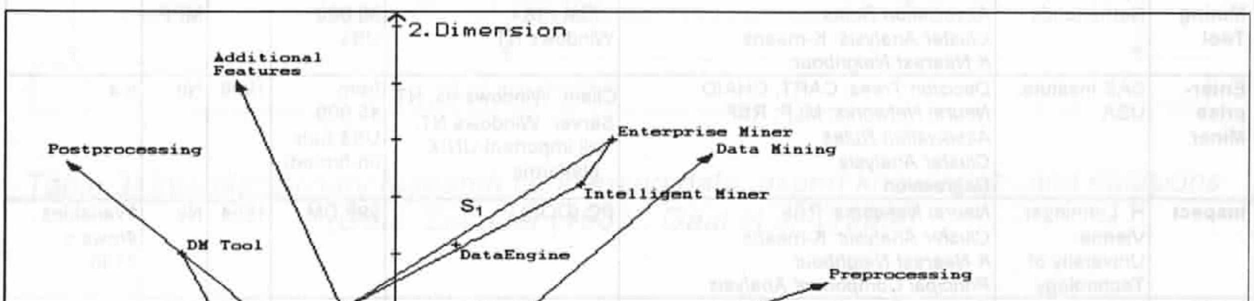

Figure 1: Positioning of data mining tools (Gaul, Säuberlich (1998))
(S_k, k=1,...,4, is the abbreviation for segment k)

Interpretations based on these straightforward clustering and scaling results could be the following:

Tools as KnowledgeSeeker and SIPINA which provide just decision trees as data mining techniques and don't offer comparable many pre- and postprocessing capabilities as other software suppliers are separated from the rest of the subsample.

Darwin and Data Mining Tool are examples where postprocessing characteristics and additional features influence their positioning in the map of Figure 1.

Clementine, Data Engine, Intelligent Miner and Enterprise Miner build a segment that more than others tries to support all main stages of the KDD process and offers more different data mining techniques than most of the competitors from the other segments under consideration.

The visualization presented with the help of clustering and positioning methods shows a trend from single-technique software products as KnowledgeSeeker to multi-task tools that try to support as many as possible main stages of the KDD process.

However, some of the data mining techniques mentioned in Table 4 have been used in the data analysis area for quite some time (see, e.g. Gaul et al. (1994) for the application of neural nets to panel and POS-scanner data) while other new research directions are not considered (see, e.g., m-mode n-way data handling as described for two-mode clustering in Gaul, Schader (1996)). And even with respect to association rules – a methodology that is closer related to the data mining discussion than others – a user has to be familiar with the basics when a non-standard situation has to be tackled as will be shown in the next section.

Analysis of Buying Histories by a Modification of Association Rules

Notation

In standard applications of association rules subsets *X* and *Y* of an interesting set are checked with the help of certain measures, e.g., confidence and support, whether rules that "associate *X* with *Y*" are of importance.

The task of an association rule algorithm is to find all association rules which fulfil prescribed bounds for support and confidence values. Since the number of sets which satisfy given bounds can be very large, corresponding algorithms use special techniques to reduce the search space. Association rule algorithms are a class of data mining techniques which can cope with large data sets in a reasonable running time. An example of such an association rule algorithm is the Apriori Algorithm by Agrawal et al. (1996). The following modifications for the analysis of brand switching behaviour are used:

For a given set of brands $B=\{p,q,...\}$ let $T=(t_1 \to t_2 \to ... \to t_j \to ...)$ denote an individual buying history, i.e., a sequence of subsequently bought brands $t_j \in B$, and $ind_T(t_j)$ [=j] the index of t_j in *T*. The symbols $\vec{\subset}$ resp. $\vec{\cup}$ are used to denote a subhistory as connected part of a history resp. a composition of (sub-)histories. For $X \vec{\subset} T$ the first resp. last brand of *X* is described by $b(X)$ (beginning of *X*) resp. $e(X)$ (end of *X*) and $l(X)$ (length of *X*) counts the pairs of subsequently bought brands. Some obvious properties are:

$$X,Y \vec{\subset} T \text{ with } ind_T(e(X)) = ind_T(b(Y)) \Rightarrow X \vec{\cup} Y \vec{\subset} T$$

$$X \vec{\cup} Y \vec{\subset} T \Rightarrow ind_T(b(X \vec{\cup} Y)) = ind_T(b(X)),\ ind_T(e(X \vec{\cup} Y)) = ind_T(e(Y))$$

$$l(X \vec{\cup} Y) = l(X) + l(Y)$$

Additionally, for $X \vec{\subset} T$, let $m(X,T,l)$ be the number of times that X appears as subhistory of $Z \vec{\subset} T$ with $ind_T(b(Z)) = ind_T(b(T))$ [=1] and $l(T) - l(Z) = l$.

Up to now the buying history of just one individual was used. Now, assume that I is a (large) set of individuals. Then

$$\vec{s}_l(X) := \sum\nolimits_{i \in I} m(X, T_i, l)$$

counts the occurrence of X in the set

$$D_l := \{Z_i \mid Z_i \vec{\subset} T_i, l(T_i) - l(Z_i) = l, ind_{T_i}(b(Z_i)) = ind_{T_i}(b(T_i)), i \in I\}$$

where

$$D_0 := \{T_i \mid i \in I\}$$

is a given set of individual buying histories. The value $\vec{s}_l(X)$ is called *l-generalized support* of X and

$$\vec{c}(X,Y) := \frac{\vec{s}_0(X \vec{\cup} Y)}{\vec{s}_{l(Y)}(X)}$$

which gives the percentage of individuals of I that have switched from X to Y, *generalized confidence* of X and Y. This notation contains normal conditional switching (e.g., Carpenter, Lehmann (1985)) from a brand p to a brand q as special case in the following way: Set $X = (p)$ (with $l(X) = 0$) and $Y = (p \to q)$ (with $l(Y) = 1$), then

$$\vec{c}((p),(p \to q)) = \frac{\text{number of occurrences of } (p \to q) \text{ in } D_0}{\text{number of occurrences of } (p) \text{ in } D_1}$$

describes the entries of the well known conditional switching matrix.

Empirical Example

Consider an empirical example where the switching behaviour of 1254 households with respect to a product category of 7 brands was recorded for a certain time period. The conditional switching matrix as depicted in Table 5 can be computed by "traditional counting" but if one is interested in what can be called "higher order associations" the number of compositions of subhistories is rapidly increasing.

from brand \ to brand	A	B	C	D	E	F	G
A	0,72784	0,05282	0,02596	0,02417	0,02865	0,02507	0,11549
B	0,05244	0,53165	0,07776	0,06148	0,09132	0,03617	0,14919
C	0,04192	0,14770	0,43114	0,08583	0,08982	0,04790	0,15569
D	0,04560	0,12541	0,07329	0,43811	0,09609	0,07492	0,14658
E	0,03625	0,12875	0,06250	0,08500	0,52375	0,03250	0,13125
F	0,05523	0,10848	0,05128	0,06706	0,05720	0,42998	0,23077
G	0,07503	0,09672	0,05041	0,05862	0,05920	0,06155	0,59848

Table 5: Conditional switching matrix (Gaul, Säuberlich (1998))

Using the just explained methodology "modified" association rules can be formulated with the help of subhistories X, Y, $\vec{s}_0(X \vec{\cup} Y)$, and $\vec{c}(X,Y)$ to get deeper insights into the buying behavior of individuals based on a sample of buying histories D_0. Table 6 shows selected results that enrich the information obtainable by traditional conditional switching considerations, e.g., the first column of Table 6 coincides with the first row of Table 5.

Rule (X, Y)	$\vec{c}(X,Y)$	$\vec{s}_0(X \vec{\cup} Y)$	Rule (X, Y)	$\vec{c}(X,Y)$	$\vec{s}_0(X \vec{\cup} Y)$	Rule (X, Y)	$\vec{c}(X,Y)$	$\vec{s}_0(X \vec{\cup} Y)$
(A), (A→A)	0,72784	813	(A→A), (A→A→A)	0,71215	381	(B), (B→E→B)	0,03812	34
(A), (A→B)	0,05282	59	(A→A→A), (A→A)	0,86788	381	(B→E), (E→B)	0,41975	34
(A), (A→C)	0,02596	29	(E), (E→E→E)	0,36926	233	(B), (B→E→E)	0,03027	27
(A), (A→D)	0,02417	27	(E→E), (E→E)	0,70606	233	(B→E), (E→E)	0,33333	27
(A), (A→E)	0,02865	32	(B→B), (B→B→B)	0,59726	218	(B), (B→G→B→B)	0,01685	12
(A), (A→F)	0,02507	28	(B→B→B), (B→B)	0,83846	218	(B), (B→E→G)	0,00897	8
(A), (A→G)	0,11549	129	(D→D), (D→D)	0,62326	134	(B→E), (E→G)	0,09877	8

Table 6: Part of the results of the modified Apriori algorithm (Gaul, Säuberlich (1998))

Conclusion

Three points of view – the application side, the side of commercial software suppliers, and the research side – were taken to approach what nowadays is labeled as data mining. Using samples from the literature concerning data mining applications and from the software market concerning data mining tools as well as a modification of association rules it was tried to find out whether data mining has emerged as a new discipline from the data analysis area that deserves special attention.

In research – as in other areas – the cycle

EXPECTATIONS → ENTHUSIASM → DISILLUSIONMENT → EXPECTATIONS

can be observed (see, e.g., the treatment of expert systems in theory and practice). The paper tries to help the reader to find out where the recent discussion concerning data mining has to be positioned.

References:

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I. (1996): Fast Discovery of Association Rules, in: *Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 277-296.

Anand, S.S., Hughes, J.G., Bell, D.A., Patrick, A.R. (1997): Tackling the Cross-Sales Problem using Data Mining, *Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining*, 1997.

Baier, D., Gaul, W. (1998): Optimal Product Positioning Based on Paired Comparisons Data, to appear in *Journal of Econometrics.*

Borok, L.S. (1997): Data Mining: Sophisticated Forms of Managed Care Modeling through Artificial Intelligence, *Journal of Health Care Finance*, Vol. 23, Iss. 3, 20-36.

Brachman, R.J., Anand, T. (1996): The Process of Knowledge Discovery in Databases: A Human-Centered Approach, in: *Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 33-51.

Carpenter, G.S., Lehmann, D.R. (1985): A Model of Marketing Mix, Brand Switching, and Competition, *Journal of Marketing Research*, Vol. 22, 318-329.

Donato, J.M., Schryver, J.C., Grady, N.W., Hinkel, G.C., Schmoyer, R.L., Leuze, M.R. (1997): Mining Credit Card Data for Decision Support, *AFCEA, First Federal Data Mining Symposium*, 1997.

Evans, R.B. (1997): A Case Study in Data Mining, *Database Programming & Design*, April 1997, 42-49.

Fayyad, U.M., Djorgovski, S.G., Weir, N. (1996 a): Automating the Analysis and Cataloging of Sky Surveys, in: *Fayyad U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 425-444.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., (1996 b): From Data Mining to Knowledge Discovery: An Overview, in: *Fayyad U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1-29.

Fürnkranz, J. Petrak, J., Trappl, R. (1997): Knowledge Discovery in International Conflict Databases, *Applied Artificial Intelligence*, Vol. 11, 91-118.

Gaul, W., Baier, D. (1994): Marktforschung und Marketing Management, Oldenbourg, München, 2nd Edition.

Gaul, W., Decker, R., Wartenberg, F. (1994): Analyse von Panel- und POS-Scanner-Daten mit Neuronalen Netzen, *Jahrbuch der Absatz- und Verbrauchsforschung*, 40. Jahrgang, Heft 3, 281 – 306.

Gaul, W., Pfeifer, D. (Eds.) (1996): From Data to Knowledge, *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer.

Gaul, W., Radermacher, F.J., Schader, M., Solte, D. (1995): Data, Expert Knowledge and Decisions: An Introduction to the Volume, Special Issue (Edited by Gaul, W., Radermacher, F.J., Schader, M., Solte, D.), *Annals of Operations Research*, Vol. 55, 1-7.

Gaul, W., Säuberlich, F. (1998): Classification and Positioning of Data Mining Tools, to appear in *Studies in Classification, Data Analysis, and Knowledge Organization*.

Gaul, W., Schader, M. (Eds.) (1988): Data, Expert Knowledge and Decisions, Springer.

Gaul, W., Schader, M. (1989): Data Analysis and Decision Support, *Applied Stochastic Models and Data Analysis*, Vol. 5, 341 – 356.

Gaul, W., Schader, M. (1994): Pyramidal Classification Based on Incomplete Dissimilarity Data, *Journal of Classification* , 11, 171 - 193

Gaul, W., Schader, M. (1996): A New Algorithm for Two-Mode Clustering, in: *Bock, H.-H., Polasek, W., Data Analysis and Information Systems, Studies in Classification, Data Analysis, and Knowledge Organization*, 15 – 23.

Hätönen, K., Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. (1996): Knowledge Discovery from Telecommunication Network Alarm Databases, *Proceedings of the 12th International Conference on Data Engineering (ICDE'96)*, New Orleans, Louisiana, February/March 1996.

Hoffman, P., Grinstein, G., Marx, K., Grosse, I., Stanley, E. (1997): DNA Visual and Analytic Data Mining, *Visualization* 1997.

Mannila, H. (1997): Methods and Problems in Data Mining, in: *Afrati, F., Kolaitis, P. (Eds.), Proceedings of the International Conference on Database Theory*. Delphi, Greece, January 1997, Springer-Verlag.

Matheus, C.J., Piatetsky-Shapiro, G., McNeill, D. (1996): Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data, in: *Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthursamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 445-463.

Mertens, P., Bissantz, N., Hagedorn, J. (1997): Data Mining im Controlling, *Zeitschrift für Betriebswirtschaft*, Vol. 67, Iss. 2, 179-201.

Sasisekharan, R., Seshadri, V., Weiss, S.M. (1996): Data Mining and Forecasting in Large-Scale Telecommunication Networks, *IEEE Expert - Intelligent Systems and their Applications*, Vol. 11, Iss. 1, 37-43.

Schader, M., Gaul, W. (Eds.) (1990): Knowledge, Data and Computer-Assisted Decisions, NATO ASI Series F: Computer and Systems Sciences, Vol. 61, Springer.

Williams, G.J., Huang, Z. (1996): A Case Study in Knowledge Acquisition for Insurance Risk Assessment using a KDD Methodology, *Pacific Rim Knowledge Acquisition Workshop*, Sydney, Australia, October 1996.

Wirth, R., Reinartz, T.P. (1996): Towards a Task Model for KDD-Processes, in: *Kodratoff, Y., Nakhaeizadeh, G. (Eds.), Workshop Notes; Statistics, Machine Learning, and Knowledge Discovery in Databases*. MLNet Familiarisation Workshop.