# Classification and Positioning of Data Mining Tools

W. Gaul, F. Säuberlich

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe, D-76128 Karlsruhe, Germany

**Abstract:** Various models for the KDD (Knowledge Discovery in Databases) process are known, which mainly differ with respect to the number and description of process activities. We present a process unification by assigning the single steps of these models to five main stages and concentrate on data mining aspects. An overview concerning data mining software tools with focus on inbuilt algorithms and additional support provided for the main stages of the KDD process is given within a classification and positioning framework. Finally, an application of a modification of an association rule algorithm is used as empirical example to demonstrate what can be expected when data mining tools are used to handle large data sets.

## 1 Knowledge Discovery in Databases and Data Mining

Because of advances in data storage technology and the explosive growth with respect to the capabilities to generate and collect data, companies are nowadays more and more aware of their "data mining" and "data warehousing" problems (to use these trendy labels for situations well known to the data analysis community for a long time). The possibility to raise enormous quantities of data (from which in practice often only a relatively small amount is actually used) has led to demands for "easy" data handling, outpaced the traditional abilities to interpret and digest such data and created a need for tools and techniques of automated and intelligent database analysis. The area of Knowledge Discovery in Databases (KDD for short) tries to cope with these problems.

In practice the terms KDD and data mining are often used synonymously while in research a differentiation of these two terms is usual. KDD denotes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al. (1996)). Thus, the term KDD describes a whole process of identifying patterns in data, whereas data mining is just a step in the KDD process consisting of the application of particular data mining algorithms that, under some acceptable computational efficiency limitations, produces a particular enumeration of patterns (Fayyad et al. (1996)).

In 1989 the term data mining was not mentioned among the database management systems research topics for which it was postulated that they would

deserve research attention in the future, whereas in 1993 data mining together with other topics already occupied the second position (Stonebreaker et al. (1993)).

As in practice software offers in the KDD area are often called data mining tools, we will also use this term, although most of the facilities in these tools try to support several main stages within the KDD process.

## 2 KDD Process Models

Table 1 shows four KDD process models, which reveal a similar structure but offer different activities within their process representations.

| | 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) | Selection | Pre-processing | Transfor-mation | 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 | Pre-processing | Pattern Extraction | Post-processing | Deployment | |

Table 1: Main stages of the KDD process

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. Because of page restrictions the description of the process models of Fayyad et al. (1996) and Mannila (1997) is restricted to the terms used for labelling the single process steps in Table 1. In the last row of this table Wirth, Reinartz (1996) are mentioned who formulate a requirement analysis step in the beginning, in which characteristics, needs and goals of the application are considered, and continue with a knowledge acquisition step, in which 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.

Of course, whenever various descriptions of an underlying phenomenon are available one can look for structural similarities. We propose a unification of the just mentioned different process representations in terms of the following five main stages also shown in Table 1: Task analysis, preprocessing, data mining, postprocessing, and deployment. In the next section, we concentrate on data mining aspects and examine data mining software tools which—despite of the generic term—try to support at least the following KDD process main stages: data mining, pre- and postprocessing.

# 3 Data Mining Software Tools

Table 2 gives an overview of 16 data mining software tools where we restrict ourselves, apart from SIPINA and KnowledgeSeeker, to so-called multi-task tools which support different data mining tasks and techniques. Besides name of the tool, company and (data mining) techniques supported, the platforms on which the software could be operated, price and year of the release of the first version are mentioned. The last two columns show, whether the software can be used on parallel environments and whether there are certain restrictions in the size of the data sets, which can be analysed. The survey on which the information contained in Table 2 (and in Table 3 as well as in Figure 1) is based was conducted up to April 1998.

The techniques supported most often are (ranked according to importance) decision trees, neural networks, cluster analysis, association rules, k nearest neighbour, and regression.

In Table 3, a subset of 12 of these 16 tools has been examined (selection criterion was availability of information) with respect to features concerning the three main steps—preprocessing, data mining and postprocessing—within the KDD process as well as with respect to additional features like visual programming, parallel environment, and platform. For the remaining tools shown in Table 2 some parts of the information mentioned in Table 3 were missing.

We aggregated the characteristics of Table 3 in a suitable manner and used multidimensional scaling (Kruskal MDS and Principal Component Analysis together with Property Fitting) and cluster analysis (Single, Complete, and Average Linkage as well as McQuitty and Ward) to obtain a solution of four segments as shown in Figure 1 (see Gaul, Baier (1994) for details with respect to the application of standard positioning and segmentation procedures). Of course, clustering would not have been necessary to reveal the structure depicted in Figure 1 but it helped to get a feeling for the task of grouping "similar" software tools. Finally, it should be mentioned that the Kruskal stress for the solution depicted in Figure 1 was very good (stress=0.003) and that the best CCC-value was obtained for Average Linkage and McQuitty (CCC=0.821).

| Name | Company | Techniques | Platform | Price | F. V. | P. E. | Restriction |
|---|---|---|---|---|---|---|---|
| **Clementine** | Integral Solutions Ltd., GB | *Decision Trees*: ID3, C4.5<br>*Neural Networks*: MLP, Kohonen<br>*Association Rules*: Apriori-Alg.<br>*Regression* | Unix: Sun SPARC, HP, Digital Alpha Unix<br>Windows NT | 15.000 £ | 1994 | No | n.a. |
| **Darwin** | Thinking Machines Corp., USA | *Decision Trees*: CART<br>*Neural Networks*: MLP<br>*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)<br>*Neural Networks*: MLP, Kohonen, Fuzzy Kohonen<br>*Cluster Analysis*: Fuzzy C-Means<br>*k Nearest Neighbour* (PlugIn)<br>*Regression* | Unix<br>Windows 95, NT | Windows: 5.990 DM<br>Unix: 11.990 DM | 1995 | No | n.a. |
| **Data Mining Tool** | Syllogic, Netherlands | *Decision Trees*: C4.5<br>*Association Rules*<br>*Cluster Analysis*: K-means<br>*K Nearest Neighbour* | Unix: Silic. Graph, IBM AIX<br>Windows NT | Unix/NT 30.000 US$ | 1996 | SMP, MPP | 50.000 rows |
| **Enterprise Miner** | SAS Institute, USA | *Decision Trees*: CART, ChAID<br>*Neural Networks*: MLP, RBF<br>*Association Rules*<br>*Cluster Analysis*<br>*Regression* | Server: Windows NT, all important Unix platforms<br>Clients: Windows 95, NT | from 45.000 US$ (unconfirmed) | 1998 | No | n.a. |
| **Inspect** | H. Lohninger, Vienna University of Technology | *Neural Networks*: RBF<br>*Cluster Analysis*: K-means<br>*K Nearest Neighbour*<br>*Principal Component Analysis* | PC (DOS) | 598 DM | 1994 | No | #variables × #rows < 8100 |
| **Intelligent Miner** | IBM, USA | *Decision Trees*: Based on ID3<br>*Neural Networks*: MLP, RBF, Kohonen<br>*Association Rules*<br>*Cluster Analysis*: Propr. algorithm based on distance measure | Server: IBM AIX, OS/400,OS/390, MVS/ESA<br>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<br>*Association Rules*<br>*Cluster Analysis*: K-means | Unix<br>PC | from 39.500 US$ | 1998 | SMP | n.a. |
| **Knowledge Seeker** | Angoss Software Corp., Canada | *Decision Trees*: CART, ChAID | Unix<br>Windows 95, NT | PC: 4.625 US$ | 1991 | No | n.a. |
| **MineSet** | Silicon Graphics, USA | *Decision Trees*: C4.5<br>*Association Rules*<br>*Simple Bayes Classifier* | Unix: Silic. Graph. Challenge & Origin | from 20.000 US$ | 1996 | No | n.a. |
| **Neovista Decision Series** | NeoVista, USA | *Neural Networks*: MLP<br>*Association Rules*<br>*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<br>*Association Rules*: Apriori-Alg. | Unix: Sun, IBM | from 12.500 US$ | 1996 | SMP, MPP | n.a. |
| **Partek** | Partek Inc., USA | *Neural Networks*: MLP<br>*Cluster Analysis*: K-means<br>*Regression*<br>*Correspondence Analysis*<br>*Principal Component Analysis* | Unix: HP 900, IBM RS/6000, Silicon Graphics, Sun Microsystems | 11.955 US$ | 1994 | SMP | n.a. |
| **Pattern Recogn. Workbench** | Unica Technologies, Inc., USA | *Neural Networks*: MLP, RBF<br>*Cluster Analysis*: K-means<br>*K Nearest Neighbour*<br>*Regression* | Windows 95, NT | from 995 US$ | 1993 | No | n.a. |
| **SIPINA** | Lab. E.R.I.C., Univ. Lyon, France | *Decision Trees*: CART, Elisee, ID3, C4.5, ChAID, SIPINA | Windows 95 | 1.000 US$ | 1997 | No | 16.384 attributes 2^32 - 1 rows |
| **XpertRule Profiler** | Attar Software, GB | *Decision Trees*: C4.5<br>*Association Rules*<br>*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; MLP = Multilayer Perceptron; RBF = Radial Basis Functions) Evaluation Date: April 1998

Table 2: Data mining software tools

Tools as KnowledgeSeeker and SIPINA, which provide only one data mining technique (decision trees) and don't offer much pre- and postprocessing capabilities are clearly separated from the other tools.

Darwin and Data Mining Tool are examples of tools, which combine some data mining features with more postprocessing capabilities.

| | | Clementine | Darwin | Data Engine | DM Tool | Enterpr. Miner | Inspect | Intell. Miner | Knowl. Seeker | Neov. DS | Partek | PR Workbench | SIPINA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Preprocessing | Missing Values | | | | | | | | | | | | |
| | Replace by function | ✓ | ✓ | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | |
| | Estimate | ✓ | | ✓ | | ✓ | | ✓ | | | | | |
| | Transformation | | | | | | | | | | | | |
| | Standardization | | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | ✓ | |
| | Scale Transform. | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| | Generate Attributes | ✓ | | ✓ | ✓ | ✓ | | ✓ | | | | ✓ | |
| | Autom. Feature Sel. | | | | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | |
| Data Mining | Decision Trees | | | | | | | | | | | | |
| | C4.5 | ✓ | | ✓ | ✓ | | | ✓ | | | | | ✓ |
| | CART | | ✓ | | | ✓ | | | ✓ | | | | ✓ |
| | ChAID | | | | | ✓ | | | ✓ | | | | ✓ |
| | Neural Networks | | | | | | | | | | | | |
| | MLP | ✓ | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | ✓ | |
| | RBF | | | | | ✓ | ✓ | ✓ | | | | ✓ | |
| | Kohonen | ✓ | | ✓ | | | | ✓ | | | | | |
| | Association Rules | ✓ | | | ✓ | ✓ | | ✓ | | ✓ | | | |
| | Cluster Analysis | | | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | |
| | K Nearest Neighbour | | ✓ | ✓ | ✓ | | ✓ | | | | | ✓ | |
| | Regression | ✓ | | ✓ | | ✓ | ✓ | | | | | ✓ | |
| Postprocessing | Export of Results | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ |
| | Visual. of Results | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | | ✓ |
| Additional Features | Visual Programming | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | |
| | Parallel Environm. | | ✓ | | ✓ | | | ✓ | | ✓ | ✓ | | |
| | Platform | | | | | | | | | | | | |
| | Unix | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | | |
| | Windows | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ |

(Scale Transform. = Scale Transformation; Autom. Feature Sel. = Automatic Feature Selection; Visual. of Results = Visualization of Results; Parallel Environm. = Parallel Environment) Evaluation Date: April 1998

Table 3: Features of selected data mining tools

Neovista Decision Series, Pattern Recognition Workbench, Inspect, and Partek provide a medium number of different data mining techniques and preprocessing capabilities.

Clementine, DataEngine, Intelligent Miner, and Enterprise Miner build the segment, which more than the others tries to support all the main stages of the KDD process and offers more different data mining techniques than the competitors under consideration.

The visualization presented shows the positioning of data mining software tools in a competitive environment and confirms an expected development: from single-technique software products as KnowledgeSeeker to multi-task

tools, which try to support the whole KDD process in an integrated environment. A further analysis could incorporate considerations with respect to application areas or prices.

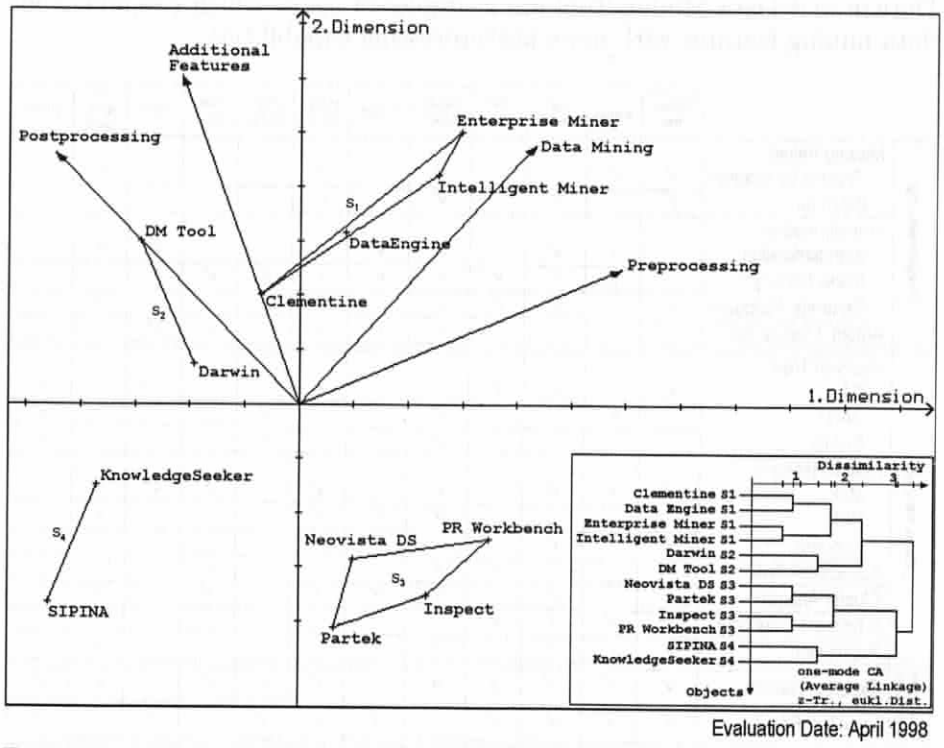


Figure 1: Clustering and positioning of data mining tools ($S_k$, $k = 1, ..., 4$, is the abbreviation for segment $k$)

# 4 Application: Generalized Brand Switching Analysis via a Modification of Association Rules

Within the standards for data mining it is often mentioned that corresponding techniques should be able to handle huge "mountains" of so-called itemsets. In this regard approaches that use association rules are of interest.

Association rule algorithms build a class of techniques used in more than half of the software tools shown in Table 2. Basic association rule algorithms are capable to handle a special form of data and—therefore—can only be used for a specific class of problems, e.g. market basket analysis. In the following we give a basic description for situations in which association rules can be

applied and formulate modifications needed for the analysis of consumer behaviour time series.

A pair $(X, Y)$ of itemsets, e.g., subsets of brands of a product category, can be viewed as starting point for building association rules in a given data base $D$ of itemsets. An association rule uses bounds for support $s(X \cup Y)$ and confidence $c(X, Y)$ measures of $X$ and $Y$ to check whether the "association of $X$ and $Y$" is meaningful (where the support $s(X \cup Y)$ gives the percentage of itemsets in $D$ which contain the itemset $X \cup Y$ and the confidence $c(X, Y)$ describes the fact that $c(X, Y)$ percent of the itemsets in $D$ that contain $X$ also contain $Y$). 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 itemsets 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). We have used the following modifications for the analysis of brand switching behaviour:

For modelling buying histories with respect to a given set of brands $\mathcal{B} = \{p, q, ...\}$ let $T = (t_1 \to ... \to t_j \to ... \to t_{n(T)})$ denote an indexed individual buying history, i.e., a sequence of subsequently bought brands $t_j \in \mathcal{B}$ where $n(T)$ counts the number of purchases described by $T$. Note that the same brand can be bought at different purchase occasions.

We call a subhistory $X$ of $T$ a *connected subhistory* if there exists an index $j(X) \in \mathbb{N}_0$ so that $X$ can be written in the form $X = (t_{j(X)+1} \to t_{j(X)+2} \to ... \to t_{j(X)+n(X)})$ with $j(X) + n(X) \leq n(T)$ and use the symbol $\vec{\subset}$ to denote such a connected subhistory. For $X \vec{\subset} T$ the first and last brand of $X$ is described by $b(X)$ (beginning of $X$) and $e(X)$ (end of $X$), respectively, and $l(X)$ $[:= n(X) - 1]$ (length of $X$) counts the pairs of subsequently bought brands.

If for connected subhistories $X, Y \vec{\subset} T$ there exist $j(X), j(Y) \in \mathbb{N}_0$ such that $j(X) \leq j(Y)$ and $j(X) + n(X) = j(Y) + k$ with $k = k(X, Y) \in \{1, ..., n(Y)\}$ we use the symbol $X \vec{\cup}_k Y = (t_{j(X)+1} \to ... \to t_{j(X)+n(X)} \to t_{j(Y)+k+1} \to ... \to t_{j(Y)+n(Y)})$ to denote the so-called *k-overlapping composition* of $X$ and $Y$. Note that $X \vec{\cup}_k Y \vec{\subset} T$.

Some obvious properties are:

$$(X \vec{\cup}_k Y) \vec{\subset} T \Rightarrow b(X \vec{\cup}_k Y) = b(X),\ e(X \vec{\cup}_k Y) = e(Y),$$
$$l(X \vec{\cup}_k Y) = l(X) + l(Y) - k + 1.$$

Additionally, for $X \vec{\subset} T$, let $m(X, T, l)$ be the number of times that $X$ appears as connected subhistory of $Z \vec{\subset} T$ with $j(Z) = 0$ 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

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

counts the occurrence of $X$ in the set

$$D_l := \{Z_i \mid Z_i \vec{\subset} T_i, l(T_i) - l(Z_i) = l, j(Z_i) = 0, i \in I\}$$

where $D_0 := \{T_i \mid i \in I\}$ is a given set of individual buying histories that corresponds with the given data base $D$ of itemsets mentioned in the general description in the beginning. The value $\vec{s}_l(X)$ is called *l-generalized support* of $X$. For $X, Y \vec{\subset} T$ with $k(X, Y) = 1$

$$\vec{c}(X, Y) := \frac{\vec{s}_0(X \vec{\cup}_1 Y)}{\vec{s}_{l(Y)}(X)},$$

can be labeled as *generalized confidence* of $X$ and $Y$ and gives the percentage of individuals of $I$ that have switched from $X$ to $Y$ (Note, that $k > 1$ is needed for a generalized version of the Apriori Algorithm.). This notation contains normal conditional switching (see, 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

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

describes the entries of the well known conditional switching matrix.

Consider an empirical example where the switching behaviour of 1254 households with respect to a product category of 7 brands {A,B,C,D,E,F,G} was recorded for a certain time period. The conditional switching matrix as depicted in Table 4 can be computed by "traditional counting" but if one is interested in what can be called "higher order associations" then 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 4: Conditional switching matrix

Using the just explained methodology "modified" association rules can be formulated with the help of subhistories $X$, $Y$, $\vec{s}_0(X \vec{\cup}_1 Y)$, and $c(X, Y)$ to get deeper insights into the buying behavior of individuals based on a sample of buying histories $D_0$. Table 5 shows selected results that enrich the information obtainable by traditional conditional switching considerations, e.g., the first column of Table 5 coincides with the first row of Table 4.

| Rule (X, Y) | $\bar{c}(X, Y)$ | $\bar{s}_0(X\dot{\cup}, Y)$ | Rule (X, Y) | $\bar{c}(X, Y)$ | $\bar{s}_0(X\dot{\cup}, Y)$ | Rule (X, Y) | $\bar{c}(X, Y)$ | $\bar{s}_0(X\dot{\cup}, Y)$ |
|---|---|---|---|---|---|---|---|---|
| (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 5: Part of the results of the modified Apriori algorithm

# 5 Conclusion

In the last years, numerous so-called data mining software tools were introduced into the market. Within the tools for which we got information, decision trees, neural networks, cluster analysis, and association rule algorithms belong to the data mining techniques supported most often. The development in this area is directed to multi-task tools which provide different techniques and solve tasks from nearly all main stages of the entire KDD process. But nevertheless—and this is a contradiction to statements of some data mining software vendors—the user has to be familiar with most of the methods and techniques in order to be able to solve his problems and to interpret the results obtained. We selected the analysis of buying histories by association rules to stress this point and to show that modifications of standard descriptions and algorithms could be necessary to solve specific analysis tasks.

## References

AGRAWAL, R., MANNILA, H., SRIKANT, R., TOIVONEN, H. and VERKAMO, A.I. (1996): Fast Discovery of Association Rules. In: *Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.) (1996): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 277-296.*

BRACHMAN, R.J. and 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.) (1996): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 33-51.*

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

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