Use of genetic algorithms for a user specific reduction of amounts of interesting association rules

Birgit Wenke

Fakultät für Informatik
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, D-85577 Neubiberg, Germany
email: birgit.wenke@unibw-muenchen.de

Abstract: The huge amounts of stored digital data, which are nowadays available in many domains contain lots of previously unknown coherences. For the user it is often difficult or unfeasible to find those coherences he is interested in without any technical support. One kind of these coherences are association rules.
This paper presents an iterative procedure which supports the user in finding these association rules he is interested in, by considering his interests without the explicit formulation of these interests by the user in advance. The procedure presents iteratively association rules to the user, who has to value each of them as interesting or uninteresting. With the help of a genetic algorithm the procedure learns interactively the interests of the user and formulates classification rules, which are used to classify the not yet presented association rules in the classes interesting and uninteresting so that only interesting classified association rules are presented to the user in the following.
The procedure was evaluated on a standard dataset and a dataset of the web2.0 application flick. The evaluation results show, that the developed procedure is useful for both standard database applications and innovative web2.0 applications. Different genetic methods and scenarios of interests were evaluated. The most interesting evaluation results will be presented in this paper.

1. Introduction

The huge amounts of digital data, which are available in many different domains due to the growing use of the internet and other digital applications contain lots of previously unknown coherences. For the user it is often difficult or unfeasible to find these coherences he is interested in without any technical support.

For example 2.5 mil. people suffer worldwide from multiple sclerosis. The cause and course of this illness are as far as possible unexplored, although a huge amount of data is available. This shows the need for new methods to find previously unknown coherences, which make it possible to formulate new research projects. This paper concentrates on

association rules which is one of several different kinds of patterns in data, generally known as data mining methods.

In this paper an iterative procedure will be presented, which supports the user in finding those associations rules of a dataset, he is interested in. The procedure consists of three main phases (Figure 1).

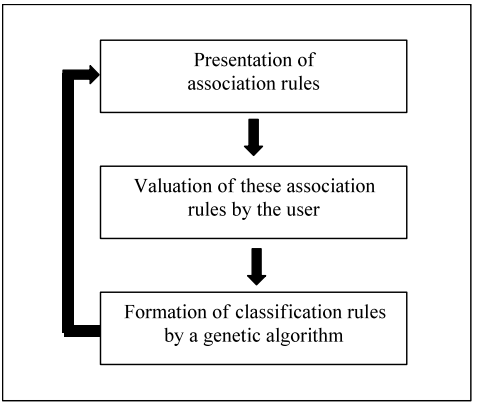

Figure 1: main phases of the developed procedure

The first phase is the presentation of about five association rules to the user. In the next phase the user has to valuate each of the presented rules if it is interesting or not for him. A genetic algorithms uses this subjective information of the user to formulate classification rules which classify the not yet presented association rules is the classes interesting or uninteresting. Starting from the second run only association rules will be presented to the user, which were classified as interesting in the previous run.

Main characteristic of this procedure is the consideration of the user's interests without an explicit formulation of these interests by this user. This is an advantage as the formulation of interests is often too difficult for the user, or the interests change by time or the user is not aware of all his interests.

The remainder of this paper is organised as follows. In the next section a brief introduction to data mining and genetic algorithms is given. Section 3 presents the most important phases of the developed procedure in detail and Section 4 presents the proceedings of the evaluation and the most interesting evaluation results. Also a summarisation of the results of the interviews with experts is given. The last section concludes with a summary.

2. Data mining and genetic algorithms

The task of knowledge discovery, especially data mining, is to discover new, potentially useful, nontrivial, unexpected and comprehensible knowledge from collected data of databases. [FPS96] Two commonly used methods are association rules, first introduced by [AIS93], and classification rules, which describe patterns hidden in the data.

In the following the data mining methods association rules and classification rules will be explained, two approaches of the literature will be presented, which support the user in finding interesting association rules and genetic algorithms will be described shortly.

2.1 Association rules

An association rule describes the co-occurrence of two sets of attribute-values, X and Y. Whenever the attribute-values of X occur in a tuple, the attribute-values of Y are likely to be also found there. Given is a set of attributes of a relational database D, $atts(D) = \{A_1, A_2, ..., A_n\}$, where each attribute consists of a set of attribute-values, $dom(D) = dom(A_1) \times dom(A_2) \times ... \times dom(A_n)$,. An association rule is an expression of the form $X \Rightarrow Y$, where X and Y are tow sets of attribute-values with $dom(X) \in dom(D)$, $dom(Y) \in dom(D)$ and $X \cap Y = \varnothing$ and which satisfy the user defined thresholds for support and confidence.

Support expresses the statistical significance of an association rule, i.e. the percentage of tuples in the database D, in which X and Y occur together.

$Sup(X \Rightarrow Y) = \frac{|X \cup Y|}{|D|}$, where |D| is the number of tuples in the database.

All itemsets which meet the user defined support threshold are called large itemsets.
Confidence is a measure of the strength of an association rule and is calculated by the conditional probability. It expresses the percentage of tuples containing X that are also containing Y.

$Conf(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|}$.

All association rules which meet the user defined confidence threshold are called valid association rules.

A vast number of different algorithms can be found in the literature for the computation of association rules. [cp. AP98, CL02, GB03, HGN02, HMGN02, NDD99] They were analysed if they consider user's interests during the computation of these rules. The result shows, that only few of them do and if they do, they need user defined queries in advance through which the mentioned problems of formulating interests arise.

2.2 Classification rules

Classification rules are used for the sorting of objects in distinct and so far unknown classes. They use attribute values of tuples given in the dataset to decide about the class memberships of these tuples. The quality of the classification rules is determined by the number of faultless classifications [cp. BS01].

At first glance there might be no big difference between association rules and classification rules. But each of this methods has a different task. The task of association rules is to describe coherences hidden inside the data whereas classifications are used for the prediction of class membership with the help of attribute values in the database [cp. F00].

2.3 Other approaches that support the user's search

The literature was analysed for other approaches that support the user in his search for interesting association rules. In the following the two most interesting approaches 'interestingness measures' and 'redundancy of association rules' will be presented and compared with the developed procedure.

Interestingness measures

Interestingness measures determine the interestingness of association rules by statistical ratios or beliefs of the user. They were analysed, if and how they consider interests of the user. Some of them do. They are called subjective measures and need user formulated interests in advance [cp. KMRTV94, LHCM00, PT98, ST95]. Thereby they are linked with the problems of formulating interests as mentioned before.

Redundancy of association rules

In the literature exist several definitions about redundancy of association rules. They all are highly non uniform. The application of different definitions of redundancy on the same dataset leads to different, sometimes even conflicting results. Moreover only few of them are generally applicable on any amount of association rules as only some of them deal with redundancy in a logical purpose. None of them consider user interests [cp. BAG99, CS02, LHM99, SA95, SLRS99].

2.4 Genetic algorithms

Genetic algorithms are stochastically, intelligent search methods that imitate the natural evolution. They are characterised by an iterative run of several phases. For each of these phases different genetic methods can be used. This leads to a vast number of method combinations [cp. N97].

An analysis of the literature was done to examine the use of genetic algorithms for the computation of association rules and classification rules. The result of this analysis shows, that the computation of classification rules is a main field of application of genetic algorithms. Lots of different approaches can be found in the literature [AMR01, EJ93, EKK04, FLF00, IF03, K94, KK05, MVFN01]. However, for the computation of association rules genetic algorithms are only used for the consideration of special cases, for example like the computation of frequent amounts of attribute values by the computation of association rules [MAR02].

3. The developed procedure in detail

In the following the most important steps of the developed procedure will be explained. Figure 2 gives a review of the procedure. Starting point of the procedure is the amount of valid association rules of the considered dataset. This amount is given and has not to be computed by the procedure. Firstly some association rules have to be selected and presented to the user. Secondly the user has to value each of these rules as interesting or uninteresting. A genetic algorithms uses this subjective information of the user to form classification rules, which are then used to classify the not yet presented association rules in the classes 'interesting' and 'uninteresting'. Starting from the second run, only those association rules can be selected for the presentation, which were classified as interesting in the previous run.

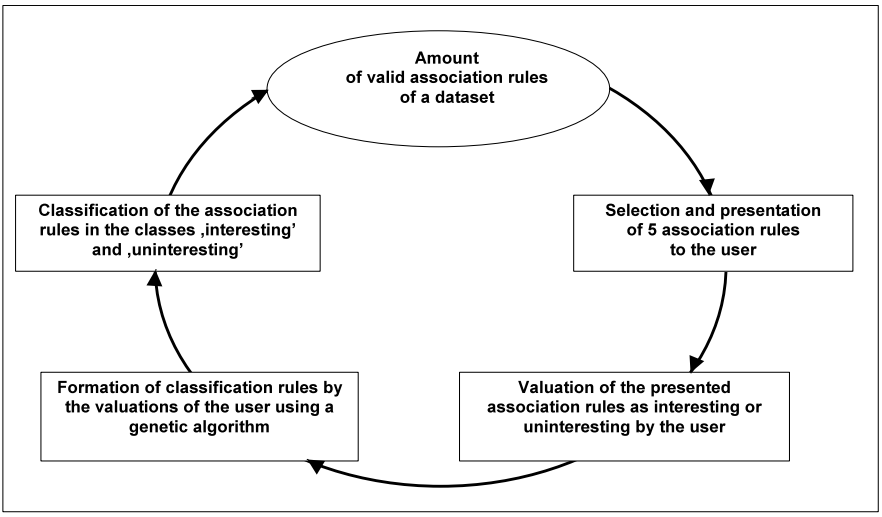

Figure 2: Review of the developed procedure

First important step is the selection of the presented association rules. Basis of this selection is the following distance measure:

$$Dist(X{:}x \Rightarrow Y{:}y, V{:}v \Rightarrow W{:}w) = |(X{:}x, Y{:}y)\ \Phi\ (V{:}v, W{:}w)| \text{ [cp. DL98)]}$$

This measure computes the distance of two association rules X⇒Y und V⇒W by the number of attribute values they are different in. This distance is computed for every association rule to the others. Those association rules will be selected for the presentation which have the highest distances.

Second important step is the valuation of the user. To illustrate the user valuation the following association rule of a dataset about patients that suffer from multiple sclerosis is used: gender:man, handicap:impaired vision, age:older than 50 years ⇒ illness:no multiple sclerosis. For example, if a user is interested in association rules that give an information about the gender of a patient, the given association rule would be interesting. Otherwise, if a user is for example interested in association rules about patients between 30 and 40 years, the given association rule would be uninteresting. This valuation has to be done for every presented association rule.

In the first run an initialisation strategy is needed to create the first classification rules, which will be the parents for the genetic operators. In the following runs the classification rules of the previous run will be used instead. The initialisation strategy of the procedure uses the valued association rules and turns them into classification rules as their classes are known due to the valuation of the user. To make these classification rules applicable on other association rules a generalisation is done by the random removal of one attribute value of each classification rule.

Now the genetic operators recombination and mutation are applied to create new classification rules. The recombination combinates the parents to form new classification rules whereas the mutation uses duplicates of the parents and alters them randomly. Together with the evaluation results of the procedure will be explained which genetic operators were used for the evaluation.

To make a selection of classification rules possible an evaluation of these rules has to be done. This is done by the sum of the following two ratios:

$$laplacefunction = \frac{tp+1}{tp+fp+2} \text{ and } completeness = \frac{tp}{tp+fn} \text{ [cp. LK00]}$$

Both ratios base on the results of the application of the classification rules on the yet presented and user-valued association rules. The laplacefunction is the percentage of association rules that were classified as interesting and are interesting for the user on all association rules that were classified as interesting. The completeness expresses the percentage of association rules that were classified as interesting and are interesting for the user on all association rules that are interesting for the user. This sum is computed for every classification rule (parents, recombination rules and mutation rules).

With these results a selection of classification rules is possible. This selection determines, which classification rules will be used for the classification of the not yet presented association rules in the classes ‘interesting’ und ‘uninteresting’ in the current run and as parents of the genetic operators in the following run. Together with the

evaluation results of the procedure will be explained which genetic operators were used for the evaluation.

4. Evaluation and interviews with experts

The evaluation of the developed procedure was done to analyse the general performance of the procedure and the influence of different genetic parameters and different scenarios of interest. In the following the datasets, used for the evaluation, the proceeding of the evaluation, the most interesting evaluation results and a summary of the interviews with experts will be presented.

4.1 Evaluation proceeding

Two highly different datasets were used for the evaluation. First dataset is the heart disease dataset. It is a standard dataset, which is available on the internet (www.liacc.up.pt/ML/statlog/datasets/heart/heart, download 18.10.2004). 4053 association rules where computed for this dataset.

The second dataset is the flickr dataset, an individual dataset of tag tuples of the web2.0 application flickr. As is data is not available as complete dataset on the internet, it first had to be collected and 5932 association rules were computed for it.

Flickr is an internet portal which allows users to store and present their photos. One important characteristic of this portal is, that every photo is tagged by his owner with up to four tags. These tags are used by the search engine of the portal. As the users are completely free in choosing their tags, for other user it is often very difficult to find those tags which lead to the photos they are interested in. To enhance this situation, the developed procedure supports the user in finding those tags, which lead to the photos he is interested in.

These tow datasets are characterised by two main differences. First difference is about the attributes. In the flickr dataset every attribute consists of exactly one attribute value as it is not possible to eliminate any tag combination, because there are not restrictions for the user's tag choices. This fact leads to about 70 attributes for the flickr dataset whereas the heart disease data dataset has only 14 attributes. Second difference is about the presentation of association rules. In the flickr application it does not make sense to present different association rules about the same amount of attribute values to the user. Therefore only one association rule for each amount of attribute values was chosen for the presentation. This fact makes the learning of the user's interests much more difficult.

The evaluation was done with 150 runs of the procedure, each with 50 iterations, in which six users with different scenarios of interests were simulated.

Evaluation criterion was the average predictive accuracy, which is the average number of presented association rules that agree to the interests of the simulated users.

4.2 Evaluation results

In the following, evaluation results of different genetic methods and different scenarios of interest will be presented for both datasets.

Genetic methods

For each of the three important genetic algorithm phases recombination, mutation and selection the two most different methods were chosen (Table 1).

	Method 1	**Method 2**
Recombination	Diagonal crossover (several parents are used for the formation of a new individual)	Single-point crossover (exactly two parents are used for the formation of a new individual)
Mutation	Mutation probability: 1	Mutation probability: ≤ 1
Selection	Stochastic universal sampling (probabilistic selection)	Truncation selection (deterministic selection)

Table 1: genetic methods

The combination of these six methods leads to eight different alternatives (Table 2) which were evaluated for both datasets (Figure 3).

Alternative	**Recombination method**	**Mutation method**	**Selection method**
1	Diagonal crossover	Mutation probability ≤ 1	Stochastic universal sampling
2	Diagonal crossover	Mutation probability ≤ 1	Truncation selection
3	Diagonal crossover	Mutation probability = 1	Stochastic universal sampling
4	Diagonal crossover	Mutation probability = 1	Truncation selection
5	Single-point crossover	Mutation probability ≤ 1	Stochastic universal sampling
6	Single-point crossover	Mutation probability ≤ 1	Truncation selection
7	Single-point crossover	Mutation probability = 1	Stochastic universal sampling
8	Single-point crossover	Mutation probability = 1	Truncation selection

Table 2: genetic method alternatives

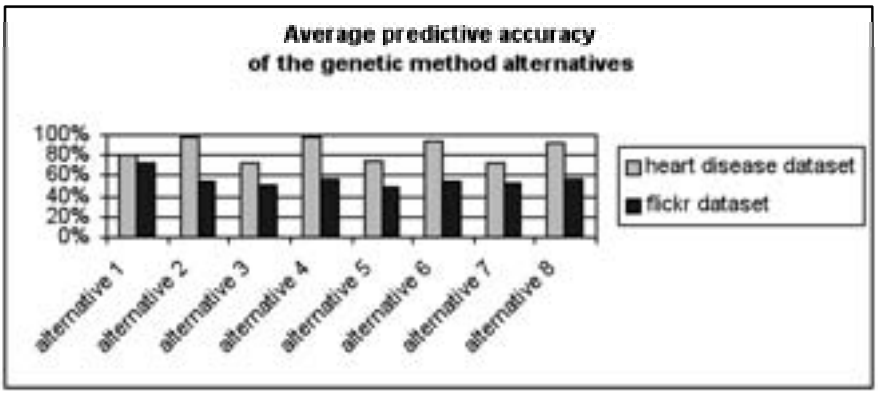

Figure 3: Evaluation results for the genetic method alternatives

The results of the heart disease dataset show, that the alternatives with the deterministic selection method truncation selection achieved the best results, whereas the flickr dataset achieved the best results with the first alternative, the recommendation of the literature (diagonal crossover, mutation probability ≤ 1 and stochastic universal sampling). This illustrates, that for a small dataset with relatively easy learning conditions like the heart disease dataset the need for stochastic elements is low as the deterministic selection leads to better results. However in case of a bigger data set with more difficult learning conditions like the flickr dataset the influence of stochastic methods is needed to achieve better results.

The comparison of the genetic method results of both datasets shows, that the results of the heart disease dataset are always better than those of the flickr dataset. This is due to the characteristics of the datasets, which were mentioned before.

Scenarios of interest

Six scenarios of interest, which differ in complexity and the number of interesting association rules were evaluated for both datasets. An scenario of interest with a low complexity is for example the interest in association rules that give an information about the age of a person, whereas the interest in association rules about men older than 50 years with an impaired vision is more complex. The numbers of interesting association rules are the numbers of association rules that agree to the different scenarios of interest.

Figure 4 shows the results for the heart disease dataset. The scenarios of interest are ordered increasingly after their complexity and the numbers of interesting association rules for each scenario are given in brackets.

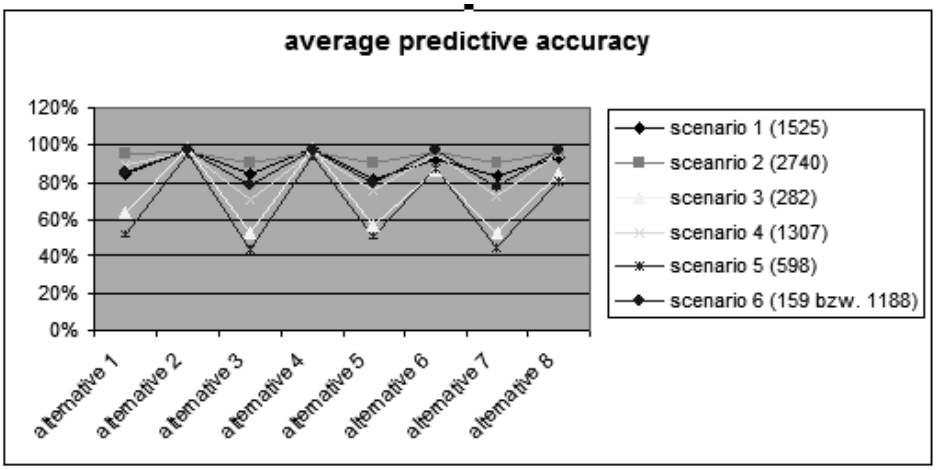

Figure 4: Results of the scenarios of interest for the heart disease dataset

Referring to the complexity of the scenarios of interest no influence on the performance can be derived by the results, because for example the most complex scenario 6 achieved always better results than scenario 3 which is less complex. Referring to the numbers of interesting association rules an influence on the performance can be derived from the results. For example scenario 2 with the highest number of interesting association rules achieved for all eight alternatives better results than scenario 3 and 5, both with much lower numbers of interesting association rules. As the results correspond not for all scenarios directly to the numbers of interesting association rules (scenario 6 has sometimes better results than scenario 1), the influence is moderate.

Figure 5 shows the results for the flickr dataset. The scenarios of interest are ordered increasingly after their complexity and the numbers of interesting association rules for each scenario are given in brackets.

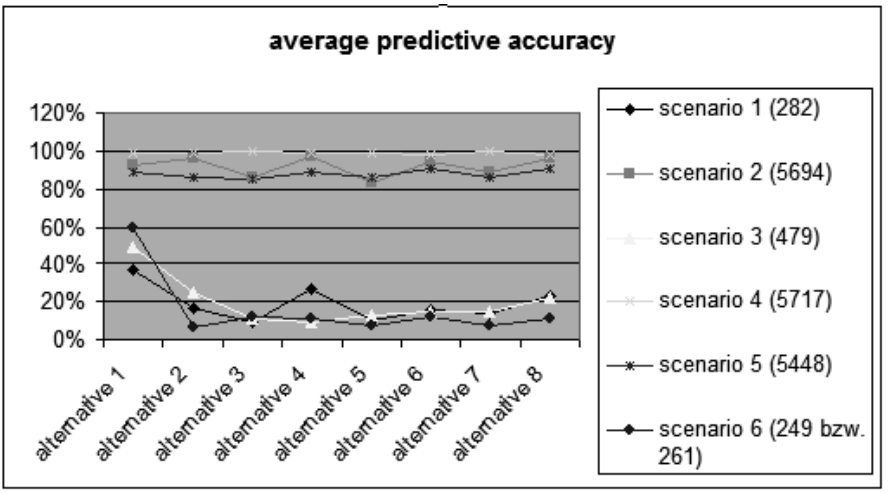

Figure 5: Results of the scenarios of interest for the flickr dataset

Referring to the complexity of the scenarios of interest no influence on the performance can be derived by the results, because for example the very complex scenario 5 achieved always better results than scenario 1 and 3 which are less complex. Referring to the numbers of interesting association rules an influence on the performance can be derived from the results. The scenarios 2, 4 and 5 with much higher numbers of interesting association rules than the other three scenarios achieved for all eight alternatives much better results. As the results correspond almost directly to the numbers of interesting association rules, the influence is strong.

The comparison of the scenarios of interest results for the complexity do not indicate an influence on the performance of the procedure for both datasets. However for the numbers of interesting association rules an influence can be derived from the results of both datasets. In case of the heart disease dataset the influence is moderate whereas in case of the flickr dataset the influence is strong.

4.3 Interviews with experts

To analyse the main field of application of the developed procedure and the relevancy of practice, four experts of medicine, statistics and biochemistry where interviewed in partial structured interviews. As main field of application the research domain was mentioned and the experts affirmed the need and relevancy of practice of the procedure.

5. Summary

The developed procedure allows the user an interactive and explorative search for previously unknown coherences hidden in huge amounts of digital data. The proceeding

of the procedure is related to browsing the internet. Main characteristic of the developed procedure is, that the user does not have to formulate his interests explicitly in advance although his interests are considered by the procedure. This is an important fact as often the formulation of interests is difficult or unfeasible for the user, the interests of the user change by and by or the user is not aware of all his interests as he does not know all the coherences, which are hidden inside the data. Huge amounts of digital data can be found in many different domains whereby the developed procedure is applicable in various fields of applications.

The evaluation results have shown that the use of genetic algorithms for an user specific reduction of amounts of interesting association rules leads to good results. The maximal predictive accuracy of the heart disease dataset is 99% whereas for the flickr dataset the maximal predictive accuracy is 73%. The analysis of different genetic methods has shown that they have an influence on the performance of the procedure and that the genetic methods which achieved the best results are dependent on the dataset they are applied on. The analysis of different interest scenarios has led to two different findings. For the degree of complexity of the interest scenarios no influence on the performance can be derived from the evaluation results of both datasets. In contrast the number of interesting association rules of an interest scenario had an influence on the performance for both datasets. For the heart disease dataset a light influence was indicated by the evaluation results as the for the flickr dataset the evaluation results showed a strong influence on the performance.

5. References

[AIS93] Agrawal, R. / Imielinski, T. / Swami, M.: Mining association rules between sets of items in very large databases in: Proc. of the ACM SIGMOD Conf. on Management of data, p. 207-216, 1993

[AMR01] Alvarez, J. L. / Mata, J. / Riquelme, J. C.: Oblic: Classification Systems using Evolutionary Algorithm, in: Proc. of the 6th Int'l. Work-Conference on Artificial and Natural Neural Networks 2001, p. 644-651

[AP98] Aggarwal, Charu C. / Philip S. Yu: Online generation of association rules, in: Proc. of the 4th Int'l. Conf. on Knowledge Discovery and Data Mining 1998, p. 129-133

[BAG99] Bayardo, Roberto J. Jr. / Agrawal, Rakesh / Gunopulos, Dimitrios: Constraint-based rule mining in large dense databases, in: Proc. of the IEEE Int'l. Conf. on Data Engineering 1999, p. 188-197

[BS01] Bäck, Thomas / Schütz, Martin: Evolutionäre Algorithmen im Data Mining, in: Handbuch Data Mining im Marketing, hrsg. v. Hippner, Hajo / Küsters, Ulrich / Meyer, Matthias / Wilde, Klaus, Braunschweig, Wiesbaden 2001, p. 403-426

[CL02] Cong, Gao / Liu, Bing: Speed-up iterative frequent itemset mining with constraint changes, in: Proc. of the 2nd IEEE Int'l. Conf. on Data Mining 2002, p. 107-114

[CS02] Cristofor, Laurentiu / Simovici, Dan: Generating an informative cover for association rules, in: Proc. of the IEEE Int. Conf. on Data Mining 2002, p. 597-600

[DL98] Dong, Gouzhu / Li, Jinyan: Interestingness of discovered association rules in terms of neighborhood-bases unexpectedness, in: Proc. of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining 1998, p. 72-86

[EJ93] Eick, C. F. / Jong, D.: Learning bayesian classification rules through genetic algorithms, in: Proc. of the 2nd Int'l. Conference on Information Knowledge Management 1993, p. 305-313

[EKK04] Eggermont, J. / Kok, J. N. / Kosters, W. A.: Genetic programming for data classification: paritioning the search space, in: Proc. of the 2004 ACM Symposium on Applied Computing 2004, p. 1001-1005

[F00] Freitas, Alex A.: Understanding the crucial difference between classification and discovery of association rules - a position paper, in: ACM SIGKDD Explorations 2(1) 2000, p. 65-69

[FLF00] Fidelis, M. V. / Lopes, H. S. / Feitas, Alex A.: Discovering comprehensible classification rules with a genetic algorithm, in: Proc. of the Congress on Evolutionary Computation 2000, p. 805-810

[FPS96] Fayyad, Usama M. / Piatetsky-Shapiro, Gregory / Smyth, Padhraic: From Data Mining to Knowledge Discovery: An Overview, in: Advances in knowledge discovery and data mining, hrsg. v..Fayyad, Usama M. et. al., Menlo Park, California /Cambridge, Massachusetts / London, England, p. 1-34, 1996

[GB03] Goethals, Bart / Bussche, Jan Van den: On supporting interactive constrained association rule mining, in: Proc. of the 2nd Int'l. Conf. on Data Warehousing and Knowledge Discovery 2000, p. 307-316

[HGN02] Hipp, Jochen / Güntzer, Ulrich / Nakhaeizadeh, Gholamreza: Data mining of association rules and the process of knowledge discovery in databases, in: Proc. of the Industrial Conf. on Data Mining 2002, p. 15-36

[HMGN02] Hipp, Jochen / Mangold, Christoph / Güntzer, Ulrich / Nakhaeizadeh, Gholamreza: Efficient rule retrieval and postponed restrict operations for association rule mining, in: Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining 2002, p. 52-65

[IF03] Isasi, P. / Fernandez, F.: Evolutionary approach to overcome initialization parameters in classification problems, in: Proc. of the 7th Int'l. Work-Conference on Artificial and Natural Neural Networks 2003, p. 254-261

[K94] Konstam, A.: N-Group classification using genetic algorithms, in: Proc. of the 1994 ACM Symposium on Applied Computing 1994, p. 212-216

[KK05] Kshetrapalapuram, K. K. / Kirley, M.: Mining classification rules using evolutionary multi-objective Algorithms, in: Proc. of the 9th Conference on Knowledge-based Intelligent Information and Engineering Systems 2005, p. 959-965

[KMRTV94] Klemettinen, Mika / Mannila, Heikki / Ronkainen, Pirjo / Toivonen, Hannu / Verkamo, A. Inkeri: Finding interesting rules from large sets of discovered association rules, in: Proc. of the 3rd Int'l. Conf. on Information and Knowledge Management 1994, p. 401-407

[LHCM00] Liu, Bing / Hsu, Wynne / Chen, Shu / Ma,Yiming: Analyzing the subjective interestingness of association rules, in: IEEE Intelligent Systems 15(5) 2000, p. 47-55

[LHM99] Liu, Bing / Hsu, Wynne / Ma, Yiming: Pruning and summarizing the discovered association rules, in: Proc. of the 5th ACM SIGKDD Int'l. Conference on Knowledge Discovery and Data Mining 1999, p. 125-134

[LK00] Liu, Juliet Juan / Kwok, James Tin-Yau: An extended genetic rule induction algorithm, in: Proc. of the Congress on Evolutionary Computation 2000, p. 458-463

[MAR02] Mata, Jacinto / Alvarez, José-Luis / Riquelme, José-Cristobal: An evolutionary algorithm to discover numeric association rules, in: Proc. of the 2002 ACM Symposium on Applied Computing 2002, p. 590-594

[MVFN01] Mendes, R. R. F. / Voznika, F. de B. / Freitas, A. A. / Nievola, J. C.: Discovering Fuzzy Classification Rules with Genetic Programming and Co-evolution, in: Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery 2001, P. 314-325

[N97] Nissen, Volker: Einführung in Evolutionäre Algorithmen: Optimierung nach dem Vorbild der Evolution, Braunschweig, Wiesbaden 1997

[NDD99] Nag, Biswadeep / Deshpande, Prasad M. / DeWitt, David J.: Using a knowledge cache for interactive discovery of association rules, in: Proc. of the 5th ACM SIGKDD Int'l. Conference on Knowledge Discovery and Data Mining 1999, p. 244-253

[PT98] Padmanabhan, Balaji / Tuzhilin, Alexander: A belief-driven method for discovering unexpected patterns, in: Proc. of the 4th Int'l. Conf. on Knowledge Discovery and Data Mining 1998, P. 94-100

[SA95] Srikant, Ramakrishnan / Agrawal, Rakesh: Mining generalized association rules, in: Proc. of the 21st Int'l. Conf. on Very Large Data Bases 1995, P. 407-419

[SLRS99] Shah, Devavrat / Lakshmanan, Laks V. S. / Ramamritham, Krithi / Sudarshan, S.: Interestingness and pruning of mined patterns, in: Proc. of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 1999

[ST95] Silberschatz, Avi / Tuzhilin, Alexander: On subjective measures of interestingness in knowledge discovery, in: Proc. of the 1st Int'l. Conf. on Knowledge Discovery and Data Mining 1995, p. 275-281