Evaluation of Association Rule Quality Measures through Feature Extraction*

José L. Balcázar[1] and Francis Dogbey[2]

[1] Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Barcelona, Spain
jose.luis.balcazar@upc.edu
[2] Advanced Information Technology Institute
Ghana-India Kofi Annan Centre of Excellence in ICT
Accra, Ghana
francisd@aiti-kace.com.gh

Abstract. The practical success of association rule mining depends heavily on the criterion to choose among the many rules often mined. Many rule quality measures exist in the literature. We propose a protocol to evaluate the evaluation measures themselves. For each association rule, we measure the improvement in accuracy that a commonly used predictor can obtain from an additional feature, constructed according to the exceptions to the rule. We select a reference set of rules that are helpful in this sense. Then, our evaluation method takes into account both how many of these helpful rules are found near the top rules for a given quality measure, and how near the top they are. We focus on seven association rule quality measures. Our experiments indicate that multiplicative improvement and (to a lesser extent) support and leverage (a.k.a. weighted relative accuracy) tend to obtain better results than the other measures.

Keywords: Association rules, Feature Extraction, Prediction, Support, Confidence, Lift, Leverage, Improvement.

1 Introduction and Preliminary Definitions

Association rules have the syntactic form of implications, $X \to A$, where X is a set of items and A is an item. Although several works allow for several items in the right-hand side (often under the term *partial implications*, e.g. [20]), here, like in many other references, we follow [2,6] and restrict ourselves to single-item consequents. Semantically, association rule $X \to A$ is expected to mean that A tends to appear when X appears. However, there are many different formalizations for such a relaxed implication; this issue applies both to supervised learning and in the context of associations, see [13,15,17,18,24].

* This work has been partially supported by project BASMATI (TIN2011-27479-C04) of Programa Nacional de Investigación, Ministerio de Ciencia e Innovación (MICINN), Spain, and by the Pascal-2 Network of the European Union.

A. Tucker et al. (Eds.): IDA 2013, LNCS 8207, pp. 68–79, 2013.

We aim at finding ways to objectively evaluate the rule quality measures themselves. One existing approach consists of generating algorithmically datasets where certain itemsets are to be found later [27]. We explore an alternative approach to the same end. This paper is based on the following informal working hypothesis: since different rule quality measures select different association rules, we might be able to compare objectively these sets of "top" rules by combining them with other Data Mining decision processes, and evaluating their contribution. Specifically, here we relate associations with feature extraction and with predictors.

In classification problems of Machine Learning, feature selection and extraction are processes, often considered as preprocessing, through which better accuracies can be sometimes obtained. Along these processes, every observation acquires, explicitly or implicitly, additional columns, that may even fully replace the original features. It is a wide area with a large body of references: see [16] and the references there.

In our previous paper [4], we explored the average accuracy improvements (which can be negative) of predictors as new features are added to the observations. Each of these added features is Boolean, and flags exceptions to association rules. In turn, these association rules are picked as the "top-N", according to each of several popular rule quality measures. In that paper, our results indicated, essentially, very little average positive accuracy increment or, in many cases, none at all. Also, the differences among the average accuracy increments for several quality measures were rather marginal.

Hence, we focus now on individual rules that actually do improve accuracy. We propose to evaluate rule quality measures according to how many of these helpful rules are found among the top-quality rules for the measure, and how close to the top they appear. We study two variants, depending on whether it is desired to employ the new feature for actual classification tasks. We demonstrate our new score on seven quality measures, and find that multiplicative improvement and, to a lesser extent, support and leverage (a.k.a. weighted relative accuracy) tend to obtain better scores.

We must point out here that we do not wish to argue, nor even implicitly suggest, that helpfulness for classifiers, at the current status of our understanding, is to be equated with subjective interest for the user. However, our proposal combines, in an interesting, novel way a precise formal definition, leading to numeric scores, with other data mining processes that could take place on the same data.

2 Setting

Our notation is standard. The reader is assumed to be familiar with association rules (see the survey [9]), and with the C5.0 tree-based predictor, an improved version of C4.5 [23]. See also e.g. [25] and the references there.

For our development, we need our datasets both in relational and transactional formats. In transactional datasets, each observation ("row" or "transaction") consists of a set of items. In the relational case, each observation ("row")

consists of an observed value for each of a fixed set of attributes ("columns"). We use a relational and a transactional variant for each dataset. Association rules are computed on the transactional variant, then used to add extra features (new columns) to the relational variant before passing it on to the predictor. Transactional data can be seen as relational in several non-equivalent ways; here, we consider one binary column per item. Thus the association rules only use *positive* appearances from the relational table. Relational datasets are casted in transactional form through the fully standard approach of adding one item for each existing pair ⟨attribute,value⟩.

The support of an itemset in a transactional dataset is the number of transactions in which it is contained. It is used often normalized, by dividing by the total number of transactions, and then it is akin to a frequentist estimator of the probability of the itemset. Accordingly, we will denote by $p(X)$ the normalized support of itemset X. This notation is naturally extended to conditional probabilities $p(A|X) = p(AX)/p(X)$.

2.1 Rule Quality Measures

There are literally dozens of measures proposed in the literature for choosing association rules. They assign a real number as a value to each association rule. This value can be used either by thresholding, thus discarding rules that receive low value, or by priorization, by sorting the rules obtained according to their value. There are many studies of the properties of the different measures; see [15,17,18,24]. To evaluate association rule $X \to A$, most measures involve an arithmetic combination of the quantities $p(X)$, $p(A)$, and $p(XA)$. We define now the measures we are interested in.

Definition 1. *[1] The* support *of association rule* $X \to A$ *is* $s(X \to A) = p(XA)$.

Definition 2. *[1,20] The* confidence *of association rule* $X \to A$ *is* $c(X \to A) = p(A|X) = p(XA)/p(X)$.

The most basic and widely applied scheme for selection of association rules consists in imposing thresholds for the support (thereby reducing the exponentially growing exploration space of itemsets) and for the confidence.

Definition 3. *[17] The* relative confidence *of association rule* $X \to A$*, also called* centered confidence *or* relative accuracy*, is* $r(X \to A) = p(A|X) - p(A) = c(X \to A) - c(\emptyset \to A)$.

The relative confidence is, therefore, measuring additively the effect of "adding the condition" or antecedent X on the support of the consequent A.

Definition 4. *[8] The* lift *of association rule* $X \to A$ *is* $\ell(X \to A) = \frac{p(XA)}{p(X) \times p(A)}$.

Definition 5. *[17,22] The* leverage *of association rule* $X \to A$ *is* $v(X \to A) = p(XA) - p(X) \times p(A)$.

Both in lift and in leverage, if supports are unnormalized, extra factors n are necessary. A widespread criticism of lift and leverage is that they are symmetric, that is, give the same value to the rule obtained by permuting antecedent and consequent, and, therefore, fail to support the intuitive directionality of the syntax $X \to A$.

In case of independence of both sides of the rule at hand, $p(XA) = p(X)p(A)$; therefore, both lift and leverage are measuring deviation from independence: lift measures it multiplicatively, and leverage does it additively. By multiplying and dividing by $p(X)$, it is easy to check that leverage can be rewritten as $v(X \to A) = p(X)(p(A|X) - p(A)) = p(X)r(X \to A)$, and is therefore called also *weighted relative accuracy* [17].

One of the few published measures that do not depend only on the supports $p(X)$, $p(A)$, $p(XA)$, and similar quantities, but requires instead exploring a larger space, is:

Definition 6. *[5,19] The* improvement *of association rule $X \to A$, where $X \neq \emptyset$, is $i(X \to A) = \min\{c(X \to A) - c(Y \to A) | Y \subset X\}$.*

Improvement generalizes relative confidence by considering not only the alternative rule $\emptyset \to A$ but all rules where the left-hand side is some proper subset of X. The same process can be applied to lift, which is, actually, a multiplicative, instead of additive, version of relative confidence: $\ell(X \to A) = \frac{p(XA)}{p(X) \times p(A)} = \frac{c(X \to A)}{p(A)} = \frac{c(X \to A)}{c(\emptyset \to A)}$. Taking inspiration in this correspondence, we introduced in [4] a multiplicative variant of improvement that generalizes lift, exactly in the same way as improvement generalizes relative confidence:

Definition 7. *The* multiplicative improvement *of association rule $X \to A$, where $X \neq \emptyset$, is $m(X \to A) = \min\{c(X \to A)/c(Y \to A) | Y \subset X\}$.*

2.2 Datasets

A key condition we have imposed on our empirical study is to employ datasets that allow for sensible relational and transactional formulations without resorting to further algorithmics. In this sense, numeric attributes may need a discretization process and, then, it would be far from clear how to tell which empirical effects were due to the interaction between the associator and the classifier, as we wish to study, and which ones were introduced by the discretization phase. We restrict ourselves to datasets that allow for direct run of a standard associator.

We introduce now the datasets on which we run our tests. They are among the far most common benchmark datasets in association rule quality studies. All of them are publicly available [11].

CMC is short for Contraceptive Method Choice, and contains data from an Indonesian survey of demographic features, socio-economic features and contraceptives choices (predicted attribute) among married women. The dataset is made up of 1473 observations and 10 attributes. ADULT contains census information, extracted from questionnaire data. It has 48842 observations and 11

Table 1. Size, items, support, rules, and initial accuracies of C5.0

Dataset	Size	Nb. of items	Attrib. Predict	Supp.	Nb of rules	C5.0 Accur.
CMC	1473	74	Contracep. Meth.	1%	19067	50.17%
ADULT	48842	272	Income	1.18%	19840	83.49%
GERMAN	1000	1077	Customer class	11.75%	19762	72.69%
VOTES	435	50	Party	22.5%	19544	95.40%
MUSHROOM	8124	119	Cap surface	15.9%	19831	53.30%

attributes. As usual for this dataset, we predict whether the annual income of individuals is over a certain threshold (50K). We prepared the adult dataset by merging the known train and test sets. We removed the four numeric attributes fnlwgt (final weight), education-num (which is redundant), capital-gain, and capital-loss. GERMAN is a dataset on credit scoring from Germany. There, the data from clients must be used to predict whether the client is a good candidate to receive a loan. It has been suggested that prediction should be made by weighting differently (by a factor of 5) mistakes in predicting good than mistakes in predicting bad; however, in order to keep fair comparisons with all our other datasets, here we did not weigh differently the mistakes. VOTES records votes of representatives from the US Congress to a number of law proposals in 1984; the attribute to be predicted is the party the representative belongs to. Finally, MUSHROOM reports characteristics that can appear together in a sample of certain mushroom species. The usual prediction task for this dataset is whether the exemplar is edible or poisonous. However, the predictor attains full accuracy directly on the original data, which renders the task useless to evaluate association rule quality measures according to our proposal. Therefore, we only report on results predicting the sort of "cap surface" of the exemplar, given the rest of the attributes. The attributes to be predicted are two-valued for all the datasets, except for CMC and MUSHROOM, where predictions are respectively three-valued and four-valued.

Table 1 indicates the main characteristics of the chosen datasets. The *size* is the number of observations; the *number of items* is the total of different values existing for all the attributes, and, therefore, coincides with the number of items in the transactional version of the dataset. For each dataset, we impose a support constraint and compute association rules with no confidence constraint with a standard Apriori associator [6]. Since our purposes are essentially conceptual, here we are not (yet) after particularly fast algorithmics, and that associator was often fast enough. The support is set at a value appropriate for Apriori to yield between 19000 and 20000 rules in these conditions. These supports were identified manually, and are also reported in the table, together with the number of rules. We also report in the table the baseline accuracies of C5.0, for each dataset, before any extra features are added.

The feature extraction and accuracy test is developed as follows. For a given association rule, an enlarged relational version of the dataset is fed to the predictor: it contains one additional binary column, which indicates whether each row is

an exception of the association rule, that is, fulfills the antecedent of the rule but not the consequent. Only rules of confidence higher than 50% and strictly lower than 100% (in order to ensure the presence of exceptions) are recorded; also, only rules with nonempty antecedent are employed: otherwise, the additional feature would be redundant, flagging just failure of the consequent. The parameter settings of C5.0 are left at their default values. The accuracy of the predictor on the enlarged dataset is compared to the accuracy obtained on the original unexpanded dataset. All accuracies are computed by 10-fold cross-validation.

In [4], for each rule quality measure under study, the top N rules were recorded, and the accuracy test performed as just described; then, the process continued by averaging, separately by measure, the resulting accuracy changes. The average accuracy, however, changed little if we compare among them the different rule quality measures, and its ability to evaluate rule quality measures is, thus, debatable. Thus, a finer analysis is needed. This is the aim of the present paper.

3 A Score for Rule Quality Measures

This section describes our major contribution: a score for association rule quality measures in terms of usefulness for predictive tasks. Our proposal for a finer evaluation of the rule quality measures is as follows: first, one identifies a number of "good", helpful, rules, that is, rules whose corresponding new features do improve clearly the accuracy of the predictors. Then, we can evaluate how many of these good rules actually appear among the top N rules according to each of the measures. Thus, we ask ourselves which rule quality measure is able to pick these "good" rules.

Let us move on into the details. First of all, we need to fix a predictor. We use one of the most popular options, namely C5.0. Other commonly employed predictors might be studied in later work; but we note here that [4] included Naïve Bayes predictors (see our discussion of its inappropriateness there), and that numerically-based Support Vector Machines are difficult to conciliate with the categorical attributes that we look for in order to have sensible transactional versions of our datasets without resorting to additional discretizations.

Then, we require to work with a set of rules that, through our feature addition process, increase the C5.0 accuracy by a certain amount. We denote the set of selected rules as G. For each dataset, we select for G those association rules that lead to noticeable accuracy improvements, after adding the new feature flagging exceptions to the rule, as indicated above. In order to be somewhat fair, these rules are taken from the joint pool of all the top-N rule sets provided by all our measures. In our case, we will use $N = 50$. More precisely, from all these rules, 50 from each measure (with a handful of duplications), in turn,

1. the dataset is expanded with one further feature flagging the exceptions, as already indicated,
2. the predictor is run both on the expanded dataset and on the original dataset,
3. the respective accuracies are evaluated via 10-fold cross-validation, and

4. the relative improvement of accuracy (ratio among both accuracies) is determined.

This allows us to identify the set G of helpful rules: those that increase by ϵ the accuracy of the predictor upon having the new feature available. We set $\epsilon = 0.5\%$. To each quality measure, we can assign now a score related to how many of these rules from G do actually appear among the top N rules.

We observe that it is better if helpful rules are captured near the top. Hence, we wish to assign a higher score for a measure both if more helpful rules from G are captured, and if they are captured near the top. More precisely, assume that any given measure has, among its top N rules, k rules from G, and that they are in positions $a_1, \ldots, a_k$, for $1 \leq a_i \leq N$. Then, we assign to it the score

$$\frac{1}{Z_D} \sum_{i=1}^{k} (N + 1 - a_i) = \frac{1}{Z_D} (k(N+1) - \sum_{i=1}^{k} a_i)$$

In this way, hitting a helpful rule as first one ($a_i = 1$) contributes N units to the score; $a_i = 2$ contributes $N - 1$, and so on. The value $N + 1$ ensures that even the rule at the N-th place, $a_k = N$, would contribute something, if little. The value Z_D is a dataset-dependent normalization factor, defined as the highest score reached by any of the measures for that dataset. Its aim is to provide us with a way of comparing the outcomes of the experiments on different datasets along the same scale: Z_D is set in such a way that the rule quality measure that reaches highest score is scored exactly at 1. (A different option for normalization could be the maximum reachable value. We prefer this Z_D as the other option leads to very small values, harder to undertand, for all measures tested.) Then, the presence of values substantially less than 1 means that some rule quality measures are substantially worse than the best one, whereas, if all scores are near 1, it means that the dataset is such that the different measures score relatively similarly.

In a sense, this proposal corresponds to the natural "tuning" of an Area-Under-Curve (AUC) approach [7,10]. In the usual computation of AUC, binary predictions are ranked, and low-ranked predictions are expected to correlate with negative labels, and high-ranked predictions with positive labels. On the other hand, here, many of the rules in the set G may not be "ranked" at all by a given measure, if they do not appear among the top-N. Of course, we must take this fact into account. Except for this, our score can be seen actually as an area-under-curve assessment.

4 Empirical Results

We report the scores for our datasets, and provide below also the outcome of a variant of the process, in which we will disallow the access to some attributes at the time of computing association rules. In the first variant, though, the associator has no such limit.

Table 2. Scores and other parameters at 0.5% accuracy increase

Dataset	Conf.	Lift	R. conf.	Impr.	M. impr.	Supp.	Lev.	$\|G\|$	Max.	Z_D
ADULT	*0*	0.25	*0*	0.24	0.28	**1**	*0.02*	34	20	507
CMC	0.95	0.98	*0.93*	**1**	0.97	0.98	0.94	253	47	1220
VOTES	*0.22*	0.52	0.35	0.64	0.82	0.67	**1**	71	17	425
GERMAN	*0.17*	0.56	0.46	0.79	**1**	0.35	0.21	72	27	745
MUSHROOM (predict cap surface)	0.41	*0*	*0.22*	0.49	**1**	0.56	0.26	18	4	148

4.1 Associator Accesses All Attributes

Our experiments use, as indicated, a very mild threshold of at least $\epsilon = 0.5\%$ accuracy increase in determining the helpfulness of a rule; see Table 2. We also ran experiments on a quite large dataset with census data of elderly people (about 300000 transactions) but, after considerable computational effort, no rule at all was found helpful for that dataset.

Both in this table and in the next, different typefaces mark the highest and the lowest nonzero scores per dataset: boldface marks the measure scoring highest and italics the measure(s) scoring lowest; if some are zero, the lowest nonzero is also marked. We also report the value of Z_D to show how the unnormalized-sum score varies considerable among the datasets. We report as well the size of the set of helpful rules and the maximum amount of helpful rules, "Max.", found among the top 50 rules as all measures are considered. Except for one dataset, this column shows that only a small part of the top rules for each measure are actually helpful.

Generally speaking, we see that multiplicative improvement tends to attain higher values of our score, whereas both confidence and relative confidence tend to perform rather poorly; support, leverage, and, to a lesser extent, also lift and improvement offer relatively good scores. Altogether, however, each measure only catches a handful of the set of good rules.

4.2 Disallow the Associator Access to the Predicted Class

We explore now the following issue: besides the usage for ranking rule quality measures, do we intend to actually employ the new features obtained from the helpful rules to improve the performance of the classifiers?

This question is relevant because, in the affirmative case, we cannot afford to make use of the predicted attribute for the computation of the new features: upon predicting, it will be unavailable. Thus, this attribute is to be excluded from the computation of association rules.

The price is that we rely on hypothetical correlations between the class attribute and the exceptions to rules that do not involve the class attribute; these correlations may not exist. In contrast, in the previous section, exceptions to all rules, whether they involve or not the class attribute, are available. The results

Table 3. Scores at 0.5% accuracy increase (class disallowed)

Dataset	Conf.	Lift	R. conf.	Impr.	M. impr.	Supp.	Lev.	$\|G\|$	Max.	Z_D
CMC	*0.88*	0.90	0.89	0.91	0.96	0.99	**1**	261	48	1236
VOTES	0.87	0.53	*0.22*	0.90	0.53	0.76	**1**	47	9	264
GERMAN	**1**	0.59	*0.34*	0.70	0.68	0.61	*0.34*	39	8	250
MUSHROOM (predict cap surface)	0.18	0.41	*0.15*	0.56	**1**	0.57	0.57	20	9	216

in this section were therefore obtained by excluding the predicted attribute from the computation of the association rules, all the rest being the same. The results are given in Table 3. There was only one helpful rule for the ADULT dataset: $|G| = 1$; hence, it is omitted from this table.

The figures support the candidacy of leverage, multiplicative improvement, and support as occassionally good measures; however, all show some cases of mediocre performance. The outcomes are also less supportive of lift. The unreliability of relative confidence is confirmed, and the behavior of confidence is more erratic than in the previous case: top or close to the top for some cases, low and close to the minimum scores in others.

A fact we must mention is the direct effect of allowing the class attribute to appear in the associations. Comparing both tables, we see that several datasets provide a larger $|G|$ if the class is available, but not substantially larger; and it is in fact smaller for other datasets. In general, with or without access to the predicted class attribute, in most datasets, few helpful rules are captured by each measure ("Max." columns in both tables). The access to the class attribute is not, therefore, as key a point as it could be intuitively expected.

5 Conclusions and Further Work

Along a wide study of proposals to measure the relevance of association rules, we have added a novel approach. We have deployed a framework that allows us to evaluate the rule quality evaluation measures themselves, in terms of their usefulnes for subsequent predictive tasks.

We have focused on potential accuracy improvements of predictors on given, public, standard benchmark datasets, if one more Boolean column is added, namely, one that is true exactly for those observations that are exceptions to one association rule: the antecedent holds but the consequent does not. In a sense, we use the association rule as a "hint of outliers", but, instead of removing them, we simply offer direct access to this label to the predictor, through the extra column.

Of course, in general this may lead astray the predictor instead of helping it. In our previous work [4] we presented an initial analysis of the average change of accuracy, and saw that it may well be negative, and that, generally speaking,

it does not distinguish well enough among various rule quality measures. Therefore, we have concentrated on an analysis based on selected rules that actually provided accuracy increases. We defined a score for rule quality measures that represents how well a given measure is able to bring, near the top, rules that are helpful to the predictor. We account both for the number of helpful rules among the top-k, and for how close to the top they appear, by means of a mechanism akin to the AUC measure for predictor evaluation.

Our experiments suggest that leverage, support, and our recently proposed measure, multiplicative improvement, tend to be better than the other measures with respect to this evaluation score. Leverage and support are indeed known to offer good results in practice. Possibly further empirical analysis of the multiplicative improvement measure may confirm or disprove whether it has similar potential.

5.1 Related Work

We hasten to point out the important differences of association rules versus classification tasks: they differ in the consideration of locally applicable patterns versus modeling for prediction in the global dataset; in association tasks, no particular attribute is a "class" to be predicted, and the aim is closer to providing the user with descriptive intuitions, rather than to foresee future labels. An interesting related discussion is [12].

As a token consequence of the difference, we must observe that, in a context of associations, "perfect" rules of confidence 1 (that is, full implications without exceptions) are, in the vast majority of cases, useless in practice—a surprising fact for those habitued to using rules for classification, where finding a predictive rule without exceptions is often considered progress. See additional discussion in [3].

That said, the idea of evaluating associators through the predictive capabilities of the rules found has been put forward e.g. in [21]. The usage of association rules for direct prediction (where the "class" attribute is forced to occur in the consequent) has been widely studied (e.g. [26]). In [21], two different associators are employed to find rules with the "class" as consequent, and they are compared in terms of predictive accuracy. Predictive Apriori turns out to be better than plain Apriori, but neither compares too well to other rule-based predictor. (We must point out as well that this reference employs datasets with numeric attributes that are discretized, see our discussion in Section 2.2.) Our work can be seen as a natural next step, by decoupling the associator from the predictor, as they do not need to be the same sort of model, and, more importantly, conceptually decoupling the rule quality measure employed to rank the output associations from the algorithm that is actually used to construct the rules.

5.2 Future Work

A number of natural ideas to explore appear. We have evaluated only a handful of rule quality measures; many others exist. Also, other related explorations

remain open: we limited, somewhat arbitrarily, to 50 the top rules considered per measure, but smaller or larger figures may provide different intuitions. Our choice was dictated by the consideration that the understanding of 50 association rules by a human expert would require some time, and one hardly could expect this human attention span to reach much further.

Also, we only tried one popular predictor, C5.0, and we could study others; one could run as well heavy, exhaustive searches in order to find the best set of helpful rules G on which to base the score. The alternative normalization mentioned in the text, instead of Z_D, could be explored, and might end up in figures offering more clear cross-comparison among different datasets. Other evaluations of the predictor could substitute accuracy. Yet other variations worth exploring would be to flag conformance to a rule instead of being an exception to it, that is, marking those observations that support both the antecedent and the consequent of the rule; to explore partial implications, that is, association rules with more than one item in the consequent, where multiplicative improvement would partially correspond to confidence boost [3]; and to allow addition of more than one feature at a time. We hope to explore some of these avenues soon.

Additionally, we consider that our approach might be, at some point, of interest in subgroup discovery (see [14] for its connection to association rules via the common notion of closure spaces, and the further references in that paper). The new features added in our approach, in fact, do identify regions of the space (identified by the antecedent of the association rule, hence having a geometric form of Boolean hypersubcubes) where the consequent of the rule behaves quite differently than in the general population. How this relates to the other existing proposals of subgroup discovery, and whether it is efficient at all in that sense, remains an interesting topic for further research.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) SIGMOD Conference, pp. 207–216. ACM Press (1993)
2. grawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules. In: Advances in Knowledge Discovery and Data Mining, pp. 307–328. AAAI/MIT Press (1996)
3. Balcázar, J.L.: Formal and computational properties of the confidence boost in association rules. To appear in ACM Transactions on KDD (2013), `http://www.lsi.upc.edu/~balqui/`
4. Balcázar, J.L., Dogbey, F.K.: Feature extraction from top association rules: Effect on average predictive accuracy. In: 3rd EUCogIII Members Conference and Final Pascal Review Meeting (2013), `http://www.lsi.upc.edu/~balqui/`
5. Bayardo, R., Agrawal, R., Gunopulos, D.: Constraint-based rule mining in large, dense databases. In: ICDE, pp. 188–197 (1999)
6. Borgelt, C.: Efficient implementations of Apriori and Eclat. In: Goethals, B., Zaki, M.J. (eds.) FIMI, CEUR Workshop Proceedings, vol. 90. CEUR-WS.org (2003)
7. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997)

8. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Peckham, J. (ed.) SIGMOD Conference, pp. 255–264. ACM Press (1997)
9. Ceglar, A., Roddick, J.F.: Association mining. ACM Comput. Surv. 38(2) (2006)
10. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Pattern Recognition Letters 27(8), 882–891 (2004)
11. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
12. Freitas, A.A.: Understanding the crucial differences between classification and discovery of association rules - a position paper. SIGKDD Explorations 2(1), 65–69 (2000)
13. Fürnkranz, J., Flach, P.A.: ROC 'n' rule learning—towards a better understanding of covering algorithms. Machine Learning 58(1), 39–77 (2005)
14. Garriga, G.C., Kralj, P., Lavrac, N.: Closed sets for labeled data. Journal of Machine Learning Research 9, 559–580 (2008)
15. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. 38(3) (2006)
16. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
17. Lavrač, N., Flach, P.A., Zupan, B.: Rule evaluation measures: A unifying view. In: Džeroski, S., Flach, P.A. (eds.) ILP 1999. LNCS (LNAI), vol. 1634, pp. 174–185. Springer, Heidelberg (1999)
18. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research 184(2), 610–626 (2008)
19. Liu, B., Hsu, W., Ma, Y.: Pruning and summarizing the discovered associations. In: Proc. Knowledge Discovery in Databases, pp. 125–134 (1999)
20. Luxenburger, M.: Implications partielles dans un contexte. Mathématiques et Sciences Humaines 29, 35–55 (1991)
21. Mutter, S., Hall, M., Frank, E.: Using classification to evaluate the output of confidence-based association rule mining. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 538–549. Springer, Heidelberg (2004)
22. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Proc. Knowledge Discovery in Databases, pp. 229–248 (1991)
23. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
24. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4), 293–313 (2004)
25. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A.F.M., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowl. Inf. Syst. 14(1), 1–37 (2008)
26. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In: Barbará, D., Kamath, C. (eds.) SDM. SIAM (2003)
27. Zimmermann, A.: Objectively evaluating interestingness measures for frequent itemset mining. In: Li, J., Cao, L., Wang, C., Tan, K.C., Liu, B., Pei, J., Tseng, V.S. (eds.) PAKDD 2013 Workshops. LNCS, vol. 7867, pp. 354–366. Springer, Heidelberg (2013)