A Bayesian criterion for evaluating the robustness of classification rules in binary data sets

Dominique Gay and Marc Boullé

Abstract Classification rules play an important role in prediction tasks. Their popularity is mainly due to their simple and interpretable form. Classification methods combining classification rules that are interesting (w.r.t. a defined interestingness measure) generally lead to good predictions. However, the performance of rule-based classifiers is strongly dependent on the interestingness measure used (e.g. confidence, growth rate, ...) and on the measure threshold to be set for differentiating interesting from non-interesting rules ; threshold setting is a non-trivial problem. Furthermore, it can be easily shown that the mined rules are individually non-robust: an interesting (e.g. frequent and confident) rule mined from the training set could be no more confident in a test phase. In this paper, we suggest a new criterion for the evaluation of the robustness of classification rules in binary labeled data sets. Our criterion arises from a Bayesian approach : we propose an expression of the probability of a rule given the data. The most probable rules are thus the rules that are robust. Our Bayesian criterion is derived from this defined expression and allows us to mark out the robust rules from a given set of rules without parameter tuning.

1 Introduction

Among the main data mining tasks, pattern mining has been extensively studied. Association rules [Agrawal et al., 1993] are one of the most popular patterns. In binary data sets, an association rule is an expression of the form $\pi : X \rightarrow Y$, where X (the body) and Y (the consequent) are subsets of Boolean attributes. Intuitively,

D. Gay · M. Boullé
Orange Labs
TECH/ASAP/PROFiling & data mining
2, avenue Pierre Marzin
F-22307 Lannion Cédex, FRANCE
e-mail: firstname.name@orange.com

the rule π means that *"when attributes of X are observed, then attributes of Y are often observed"*. The main interest of a rule pattern is its inductive inference power: from now on, if we observe the attributes of X then we will also probably observe attributes of Y. When Y is a class attribute we talk about classification rules. In this paper, we focus on such rules $X \rightarrow c$ (concluding on a class attribute c). Classification rules seem to be favorable for classification tasks (*if an object is described by attributes of X then it is probably of class c*). Recent advances in rule mining have given rise to many rule-based classification algorithms (see, e.g., pioneering work [Liu et al., 1998] or [Bringmann et al., 2009] for a survey). Existing rule-based methods are known for their interpretable form and also to perform quite well in classification tasks. However, we may point out at least two weaknesses:

The Curse of Parameters. The choice of parameter values is crucial but not trivial. The dilemma is well-known: a high frequency threshold may lead to less rules, but also lesser coverage rate and less discriminating power. A low frequency threshold may lead to a huge amount of rules, among which some rules (with low frequency) may be spurious. The same dilemma stands when thresholding interestingness measures like confidence (i.e. an estimation of the probability $P(c \mid X)$) or growth rate (which highlights the so-called emerging patterns, i.e. those patterns that frequent in a class of the data set and barely infrequent in the rest of the data [Dong and Li, 1999]): indeed, high confidence (or growth rate) threshold values lead to strong (pure) class association rules which may be rare in real-world data or even wrong when combined with a low frequency threshold whereas "low" thresholds generate a lot of rules with limited interest. Thus, finding a trade-off between frequency and interestingness measure values is not trivial.

Instability of interestingness measures. Even if subsets of extracted rules have shown to be quite effective for predictions, it can be easily shown that highly confident or emerging rules are not individually robust. In figure 1, we compare the confidence (resp. growth rate) train values with the confidence (resp. growth rate) test values of rules extracted from `UCI breast-w` data set [Asuncion and Newman, 2007]. We observe that confidence and growth rate values of extracted rules are clearly unstable from train to test data. The same observation arises when considering lift values: when $lift \geq 2$, then there is a positive correlation between the body of the rule and the class attribute. However, this correlation is not always confirmed in test phase. Thus, confidence, growth rate and lift do not allow us to determine whether a rule is robust: a "good" rule w.r.t. confidence (or growth rate) in training phase may turn out to be weak in test phase.

In this paper, we suggest a Bayesian criterion which allows us to mark out the extracted rules that are robust. Our approach benefits from the `MODL` framework [Boullé, 2006], provides a parameter-free criterion and does not need any wise thresholding. Notice that this paper is the extended English version of the French paper [Gay and Boullé, 2011] presented at EGC 2011 [Khenchaf and Poncelet, 2011].

The rest of the paper is organized as follows: section 2 briefly recalls some needed definitions and the main concepts of the `MODL` approach. Then, we describe

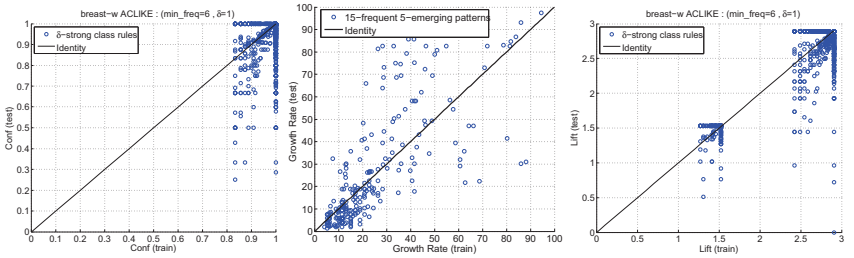

Fig. 1 Comparison of confidence (resp. growth rate and lift) values for classification rules in a train-test experiment: 50% train / 50% test for `breast-w` data set.

our extension of the `MODL` approach for classification rules and a Bayesian criterion for evaluating the robustness of rules. Section 3 reports the experiments we led to validate the proposed criterion. We, then discuss further related work in section 4. Finally, section 5 briefly concludes and opens several perspectives for future work.

2 From classification rules to `MODL` rules

Definitions. Let $r = \{\mathscr{T}, \mathscr{I}, \mathscr{C}, \mathscr{R}\}$ be a binary labeled data set, where $\mathscr{T}$ is a set of objects, $\mathscr{I}$ a set of Boolean attributes, $\mathscr{C}$ a set of classes and $\mathscr{R} : \mathscr{T} \times \mathscr{I} \mapsto \{0,1\}$ a binary relation such that $\mathscr{R}(t,a) = 1$ means object t contains attribute a. Every object $t \in \mathscr{T}$ is labeled by a unique class attribute $c \in \mathscr{C}$. A *classification rule* π in r is an expression of the form $\pi : X \to c$ where $X \subseteq \mathscr{I}$ is an itemset (i.e., a set of attributes), and $c \in \mathscr{C}$ a class attribute. The *frequency* of itemset (i.e. a set of attributes) X in r is $freq(X,r) = |\{t \in \mathscr{T} \mid \forall a \in X : \mathscr{R}(t,a)\}|$ and the *frequency* of π is $freq(\pi,r) = freq(X \cup \{c\})$. The *confidence* of π in r is $conf(\pi,r) = freq(\pi,r)/freq(X,r)$. The *growth rate* of π is $GR(\pi,r) = freq_r(X,r_c)/freq_r(X,r \setminus r_c)$ where r_c is the data set r restricted to objects of class c ($\mathscr{T}_c$) and $freq_r$ stands for *relative frequency* (i.e. $freq_r(X,r_c) = freq(X,r_c)/|\mathscr{T}_c|$).

The pioneering works in classification based on association rules (i.e. the `CBA`-like methods, e.g., [Dong et al., 1999, Li et al., 2001, Liu et al., 1998]) state that a rule is interesting for classification if its frequency and confidence (or growth rate) exceed user-defined thresholds. Setting good thresholds may be a hard task for an end-user, therefore low thresholds are arbitrarily set – generating a huge number of rules. Then, a subset of extracted rules is selected in a post-processing phase w.r.t. coverage, redundancy, correlation (e.g. by choosing the k best rules or using the χ^2 test). Therefore, other non-trivial parameter tuning skills are needed.

In this paper, we suggest to follow the `MODL` approach to evaluate classification rules. The `MODL` approach, already used for values grouping [Boullé, 2005], discretization [Boullé, 2006], regression [Hue and Boullé, 2007] or decision trees [Voisine et al., 2010], bets on a trade-off between, *(i)* the fineness of the predictive

information provided by the model and *(ii)* the robustness, in order to obtain a good generalization of the model. In our context, from a MODL point of view, a model is classification rule. To choose the best rule model, we use a Bayesian approach: we look for maximizing $p(\pi \mid r)$ the posterior probability of a rule model π given the data r. Applying the Bayes theorem and considering the fact that the probability $p(r)$ is constant for a given classification problem, then the expression $p(\pi) \times p(r \mid \pi)$ is to be maximized ; where $p(\pi)$ is the prior probability of a rule and $p(r \mid \pi)$, the likelihood, is the conditional probability of the data given the rule model π. Thus, the rule π maximizing this expression, is the most probable rule arising from the data. Our evaluation criterion is based on the negative logarithm of $p(\pi \mid r)$, which we call the *cost* of the rule:

$$c(\pi) = -\log(p(\pi) \times p(r \mid \pi))$$

In order to compute the prior probability $p(Rule)$ of the MODL criterion, we propose a definition of a classification rule based on a hierarchy of parameters that uniquely identifies a given rule.

Standard Classification Rule Model. A MODL rule (also called *standard classification rule model* (SCRM)) is defined by:

- the constituent attributes of the rule body
- for each attribute of the rule body, the value (0 or 1) that belongs to the body
- the distribution of classes inside and outside of the body

The two last key points of the SCRM definition lead us to a notion of rule that extend the "classical" association and classification rule. Indeed, for a given binary attribute a, the values 0 and 1 are two possible values belonging to the body. This may be related to the notion of rules with negations of attributes in their body (see [Antonie and Zaïane, 2004]). SCRM is also related to the recently introduced *distribution rule* [Jorge et al., 2006]. The consequent of such a rule is a probabilistic distribution over the classes (instead of being a class value). The following example illustrates these two differences.

Example of SCRM. Let us consider the rule $\pi : (A_1 = 0) \wedge (A_2 = 1) \wedge (A_4 = 1) \rightarrow (P_{c_1} = 0.9, P_{c_2} = 0.1)$. Describing the body of such a rule consists in choosing the attributes involved in the body, then choosing the values (0 or 1) of the involved attributes. Notice that a classification rule with negations might be trivially derived from a SCRM using the class with maximum probability as the consequent. For example, $\pi : (A_1 = 0) \wedge (A_2 = 1) \wedge (A_4 = 1) \rightarrow c_1$.

To formally define our evaluation criterion we will use the following additional notations:

Notations. Let r be a binary labeled data set with N objects, m binary attributes and J classes. For a SCRM, $\pi : X \rightarrow (P_{c_1}, P_{c_2}, \ldots, P_{c_J})$ such that $|X| = k \leq m$, we will use the following notations:

- $X = \{x_1, \ldots, x_k\}$: the constituent attributes of the rule body ($k \leq m$)
- $i_{x_1}, \ldots, i_{x_k}$: the indexes of binary values involved in the rule body
- $N_X = N_{i_{x_1} \ldots i_{x_k}}$: the number of objects in the body $i_{x_1} \ldots i_{x_k}$
- $N_{\neg X} = N_{\neg i_{x_1} \ldots i_{x_k}}$: the number of objects outside of the body $i_{x_1} \ldots i_{x_k}$
- $N_{Xj} = N_{i_{x_1} \ldots i_{x_k} j}$: the number of objects of class j in the body $i_{x_1} \ldots i_{x_k}$
- $N_{\neg Xj} = N_{\neg i_{x_1} \ldots i_{x_k} j}$: the number of objects of class j outside of the body $i_{x_1} \ldots i_{x_k}$

`MODL` hierarchical prior. We use the following distribution prior on SCRM models, called the `MODL` hierarchical prior, to define the prior $p(\pi)$.

- *(i)* the number of attributes in the rule body is uniformly distributed between 0 and m
- *(ii)* for a given number k of attributes, every set of k constituent attributes of the rule body is equiprobable
- *(iii)* for a given attribute value, belonging to the body or not are equiprobable
- *(iv)* the distributions of class values in and outside of the body are equiprobable
- *(v)* the distributions of class values in and outside of the body are independent

Thanks to the definition of the model space and its prior distribution, we now apply the Bayes theorem to express the prior probabilities of the model and the probability of the data given the model (i.e. $p(\pi)$ and $p(r \mid \pi)$).
The prior probability $p(\pi)$ of the rule model is:

$$p(\pi) = p(X) \times \prod_{1 \leq l \leq k} p(i_{x_l}) \times \prod_{i \in \{X, \neg X\}} p(\{N_{ij}\} \mid N_X, N_{\neg X})$$

Firstly, we consider $p(X)$ (the probability of having to the attributes of X in the rule body). The first hypothesis of the hierarchical prior is the uniform distribution of the number of constituent attributes between 0 and m. Furthermore, the second hypothesis says that every set of k constituent attributes of the rule body is equiprobable. The number of combinations $\binom{m}{k}$ could be a natural way to compute this prior; however, it is symmetric. Beyond $m/2$, adding new attributes make the selection more probable. Thus, adding irrelevant variables is favored, provided that this has an insignificant impact on the likelihood of the model. As we prefer simpler models, we suggest to use the number of combinations with replacement $\binom{m+k-1}{k}$. Using the two first hypothesis, we have:

$$p(X) = \frac{1}{m+1} \cdot \frac{1}{\binom{m+k-1}{k}}$$

For each attribute x part of the body of the rule, the value involved in the body has to be chosen from $\{0,1\}$. Thus we have $p(i_x) = 1/2$ (considering hypothesis *(iii)*).
Now considering hypothesis *(iv)* and *(v)*, enumerating the distributions of the J classes in and outside of the body turns into a combinatorial problem:

$$p(\{N_{Xj}\} \mid N_X, N_{\neg X}) = \frac{1}{\binom{N_X+J-1}{J-1}}$$

$$p(\{N_{\neg X\,j}\} \mid N_X, N_{\neg X}) = \frac{1}{\binom{N_{\neg X}+J-1}{J-1}}$$

Concerning the likelihood term, the probability of the data given the model is the probability of observing the data inside and outside of the rule body (with resp. N_X and $N_{\neg X}$ objects) given the multinomial distribution defined for N_X and $N_{\neg X}$. We have:

$$p(r \mid \pi) = \frac{1}{\frac{N_X!}{\prod_{j=1}^{j=J} N_{X,j}!}} \cdot \frac{1}{\frac{N_{\neg X}!}{\prod_{j=1}^{j=J} N_{\neg X,j}!}}$$

We now have a complete definition of the cost a MODL rule (SCRM) π:

$$c(\pi) = \log(m+1) + \log\binom{m+k-1}{k} + k\log(2) \tag{1}$$

$$+ \log\binom{N_X+J-1}{J-1} + \log\binom{N_{\neg X}+J-1}{J-1} \tag{2}$$

$$+ \left(\log N_X! - \sum_{j=1}^{j=J} \log N_{X,j}!\right) + \left(\log N_{\neg X}! - \sum_{j=1}^{j=J} \log N_{\neg X,j}!\right) \tag{3}$$

The cost of the rule is made of negative logarithms of probabilities ; according to [Shannon, 1948], this transformation links probabilities with code length. Thus, $c(\pi)$ might be seen as the ability of a MODL rule to encode the classes given the attributes. The first line stands for the choice of the number of attributes, the attributes and the values involved in the rule body. The second line corresponds to the class distribution in and outside of the body. The two last lines stand for the likelihood (the probability of observing the data given the rule).

Intuitively, rules with low MODL cost are the most probable and thus the best ones. Notice that $c(\pi)$ is smaller for lower k values (cf eq. 1), i.e. rules with shorter bodies are more probable thus preferable. Consequently, frequent rules are more probable than non-frequent ones – that meets the obvious fact. From $c(\pi)$ expression again (two last lines), the notion of pureness (fineness) arises: the stronger rules are cheaper w.r.t. c, thus are the best ones. Since the magnitude of the MODL cost of rules depends on the size of the data set (i.e. the number of objects N and the number of attributes m), we define a normalized criterion (noted *level*[1]) to compare two MODL rules:

$$level(\pi) = 1 - \frac{c(\pi)}{c(\pi_\emptyset)}$$

where $c(\pi_\emptyset)$ is the MODL cost for the default rule (i.e. with empty body). Intuitively, $c(\pi_\emptyset)$ is the coding length of the classes when no information is used from the attributes. The cost of the default rule $\pi_\emptyset$ is formally:

[1] The *level* may also be seen as a compression rate.

$$c(\pi_\emptyset) = \log(m+1) + \log\binom{N+J-1}{J-1} + \log N! - \sum_{j=1}^{j=J} \log N_j!$$

That way, for a given rule π, if $level(\pi) = 0$ then π has the same cost as $\pi_\emptyset$; thus π is not more probable than the default rule. When $level(\pi) < 0$, then using π to explain the data is more costly than using the empty rule. In other words, π is less probable than $\pi_\emptyset$ and will not be considered as interesting. The cases where $0 < level(\pi) \leq 1$ highlight the interesting classification rules π. Indeed, rules with lowest cost (and high *level*) are the most probable and show correlations between the rule body and the class attribute. Notice that $level(\pi) = 1$ is the particular case where π (on its own) is sufficient to exactly characterize the class distribution.

We argue that the level allows us to identify the robust and interesting classification rules. In the following, we lead several experiments to support our point of view.

3 Experimentations

In this section, we lead several experiments to show *(i)* that confidence and growth rate are generally unstable from train to test phase and thus are not good candidates to capture the robustness of classification rules ; *(ii)* that, conversely, the *level* is stable in the same experimental conditions and *(iii)* that the *level* allows us to naturally identify robust and interesting rules.

3.1 Experimental protocol

In our experiments, we use seven `UCI` data sets [Asuncion and Newman, 2007] and a real-world data set (meningite) [François et al., 1992]. A brief description of these data sets is given in table 1.

Data set	#Objects	#Attributes	#classes and distribution
breast-w	699	9	458/241
credit-a	690	15	307/383
credit-g	1000	21	700/300
diabetes	768	8	500/268
meningite	329	23	245/84
sonar	208	60	97/111
tic-tac-toe	958	9	626/332
vote	435	17	267/168

Table 1 Experimental data sets description

The train-test experiments consist in dividing a data set in two (almost) equal class-stratified parts. One part is for training and mining frequent-confident (or emerging) rules, the other part is for evaluating the evolution of confidence and growth rate values on the test set. Since we do not provide an extractor of MODL rules in this preliminary work, we compute the value of our MODL criterion for the extracted confident (or emerging) rules on the training and test set for comparison. We use AClike prototype [Boulicaut et al., 2003] to mine frequent-confident classification rules: in fact, AClike mines γ-frequent δ-free itemsets that are bodies of rules π with $conf(\pi, r) \geq 1 - \delta/\gamma$. We also use consepminer prototype [Dong and Li, 1999, Zhang et al., 2000] to mine γ-frequent ρ-emerging patterns.

3.2 Experimental results

Original data sets. In figures 2 and 3, we report scatter plots for the study of the evolution (from train set to test set) of confidence values of extracted rules. We also compare the values of the MODL criterion. As expected, for all data sets, we observe that confidence is unstable from train to test: indeed, a highly confident rule in train may have low confidence in test (see the points far from the identity line). Conversely, the MODL level values of extracted rules are rather stable in the train-test experiments (see the points close to the identity line). A similar experimentation is reported in figures 4, 5 and the same conclusions stand: growth rate values are unstable in a train-test experiment whereas MODL level values of extracted emerging pattern remain stable.

These experiments show that it could be risky to rely on confidence or growth rate values to make predictions since they do not capture the notion of robustness. The stability of the MODL level is a sign of robustness; in the following experiments, we show that patterns with negative level values are non-significant and the ones with positive level values are patterns of interest.

Noisy data sets. In order to simulate the presence of class-noise in the breast-w data set, we add uniform noise in the class attribute using the AddNoise function of WEKA [Witten and Frank, 2005] – with various ratio: 20% and 50% amount of noisy class labels. We then proceed the train-test experiments on each artificially noisy data set. For each amount of noise (see in figure 6), classical extractors (frequent-confident rules and emerging patterns miners) succeed in outputing a set of "potentially" interesting patterns – notice that less rules arise from the most noisy contexts. However, once again the train-test experiments show the instability of classical measures. Moreover, the instability is emphasized in noisy contexts; indeed, most of the points (rules) in the scatter plots (and all rules for 50% of noise) are under the identity line, which means confidence and growth rate are wrongly optimistic and may lead to bad predictions. As an example, several rules confidence fall under 0.5 in the test set – which is contradictory.
The *level* criterion of extracted patterns is still stable in noisy contexts. Notice that

most of the confident or emerging rules in noisy contexts has a negative level. As we mentioned above, a rule with a negative level is less probable than the default rule and thus is not statistically significant, i.e. not interesting. In the last experiments, we show that a positive level indicates that a rule is interesting.

Patterns with positive level. In figures 7 and 8, we report the train and test values of a class-entropy-based measure μ (defined below) for the extracted rules π:

$$\mu(\pi) = N \times (Ent(\pi_\emptyset) - Ent(\pi))$$

μ measures the difference between the conditional entropies of the null rule model (default rule) and a given rule π. The higher μ, the more interesting π is. μ may be seen as the number of bits saved when compressing the data using π instead of using $\pi_\emptyset$. In figures 7 and 8, we highlight the rules with a positive MODL level (red 'o'). As expected, rules with a positive level are generally the most interesting, i.e. with higher μ values. Consequently, rules with a negative level (blue '+') value are located in the southwest of the graphs, with low μ values.

4 Related Work & discussions

The MODL approach [Boullé, 2005, Boullé, 2006] and the *level* criterion are at the crossroads of Bayes theory, Minimum Description Length principle (MDL [Grünwald, 2007]) and Kolmogorov complexity [Li and Vitányi, 2008].

About MDL. In [Siebes et al., 2006], the authors develop a MDL-based pattern mining approach. The authors look for itemsets that provides a good compression of the data. The link between probability and codes allow them to rewrite the code length of an item set I as $-log(P(I))$. Thus, the best item sets have shortest codes. In [van Leeuwen et al., 2006], an extension for classification purpose is suggested. The two main differences with the MODL approach are : *(i)* the use of the MODL hierarchical prior implies a different way of coding information ; *(ii)* in [van Leeuwen et al., 2006], authors look for a set of patterns to compress the data whereas our MODL criterion is defined for *one* rule.
Notice that another recent work embraces the MDL principle for classification rule discovery: in [Suzuki, 2009], the author suggests an extended version of MDL to integrate user knowledge (in the form of a partial decision list). The code length cl of the partial decision list L to be discovered from data D is extended with the user knowledge K and serves as a subjective interestingness measure: $cl(L) \equiv -\log P(L) - \log P(D \mid L) - \log P(K \mid L)$.

About robustness. The *level* criterion has shown to be stable. Thus, we may rely on classification rules with positive *level* values since the interestingness of the rules will be confirmed in a test phase. The notion of robustness has been studied recently:

$X \to c$	c	$\neg c$	$\sum$
X	$freq(Xc,r)$	$freq(X\neg c,r)$	$freq(X,r)$
$\neg X$	$freq(\neg Xc,r)$	$freq(\neg X\neg c,r)$	$freq(\neg X,r)$
$\sum$	$\lvert c\rvert$	$\lvert\neg c\rvert$	N

Table 2 Contingency table for a classification rule $X \to c$

in [Le Bras et al., 2010], the authors suggest a new notion of robustness dependent on an interestingness measure μ and a threshold μ_{min}. Starting from the observation that a rule can be characterized by a $\mathbb{R}^3$-vector of three values of its contingency table (e.g. the frequency of the body, the frequency of the target and the number of counterexamples; see figure 2), the authors define the robustness of rule π as the normalized Euclidean distance $rob(\pi,\mu_{min}) = ||\pi - \pi^*||_2/\sqrt{3}$ between π and a *limit rule* π^* (i.e. a rule minimizing $g(\pi') = ||\pi' - \pi_{min}||_2$ where π_{min} is such that $\mu(\pi_{min}) = \mu_{min}$). In such framework, comparing two rules in terms of robustness does not need any thresholding, however for filtering purpose (e.g., selection of a subset of robust rules) another non-trivial parameter (*rob*) has to be set (in addition with frequency and the current measure thresholds).

About redundancy. A classification rule $\pi_2 : Y \to c_i$ is said to be redundant w.r.t. $\pi_1 : X \to c_j$ if $c_i = c_j$, $X \subseteq Y$ and π_1 and π_2 brings (almost) the same class-discriminating power (w.r.t. an interestingness measure) – a redundant rule should be pruned. Consider two itemsets X and Y such that $X \subseteq Y$ and $freq(X,r) = freq(Y,r)$, then for a given interestingness measure m based on frequency, we have $m(X) = m(Y)$ thus some redundancy. It is common to consider support equivalence class to group itemsets having the same support (and frequency). The unique longest itemset (w.r.t. set inclusion) is the closed itemset [Pasquier et al., 1999] and the smallest ones are called the free itemsets [Boulicaut et al., 2003]. In state-of-the-art pattern-based methods for classification purpose, the intuition tells that free itemsets should be preferred [Baralis and Chiusano, 2004]. This intuition is confirmed by our *level* criterion. Indeed, if Y is a closed itemset and X a free itemset from the same support equivalence class, then $c(\pi_2 : Y \to c_i) \geq c(\pi_1 : Y \to c_i)$ since the number of attributes favors π_1 (line 1-2); and π_1 should be preferred. The main idea is translated in the following proposition (the proof is almost direct when one observes that only the terms of the cost expression that involve parameter k imply a difference of *level* between X and Y):

Proposition 1. *Let X and Y be two itemsets such that $X \subset Y$ and $freq(X,r) = freq(Y,r)$. X is preferable to Y according to the level criterion; i.e., $level(X) > level(Y)$.*

5 Conclusion & perspectives

In this paper, we have presented a new Bayesian criterion for the evaluation of classification rules in binary data sets. Based on the MODL approach (and the MDL principle), the new criterion overcomes two well-known drawbacks of existing approaches (using a frequency-confidence or growth rate framework): the non-trivial tuning of interestingness measure threshold and the non-stability of interestingness measure values from train to test phase. Our new criterion, the MODL level, promotes a trade-off between fineness and reliability and allows us to easily distinguish interesting rules (with a positive level value) from non-significant rules (with a negative level value) without parameter tuning. Furthermore, the criterion is shown to be robust and gives a true idea of the prediction power of extracted patterns. The experiments we led on UCI data sets confirm both the relevancy and robustness of the criterion. In this preliminary work, we use the MODL criterion in a post-processing step to select interesting and robust rules from a large set of confident or emerging rules. The next step is a constructive approach for mining classification rules with positive MODL level values. Since the MODL approach is also suitable for continuous and nominal attributes as well, another step will be the extension towards quantitative association rules by considering discretization and values grouping.

References

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings ACM SIGMOD'93*, pages 207–216.

[Antonie and Zaïane, 2004] Antonie, M.-L. and Zaïane, O. R. (2004). An associative classifier based on positive and negative rules. In *DMKD'04*.

[Asuncion and Newman, 2007] Asuncion, A. and Newman, D. (2007). UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[Baralis and Chiusano, 2004] Baralis, E. and Chiusano, S. (2004). Essential classification rule sets. *ACM Transactions on Database Systems*, 29(4):635–674.

[Boulicaut et al., 2003] Boulicaut, J.-F., Bykowski, A., and Rigotti, C. (2003). Free-sets : A condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1):5–22.

[Boullé, 2005] Boullé, M. (2005). A bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 6:1431–1452.

[Boullé, 2006] Boullé, M. (2006). MODL: A bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165.

[Bringmann et al., 2009] Bringmann, B., Nijssen, S., and Zimmermann, A. (2009). Pattern-based classification: A unifying perspective. In *LeGo'09 workshop co-located with EMCL/PKDD'09*.

[Dong and Li, 1999] Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings KDD'99*, pages 43–52. ACM Press.

[Dong et al., 1999] Dong, G., Zhang, X., Wong, L., and Li, J. (1999). CAEP : Classification by aggregating emerging patterns. In *Proceedings DS'99*, volume 1721 of *LNCS*, pages 30–42. Springer.

[François et al., 1992] François, P., Crémilleux, B., Robert, C., and Demongeot, J. (1992). MENINGE: a medical consulting system for child's meningitis study on a series of consecutive cases. *Artificial Intelligence in Medecine*, 4(4):281–292.

[Gay and Boullé, 2011] Gay, D. and Boullé, M. (2011). Un critère bayésien pour évaluer la robustesse des règles de classification. In *EGC'11*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 539–550. Hermann-Éditions.

[Grünwald, 2007] Grünwald, P. (2007). *The minimum description length principle*. MIT Press.

[Hue and Boullé, 2007] Hue, C. and Boullé, M. (2007). A new probabilistic approach in rank regression with optimal bayesian partitioning. *Journal of Machine Learning Research*, 8:2727–2754.

[Jorge et al., 2006] Jorge, A. M., Azevedo, P. J., and Pereira, F. (2006). Distribution rules with numeric attributes of interest. In *PKDD'06*, pages 247–258.

[Khenchaf and Poncelet, 2011] Khenchaf, A. and Poncelet, P., editors (2011). *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*. Hermann-Éditions.

[Le Bras et al., 2010] Le Bras, Y., Meyer, P., Lenca, P., and Lallich, S. (2010). A measure of robustness of association rules. In *ECML/PKDD'10*, volume 6322 of *LNCS*, pages 227–242. Springer.

[Li and Vitányi, 2008] Li, M. and Vitányi, P. M. B. (2008). *An Introduction to Kolmogorov Complexity and Its Applications (3rd edition)*. Springer.

[Li et al., 2001] Li, W., Han, J., and Pei, J. (2001). CMAR: Accurate and efficient classification based on multiple class-association rules. In *Proceedings ICDM'01*, pages 369–376. IEEE Computer Society.

[Liu et al., 1998] Liu, B., Hsu, W., and Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings KDD'98*, pages 80–86. AAAI Press.

[Pasquier et al., 1999] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.

[Siebes et al., 2006] Siebes, A., Vreeken, J., and van Leeuwen, M. (2006). Item sets that compress. In *SIAM DM'06*.

[Suzuki, 2009] Suzuki, E. (2009). Negative encoding length as a subjective interestingness measure for groups of rules. In *PAKDD'09*, pages 220–231.

[van Leeuwen et al., 2006] van Leeuwen, M., Vreeken, J., and Siebes, A. (2006). Compression picks item sets that matter. In *PKDD'06*, pages 585–592.

[Voisine et al., 2010] Voisine, N., Boullé, M., and Hue, C. (2010). A bayes evaluation criterion for decision trees. In *Advances in Knowledge Discovery and Management [Best of EGC 2009]*, volume 292 of *Studies in Computational Intelligence*, pages 21–38. Springer.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques (2nd edition)*. Morgan Kaufmann.

[Zhang et al., 2000] Zhang, X., Dong, G., and Ramamohanarao, K. (2000). Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *KDD'00*, pages 310–314.

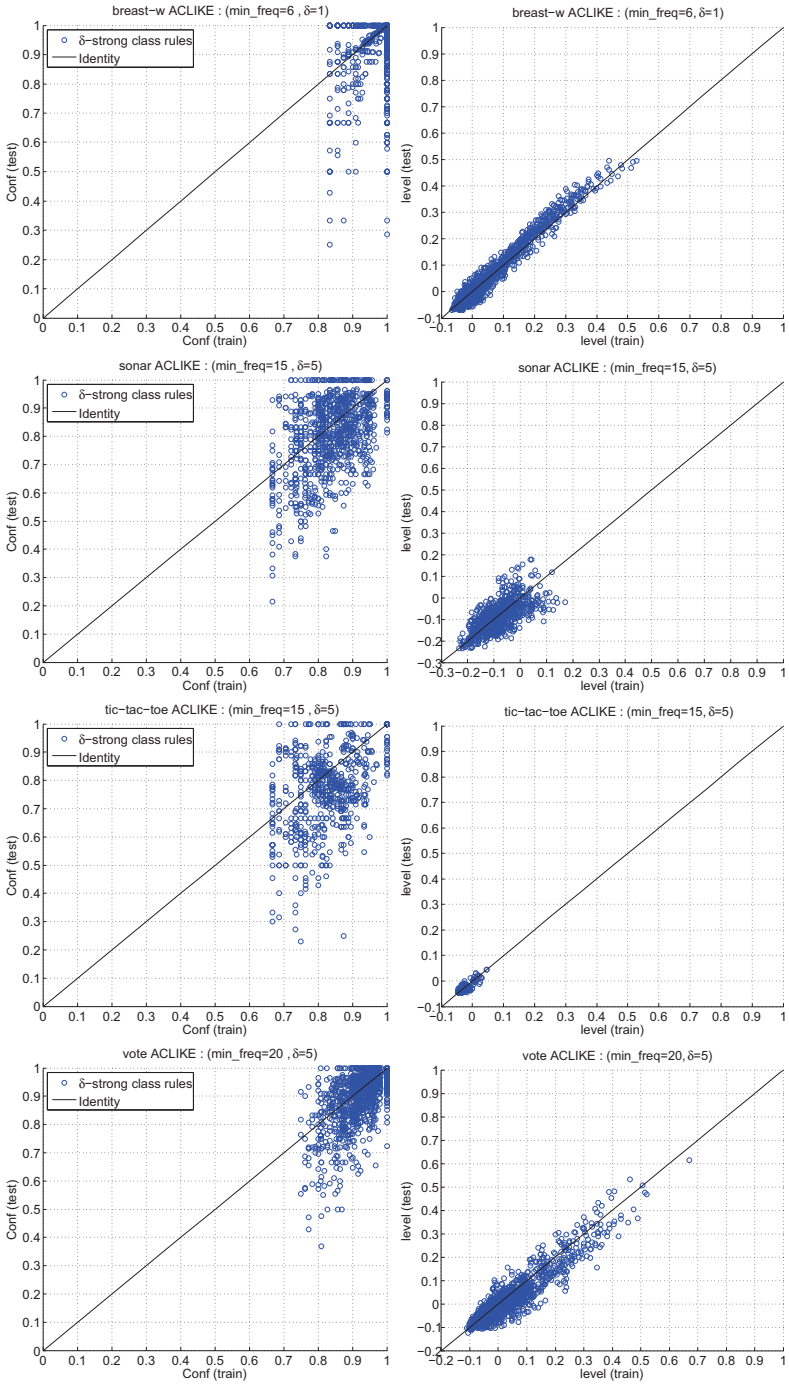

Fig. 2 Comparison of *confidence* and *level*: train values vs test values. Confidence is unstable from train to test phase while *level* values are clearly stable (points close to the identity line) – ensuring the robustness of the criterion.

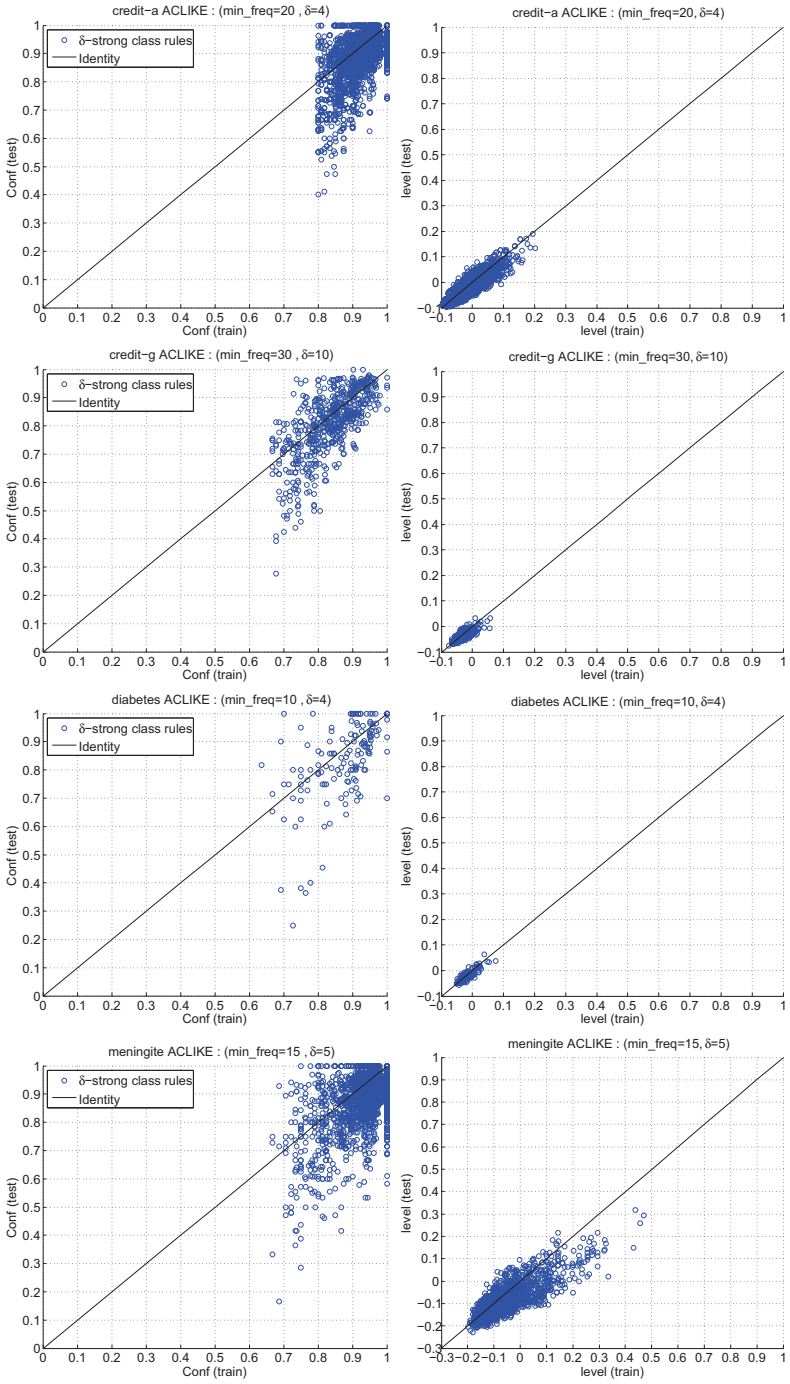

Fig. 3 Comparison of *confidence* and *level*: train values vs test values

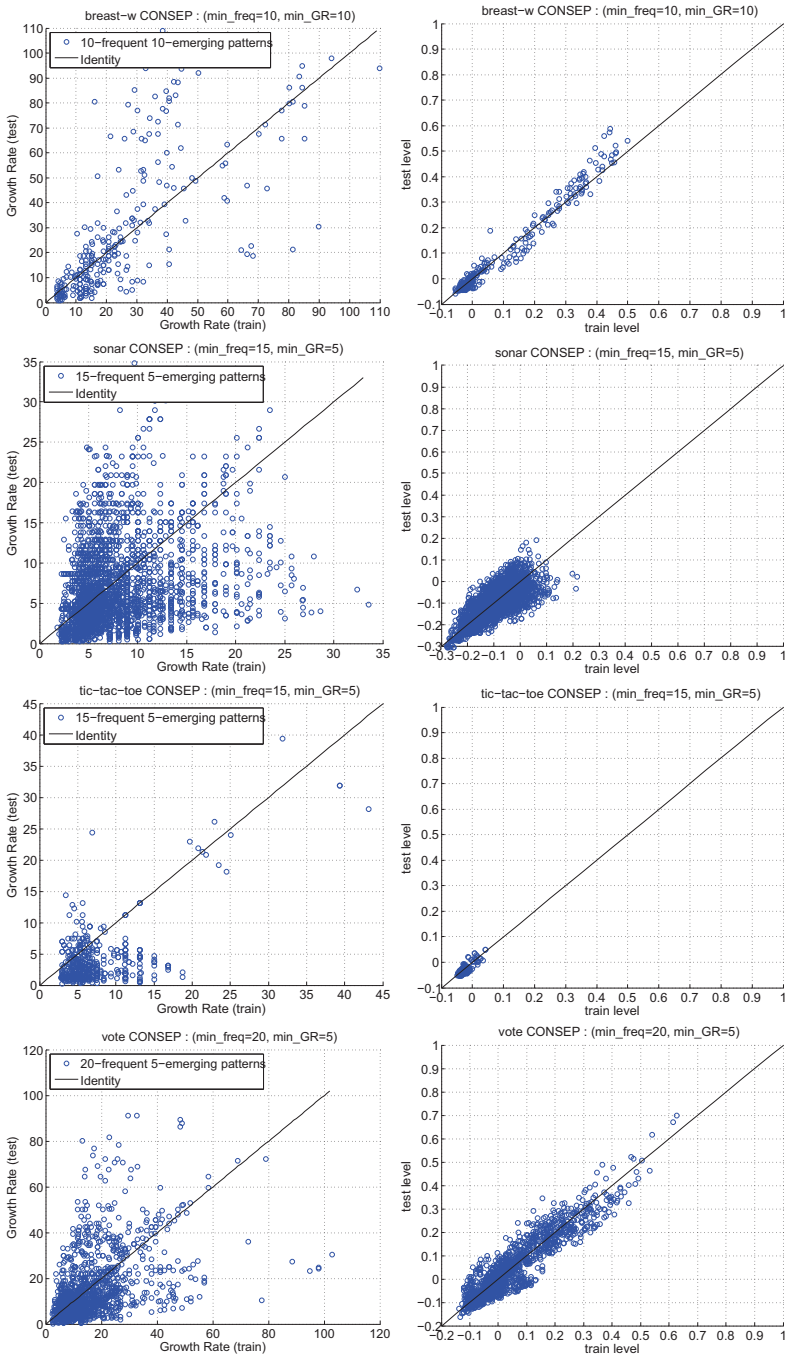

Fig. 4 Comparison of *GR* and *level*: train values vs test values. Growth rate shows instability in train-test experiments while *level* still remains stable.

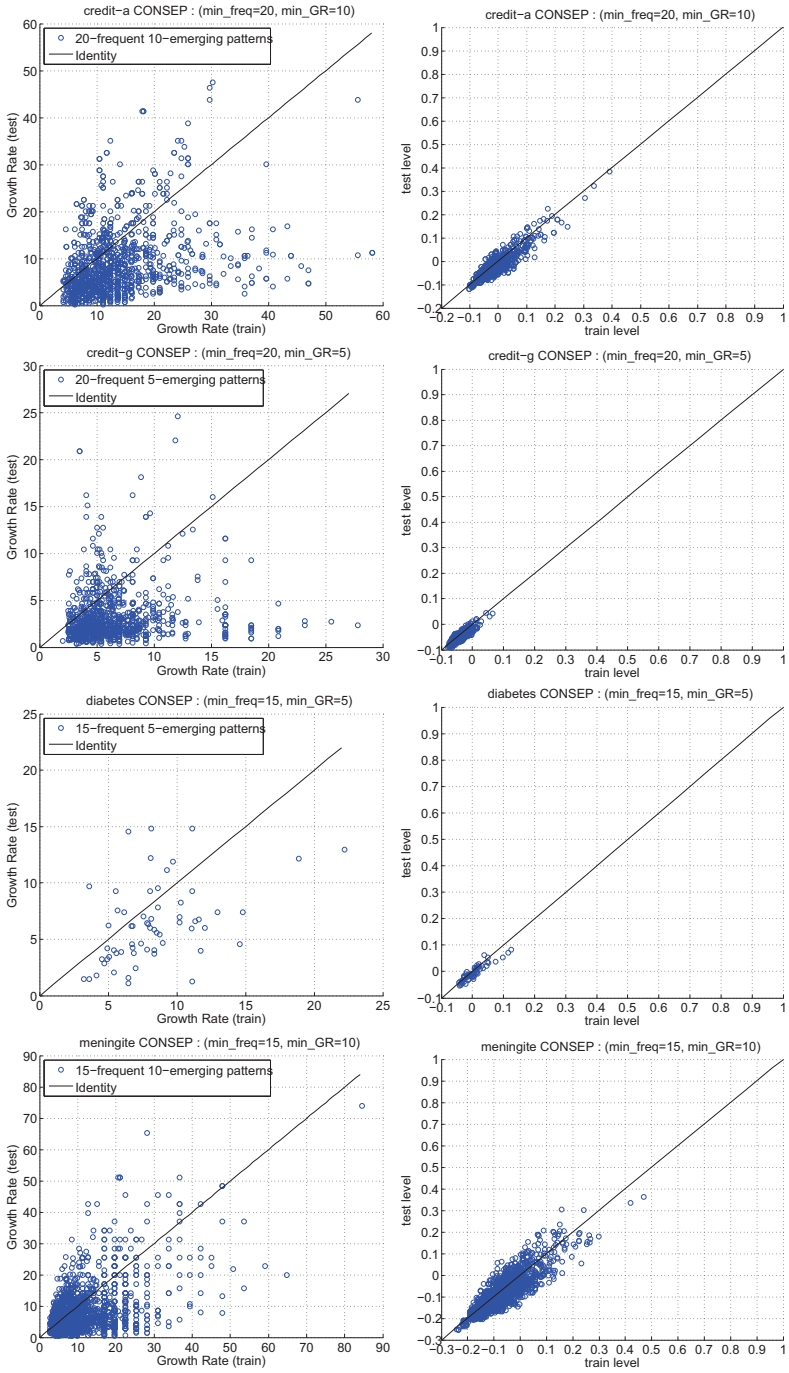

Fig. 5 Comparison of *GR* and *level*: train values vs test values

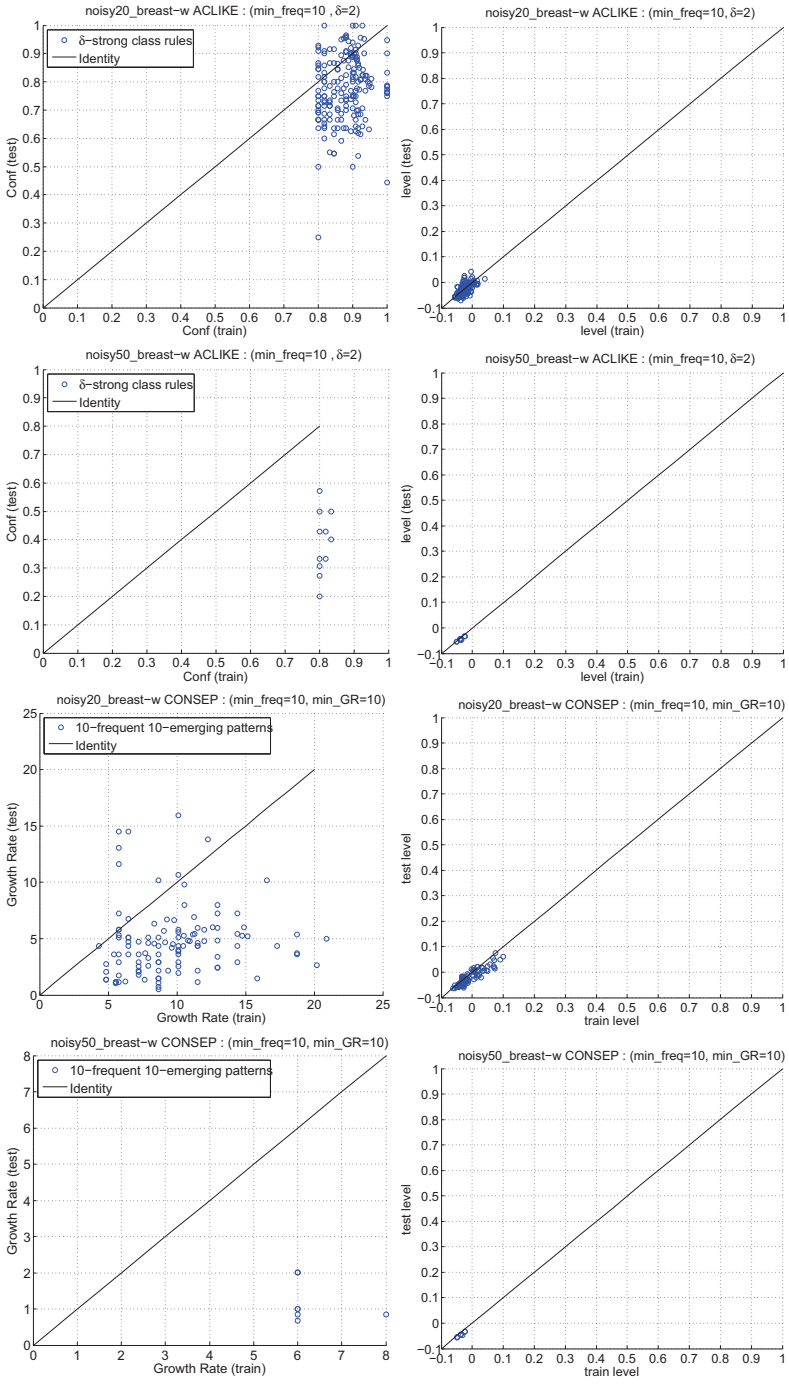

Fig. 6 Comparison *GR*, *confidence* and *level* in artificially noisy `breast-w` data set: train values vs test values. Potentially interesting rules w.r.t. confidence (or growth rate), that are actually 'wrong' in highly noisy environment, have a negative *level* value.

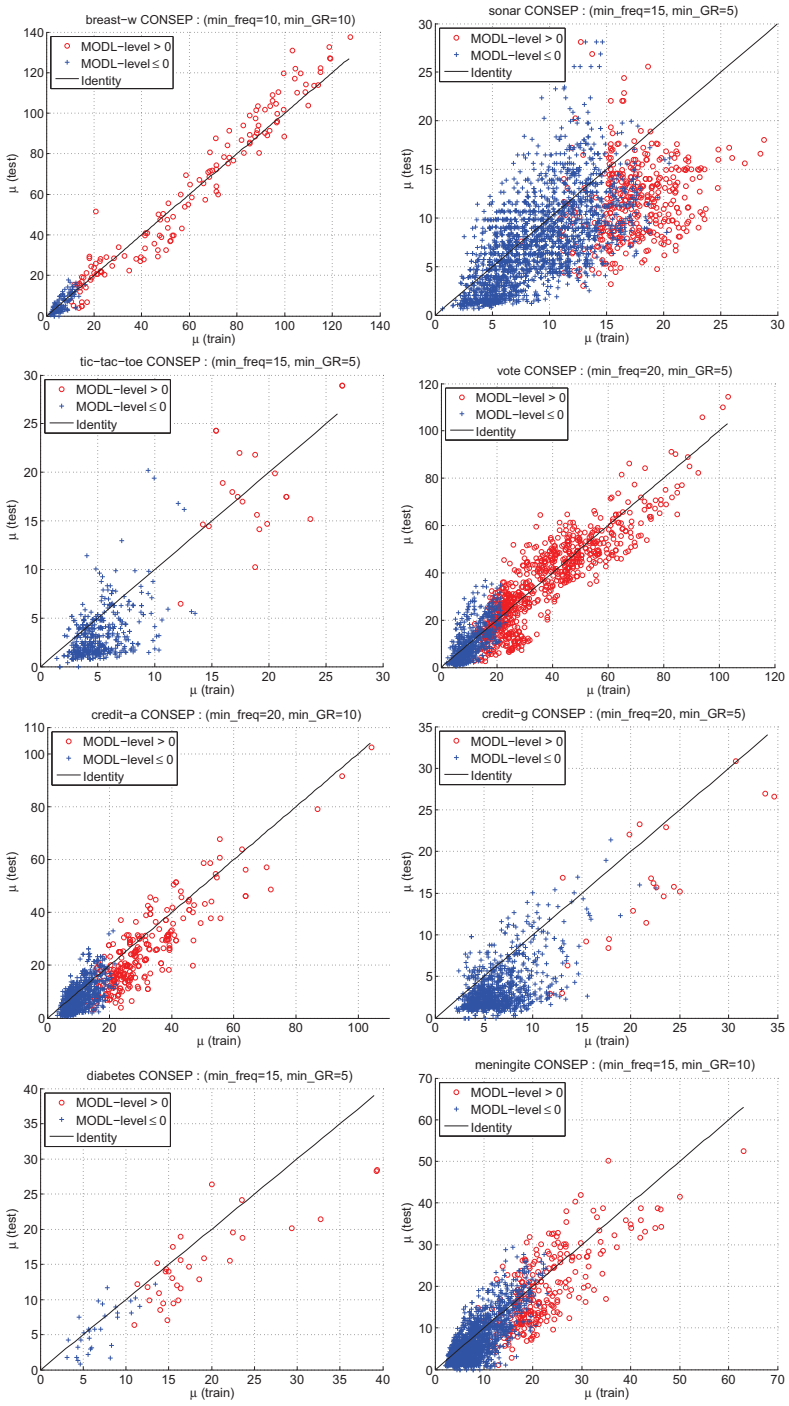

Fig. 7 Comparison μ : train values vs test values of emerging rules. The best rules (i.e. the most probable ones with a positive *level* value, red 'o') are generally located at the north-east of the graph whereas non-robust one (with negative *level* value, blue '+') are close to the origin.

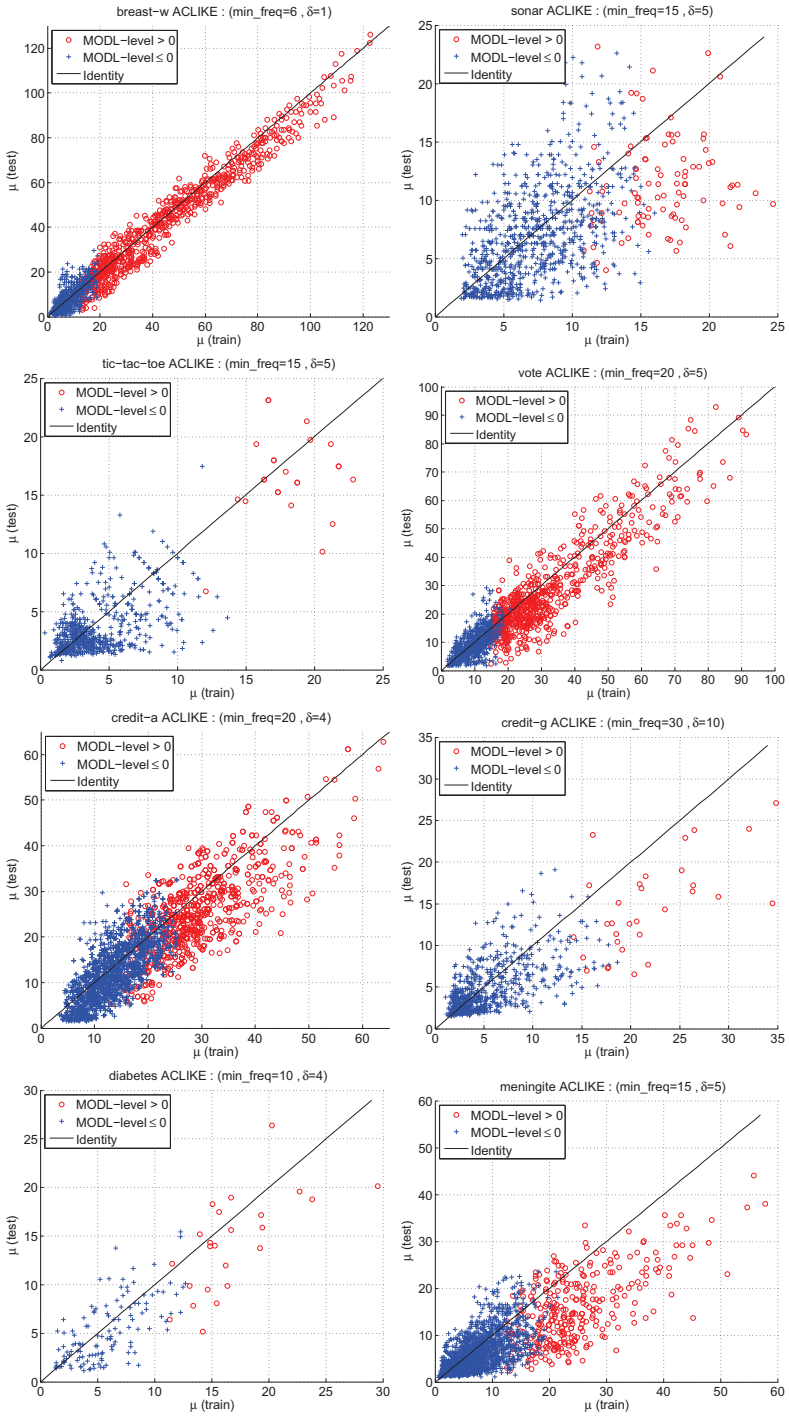

Fig. 8 Comparison μ : train values vs test values of confident rules