Comparing Rule Measures for Predictive Association Rules [⋆]

Paulo J. Azevedo[1] and Alípio M. Jorge[2]

[1] Departamento de Informática, Universidade do Minho, Portugal pja@di.uminho.pt
[2] LIAAD, Faculdade de Economia, Universidade do Porto, Portugal
amjorge@fep.up.pt

Abstract. In this paper we study the predictive ability of some association rule measures typically used to assess descriptive interest. Such measures, namely conviction, lift and χ^2 are compared with confidence, Laplace, mutual information, cosine, Jaccard and ϕ-coefficient. As prediction models, we use sets of association rules generated as such. Classification is done by selecting the best rule, or by weighted voting (according to each measure). We performed an evaluation on 17 datasets with different characteristics and conclude that conviction is on average the best predictive measure to use in this setting.

1 Introduction

Association rule mining is a technique primarily used for exploratory data mining. In such a setting it is useful to discover relations between sets of variables, which may represent products in an on-line store, disease symptoms, keywords, demographic characteristics, to name a few. To guide the data analyst identifying interesting rules, many objective interestingness rule measures have been proposed in the literature [17]. Although these measures have descriptive aims, we will evaluate their use in predictive tasks. One of these measures, conviction, will be shown as particularly successful in classification.

Classification based on association rules has been proved as very competitive [12]. The general idea is to generate a set of association rules with a fixed consequent (involving the class attribute) and then use subsets of these rules to classify new examples. This approach has the advantage of searching a larger portion of the rule version space, since no search heuristics are employed, in contrast to decision tree and traditional classification rule induction. The extra search is done in a controlled manner enabled by the good computational behavior of association rule discovery algorithms. Another advantage is that the produced rich rule set can be used in a variety of ways without relearning, which can be used to improve the classification accuracy [8].

In this work, we study the predictive power of many of the known interestingness measures. Although this is done in terms of association rule based classification, the results can be potentially useful to other classification settings.

[⋆] Supported by Fundação Ciência e Tecnologia, Project Site-o-matic, FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

We start by describing in detail the classification approach used, and define each of the used measures, previously appraised for descriptive data mining tasks in [10][17]. We perform a thorough experimental validation and study the results using Best Rule and Weighted Voting prediction implemented in the *CAREN* system [1].

1.1 Classifying with Association Rules

The classification approach we describe in this paper consists in obtaining a classifier, or a discriminant model M, from a set of association rules. The rules are generated from a particular propositional data set, involve categorical and numerical $< attribute = value >$ pairs in the antecedent and a class value in the consequent. We want the model M to be successful in the prediction of the classes of unseen cases taken from the same distribution as D. A Bayesian view of the success of a classifier defines that the optimal classifier M_{Bayes} maximizes the probability of predicting the correct class value for a given case x [7].

Previous work on classification from association rules has confirmed the predictive power of confidence [12]. In this paper we provide empirical indication that another measure, *conviction*, tends to obtain better results.

2 The Measures

In this section we describe the measures used in this work. Let us first introduce some notation. Let r be a rule of the form $A \rightarrow C$ where A and C are sets of items. In a classification setting, each item in A is a pair $< attribute = value >$, and C has one single pair $< class\_attribute = class\_value >$. We are assuming that the rule was obtained from a a dataset D. The size of D is N.

In an association rule framework, *Confidence* is a standard measure, and is defined as:

$$conf(A \rightarrow C) = \frac{sup(A \cup C)}{sup(A)} \tag{1}$$

Confidence ranges from 0 to 1. Confidence is an estimate of $\Pr(C \mid A)$, the probability of observing C given A. After obtaining a rule set, one can immediatly use confidence as a basis for classifying one new case x. Of all the rules that apply to x (i.e., the rules whose antecedent is true in x), we choose the one with highest confidence. This loosely follows the optimal Bayes classifier. In the extreme case where there is exactly one rule that covers each new case, the Best Rule-confidence classifier concides with the optimal Bayes. In general, the BR_{conf} classifier approximates the M_{Bayes} but ignores the combined effect of evidence. For rules with the same confidence, the rule with the highest support is preferred. The rationale is that the estimate for confidence is more reliable.

Another measure sometimes used in classification is *Laplace*. It is a confidence estimator that takes support into account, becoming more pessimistic as the support of A decreases. It ranges within $[0, 1[$ and is defined as:

$$lapl(A \rightarrow C) = \frac{sup(A \cup C) + 1}{sup(A) + 2} \tag{2}$$

Confidence alone (or Laplace) may not be enough to assess the descriptive interest of a rule. Rules with high confidence may occur by chance. Such spurious rules can be detected by determining whether the antecedent and the consequent are statistically independent. This inspired a number of measures for association rule interest. One of them is *Lift*, defined as:

$$lift(A \rightarrow C) = \frac{conf(A \rightarrow C)}{sup(C)} \tag{3}$$

Lift measures how far from independence are A and C. It ranges within $[0, +\infty[$. Values close to 1 imply that A and C are independent and the rule is not interesting. Values far from 1 indicate that the evidence of A provides information about C. Lift measures co-occurrence only (not implication) and is symmetric with respect to antecedent and consequent. Lift is also used to characterize the classification rules derived by C4.5 [15].

Conviction is another measure proposed in [4] to tackle some of the weaknesses of confidence and lift. Unlike lift, conviction is sensitive to rule direction ($conv(A \rightarrow C) \neq conv(C \rightarrow A)$). Conviction is somewhat inspired in the logical definition of implication and attempts to measure the degree of implication of a rule. Conviction is infinite for logical implications (confidence 1), and is 1 if A and C are independent. It ranges along the values $0.5, ..., 1, ...\infty$. Like lift, conviction values far from 1 indicate interesting rules. It is defined as:

$$conv(A \rightarrow C) = \frac{1 - sup(C)}{1 - conf(A \rightarrow C)} \tag{4}$$

According to [4], conviction intuitively captures the notion of *implication rules*. Logically, $A \rightarrow C$ can be rewritten as $\neg(A \wedge \neg C)$. Then, one can measure how $(A \wedge \neg C)$ deviates from independence and take care of the outside negation. To cope with this negation, the ratio between $sup(A \cup \neg C)$ and $sup(A) \times sup(\neg C)$ is inverted. Unlike confidence, the support of both antecedent and consequent are considered in conviction.

Within the association rules framework, the *Leverage* measure was recovered by Webb for the Magnus Opus system [18]. It had been previously proposed by Piatetsky-Schapiro [14]. In [10] is called *novelty*. The idea is to measure how much more counting is obtained from the co-occurrence of the antecedent and consequent from the expected, i.e., from independence. It ranges within $[-0.25, 0.25]$ and is defined as:

$$leve(A \rightarrow C) = sup(A \cup C) - sup(A) \times sup(C) \tag{5}$$

The definite way for measuring the statistical independence between antecedent and consequent is the χ^2 test. The test's statistic can be used as a rule measure.

$$\chi^2(A \rightarrow C) = N \times \sum_{X \in \{A, \neg A\}, Y \in \{C, \neg C\}} \frac{(sup(X \cup Y) - sup(X).sup(Y))^2}{sup(X) \times sup(Y)} \quad (6)$$

where N is the database size.

As stated in [3], χ^2 does not assess the strength of correlation between antecedent and consequent. It only assists in deciding about the independence of these items which suggests that the measure is not feasible for ranking purposes. Our results will corroborate these claims.

The following measures, rather then indicating the absence of statistical independence between A and C, measure the degree of overlap between the cases covered by each of them. The *Jaccard* coefficient takes values in $[0, 1]$ and assesses the distance between antecedent and consequent as the fraction of cases covered by both with respect to the fraction of cases covered by one of them. High values indicate that A and C tend to cover the same cases.

$$jacc(A \rightarrow C) = \frac{sup(A \cup C)}{sup(A) + sup(C) - sup(A \cup C)} \quad (7)$$

Cosine is another way of measuring the distance between antecedent and consequent when these are viewed as two binary vectors. The value of one indicates that the vectors coincide. The value of zero only happens when the antecedent and the consequent have no overlap. It ranges along $[0, 1]$ and is defined as:

$$cos(A \rightarrow C) = \frac{sup(A \cup C)}{\sqrt{sup(A) \times sup(C)}} \quad (8)$$

Also the ϕ-coefficient can be used to measure the association between A and C. It is analogous to the discrete case of the Pearson correlation coeficient [17].

$$\phi(A \rightarrow C) = \frac{leve(A \rightarrow C)}{\sqrt{(sup(A) \times sup(C)) \times (1 - sup(A)) \times (1 - sup(C))}} \quad (9)$$

ϕ ranges within $[-1, 1]$. It is one when the antecedent and the consequent cover the same cases and -1 when they cover opposite cases. In [16] it is shown the relation between ϕ and the χ^2 statistics. i.e. that $\phi^2 = \frac{\chi^2}{N}$ being N the database size.

The last measure considered is information-based. Many classification algorithms use similar measures to assess rule predictiveness. This is the case of CN2 [5] which uses *Entropy.* In this paper we make use of the *Mutual Information.* Mutual information measures the amount of reduction in uncertainty of the consequent when the antecedent is known [16].

$$MI(A \rightarrow C) = \frac{\sum_i \sum_j sup(A_i \cup C_j) \times log(\frac{sup(A_i \cup C_j)}{sup(A_i) \times sup(C_j)})}{min(\sum_i -sup(A_i) \times log(sup(A_i)), \sum_j -sup(C_j) \times log(sup(C_j)))} \quad (10)$$

where $A_i \in \{A, \neg A\}$ and $C_j \in \{C, \neg C\}$. MI ranges over $[0, 1]$.

Notice that measures lift, leverage, χ^2, Jaccard, cosine, ϕ and MI are symmetric, whereas confidence, Laplace and conviction are asymmetric. We will see that this makes all the difference in terms of predictive performance. Other measures could have been considered, but this study focuses mainly on the ones mostly used in association rule mining.

2.1 Ordering Rules

The prediction given by the best rule is the best guess we can have with one single rule. When the best rule is not unique we can break ties maximizing support [12]. A kind of best rule strategy, combined with a coverage rule generation method, provided encouraging empirical results when compared with state of the art classifiers on some datasets from UCI [13]. Our implementation of Best Rule prediction follows closely the rules ordering described in CMAR [11]. Thus, having R_1 earlier than R_2 is defined as:

$$\begin{aligned}
&R_1 \prec R_2 \quad if\\
&\qquad metric(R_1) > metric(R_2) \quad or\\
&\qquad metric(R_1) == metric(R_2) \wedge sup(R1) > sup(R2) \quad or\\
&\qquad metric(R_1) == metric(R_2) \wedge sup(R1) == sup(R2) \wedge ant(R1) < ant(R2).
\end{aligned}$$

where *metric* is the used interest measure and *ant* is the length of the antecedent.

2.2 Voting

Apart from Best Rule strategies that select the prediction of the rule from the top of the rules rank, one can make use of a different strategy that allows all firing rules to contribute to the final prediction. These strategies combine the rules $F(x)$ that fire upon a case x. A *simple voting* strategy takes all the rules in $F(x)$, groups the rules by antecedent, and for each antecedent x' obtains the class corresponding to the rule with highest confidence. We will denote the class voted by an antecedent x' with a binary function $vote(x', g)$ which takes the value 1 when x' votes for g, and 0 for the other classes.

$$prediction_{sv} = arg \max_{g \in G} \sum_{x' \in antecedents(F(x))} vote(x', g) \tag{11}$$

2.3 Weighted voting

This strategy is similar to voting, but each vote is multiplied by a factor that quantifies the quality of the vote [9]. In the case of association rules, this can be done using one of the above defined measures.

$$prediction_{wv} = arg \max_{g \in G} \sum_{x'} vote(x', g). \max metric(x' \to g) \tag{12}$$

Caren implements these and other prediction strategies efficiently by keeping in an appropriate data structure [2].

3 Experiments

In our experiments, we have tested the effects of each measure on a number of benchmark datasets. For that, we ran CAREN using the “Best Rule” and “Weighted Voting” approaches. For each of these approaches we used a different variant for each rule measure. For reference we show the results of the *rpart* and the *c4.5* TDIDT algorithms. Evaluation was performed by running stratified 10 fold cross-validation and measuring the error rate of each variant on each dataset (Table 2). From these results we derive the algorithm ranking for each dataset (Table 3): the smallest error rate gets rank 1, the second rank 2 and so on. In the case of a draw in rank n, the algorithms get rank $n.5$. From these partial rankings we calculate the mean rank of each algorithm or variant.

In table 1 we describe the datasets used for evaluation. These sets have varied sizes, number of total and numerical attributes and number of classes and were obtained from the UCI repository [13].

Since the existence of minority classes may be important to explain the results of the approaches, we have also measured, for each dataset, class balancing using two measures ranging in $(0, 1]$. One is normalized Gini, defined as $\sum_i {p_i}^2/(1 - nclasses^{-1})$, where p_i is the proportion of class i. The other is normalized entropy, $-\sum_i p_i log_2(p_i)/log_2(nclasses)$. Both measures have the value of 1 when the classes are balanced and tend to 0 if the weight of a dominant class increases. Both measures are undefined when there is only one class. In general, high values mean balanced classes and low values mean unbalanced classes. The two measures are not very different in value. For these 17 datasets, the correlation between normalized Gini and normalized entropy is above 0.95.

Table 1. Datasets used for the empirical evaluation

Dataset	nick	#examples	#classes	#attr	#numerics	norm. Gini	norm. entropy
australian	aus	690	2	14	6	0.99	0.99
breast	bre	699	2	9	8	0.90	0.93
pima	pim	768	2	8	8	0.91	0.93
yeast	yea	1484	10	8	8	0.86	0.75
flare	fla	1066	2	10	0	0.61	0.70
cleveland	cle	303	5	13	5	0.81	0.80
heart	hea	270	2	13	13	0.99	0.99
hepatitis	hep	155	2	19	4	0.66	0.73
german	ger	1000	2	20	7	0.84	0.88
house-votes	hou	435	2	16	0	0.95	0.96
segment	seg	2310	7	19	19	1.00	1.00
vehicle	veh	846	4	18	18	1.00	1.00
adult	adu	32561	2	14	6	0.73	0.80
lymphography	lym	148	4	18	0	0.71	0.61
sat	sat	6435	6	36	36	0.97	0.96
shuttle	shu	58000	7	9	9	0.41	0.34
waveform	wav	5000	3	21	21	1.00	1.00

Since CAREN does not directly process numerical attributes, we have preprocessed these using CAREN’s implementation of Fayyad and Irani’s supervised discretization method [6]. For association rule construction, Minimal support was set to 0.01 or 10 training cases. The only exception was the *sat* dataset, where we used 0.02 for computational reasons. Minimal improvement was 0.01

and minimal confidence 0.5. We have also used the χ^2 filter to eliminate potentially trivial rules. C4.5 and rpart were ran using the original raw data.

Table 2. Error rates (in percent) for rpart, c4.5 and the different CAREN variants.

	aus	bre	pim	yea	fla	cle	hea	hep	ger	hou	seg	veh	adu	lym	sat	shu	wav
rpart	16.23	6.15	24.72	43.27	17.73	46.16	20.00	26.00	25.20	4.87	8.31	31.76	15.55	25.27	19.04	0.53	26.64
c4.5	13.92	5.00	24.36	44.27	17.44	50.04	21.09	21.32	30.20	3.25	3.21	25.96	13.61	23.07	13.97	0.05	22.73
BR.conf	14.21	4.57	22.78	41.27	19.14	45.70	18.52	19.99	28.50	8.11	9.91	38.74	14.81	17.29	19.75	0.47	17.40
BR.lift	36.09	18.88	41.66	44.57	21.66	44.09	31.48	47.51	63.80	46.67	9.91	43.92	36.49	47.75	34.51	21.74	29.82
BR.conv	14.35	4.28	22.38	42.07	19.78	44.09	18.89	16.78	26.70	8.11	9.91	38.63	14.27	18.00	19.39	0.45	17.42
BR.chi	32.89	14.02	33.71	45.10	21.66	44.09	27.04	47.51	63.30	40.95	31.77	46.21	36.49	47.08	32.43	20.59	21.04
BR.lapl	14.22	5.86	25.25	44.29	18.86	45.70	17.04	20.58	28.90	6.48	10.61	39.22	15.70	18.00	19.86	0.47	17.34
BR.lev	14.51	11.58	30.71	48.13	18.85	44.09	21.85	21.68	29.30	5.55	36.58	48.82	20.10	27.40	47.64	1.64	26.82
BR.jacc	14.51	19.45	34.89	44.81	18.86	45.70	38.15	20.58	30.00	5.55	34.03	48.81	24.08	34.32	42.14	15.58	26.74
BR.cos	14.51	23.17	34.89	55.80	18.86	45.70	44.44	20.58	30.00	5.55	33.94	49.75	24.08	43.62	42.75	15.58	32.16
BR.phi	14.51	6.85	27.85	44.55	19.22	44.09	18.89	30.27	30.70	5.55	33.12	49.06	17.91	28.69	36.25	6.45	26.46
BR.MI	25.81	5.85	24.20	46.14	17.90	44.09	27.04	19.99	29.90	15.88	14.50	37.90	17.41	45.76	35.29	6.04	28.52
Voting.conf	16.24	3.57	22.64	42.61	18.47	45.70	17.41	16.27	24.10	13.35	15.11	35.43	16.35	27.37	35.17	2.85	17.28
Voting.lift	15.37	3.29	25.63	42.01	19.50	45.70	16.67	17.14	27.50	14.73	15.11	35.55	20.02	32.19	33.43	9.20	17.24
Voting.conv	18.40	4.72	22.77	41.87	18.56	45.70	19.63	17.45	26.50	13.80	20.22	36.04	15.07	22.50	23.17	0.60	28.42
Voting.chi	16.10	3.43	25.63	42.88	18.94	45.70	17.04	15.22	25.70	13.81	15.89	36.03	18.42	32.19	35.79	3.60	17.12
Voting.lapl	16.39	3.57	22.77	42.41	18.38	45.70	17.41	16.27	23.90	13.35	15.15	35.55	16.41	27.37	35.17	2.84	17.22
Voting.Lev	15.82	4.58	24.72	46.04	18.56	45.70	17.78	14.12	24.40	14.50	22.34	39.00	17.04	32.72	36.60	2.69	19.72
Voting.Jacc	17.41	4.87	24.07	43.34	18.19	45.70	17.78	15.53	24.40	14.26	21.52	38.89	18.14	32.10	35.04	2.55	19.96
Voting.Cos	16.98	4.29	23.29	43.41	17.91	45.70	17.04	14.86	24.40	13.57	18.40	38.30	16.50	30.72	35.21	2.14	18.30
Voting.Phi	15.52	3.43	24.59	42.68	18.93	45.70	17.41	15.88	26.00	14.50	18.31	38.17	18.03	30.72	34.50	5.54	17.98
Voting.MI	77.84	49.08	34.36	45.03	18.66	45.70	73.33	80.09	66.00	80.63	97.71	74.69	34.88	93.25	84.38	40.84	90.90

Table 3. Ranks

	mean	aus	bre	pim	yea	fla	cle	hea	hep	ger	hou	seg	veh	adu	lym	sat	shu	wav
BR.conv	6.35	4	6	1	4	20	3.5	11.5	8	10	8.5	4	11	2	2.5	3	2	7
BR.conf	7.18	2	8	5	1	17	13.5	10	11.5	12	8.5	4	12	3	1	4	3.5	6
Voting.conf	7.44	14	4.5	2	6	7	13.5	6	6.5	2	10.5	8.5	3	7	7.5	12.5	12	4
Voting.lapl	7.47	15	4.5	3.5	5	6	13.5	6	6.5	1	10.5	10	4.5	8	7.5	12.5	11	2
c4.5	7.65	1	12	9	12	1	22	15	16	18	1	1	1	1	5	1	1	13
rpart	8.74	13	15	11.5	9	2	21	14	18	6	2	2	2	5	6	2	5	15
BR.lapl	8.79	3	14	13	13	13	13.5	3	14	13	7	6	15	6	2.5	5	3.5	5
Voting.Cos	9	16	7	6	11	4	13.5	3	2	4	12	13	10	9	11.5	14	8	9
Voting.conv	9.38	18	10	3.5	2	8.5	13.5	13	10	9	13	14	7	4	4	6	6	18
Voting.chi	10	12	2.5	14.5	8	16	13.5	3	3	7	14	11	6	15	14.5	16	13	1
Voting.Phi	10	10	2.5	10	7	15	13.5	6	5	8	16.5	12	9	13	11.5	9	14	8
Voting.lift	10.03	9	1	14.5	3	19	13.5	1	9	11	18	8.5	4.5	16	14.5	8	17	3
Voting.Jacc	10.65	17	11	7	10	5	13.5	8.5	4	4	15	15	13	14	13	11	9	11
Voting.Lev	11.56	11	9	11.5	19	8.5	13.5	8.5	1	4	16.5	16	14	10	16	18	10	10
BR.MI	13.15	19	13	8	20	3	3.5	17.5	11.5	15	19	7	8	11	19	15	15	19
BR.phi	13.82	6.5	16	16	14	18	3.5	11.5	19	19	4.5	18	20	12	10	17	16	14
BR.lev	14.03	6.5	17	17	21	11	3.5	16	17	14	4.5	21	19	17	9	21	7	17
BR.jacc	15.97	6.5	20	20.5	16	13	13.5	20	14	16.5	4.5	20	18	18.5	17	19	18.5	16
BR.cos	16.97	6.5	21	20.5	22	13	13.5	21	14	16.5	4.5	19	21	18.5	18	20	18.5	21
BR.chi	17.15	20	18	18	18	21.5	3.5	17.5	20.5	20	20	17	17	21.5	20	7	20	12
BR.lift	17.47	21	19	22	15	21.5	3.5	19	20.5	21	21	4	16	21.5	21	10	21	20
Voting.MI	20.21	22	22	19	17	10	13.5	22	22	22	22	22	22	20	22	22	22	22

4 Discussion

The first observation is that conviction gets the best mean rank. This confirms the experiments in [8] which motivated the present study of this measure. We observe that, using a t-test with 5% significance, conviction has 2 out of 17 significant wins over confidence and 3 over Laplace (and loses none). This seems to be a marginal but consistent advantage. The advantage of conviction is not observed for the Voting strategy.

The second observation is that the other 7 measures do not produce competitive classifiers. Notice that these are the symmetric rule interest measures.

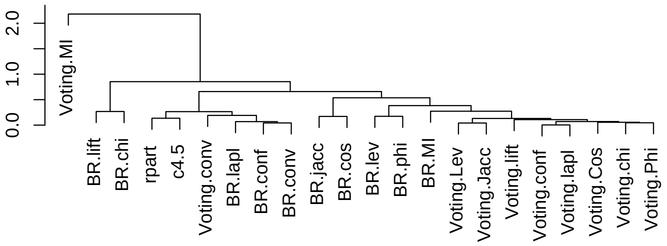

Fig. 1. Clustering measures and strategies using complete linkage and Euclidean distance

Using the error rate arrays, we can group the pairs strategy-measure using hierarchical clustering. The distance between strategies was measured using plain Euclidean distance and the clusters were aggregated using complete linkage. The obtained clustering (Fig. 1) indicates the predictive proximity of confidence, conviction and Laplace (the three symmetric measures employed), when used with best rule. Almost all Voting strategies are clustered together, except for Voting.conv (which is clustered with the top performing best rule approaches) and Voting.MI which performs particularly poorly. Other expected pairs of measures also cluster together. These are Jaccard and cosine (for best rule), lift and χ^2, leverage and ϕ and also rpart and C4.5.

4.1 Discriminating meta features

We now try to study if some features of the data set (meta features) may indicate whether to use either conviction, confidence or laplace. As meta features we have selected *nclasses*, the number of classes, *n.entropy*, normalized class entropy, which measures the balance of class distribution, and the *nexamples*, the number of examples. Given the relatively small number of datasets, we will perform a graphical exploration using 2-dimensional xy-plots. The first exploration investigates the importance of the number of classes and class distribution. Since both class distribution measures are very similar, we have used only normalized class entropy.

In Fig. 2 we can see which of the three measures performed better with a best rule approach. The chart represents the datasets in the “number of classes” × “class distribution” space. Each dataset is represented by the measure that performed better within the considered pool. Conviction is represented by “V”, confidence by “F” and Laplace by “L”. When there is a tie between the two best ones the dataset point is signaled by an “X”.

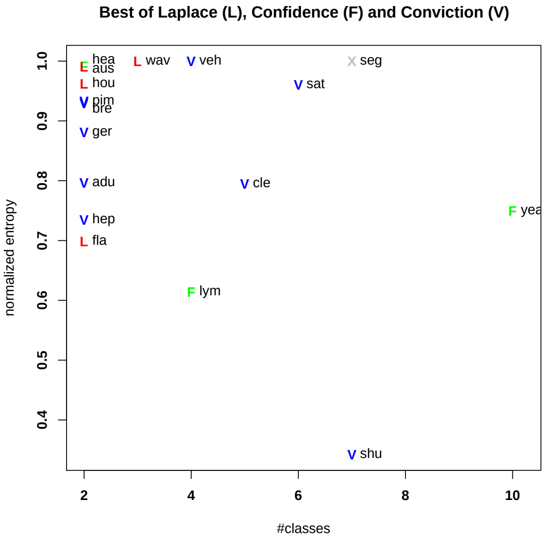

Fig. 2. Meta exploration of the results. The chart plots the datasets in a two dimensional space. Each dataset is represented by the symbol corresponding to the best performing measure. Ties are represented by “X”.

As we observe, there is no clear pattern explaining the success of each measure. Conviction dominates in general. Confidence is successful with the yeast dataset (10 classes, relatively unbalanced), and with two other. Laplace has a visible but not dominating presence in datasets with 2 or 3 classes. Regarding class balancing, there is no visible tendency. We observe, however, that the most unbalanced dataset is won by conviction.

4.2 Comparing Rule Ranking

Confidence and conviction differ in the rule ranking they produce when rules of different classes are involved. For rules of the same class, or of classes with the same support, conviction preserves the ordering given by confidence. This is because the expression of conviction (Eq. 4) has the same numerator for rules with classes with the same support. The denominator increases when confidence of the rule decreases and vice-versa.

The datasets considered are never exactly balanced (The values of normalized entropy and Gini shown in table 1 are 1.00 only because of rounding). In Fig.

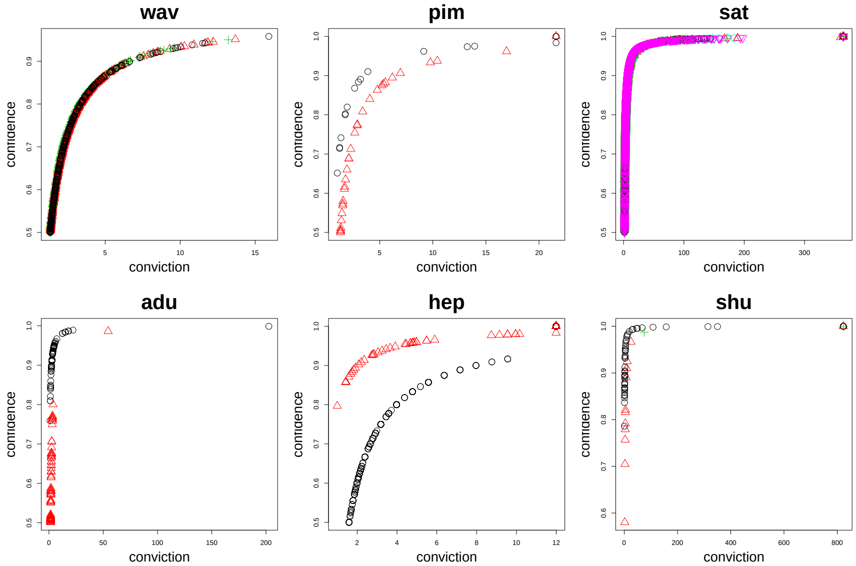

Fig. 3. Visualization of the comparison of the rule rankings produced by conviction and confidence. Each point represents one rule generated for the respective dataset. Different rule classes are represented with different characters.

3 we can compare the rule rankings provided by confidence and conviction for some datasets. These charts were built by generating a single rule set for each dataset with 80% of the examples, and plotting each rule on a 2 dimensional space defined by confidence and conviction.

For the almost balanced datasets (segment, vehicle and waveform), the ranking is practically preserved. This explains the almost equal error values of BR.conf and BR.conv (Table 2). For datasets with normalized entropy above 0.9 and two classes (australian, pima, heart and house-votes) we can observe two different curves, one for each class (each class is represented with a different marker). Despite the small difference between class supports, the different effects of conviction and confidence are quite visible in Fig. 3. However, in these cases, the difference in error is still very small (p-value is around 0.5). The sat dataset has a high entropy but has 6 classes. This makes the p-value of the error comparison drop to 0.14 with advantage to conviction.

The two significant wins of BR.conv over BR.conf are in datasets adult and hepatitis. These are datasets with mid-range entropy and 2 classes. In the case of hepatitis we can see that confidence tends to rank first the rules of one of the classes, whereas conviction tends to interleave rules of the two classes. The dataset with lowest entropy (highly unbalanced) is shuttle. It has 7 classes, but only rules for 3 of the classes were derived. For this dataset, conviction has advantage over confidence with a p-value of 0.067.

In summary, what we observe in Fig 3 is that conviction favours rules with less frequent classes, and ranks the rules differently from confidence. This is relatively innocuous for most of the datasets, although more frequently advantageous to conviction. When there is a significant difference it is in favour of BR.conv.

5 Conclusion

Despite having been introduced as a measure of interestingness for descriptive data mining, mainly to overcome some limitations of lift (a.k.a. interest) and confidence, conviction proves to be effective for predictive tasks. We have compared conviction with a few different measures on 17 datasets and concluded that it shows a systematic advantage over them when used with a best rule association based classifier. Compared to confidence, conviction favours low frequency classes and produces different rule orderings. This is mainly visible with unbalanced datasets. Besides conviction, confidence and Laplace, all the other yielded uninteresting results for best rule.

In the case of Voting, different results were obtained. Confidence and Laplace ranked relatively high, whereas conviction ranked amid the other measures. However Voting.conviction clustered close to the top performing best rule approaches. The negative results of conviction with the voting strategy may be due to the fact that rule ordering is diluted by the combined effect of voting rules. Another difficulty in using conviction with a voting strategy may be related with its overly stretched value range.

The symmetry of measures with respect to antecedent and consequent seems to make all the difference. Confidence, conviction and Laplace (asymmetric) obtained superior results, whereas symmetric measures had poor predictive performance.

For future work it would be worthwhile to study the effect of combining different measures to produce ensembles of classifiers. This way we could use a single set of learned rules to build an ensemble of models, each corresponding to the ordering yielded by a different measure.

References

1. P. J. Azevedo. CAREN - A java based apriori implementation for classification purposes. Technical report, Universidade do Minho, Departamento de Informática, June 2003.
2. P. J. Azevedo. A data structure to represent association rules based classifiers. Technical report, Universidade do Minho, Departamento de Informática, 2005.
3. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Peckham, editor, *SIGMOD Conference*, pages 265–276. ACM Press, 1997.
4. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In J. Peckham, editor, *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, 13–15 June 1997.

5. P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261–283, 1989.
6. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1029, 1993.
7. T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, August 2001.
8. A. Jorge and P. J. Azevedo. An experiment with association rules and classification: Post-bagging and conviction. In A. G. Hoffmann, H. Motoda, and T. Scheffer, editors, *Discovery Science*, volume 3735 of *Lecture Notes in Computer Science*, pages 137–149. Springer, 2005.
9. I. Kononenko. Combining decisions of multiple rules. In *AIMSA*, pages 87–96, 1992.
10. N. Lavrac, P. A. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In S. Dzeroski and P. A. Flach, editors, *ILP*, volume 1634 of *Lecture Notes in Computer Science*, pages 174–185. Springer, 1999.
11. W. Li, J. Han, and J. Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In N. Cercone, T. Y. Lin, and X. Wu, editors, *ICDM*, pages 369–376. IEEE Computer Society, 2001.
12. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD '98: Proceedings of the fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 80–86, New York, NY, USA, 1998. ACM Press.
13. C. J. Merz and P. Murphy. Uci repository of machine learning database. http://www.cs.uci.edu/∼mlearn, 1996.
14. G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.
15. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
16. P. Tan and V. Kumar. Interestingness measures for association patterns: A perspective. Technical Report TR00-036, Department of Computer Science, University of Minnesota, 2000.
17. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measures for association patterns. In *Proceedings of the 2002 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1–12, Alberta, Canada, 07 2002.
18. G. I. Webb. Efficient search for association rules. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–107, New York, NY, USA, 2000. ACM Press.