An Overview of Alternative Rule Evaluation Criteria and Their Use in Separate-and-Conquer Classifiers

Fernando Berzal, Juan-Carlos Cubero, Nicolás Marín, and José-Luis Polo

Department of Computer Science and Artificial Intelligence,
University of Granada, 18071 Granada, Spain
{fberzal, JC.Cubero, nicm, jlpolo}@decsai.ugr.es

Abstract. Separate-and-conquer classifiers strongly depend on the criteria used to choose which rules will be included in the classification model. When association rules are employed to build such classifiers (as in ART [3]), rule evaluation can be performed attending to different criteria (other than the traditional confidence measure used in association rule mining). In this paper, we analyze the desirable properties of such alternative criteria and their effect in building rule-based classifiers using a separate-and-conquer strategy.

1 Introduction

The aim of any classification algorithm is to build a classification model given some examples of the classes we are trying to model. The model we obtain can then be used to classify new examples or simply to achieve a better understanding of the available data. Rule-based classifiers, in particular, can be built using a "Separate and Conquer" strategy (as in decision lists) or a "Divide and Conquer" approach (as in decision trees). In the former case, the criteria used to evaluate the rules that will be included in the classification model are of the utmost importance.

In this paper, we analyze alternative rule evaluation criteria and study their actual impact in "separate and conquer" classification models built with ART [3], which is a generalized "separate and conquer" algorithm that builds decision lists that can also be viewed as degenerate, polythetic decision trees.

2 Alternative Rule Evaluation Criteria

Even though accurate classification models can be built using standard association rules [3] [1] [17] [11] [14] [18] [15] [7] [8] [12] [13] [10] [19], it is clear that confidence is not the best measure for a classification rule. Alternative rule evaluation measures might provide better insight into the capability of a given association rule to classify data. Before delving into the details of existing measures, we should review what we consider a good classification rule. An association rule $A \Rightarrow C$ will be useful in classification problems when the logical implication $A \rightarrow C$ is valid:

F. Esposito et al. (Eds.): ISMIS 2006, LNAI 4203, pp. 591–600, 2006.

– C happens more often when A holds.
– $\neg C$ should be less frequent when A holds.

We might also be interested in verifying the validity of $\neg C \rightarrow \neg A$, which is mathematically equivalent to $A \rightarrow C$. Hence, a potentially useful rule should also verify the following properties:

– $\neg A$ happens more often when $\neg C$.
– A should occur less frequently when C does not hold.

Although the latter two properties might be interesting from a logical point of view, they do not directly affect classification accuracy, where we are interested only in determining the class C given A. Even then, those properties might be desirable to improve the understandability of the classification model obtained.

The following paragraphs discuss some rule evaluation measures and their properties in the context of classification problems.

Confidence
Even when the rule support might be of interest during the association discovery process from a purely tactical point of view, the rule confidence is what finally determines the rule validity. Its meaning is easy to grasp, and it can be defined as follows: $conf(A \Rightarrow C) = P(C|A) = P(A \cap C)/P(A)$. Confidence is used to measure the association between the antecedent A and the consequent C: the higher the confidence of $A \Rightarrow C$, the lower the confidence of $A \Rightarrow \neg C$, since $P(C|A)+P(\neg C|A) = 1$. However, the rule confidence cannot be used to establish a causality relationship between antecedent and consequent. It cannot be used to perform inferences since it does not take into account the validity of $\neg C \Rightarrow \neg A$. A variation of the confidence does, hence its name: Causal confidence [9].

Causal Confidence
The causal confidence measure considers the confidence of the rule $A \rightarrow C$ and the confidence of its counterpart $\neg C \rightarrow \neg A$. Thus, the definition is given by $conf_{causal}(A \Rightarrow C) = \frac{1}{2}(P(C|A) + P(\neg A|\neg C))$. The average is used just to normalize the measure so that its values are always in the interval $[0, 1]$. Unfortunately, this measure could have a high value even when the implication $A \rightarrow C$ does not have a representative support but its counterpart $\neg C \rightarrow \neg A$ does. Therefore, causal confidence is not adequate for classification problems.

Causal Support
The idea of causal confidence can also be applied to the support measure in order to obtain the causal support: $support_{causal}(A \Rightarrow C) = P(A \cap C) + P(\neg A \cap \neg C)$. As before, even when $P(A \cap C)$ is really low, $P(\neg A \cap \neg C)$ can be high, causing causal support to be high (even when the rule might not be too useful in a classification problem). Consequently, causal support is not a good choice to evaluate rules in classification problems.

Confirmation
Another measure, called confirmation, takes into account when the antecedent holds but not its consequent, what might be used to highlight a rule depending

on the commonness of its antecedent. $K(A \Rightarrow C) = P(A \cap C) - P(A \cap \neg C)$. Since $P(A \cap \neg C) = P(A) - P(A \cap C)$, confirmation is reduced to $P(A \cap C) - P(A \cap \neg C) = 2P(A \cap C) - P(A)$. Since $P(A \cap C)$ is the rule support and $P(A)$ is the support of the antecedent, confirmation is no more than $K(A \Rightarrow C) = support(A \Rightarrow C) - support(A)$. Hence

$$K(A \Rightarrow C) \leq support(A \Rightarrow C) \tag{1}$$

The second term just penalizes the support of the rule according to the support of the antecedent. In other words, given two equally common rules, confirmation would select as most interesting the rule whose antecedent is less common. That might be useful in the knowledge discovery process, although it does not help in classification problems because it does not take into account relationship between the rule antecedent and the rule consequent. This measure has been used to obtained three new criteria to evaluate rules:

- **Causal confirmation**, takes into account $\neg C \rightarrow \neg A$ and can be reduced to $K_{causal}(A \Rightarrow C) = support_{causal}(A \Rightarrow C) - support(A) + support(C)$. As the standard confirmation, this measures varies depending upon the support of A. On the other hand, the more common the class C is, the higher causal confirmation the rule will have, what certainly makes classification difficult (especially when the class distribution is skewed). Apart from this fact, causal support is not suitable for classification problems.
- **Confirmed confidence** is obtain from the standard confidence when we try to measure the quality of the rule $A \Rightarrow \neg C$ (that is, when the antecedent holds but not the consequent): $conf_{confirmed}(A \Rightarrow C) = P(C|A) - P(\neg C|A)$. However, since $P(\neg C|A) = 1 - P(C|A)$, confirmed confidence can be reduced to $conf_{confirmed}(A \Rightarrow C) = 2 \cdot conf(A \Rightarrow C) - 1$. Therefore, if we are using this measure just to rank the candidate rules which might be part of a classification model, the rule confirmed confidence is completely equivalent to the rule confidence (only that the confirmed confidence is defined over the $[-1, 1]$ interval).
- Even a **causal confirmed confidence** can be concocted from the two previous measures, but no interesting results can be expected when dealing with classification problems.

Conviction

Conviction [4] was introduced as an alternative to confidence to mine association rules in relational databases (implication rules using their authors' nomenclature). $conviction(A \Rightarrow C) = \frac{P(A)P(\neg C)}{P(A \cap \neg C)}$. Similar in some sense to confirmation, conviction focuses on $A \Rightarrow \neg C$:

$$conviction(A \Rightarrow C) = \frac{support(\neg C)}{conf(A \Rightarrow \neg C)} \tag{2}$$

The lower $P(A \cap \neg C)$, the higher the rule conviction, what makes conviction ideal for discovering rules for uncommon classes. However, the conviction domain is

not bounded and it is hard to compare conviction values for different rules. This constrains the usability of the conviction measure in classification models such as ART [3], since no heuristics to automatically settle a conviction threshold can be devised.

Later in this paper, another bounded measure will be analyzed which is completely equivalent to conviction in the situations of interest for classification problems: the well-known Shortliffe and Buchanan's certainty factors. When there are at lest two classes (i.e. $support(C) < 1$) and the rule improves classifier accuracy (i.e. $CF(A \Rightarrow C) > 0$), certainty factors can be defined as

$$CF(A \Rightarrow C) = 1 - \frac{1}{conviction(A \Rightarrow C)}$$

This allows us to substitute conviction for another measure whose domain is bounded. Further properties of the certainty factor will be discusses in 2.

Interest
A rule interest measure was defined in [16] as $interest(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)P(C)} = \frac{P(C|A)}{P(C)}$. The more common A and C, the less interest the rule will have, which is certainly useful to guide the knowledge discovery process. Among its properties, its symmetry stands out: the interest of $A \Rightarrow C$ equals to the interest of $C \Rightarrow A$. As happened with conviction, its domain is not bounded, what might make its interpretation harder in a classification model. In some sense, it can be considered complementary to conviction if we take into account the following equality and compare it with equation 2, although interest focuses on the association $A \Rightarrow C$ while conviction focuses on $A \Rightarrow \neg C$::

$$interest(A \Rightarrow C) = \frac{confidence(A \Rightarrow C)}{support(C)} \tag{3}$$

Dependency
The following measure is the discrete equivalent of correlation in continuous domains. $dependency(A \Rightarrow C) = |P(C|A) - P(C)|$. In classification problems, it is not suitable since its value is high for common classes even when the corresponding rule confidence is minimum: $dependency(A \Rightarrow C) = |conf(A \Rightarrow C) - support(C)|$. A causal variation [9] of this measure can also be defined as follows, although it is not useful in classification problems either. Bhandari's attribute focusing measure is also derived from the dependency measure above, as the following expression shows $Bhandari(A \Rightarrow C) = support(A) \cdot dependency(A \Rightarrow C)$. Therefore, it is of no use in classification problems.

Hellinger's Divergence
Hellinger's divergence was devised to measure the amount of information a rule provides [5] and it can be viewed as a distance measure between a priori and a posteriori class distributions:

$$H(A \Rightarrow C) = \sqrt{P(A)}[(\sqrt{P(A \cap C)} - \sqrt{P(C)})^2 - (\sqrt{1 - P(A \cap C)} - \sqrt{1 - P(C)})^2]$$

This measure has been used in classifiers before and will be evaluated in Section 3.

Certainty Factors
Certainty factors were introduced by Shortliffe and Buchanan to represent uncertainty in the MYCIN expert system. Its use in association rule mining has been proposed in [2]. The certainty factor of a rule $A \Rightarrow C$ is defined as

$$CF(A \Rightarrow C) = \frac{conf(A \Rightarrow C) - support(C)}{1 - support(C)}$$

when $conf(A \Rightarrow C) > support(C)$,

$$CF(A \Rightarrow C) = \frac{conf(A \Rightarrow C) - support(C)}{support(C)}$$

when $conf(A \Rightarrow C) < support(C)$, and

$$CF(A \Rightarrow C) = 0$$

when $conf(A \Rightarrow C) = support(C)$. This rule evaluation measure can be viewed as the variation degree of the probability of C when A holds. The larger a positive certainty factor, the smaller the decrease of the probability of C not being when A holds. In extreme situations, the rule confidence determines its certainty factor:

$$conf(A \Rightarrow C) = 1 \Rightarrow CF(A \Rightarrow C) = 1$$

$$conf(A \Rightarrow C) = 0 \Rightarrow CF(A \Rightarrow C) = -1$$

Certainty factor take into account the probability of C apart from the rule confidence. They also verify an interesting property when they are positive (which is when the rules are useful for classification):

$$CF(A \Rightarrow C) = CF(\neg C \Rightarrow \neg A) \tag{4}$$

In classification problems, we could face different situations where certainty factors behavior are at their best:

1. If our problem includes a skewed class distribution, and two candidate rules hold the same confidence value but correspond to classes of different frequency, the rule corresponding to the less common class has a higher certainty factor:

$$conf(A \Rightarrow C) = conf(B \Rightarrow D) \leq 1 \ , \ support(C) > support(D) \rightarrow$$

$$\rightarrow CF(A \Rightarrow C) \leq CF(B \Rightarrow D)$$

2. Under some circumstances, comparing certainty factors is equivalent to comparing confidence values. $CF(A \Rightarrow C_1) > CF(B \Rightarrow C_2)$ can be reduced to $conf(A \Rightarrow C_1) > conf(B \Rightarrow C_2)$ when

- Both rules refer to the same class c_k.
- Both rules correspond to classes with the same probability: $support(c_1) = support(c_2)$.

3. Finally, there exist situations where higher certainty factors do not correspond to a higher confidence values. Given two rules so that $CF(A \Rightarrow C) > CF(B \Rightarrow D)$:
   - When C is more common than D:

$$conf(A \Rightarrow C) > K \cdot conf(B \Rightarrow D) \quad , \quad K = \frac{support(\neg C)}{support(\neg D)} < 1$$

   - When C is less common than D:

$$conf(A \Rightarrow C) > conf(B \Rightarrow D) - \Delta \ , \ \Delta = \frac{support(D) - support(C)}{support(\neg C)} > 0$$

In summary, even though certainty factors are intimately linked to confidence values, a higher certainty factor does not imply a higher confidence value:

$$CF(A \Rightarrow C) > CF(B \Rightarrow D) \Rightarrow conf(A \Rightarrow C) > conf(B \Rightarrow D)$$

The relative frequency of each class determines the higher certainty factors. Let us suppose that we we have two association rules: $A \Rightarrow c_1$ with 2% support and 50% confidence, and $B \Rightarrow c_2$ with 50% support and 95% confidence. The certainty factors of such rules would be 3/8 and 3/4 if c_1 has a 20% support and c_2 has a 80% support. However, if the class distribution varies, being now 10% of c_1 and 90% of c_2, when certainty factors would be 4/9 and 1/2, keeping their relative order. However, if the class distribution is inverted and now c_1 has a 94% support while c_2 only has a 6% support, when certainty factors would become 46/94 and 1/6, being the second lower than the first!

Situations such as the one described in the paragraph above should be used a a warning sign when comparing certainty factors to choose the rules to include in a classification model. Sometimes they might be useful, as when they prefer rules corresponding to uncommon classes, although you should also expect shocking difference between classification models when class distributions change.

2.1 A Usefulness Constraint

Certainty factor properties suggest an additional pruning step when considering association rules in order to build classification models. When you build such models, only association rules with a positive certainty factor are really useful, since they increase our knowledge and improve classifier accuracy. By definition, a positive certainty factor is obtained for a rule $A \Rightarrow C$ when $conf(A \Rightarrow C) > support(C)$. This constraint indicates that the use of the rule $A \Rightarrow C$ improves the classification model which would result from using a default class (at least with respect to C in the training set). That idea can be used to prune association rules which do not verify the above requirement, which can be expressed as $P(A \cap C) > P(A) \cap P(C)$. That rule pruning can be viewed as a 'usefulness criterion' which reduces the number of candidate association rules which can become part of the classification model

3 Experimental Results Using ART

Some experiments have been perform to check the influence of the rule evaluation measure on the process of building a classification model. We have tested different criteria using the ART [3] classification model as a test bed, since ART does not requires any specific measure to evaluate the rules the ART classifier is built from. Our experiments try to estimate the suitability of the measures proposed in the previous section, using 10-folded cross validation and the same datasets which were used in [3]. In all our experiments with ART, we used a 5% minimum relative support threshold and the automatic threshold selection heuristics described in [3]. Table 1 summarizes our results.

The following observations can be made from the results we have obtained:

- The usefulness criterion proposed in 2.1 consistently improves ART classification accuracy. Moreover, it improves accuracy without modifying the evaluation measure used during the association rule mining process (that is, the rule confidence). However, this increased accuracy comes at a cost: the increased complexity of the resulting classifier. The resulting ART tree has more leaves and, therefore, training time is somewhat higher since the training dataset must be scanned more times to build the classifier. This result is just an incarnation of the typical trade-off between classifier accuracy and classifier complexity.
- Certainty factors do not improve ART overall performance, probably due to some of their counterintuitive properties (see section 2).
- As we could expect from the properties analyzed in the previous section, the use of conviction achieves results which are similar to the results obtained by using certainty factors. From a classification point of view, conviction and certainty factors are equivalent when it is interesting to include a given association rule in the classification model, as was mentioned in section 2.
- The use of the interest measure (section 2) leads to the results we could expect in ART: since this measure is not bounded, the automatic threshold selection criterion in ART does not work properly. Perhaps, the additive tolerance margin might be replaced by a multiplicative tolerance factor. Even then, the definition of the interest measure makes it difficult to establish an initial desirable interest value. Such value might depend on the particular problem and, therefore, the use of the interest measure is not practical in ART. Bounded measures, such as confidence or certainty factors, will be preferred.
- The same rationale applies to Hellinger's divergence (section 2). Even when its range is bounded, the interpretation and comparison of Hellinger's divergence values make this measure impractical in classification models such as ART. In fact, no acceptable classification models were built by ART because it is hard to establish an initial desirable value for Hellinger's divergent (a prerequisite to make use of ART automatic parameter setting).

Table 1. Experiment summary using different rule evaluation criteria

	Confidence	Usefulness	CF	Conviction	Interest	Hellinger
Accuracy (10-CV)	79.22%	83.10%	77.67%	77.79%	53.71%	48.05%
Training time	18.5s	22.4s	12.7s	10.7s	2.9s	1.8s
Tree topology						
- Leaves	36.9	50.0	28.5	30.0	4.2	1
- Internal nodes	18.3	25.2	16.2	17.3	2.6	0
- Average depth	7.41	8.71	7.39	7.33	2.12	1.00
I/O operations						
- Records	36400	50100	33100	26700	20300	7000
- Scans	63	89	55	59	11	3

Table 2. Datasets used in our experiments (from the UCI Machine Learning Repository)

Dataset	*Records*	*Attr*	*Classes*
AUDIOLOGY	226	70	24
CAR	1728	7	4
CHESS	3196	36	2
HAYES-ROTH	160	5	3
LENSES	24	6	3
LUNG CANCER	32	57	3
MUSHROOM	8124	23	2
NURSERY	12960	9	5
SOYBEAN	683	36	19
SPLICE	3175	61	3
TICTACTOE	958	10	2
TITANIC	2201	4	2
VOTE	435	17	2

4 Conclusions

In summary, from all the measures we discussed above, only certainty factors and the so-called usefulness criterion are good alternatives to confidence when building ART classification models. Conviction also achieves good results, although certainty factors are preferred since they are equivalent to conviction in the cases the classification process is more interested in (and certainty factors are bounded).

Despite the experimental results, it should be noted that all rule evaluation criteria have their home grounds and their use might be suitable depending on what the user intends to obtain. The availability of a wide variety of rule evaluation measures is a good signal, since it provides us a toolkit to draw on.

Moreover, the use of a measure or another does not affect the computational cost of the rule discovery process. When building classification models such as ART, that cost is proportional to the classification model complexity. Therefore, the study of alternative rule evaluation measures keeps its interest. Such measures are just criteria at our disposal which can be used to guide the knowledge discovery process according to the particular goals and needs of a given problem.

References

1. K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, USA*, pages 115–118, August 14–17 1997.
2. Fernando Berzal, Ignacio Blanco, Daniel Sanchez, and M.A. Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.
3. Fernando Berzal, Juan Carlos Cubero, Daniel Sánchez, and J.M. Serrano. Art: A hybrid classification model. *Machine Learning*, 54(1):67–92, January 2004.
4. S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD international conference on Management of Data, Tucson, Arizona, USA*, pages 255–264, May 11-15 1997.
5. Lee C.-H. and Shim D.-G. A multistrategy approach to classification learning in databases. *Data & Knowledge Engineering*, 31:67–93, 1999.
6. W. Cohen. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA USA*, pages 115–123, July 1995.
7. G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, San Diego, CA USA*, pages 43–52, August 15–18 1999.
8. G. Dong, X. Zhang, L. Wong, and J. Li. Caep: Classification by aggregating emerging patterns. In *Proceedings of the Second International Conference on Discovery Science, Tokyo, Japan*, pages 30–42, 1999.
9. Y. Kodratoff. Comparing machine learning and knowledge discovery in databases. *Lecture Notes in Artificial Intelligence, LNAI*, 2049:1–21, 2001.
10. Wenmin Li, Jiawei Han, and Jian Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01), San Jose, California, USA*, pages 208–217, November 29 - December 02 2001.
11. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, USA*, pages 80–86, August 27–31 1998.
12. B. Liu, M. Hu, and W. Hsu. Intuitive representation of decision trees using general rules and exceptions. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas*, pages 30–42, July 30 - August 3 2000.
13. B. Liu, M. Hu, and W. Hsu. Multi-level organization and summarization of the discovered rule. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA USA*, pages 208–217, August 20 - 23 2000.

14. B. Liu, Y. Ma, and C.K. Wong. Improving an association rule based classifier. In *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France*, pages 80–86, September 13–16 2000.
15. D. Meretakis and B. Wuthrich. Extending naive bayes classifiers using long itemsets. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, San Diego, CA USA*, pages 165–174, August 15–18 1999.
16. C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 22(2):39–68, 1998.
17. K. Wang, S. Zhou, and Y. He. Growing decision trees on support-less association rules. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA USA*, pages 265–269, August 20–23 2000.
18. K. Wang, S. Zhou, and S.C. Liew. Building hierarchical classifiers using class proximity. In *Proceedings of 25th International Conference on Very Large Data Bases (VLDB'99), Edinburgh, Scotland, UK*, pages 363–374, September 7–10 1999.
19. Xiaoxin Yin and Jiawei Han. Cpar: Classification based on predictive association rules. In *Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA*, pages 208–217, May 1-3 2003.