CAR-NF⁺ : An Improved Version of CAR-NF Classifier

Raudel Hernández-León¹, José Hernández-Palancar¹,
Jesús Ariel Carrasco-Ochoa², and José Francisco Martínez-Trinidad²

¹ Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), 7a # 21812
e/ 218 and 222, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba
`{rhernandez,jpalancar}@cenatav.co.cu`

² Computer Science Department
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro No. 1, Sta. María Tonantzintla, Puebla, CP:72840, México
`{ariel,fmartine}@ccc.inaoep.mx`

Abstract. In this paper, we propose two improvements to CAR-NF classifier, which is a classifier based on Class Association Rules (CARs). The first one, is a theoretical proof that allows selecting the minimum Netconf threshold, independently of the dataset, that avoids ambiguity at the classification stage. The second one, is a new coverage criterion, which aims to reduce the number of non-covered unseen-transactions during the classification stage. Experiments over several datasets show that the improved classifier, called CAR-NF⁺, beats the best reported classifiers based on CARs, including the original CAR-NF classifier.

Keywords: Data mining, Supervised classification, Class association rules.

1 Introduction

Associative Classification is a well-known data mining technique that aims to build a classifier based on CARs to determine the class of unseen transactions [1]. A classifier based on CARs usually consists in computing an ordered CAR list l , and applying a mechanism for classifying unseen transactions using l [1,2,6,10,14,15]. Nowadays, Associative Classification is been used in real world applications, such as detection of breast cancer [11], prediction of protein-protein interaction types [12] and prediction of consumer behavior [13], among others.

In Associative Classification, similar to Association Rule Mining (ARM), a set of items $I = \{i_1, \dots, i_n\}$, a set of classes C , and a set of labeled transactions D , are given. Each transaction in D is represented by a set of items $X \subseteq I$ and a class $c \in C$. The Support of an itemset $X \subseteq I$ is the fraction of transactions in D containing X (see Eq. 1). A CAR is a expression $X \Rightarrow c$, where $X \subseteq I$ and $c \in C$. The interest of a CAR $X \Rightarrow c$ is usually evaluated using the Support and Confidence measures (see Eq. 2 and Eq. 3), where Support is the fraction of transactions in D containing $X \cup \{c\}$ and Confidence is the probability of finding

c in transactions containing X .

$$Sup(X) = \frac{|D_X|}{|D|} \quad (1)$$

where D_X is the set of transactions in D containing X and $|\cdot|$ represents the cardinality.

$$Sup(X \Rightarrow c) = Sup(X \cup \{c\}) \quad (2)$$

$$Conf(X \Rightarrow c) = \frac{Sup(X \Rightarrow c)}{Sup(X)} \quad (3)$$

Several authors have mentioned some drawbacks of these measures [3,8,15]. In order to overcome these drawbacks, in [15], the authors introduced the use of the Netconf measure (see Eq. 4) instead of Support and Confidence for computing the CARs, and additionally, they proved that a Netconf threshold value of 0.5 avoids ambiguity at classification stage. However, they did not prove that, independently of the dataset, 0.5 is the minimum Netconf threshold value that avoids ambiguity at the classification stage. Two CARs are ambiguous if they have the same antecedent implying different classes.

$$Netconf(X \Rightarrow Y) = \frac{Sup(X \Rightarrow Y) - Sup(X)Sup(Y)}{Sup(X)(1 - Sup(X))} \quad (4)$$

At the classification stage, in order to classify an unseen transaction t , the best rules covering t are selected. In previous works [1,2,15], a CAR $X \Rightarrow c$ satisfies or covers a transaction t iff the transaction t contains all the items of X . Furthermore, in these works when no rule covers a transaction, the default class (majority class in all cases) is assigned, which could influence positively or negatively the results, hiding the real classification accuracy.

In this paper, we introduce and prove a new proposition, which guarantees to use the minimum Netconf threshold that avoids ambiguity at the classification stage, independently of the dataset. Additionally, we propose a relaxed coverage criterion, which allows to reduce the number of non-covered unseen-transactions. The relaxed coverage criterion together with the use of the introduced proposition allows CAR-NF⁺ to have better accuracy than HARMONY, DDPMine and CAR-NF, the best classifiers based on CARs reported in the literature.

2 Our Proposal

In this section we introduce the improvements made to CAR-NF classifier. As was mentioned above, in [15], the authors introduced and proved the following proposition:

Proposition 1. *Let $C = \{c_1, c_2, \dots, c_m\}$ be the set of predefined classes, for each itemset X we can obtain at most one rule $X \Rightarrow c_k$ ($c_k \in C$) with Netconf value greater than 0.5.*

Based on this proposition, if we set the Netconf threshold to 0.5, we can avoid ambiguity at the classification stage. However, in [15] it was not proved that, for all datasets, 0.5 is the minimum Netconf threshold value that avoids ambiguity at classification stage. In this paper, we propose and prove a more robust proposition that guarantees, no matter the dataset, the minimum Netconf threshold value that must be used for avoiding ambiguity at the classification stage:

Proposition 2. *Let X be an itemset and $C = \{c_1, c_2, \dots, c_m\}$ be the set of predefined classes, independently of the dataset, we have that:*

- a) *If $|C| > 2$ then 0.5 is the minimum Netconf threshold value that avoids ambiguity in CARs having X as antecedent.*
- b) *If $|C| = 2$ then 0 is the minimum Netconf threshold value that avoids ambiguity in CARs having X as antecedent.*

Proof.

- a) If $|C| > 2$, we just have to find a dataset containing an itemset X where any Netconf value smaller than 0.5 does not avoid ambiguity in CARs having X as antecedent. Suppose a dataset D having m classes $\{c_1, c_2, \dots, c_m\}$ such that $|D_{\{c_1\}}| = |D_{\{c_2\}}|, |D_{\{c_3\}}| = \dots = |D_{\{c_m\}}| = 1$, and containing the itemset X in all transactions of classes c_1 and c_2 , and only in these transactions (see Table 1).

Table 1. Dataset D used for the proof of Proposition 2

Id_transaction	Itemset	Class
t_1	$\dots X \dots$	c_1
$\dots$	$\dots$	$\dots$
t_n	$\dots X \dots$	c_1
t_{n+1}	$\dots X \dots$	c_2
$\dots$	$\dots$	$\dots$
t_{2n}	$\dots X \dots$	c_2
t_{2n+1}	$\dots Y \dots$	c_3
$\dots$	$\dots$	$\dots$
t_{2n+m-2}	$\dots Y \dots$	c_m

In the dataset D , for all Netconf threshold $\alpha < 0.5$ we have two CARs $X \Rightarrow c_1$ and $X \Rightarrow c_2$ having Netconf values greater than α (see Eq. 5).

$$\begin{aligned}
 \text{Netconf}(X \Rightarrow c_1) = \text{Netconf}(X \Rightarrow c_2) &= \frac{\frac{n}{2n+m-2} - \frac{2n}{2n+m-2} \frac{n}{2n+m-2}}{\frac{2n}{2n+m-2} (1 - \frac{2n}{2n+m-2})} \\
 &= \frac{\frac{n}{2n+m-2} (1 - \frac{2n}{2n+m-2})}{\frac{2n}{2n+m-2} (1 - \frac{2n}{2n+m-2})} = 0.5 > \alpha
 \end{aligned} \tag{5}$$

Therefore, we can conclude that when we have more than two classes, 0.5 is the minimum Netconf threshold value that avoids CAR ambiguity at classification stage, independently of the dataset.

b) If $|C| = 2$, we first will show that 0 avoids ambiguity by proving that:

$$Netconf(X \Rightarrow c_1) + Netconf(X \Rightarrow c_2) = 0 \quad (6)$$

Since we have only two classes (c_1 and c_2), and each transaction has one and only one class, the following statements are satisfied:

$$\sum_{i=1}^2 Sup(X \Rightarrow c_i) = Sup(X) \quad (7)$$

$$\sum_{i=1}^2 Sup(\{c_i\}) = 1 \quad (8)$$

From equations 4, 7 and 8:

$$\begin{aligned} \sum_{i=1}^2 Netconf(X \Rightarrow c_i) &= \frac{\sum_{i=1}^2 Sup(X \Rightarrow c_i) - \sum_{i=1}^2 Sup(X) Sup(\{c_i\})}{Sup(X)(1 - Sup(X))} \\ &= \frac{\sum_{i=1}^2 Sup(X \Rightarrow c_i) - Sup(X) \sum_{i=1}^2 Sup(\{c_i\})}{Sup(X)(1 - Sup(X))} \quad (9) \\ &= \frac{Sup(X) - Sup(X) * 1}{Sup(X)(1 - Sup(X))} = 0 \end{aligned}$$

Therefore, $Netconf(X \Rightarrow c_1) = Netconf(X \Rightarrow c_2) = 0$ or both Netconf values have different signs. In both cases, we can obtain at most one CAR with Netconf value greater than 0, avoiding the ambiguity.

On the other hand, for all Netconf threshold $\alpha < 0$ we can find sets of transactions (see Table 2) where $Netconf(X \Rightarrow c_1) = Netconf(X \Rightarrow c_2) = 0$ and hence, greater than α (see Eq. 10).

$$Netconf(X \Rightarrow c_1) = Netconf(X \Rightarrow c_2) = \frac{\frac{n}{2n+2} - \frac{2n}{2n+2} \frac{n+1}{2n+2}}{\frac{2n}{2n+2} (1 - \frac{2n}{2n+2})} = 0 \quad (10)$$

Table 2. Set of transactions with only two classes, where CARs $X \Rightarrow c_1$ and $X \Rightarrow c_2$ have a Netconf value equal to 0

Id_transaction	Itemset	Class
t_1	$\cdots X \cdots$	c_1
$\cdots$	$\cdots$	$\cdots$
t_n	$\cdots X \cdots$	c_1
t_{n+1}	$\cdots Y \cdots$	c_1
t_{n+2}	$\cdots X \cdots$	c_2
$\cdots$	$\cdots$	$\cdots$
t_{2n+1}	$\cdots X \cdots$	c_2
t_{2n+2}	$\cdots Y \cdots$	c_2

Therefore, we can conclude that when we have only two classes, 0 is the minimum Netconf value that avoids CAR ambiguity at classification stage, independently of the dataset.

□

Taking into account the previous proposition, we propose to compute the CARs using a Netconf threshold value equal to 0 when we have only two classes, and a Netconf threshold value equal to 0.5 when we have more than two classes.

As it was mentioned in section 1, in previous works, a CAR r covers a transaction t if the antecedent of the rule r is a subset of t (exact coverage). Using this criterium, for the set of interesting CARs shown in Table 3, if you want to classify the transactions $\{i_2, i_3\}$ or $\{i_2, i_3, i_4\}$ then the majority class will be assigned (or the classifier refuses to classify) because, using exact coverage, these transactions will not be covered by any CAR of Table 3 although $\{i_2, i_3\}$ and $\{i_2, i_3, i_4\}$ appear in the antecedents of some rules.

Table 3. Example of a set of interesting CARs

CAR	Netconf
$\{i_1\} \Rightarrow c$	0.53
$\{i_1, i_2\} \Rightarrow c$	0.53
$\{i_1, i_2, i_3\} \Rightarrow c$	0.55
$\{i_1, i_2, i_3, i_4\} \Rightarrow c$	0.57
$\{i_5\} \Rightarrow c$	0.52
$\{i_5, i_6\} \Rightarrow c$	0.54
$\{i_5, i_6, i_7\} \Rightarrow c$	0.54

In order to reduce the number of non-covered unseen-transactions for those cases where the exact coverage does not work, we propose to relax the coverage criterion as follows; given a CAR $X \Rightarrow c$, with $|X| = n \geq 2$, it will cover a transaction t if there is an itemset $X' \subset t$ such that $X' \subset X$ and $|X'| = n - 1$.

Using the relaxed coverage criterion, the transactions $\{i_2, i_3\}$ and $\{i_2, i_3, i_4\}$ of our example would be covered by the CARs $\{i_1, i_2, i_3\} \Rightarrow c$ and $\{i_1, i_2, i_3, i_4\} \Rightarrow c$ respectively. This coverage criterion allows reducing the number of non-covered unseen-transactions, which could positively influence the results. The CAR-NF⁺ classifier is the result of applying the Proposition 2 jointly with the relaxed coverage criterion in the classification phase. Similar to the CAR-NF classifier, CAR-NF⁺ classifies an unseen transaction t applying the “Best K rules” satisfaction mechanism [15]. We first use the CARs that exactly cover t ; if no CAR exactly covers t then CAR-NF⁺ applies the relaxed coverage criterion. The class with highest average Netconf is assigned to t . If there is a tie, one of the tied classes is randomly assigned. If no rule covers t (exactly or not), CAR-NF⁺ refuses to classify t , and such abstentions are counted as errors.

3 Experimental Results

In this section, we present the results of our experimental comparison between CAR-NF⁺ and the main classifiers based on CARs reported in the literature: HARMONY [6], DDPMine [10] and CAR-NF[15]. The codes of DDPMine and CAR-NF were provided by their authors and for HARMONY, we used the accuracy values reported in [6]. All our experiments were done using ten-fold cross-validation reporting the average over the ten folds. The same folds were used for

Table 4. Classification accuracy

Dataset	HARMONY	DDPMine	CAR-NF	CAR-NF ⁺
adult	81.90	82.82	83.42	84.50
anneal	91.51	90.86	93.43	95.38
breast	92.42	86.53	85.26	85.43
connect4	68.05	67.80	62.18	62.18
dermatology	62.22	63.42	78.78	79.66
ecoli	63.60	64.25	82.36	84.01
flare	75.02	77.10	86.31	86.45
glass	49.80	53.61	67.89	68.92
heart	56.46	57.19	56.79	57.34
hepatitis	83.16	82.29	85.87	87.02
horseColic	82.53	81.07	83.25	83.56
ionosphere	92.03	93.25	84.34	86.02
iris	93.32	94.03	96.67	96.67
led7	74.56	73.98	74.53	75.88
letRecog	76.81	76.12	71.14	73.42
mushroom	99.94	100.00	99.52	99.52
pageBlocks	91.60	93.24	92.44	94.93
penDigits	96.23	97.87	78.04	78.32
pima	72.34	75.22	77.65	78.53
waveform	80.46	83.83	74.68	75.22
Average	79.20	79.72	80.73	81.65

all classifiers. All the tests were performed on a PC with an Intel Core 2 Duo at 1.86 GHz CPU with 1 GB DDR2 RAM, running Windows XP SP2.

Similar to other works [1,2,6,10,15], we used several datasets, specifically 20 datasets also reported in [6,10,15]. The chosen datasets were originally taken from the UCI Machine Learning Repository [9], and their numerical attributes were discretized by Frans Coenen using the LUCS-KDD [4] discretized/normalized ARM and CARM Data Library.

In Table 4, the results show that CAR-NF⁺ yields an average accuracy higher than all other evaluated classifiers, having a difference of 0.92% with respect to the classifier in the second place (CAR-NF). The used datasets have between 2 and 26 classes [15]. In order to show the accuracy of CAR-NF⁺ on the different datasets, we grouped them according to the number of classes, in three groups: (1) datasets with 2 or 3 classes; (2) datasets with 4,5, 6 or 7 classes and (3) datasets with more than 7 classes. In Table 5, we report the average accuracy of CAR-NF⁺ and CAR-NF (the second best) classifiers over these three groups of datasets. From Table 5, we can see that CAR-NF⁺ obtained the best results, independently of the number of classes in the datasets.

Table 5. Average accuracy of CAR-NF and CAR-NF⁺ over datasets grouped by number of classes

# classes	CAR-NF	CAR-NF ⁺
2 or 3	85.63	86.27
4, 5, 6 or 7	75.25	76.40
more than 7	78.48	79.62

In Table 6, we show the percent of abstentions and the accuracy of CAR-NF⁺ using exact coverage and relaxed coverage. When we used relaxed coverage, CAR-NF⁺ had, in average, 1.26% less abstentions than using exact coverage and its average accuracy was improved in 0.82%. Additionally, in Table 6, we highlight the datasets with only two classes in order to show how using 0 as Netconf threshold improves the accuracy results (see the accuracy of the original CAR-NF classifier, in Table 4, for the highlighted datasets).

Table 6. Percentage of abstentions (%Abst.) and accuracy (Acc.) of CAR-NF⁺ using exact coverage (EC) and relaxed coverage (RC)

Dataset	CAR-NF ⁺ (EC)		CAR-NF ⁺ (RC)	
	%Abst.	Acc.	%Abst.	Acc.
adult	1.55	83.76	0.44	84.50
anneal	3.46	93.43	0.99	95.38
breast	6.84	85.32	5.56	85.43
connect4	0.86	62.18	0.85	62.18
dermatology	5.16	78.78	3.34	79.66
ecoli	1.98	82.36	0.33	84.01
flare	1.20	86.31	1.04	86.45
glass	4.15	67.89	2.60	68.92
heart	5.13	56.79	3.67	57.34
hepatitis	2.87	86.45	0.72	87.02
horseColic	2.72	83.25	1.81	83.56
ionosphere	5.70	85.12	3.80	86.02
iris	1.48	96.67	0.74	96.67
led7	2.19	74.53	0.49	75.88
letRecog	3.01	71.14	0.56	73.42
mushroom	0.27	99.52	0.23	99.52
pageBlocks	3.19	92.44	2.50	94.93
penDigits	6.19	78.04	5.70	78.32
pima	4.63	77.91	3.04	78.53
waveform	1.96	74.68	0.96	75.22
Average	3.23	80.83	1.97	81.65

Table 7. Pairwise comparison between all evaluated classifiers. Each cell (i, j) contains the number of datasets where the classifier i significantly Win/Lose to the classifier of column j , over the 20 selected datasets.

	HARMONY	DDPMine	CAR-NF	CAR-NF ⁺
HARMONY		2/7	4/11	4/13
DDPMine	7/2		5/9	5/12
CAR-NF	11/4	9/5		0/7
CAR-NF ⁺	13/4	12/5	7/0	

Finally, in order to determine if the results shown in Table 4 are statistically significant, we performed a pairwise comparison between all tested classifiers. Each cell (i, j) in Table 7 contains the number of datasets where the classifier of row i significantly Win/Lose to the classifier of column j . We detected ties using a one-tailed T-Test [7] with significance level of 0.05. The results in the pairwise comparison reveal that CAR-NF⁺ beats in accuracy all other evaluated classifier, including CAR-NF, over most of the tested datasets.

4 Conclusions

In this paper, an improved version of the CAR-NF classifier, called CAR-NF⁺, was presented. CAR-NF⁺ combines the use of an appropriated Netconf threshold value that avoids ambiguity at classification stage with a novel relaxed coverage criterion, which reduces the number of non-covered unseen-transactions. Experiment results show that CAR-NF⁺ beats the best classifiers based on CARs reported in the literature, including the original CAR-NF classifier.

References

1. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings of the KDD, New York, NY, USA, pp. 80–86 (1998)
2. Li, W., Han, J., Pei, J.: CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings of the ICDM, pp. 369–376 (2001)
3. Berzal, F., Blanco, I., Sánchez, D., Vila, M.A.: Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis* 6(3), 221–235 (2002)
4. Coenen, F.: The LUCS-KDD discretised/normalised ARM and CARM Data Library (2003), <http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN>
5. Ahn, K.I., Kim, J.Y.: Efficient Mining of Frequent Itemsets and a Measure of Interest for Association Rule Mining. *Information and Knowledge Management* 3(3), 245–257 (2004)
6. Wang, J., Karypis, G.: HARMONY: Efficiently mining the best rules for classification. In: Proceedings of SDM, pp. 205–216 (2005)
7. Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7, 1–30 (2006)
8. Steinbach, M., Kumar, V.: Generalizing the notion of confidence. *Knowl. Inf. Syst.* 12(3), 279–299 (2007)
9. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository (2007), <http://www.ics.uci.edu/~mllearn/MLRepository.html>
10. Cheng, H., Yan, X., Han, J., Philip, S.: Yu: Direct Discriminative Pattern Mining for Effective Classification. In: Proceedings of the ICDE, pp. 169–178 (2008)
11. Karabatak, M., Ince, M.C.: An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.* 36, 3465–3469 (2009)
12. Park, S.H., Reyes, J.A., Gilbert, D.R., Kim, J.W., Kim, S.: Prediction of protein-protein interaction types using association rule based classification. *BMC Bioinformatics* 10(1) (2009)
13. Bae, J.K., Kim, J.: Integration of heterogeneous models to predict consumer behavior. *Expert Syst. Appl.* 37, 1821–1826 (2010)
14. Hernández, R., Carrasco, J.A., Martínez, J.: Fco. and Hernández, J.: Classifying using Specific Rules with High Confidence. In: Proceedings of the MICAI, pp. 75–80 (2010)
15. Hernández, R., Carrasco, J.A., Martínez, J.F., Hernández, J.: CAR-NF: A Classifier based on Specific Rules with High Netconf. *Intelligent Data Analysis* 16(1) (2012)