Class-Association-Rules Pruning by the Profitability-of-Interestingness Measure: Case Study of an Imbalanced Class Ratio in a Breast Cancer Dataset

Peera Liewlom
Kasetsart University, Thailand
Email: peera.li@ku.th

***Abstract*—The Association Rules Discovery is a technique widely used for various objectives. One is for Classification Based on Associations (CBA) with Class Association Rules (CARs). The number of rules discovered from data is extremely high with exponential numbers related to item types in the data. Thus, pruning uninteresting rules is a very important task with this technique. In the traditional technique, minimum Support and minimum Confidence are the main interestingness measures defined by the user for pruning tasks. However, some interesting rules have low Support or Confidence and are pruned at the same time as uninteresting rules. This problem usually occurs with an imbalanced Class ratio in a dataset such as the Scientific or Health dataset, positive-Class CARs usually have a much smaller number than negative-Class CARs. Positive-Class CARs are usually found to have low Support or Confidence that need trust in use without uninteresting rules. In this paper, we describe this problem in relation to a breast cancer dataset, and use a pruning task to discover interesting positive CARs even with low Support or Confidence. We propose a new measure called the Profitability-of-Interestingness Measure (PoI) to prune positive-class CARs from the dataset. Performance is measured by accuracy, precision, and recall. The results show that Pruned CARs have similar accuracy to traditional CARs. A comparison of the same rules for CBA Classifiers shows that Pruned CARs offer more precision than traditional CARs. The Pruned CARs set is more concise and easier to understand because of the lower number of confusing rules.**

***Index Terms*—association rules pruning, class association rules, interestingness measure, profitability of interestingness**

I. Introduction

Association Rule Discovery [1], [2] is one of the important techniques in Descriptive Data Mining. The rule is formed as {antecedent items} ⇒ {consequence items} so that it is easy to understand what items are antecedents and related to items as a consequence. This technique can be implemented with Predictive Data Mining [3] on the Scientific or Health datasets to easily understand the relations between Class and items, or attributes, in the form {items} ⇒ {Class}.

However, Scientific or Health datasets usually have an imbalanced Class ratio in datasets, a positive class ratio usually has a small number. When we use traditional measures such as Support and Confidence for discovering rules with these datasets, the rules discovered lead to the discovery of rules with a negative class. Rules with a positive Class usually have low support, including uninteresting rules that have low support too. However, the quality of rules with a positive Class is very important in a scientific dataset for understanding scientific reasoning. We need a new measure for pruning uninteresting rules with the same minimum Support or minimum Confidence as interesting rules even with very low Support or Confidence. The problem and our solution are explained in detail next.

The main concept in discovering rules is the implementation of many interestingness measures. For traditional measures [1], [2], Support is for discovering Frequent Item Sets, and Confidence is for discovering Association Rules from a Frequent Item Set. Other popular measures are conviction [4], lift [5] etc.

Special rules having only one "Class" on the right hand side of the rule are called Class Association Rules or CARs [3], which are used for classification with a technique called Classification Based on Association Rules or CBA.

In this paper, we focus on solving the major problems of CAR pruning or CAR filtering. An extreme number of rules can be pruned with higher minimum Support and higher minimum Confidence. Nevertheless, in a dataset with an Imbalanced Class Ratio, rules with smaller ratios may disappear, though having low minimum Support and low minimum Confidence. There are problems in many medical datasets that have a smaller ratio for Positive Class. The ratio of patients is usually smaller than for usual cases. So most CARs with an Imbalanced Class Ratio for medical datasets are Negative Classifiers because the CBA method has important rules requiring more Confidence and Support to be implemented first.

However, discovering and pruning Positive-Class CARs are still important issues in the real world. We

Manuscript received November 6, 2020; revised April 22, 2021.

doi: 10.12720/jait.12.3.246-252

scope the problem so that only Positive-Class CARs are implemented first in the CBA Classifier. So all Negative-Class CARs can be replaced by one rule, a default rule as the last order of the CBA Classifier. Details of Negative-Class CARs are omitted in this paper.

Previous CARs Pruning was implemented by a measure called eLift [6], or extended Lift. Each CAR is compared with Extended CARs that have more one member in an item set on the left-hand side of the rule. Extended CARs or longer rules are pruned when these rules have a lower Lift value. However, eLift can be reduced to calculate the Confidence value. Extended CARs are pruned when these rules have lower Confidence values. So both the pruned CAR set and the unpruned CAR set with eLift give the same CBA Classifier because of the implementation of a higher Confidence rule first that prunes all Extended CARs with low confidence via the anti-monotone principle [7], which is enabled in a case where rules have the same Class item set on the right-hand side of rules.

In our previous paper, we developed a new measure called the Profitability of Interestingness or PoI [8]. This measure is only used to represent the tree structure of CARs. However, the previous work did not define the measure and also did not prove its performance. In this paper, we found that PoI can be used to prune CARs for the CBA classifier. Pruned CARs using PoI give accuracy as good as traditional CARs, and it give better performance in some ways, as explained in part VI. So, we conclude that the PoI measure can be used for pruning CARs.

To prove the PoI measure for pruning Positive-Class CARs, we select a Breast Cancer Dataset [9]. This small medical dataset is an imbalanced class ratio dataset that is suitable to prove the research problem. Moreover, this medical dataset can be trusted more than a behaviour dataset, such as buying behaviour datasets. The maximum number of rules generated from this dataset is 64,743 of which 25,103 are Positive-Class Rules.

In part II, we detail related work. In part III, we explain the definitions and details of Association Rules, CARs, and CBA. In part IV, we show the analytical aspect of the research problem in the Breast Cancer Dataset and explain why we select the PoI measure for pruning the CARs. In part V, we explain the implementation of PoI for pruning Positive-Class CARs. Then we detail the framework to test the quality of these pruned rules via the performance of the classification in comparison to the traditional rules. In part VI, we show the results of this process and conduct a discussion. Lastly, we show our conclusion in part VII.

II. Related Works

The pruning task in traditional methods [1], [2] uses interestingness measures, i.e. Support and Confidence. Many uninteresting rules are pruned in two steps: in step1, generate Candidate Item Sets that have Support that is less than minimum Support; in step2, generate Candidate Rules that have Confidence that is less than minimum Confidence. This pruning uses the characteristics of the Lattice structure to describe anti-monotone [7]. These characteristics consider each pair of a subset-superset of item sets. The process to generate a superset of item sets is called the specialization of item sets, the other is called the generalization of item sets.

Thus, lattice pruning can be applied with many methods, except the traditional method that uses minimum Support and minimum Confidence. Some papers prune sensitive items via rule discovering, called association rule hiding [10], [11], by defining sensitive items and then control the support for them. One paper [6] developed a measure called elift to hide discrimination items.

Some papers [12], [13] use these characteristics to prune a dataset via a taxonomy, e.g. sales data for food-fast food-burgers. These papers prune uninteresting item sets by defining many values for minimum support for multi-level frequent item sets. This is suitable for discovering interesting rules with low support at a low level. However, the rules for different levels may conflict [14].

Pruning using these techniques focuses on defining or adjusting the Support for items or item sets. Thus, uninteresting rules with high Support or Confidence may still exist.

In our previous work [8], we used a measure called Profitability-of-Interestingness (PoI) only to represent an item set tree. In this paper, we assign this measure to prune uninteresting rules that exceed minimum Support and minimum Confidence. We test the quality of these pruned rules via the performance of the classification. The results show the ability to prune uninteresting rules that exceed minimum Support and minimum Confidence. Related definitions for and details of the new pruning measure are described in the following.

III. Definitions

A. Association Rules

Association Rules are rules representing the relation of items or item sets in a dataset. In a database view, an item is the attribute value in each record, and an item set is a set of items that have no redundant members in the set.

The Association Rule consists of an item set on the left-hand side of the rule (or {LHS}) and an item set on the right-hand side of the rule (or {RHS}) in the form:

$$\{\text{LHS}\} \Rightarrow \{\text{RHS}\}$$

B. Support

The Support for a rule is the ratio of the number of records for all members of the rule to all the records in a dataset. Support can be calculated by [1], [2] or by the formula:

The number of ({LHS} $\cup$ {RHS}) records / the number of all records in the dataset.

C. Confidence

The Confidence of a rule is the ratio of the number of {LHS} records and {RHS} records to the number of

{LHS} records. Confidence can be calculated by [1], [2] or by the formula:

The number of ({LHS} ∪ {RHS}) records / the number of {LHS} records.

D. Interestingness Measures for Discovering Association Rules

Association Rules are rules that have Confidence more than or equal to minimum Confidence, abbreviated as minConf, and they have the Support of a rule more than or equal to minimum Support, abbreviated as minSup. Both minConf and minSup are defined by the user.

E. Class Association Rules

Class Association Rules (CARs) [3] are Association Rules that have only one Class item member on the {RHS}.

F. Classification Based on Association Rules

Classification Based on Association Rules (CBA) [3] is a method for forming a rule-based classifier from CARs with the format <r1, r2, r3,.., rn, default rule>. This classifier is implemented from r1 to rn in order to classify unclassed objects. Each Class of unclassed objects is classified by Class item on the {RHS} of the first rule where the {LHS} matches the object. If no rules are implemented, then the default rule is selected to classify the class of objects that has the most probability. The method to create a CBA classifier is detailed in Fig. 1.

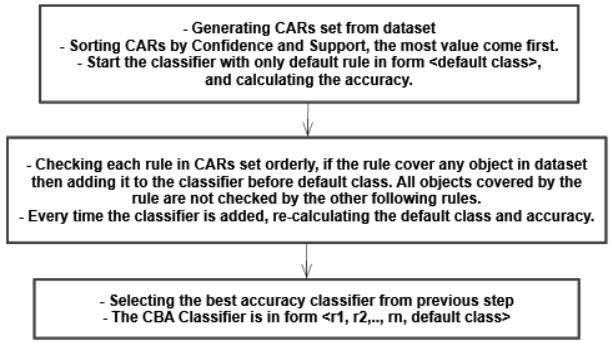

Figure 1. Summary of CBA classifier creation.

Fig. 1 demonstrates CBA Classifier Creation. This method starts by discovering all the CARs for a training dataset, then sorting all CARs by Confidence and Support from a large number to a small number. Each rule is compared with all records in the training dataset. A record will be marked as matched when it matches a rule, and unable to be compared with remaining rules. Rules are added that match any record to the classifier and end the classifier with a default rule. Every time we find there is implementation of a rule in a CARs set, we get a CBA classifier. Last, we select the classifier that has the greatest accuracy.

IV. Problems of CARs from Breast Cancer Dataset

Breast Cancer is a real dataset collected for using with a data-mining program called WEKA [9]. This dataset is an imbalance class dataset with 286 records where the Positive Class has only 85 records. However, the maximum number of CARs is 64,743 rules, and the maximum number of Positive-Class CARs is 25,103 rules. The Positive Class is Class=recurrence-events, abbreviated as Class+, and the Negative Class is Class=no-recurrence-events, abbreviated as Class-.

Determining the conditions for the dataset used in this comparative research must be done with caution in order to display the results clearly. We begin by eliminating conflict rules where the {LHS} of rules is the same but give different classes on the {RHS} of rules. We define minConf = 0.501 (greater than 0.5), leaving 49,586 CARs, and we found CARs of Class+ without conflict rules = 15,157 rules.

We then analyze the effects of various values of minConf that affect the discovery of CARs, both Class+ CARs, and Class- CARs. We avoid interference by Support by defining minSup=0. The results are shown in Table I.

Next, we analyze the effects of various minSup that affect the discovery of CARs too. We avoid interference by Confidence by defining the least minConf=0.501. The results are shown in Table II.

TABLE I. Numbers of Class+ CARs and Class- CARs at Various Minimum Confidence Levels from the Breast Cancer Dataset

minConf	CARs number	Class+ CARs number	Class- CARs- number	Ratio of Class+ CARs number / Class- CARs number
0.501	49,586	15,157	34,429	0.4402
0.6	49,233	15,045	34,188	0.4401
0.7	45,639	13,937	31,702	0.4396
0.8	43,626	13,676	29,950	0.4566
0.9	41,820	13,537	28,283	0.4786

From Table I, it is clear that every minConf value makes little difference to the numbers of Class+ CARs and Class- CARs. The ratio of Class+ CARs number to Class- CARs number is around 0.43–0.48.

TABLE II. Numbers of Class+ CARs and Class- CARs at Various Levels of Minimum Support from the Breast Cancer Dataset

minSup	Class+ CARs number	Class- CARs- number	Ratio of Class+ CARs number / Class- CARs number
0.000	15157	34,429	0.4402
0.005	3562	14112	0.2524
0.010	1228	8074	0.1521
0.015	324	4165	0.0778
0.020	190	3174	0.0599
0.025	95	2105	0.0451
0.030	67	1803	0.0372
0.035	38	1328	0.0286
0.040	30	1146	0.0262

0.045	25	1022	0.0245
0.050	17	800	0.0212
0.100	2	261	0.0077
0.150	1	116	0.0086
0.200	0	60	0.0000

However, from Table II, the results clearly show that even a small increase in minSup value has a great effect on the ratio of Class+ CARs number to Class- CARs number. These characteristics lead to creating a Negative Classifier in CBA. For example, the Ratio of Class+ CARs number / Class- CARs number is reduced from 0.440 at minSup = 0 to 0.152 at minSup = 0.01, and is only 0.009 when minSup = 0.15 (leaving only 1 record of Class+ CARs). Initially, we made the assumption that characteristics are caused by rules generating the lattice structure [7]. The increasing numbers of rules for the lattice start from different Support values. So, the lag for the lattice structure in the same dataset gives these characteristics.

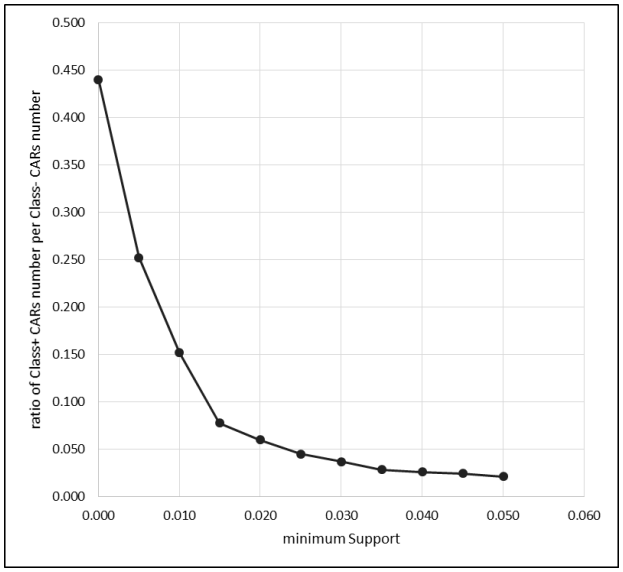

Figure 2. Ratio of class+ CARs number to class- CARs number at various levels of minimum support from the breast cancer dataset.

In Fig. 2, we show the characteristics of these unbalanced rule numbers in a graph view. CARs pruning in this paper must be implemented carefully. We select minSup=0.035 as the last point in Fig. 2. before the ratio of Class+ CARs number to Class- CARs number increases markedly.

The basic concepts of Association Rules pruning are various. Some concepts focus on the development of an interestingness measure as in [4], [5] or compare performance as in [15]. Some concepts focus on reducing the elements of item sets for rule generating as in [16]. Some concepts focus on pruning the most redundant rules as in [6]. The concept of CARs pruning in this paper is to prune the most redundant CARs that are not affected by the interestingness measure. So, to identify the most redundant CARs the focus is on CARs having a redundant sub-item set. For example, {a,b} ⇒ {Class=X} and {a,b,c} ⇒ {Class=X} are the most redundant sub-item set rules, and the members on the {LHS} differ by only 1 member. The rule {a,b} ⇒ {Class=X} is a general rule compared with the others. So, the rule {a,b,c} ⇒ {Class=X} is a specific rule compared with the first rule. If a specific rule has a small interestingness value compared with others, then prune this rule.

Usually, using the interestingness measure as Support is based on the anti-monotone principle [7], but Confidence is not part of this principle. So, we can use eLift [6] as a measure for CARs pruning because the reduced form of eLift is only compared to Confidence in most redundant CARs, and so a specific (or longer) rule having less Confidence is pruned.

However, CARs pruning with eLift gives the same CBA classifier as defined in Fig. 1, because the classifier determines the rule with the most Confidence first. A general rule with greater Confidence is determined before a specific rule with less Confidence. All records covered by a specific rule are covered by a general rule, and by the anti-monotone principle too. So, a specific rule does not cover any records remaining from the general rule implemented before. Specific rules with less Confidence than traditional CARs are pruned from a classifier that gives the same result as a specific rule pruned by eLift. So, both cases of traditional CARs and pruned CARs using eLift give the same CBA Classifier.

In this paper, we propose a new measure for pruning the most redundant CARs. We know that Support has more effect than Confidence, but Confidence is important in creating a CBA classifier. Both measures should be used as a single measure for CARs pruning. In detail, a specific rule that extends an item set from a general rule usually has less Support while we expect more Confidence in a specific rule. So we can use the concept of the "Profitability of Interestingness" measure, abbreviation PoI, from our previous paper [8] whereby an Increasing Ratio of Confidence should make more profit than a Decreasing Ratio of Support that costs less. Although specific rules have more Confidence value, the Support value is greatly decreased when compared to a general rule. These specific rules should be pruned by cost loss over profit.

In this paper, we found that PoI can be used for CARs pruning, even with an Imbalance Class Ratio Dataset. The details of PoI for CARs pruning and a framework for comparisons with traditional CARs are described in the next part.

V. Profitability of the Interestingness Measure and a Framework for Comparisons with Traditional CARs

The basic concept of CARs pruning is to prune specific CARs by some measure. Specific CARs are CARs having more {LHS} members than general CARs with only one member, and where the {LHS} of general CARs is a sub-item set of the {LHS} of specific CARs. We propose a new measure called Profitability of Interestingness or PoI to prune specific CARs in this paper.

Each pair of a general Rule and a specific Rule is measured by PoI. PoI is the profit from the Increasing Ratio of Confidence over the cost loss from the Decreasing Ratio of Support, so a specific rule should gain more profit in Confidence than cost loss in Support comparing with the general rule. A PoI that gives a value greater than or equal to 0 is a mean "profit", otherwise a "cost loss". Specific CARs with a "cost loss" should be pruned from the CARs set. So, we define an Increasing Ratio of Confidence as (1), a Decreasing Ratio of Support as (2), and a PoI as (3):

$$Increasing\ Ratio\ of\ Confidence = \frac{(Confidence_{s-rule} - Confidence_{g-rule})}{Confidence_{g-rule}} \quad (1)$$

$$Decreasing\ Ratio\ of\ Support = \frac{(Support_{g-rule} - Support_{s-rule})}{Support_{g-rule}} \quad (2)$$

$$Profitability\ of\ Interestingness = (1) - (2), \begin{cases} If\ PoI \geq 0, mean \text{ profit.} \\ If\ Poi < 0, mean \text{ cost loss.} \end{cases} \quad (3)$$

From (1) and (2), $Confidence_{s\text{-}rule}$ is the Confidence in a specific rule in the CARs set; $Confidence_{g\text{-}rule}$ is the Confidence in a general rule in the CARs set; $Support_{s\text{-}rule}$ is the Support for a specific rule in the CARs set; and $Support_{g\text{-}rule}$ is the Support for a general rule in the CARs set.

One of the problems of implementing PoI is explained in detail as follow. In case A there is specific rule for B and C, or B and C are a general rule of A. The pair (A, B) give PoI = "profit", but the pair of (A, C) give PoI = "cost loss". The question is "A should be pruned or not?", which is answered using the experimental results in this paper.

Therefore, the performance of CBA Classifiers with various CARs involves the comparison of the cases. Case I - a classifier created from a traditional CARs set; case II - a classifier created from a pruned CARs set involving pruning each specific rule having a "cost loss" from some pairs of it and its general rules; and case III - a classifier created from pruned CARs where pruning involves a specific rule for a "cost loss" from all pairs of it and its general rules.

The dataset we choose in this paper is The Breast Cancer Dataset [9], a medical dataset with an imbalanced class ratio. We focus on the Positive Class that has more interest than the Negative Class. The effect of pruning any rule from the compact rules set is clear. Therefore, we choose the whole dataset as a training dataset in order to compare the efficiency of traditional CARs with pruned CARs so as to avoid the uncertainty of a randomized training dataset, as there is high sensitivity in small datasets. This training dataset type is more standardized, especially when continuing with research in this paper. All three cases have the same conditions, minSup = 0.035 and minConf=0.501, with the reasons as specified in part IV. A Framework for generating CARs and pruned CARs for comparisons is described as Fig. 3. Then all three types of CARs are used to generate the CBA classifier described in definitions *F*, for comparisons of accuracy, precision and recall.

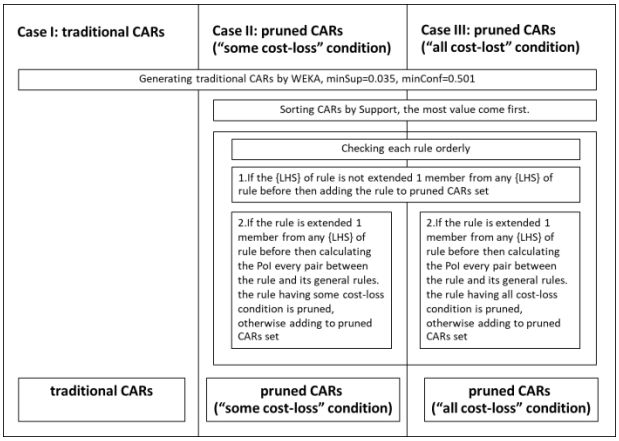

Figure 3. Framework for generating CARs and pruned CARs to compare the performance of CBA classifiers.

VI. Results and Discussion

We generated 38 rules for Positive-Class CARs at minSup=0.035 and minConf=0.501 from the Breast Cancer Dataset from WEKA. All rules with Support and Confidence are shown in Table III.

The 38 CARs are used to create CBA classifiers in three cases. Case I: traditional CARs, this case uses 25 rules for the CBA classifier. Case II: pruned CARs (with "some cost-loss" condition), which uses only 9 rules for the CBA classifier. Case III: pruned CARs (with "all cost-loss" conditions), which uses just 15 rules for the CBA classifier. The performance of all three cases is shown in Table IV, blank cells are pruned by PoI, like the steps in Fig. 3, or by CBA steps, as in definition *F*.

From Table IV, the CBA classifier with six pruned CARs from case III gives the best accuracy of 0.766, equal to the CBA classifier with the same CARs set, the CBA classifier with pruned CARs from case II gives a slightly lower accuracy of 0.762 but only uses four CARs in the classifier. The maximum precision and maximum recall are equal in the first rule and the last rule in all cases.

TABLE III. CARs with Class=Recurrence-Events from the Breast Cancer Dataset at Minimum Support = 0.035 and Minimum Confidence = 0.501

CARs No.	{LHS} of CARs with Class=recurrence-events	Confi-dence	Support
1	{node-caps=yes,deg-malig=3, breast=left, irradiat=yes}	1.000	0.038
2	{node-caps=yes, deg-malig=3, breast=left}	0.889	0.056
3	{deg-malig=3, breast=left, irradiat=yes}	0.875	0.049
4	{node-caps=yes,deg-malig=3, breast-quad=left_low}	0.846	0.038
5	{node-caps=yes, breast=left, irradiat=yes}	0.824	0.049
6	{menopause=premeno, node-caps=yes, deg-malig=3}	0.813	0.045
7	{node-caps=yes, deg-malig=3, irradiat=yes}	0.786	0.038
8	{node-caps=yes, deg-malig=3}	0.767	0.080
9	{node-caps=yes, deg-malig=3, irradiat=no}	0.750	0.042
10	{menopause=premeno, node-caps=yes,	0.733	0.038

	irradiat=no}		
11	{menopause=premeno, node-caps=yes, breast=left}	0.733	0.038
12	{node-caps=yes, breast=left}	0.700	0.073
13	{menopause=premeno, deg-malig=3, breast=left}	0.700	0.049
14	{menopause=premeno, deg-malig=3, irradiat=no}	0.667	0.049
15	{menopause=ge40,breast=left, irradiat=yes}	0.667	0.042
16	{node-caps=yes, breast=left, breast-quad=left_low}	0.667	0.042
17	{deg-malig=3, irradiat=yes}	0.655	0.066
18	{tumor-size=25-29, deg-malig=3}	0.647	0.038
19	{deg-malig=3, breast=left, breast-quad=left_low}	0.625	0.052
20	{menopause=premeno, deg-malig=3}	0.622	0.080
21	{node-caps=yes, breast-quad=left_low}	0.619	0.045
22	{tumor-size=30-34, deg-malig=3}	0.615	0.056
23	{age=40-49, node-caps=yes}	0.611	0.038
24	{breast=left, irradiat=yes}	0.600	0.073
25	{breast=left, breast-quad=left_low, irradiat=yes}	0.600	0.042
26	{deg-malig=3, breast-quad=left_low}	0.594	0.066
27	{tumor-size=30-34, deg-malig=3, irradiat=no}	0.579	0.038
28	{node-caps=yes, irradiat=yes}	0.571	0.056
29	{menopause=premeno, node-caps=yes}	0.563	0.063
30	{deg-malig=3, breast=left}	0.560	0.098
31	{node-caps=yes}	0.554	0.108
32	{menopause=ge40, irradiat=yes}	0.552	0.056
33	{deg-malig=3, breast=right, irradiat=no}	0.545	0.042
34	{age=40-49, deg-malig=3}	0.542	0.045
35	{menopause=ge40, node-caps=yes}	0.542	0.045
36	{node-caps=yes, irradiat=no}	0.536	0.052
37	{breast-quad=left_low, irradiat=yes}	0.536	0.052
38	{deg-malig=3}	0.529	0.157

For all eight CARs, all classifiers give the same accuracy, precision and recall. However, case II uses only four CARs while the others use seven CARs.

To consider all 38 CARs, this rule {deg-malig=3} ⇒ {Class=recurrent-events} is pruned by the CBA classifier with traditional CARs, case I, but rules with the item {deg-malig=3} in the classifier in this case number 15! While case III has just eight rules with {deg-malig=3}, and case II has only four rules with {deg-malig=3}. The explanation of case I is more difficult, why prune rule 38 with {deg-malig=3}? But many rules in the CARs set have {deg-malig=3} as a component of them.

TABLE IV. PERFORMANCE OF CBA CLASSIFIERS WITH TRADITIONAL CARS AND PRUNED CARS IN THREE CASES

CARs Number	case I			case II			case III		
	accu-racy	preci-sion	recall	accu-racy	preci-sion	recall	accu-racy	preci-sion	recall
1	0.741	**1.000**	0.129	0.741	**1.000**	0.129	0.741	**1.000**	0.129
2	0.752	0.889	0.188				0.752	0.889	0.188
3	0.755	0.826	0.224	0.745	0.875	0.165	0.755	0.826	0.224
4	0.755	0.800	0.235				0.755	0.800	0.235
5	0.755	0.742	0.271	0.745	0.773	0.200	0.755	0.742	0.271
6	**0.766**	0.737	0.329				**0.766**	0.737	0.329
7									
8	0.762	0.707	0.341	**0.762**	0.707	0.341	0.762	0.707	0.341
9									
10	0.759	0.667	0.376				0.759	0.667	0.376
11									
12	0.752	0.640	0.376						
13	0.748	0.607	0.435						
14	0.748	0.603	0.447						
15	0.745	0.588	0.471						
16							0.752	0.640	0.376
17	0.745	0.577	0.529						
18	0.738	0.561	0.541						
19	0.731	0.545	0.565						
20									
21									
22	0.734	0.548	0.600						
23	0.727	0.536	0.612						
24	0.713	0.514	0.635	0.745	0.611	0.388	0.734	0.571	0.424
25									
26	0.717	0.519	0.647						
27									
28	0.710	0.509	0.659						
29									
30	0.685	0.479	0.659						
31	0.675	0.467	0.659	0.710	0.514	0.447	0.710	0.514	0.447
32	0.661	0.452	0.671	0.699	0.494	0.482	0.699	0.494	0.482
33	0.654	0.446	0.682				0.689	0.479	0.541
34									
35									
36									
37	0.650	0.444	**0.694**	0.696	0.489	0.506	0.685	0.475	0.565
38				0.650	0.444	**0.694**	0.650	0.444	**0.694**

Moreover, when we compare the accuracy and precision of classifiers in order with the same rules, we find that both pruned CARs usually give better accuracy and precision.

VII. CONCLUSION

These results above lead to uninteresting CARs-pruning ability using PoI. The objective for using this measure is different from Association Rule Hiding or Mining for multi-level frequent item sets that are needed to define or adjust the Support to prune uninteresting rules. Our measure uses existing values of Support and Confidence for rules to prune uninteresting ruled even when their Support and Confidence exceed minimum Support and minimum Confidence. The quality of the pruned CARs is proved by the good performance of CBA compared to a CBA with traditional CARs, while pruned CARs use fewer rules. These results increase the trust in using low Support or low Confidence rules discovered from a scientific dataset or a dataset with an imbalanced Class ratio.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGMENT

We would like to offer special thanks to our friend, Dr. Jaruwat Pailai, for his help with this paper.

REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM Sigmod Record*, vol. 22, pp. 207-216, 1993.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases*, 1994, pp. 487-499.

[3] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining.," in *Proc. the fourth International*

Conference on Knowledge Discovery and Data Mining, 1998, pp. 80-86.
[4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proc. the ACM SIGMOD International Conference on Management of Data*, 1997, pp. 255-264.
[5] M. J. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, ACM, 1997.
[6] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 2, p. 9, 2010.
[7] P. N. Tan, M. Steinbach, and V. Kumar, "Association analysis: Basic concepts and algorithms," in *Introduction to Data Mining*, Boston, MA: Addison-Wesley, 2005.
[8] P. Liewlom, "Representation of Class Association Rules (CARs) with itemsets tree plot: Case study of breast cancer dataset," in *Proc. the 9th National Conference on Information Technology*, Thailand, 2017, pp. 99-104.
[9] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," in *Proc. Australian New Zealand Intelligent Information Systems Conference*, 1994.
[10] Y. K. Jain, V. K. Yadav, and G. S. Panday, "An efficient association rule hiding algorithm for privacy preserving data mining," *International Journal on Computer Science and Engineering*, vol. 3, no. 7, pp. 2792-2798, 2011.
[11] K. S. Rao, V. N. Mandhala, D. Bhattacharyya, and T. Kim, "An association rule hiding algorithm for privacy preserving data mining," *International Journal of Control and Automation*, vol. 7, no. 10, pp. 393-404, 2014.
[12] M. S. Gouider and A. Farhat, "Mining multi-level frequent itemsets under constraints," arXiv preprint arXiv:1012.5546, 2010.
[13] K. Sriphaew and T. Theeramunkong, "A new method for finding generalized frequent itemsets in generalized association rule mining," in *Proc. the Seventh International Symposium on Computers and Communications*, 2002, pp. 1040-1045.
[14] L. Cagliero, T. Cerquitelli, P. Garza, and L. Grimaudo, "Misleading generalized itemset discovery," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1400-1410, 2014.
[15] S. Pitakchonlasup and A. Sapsomboon, "A comparison of the efficiency of applying association rule discovery on software archive using support-confidence model and support-new confidence model," *International Journal of Machine Learning and Computing*, vol. 2, no. 4, p. 517, 2012.
[16] W. A. AlZoubi, "Mining medical databases using graph based association rules," *International Journal of Machine Learning and Computing*, vol. 3, no. 3, p. 294, 2013.

Peera Liewlom received his BSc degree in Chemistry from Mahidol University Thailand in 1992. He received his master degree in Information Technology from Mahidol University Thailand in 1999, and Ph.D. in Computer Engineering from Kasetsart University Thailand in 2007.
He is currently an assistant professor in Information Technology at The Kasetsart University in Thailand. His research interests are Data Mining, Classification, and Information Management.