The Compact Fuzzy Association Rules For Data Classification

Phaichayon Kongchai*, Keerachat Suksut, Rattaphong Sutamma, Sak Phoemhansa, Nittaya Kerdprasop and Kittisak Kerdpraparop

Data Engineering Research Unit, School of Computer Engineering, Suranaree University of Technology 111 University Avenue, Nakhon Ratchasima 30000 THAILAND

*Corresponding Author: Zaguraba_i@hotmail.com

Abstract

Data classification mining is a method to find data generalization in a form of rules then used these rules to predict some unknown value in the future data. But in actual applications, the rules may be of low accuracy and the number of rules may be so overwhelmed that users could not efficiently apply them. Therefore, this research proposes the development of data classification algorithm with compact fuzzy association rules to optimize accuracy and interpretability of the model. To evaluate the performance of the proposed method, this research will compare accuracy of the classification model and the number of rules against 9 different data classification algorithms. The results showed that our CCFAR algorithm is comparable in terms of accuracy. When considering both accuracy and size of model, our algorithm is the best one.

Keywords: Data Classification, Associative Classification, Fuzzy Classification Association Rule, Fuzzy Set.

1. Introduction

Data classification technique is one of a data mining widely used to predict the future data by inducing the model from sophisticated data. For instance, in a medical science the model is used to predict a patient who is risky of having a breast cancer (1). In a commercial bank, the model is used to screen credit requests and evaluate the credit rating of consumer. For its potential benefits, many researches have long been concentrating on improving the efficiency of data classification by proposing algorithms like C4.5(2) and OneR(3).

But these algorithms have a low predictive accuracy. Therefore, Liu and Ma(4) proposed a new method call the "associative classification" which is a combination between the association rule mining technique and the data classification technique. Their proposed algorithm was called the Classification based on Association Rules (CBA). The CBA has been tested its predictive performance by comparing with the C4.5, and it turns out that has high accuracy than C4.5 algorithm. This CBA has drawn attention from many researches to develop algorithm based on associative classification technique, such as GARC(5) algorithm, OAC(6) algorithm, and many more. However, the inherent problem of associative classification is that the association process can handle only symbolic and binary values. It cannot process continuous values. To solve this problem, we propose to use a fuzzy set to transform the continuous value to the degree of membership(7,9).

In this paper, we present the idea and the development of an algorithm called data classification with compact fuzzy association rules (CCFAR). Our main focus is to optimize both an accuracy and interpretability of the model. In addition, we applied the concept of OneR(3) algorithm to select the best rules and reduce the number of rules in the final result.

2. Related Works

Classification task is the mainstream of many researches are concentrating on developing algorithms for increasing the accuracy and reducing the number of rules. We can summarize the researches that are related to our works into 3 groups, that are group of data classification, group of associative classification, and group of fuzzy...
associative classification. The details of 3 groups are as follows:

Firstly, the group of data classification is the initiative concept of data mining to classify target variable with some related features. Examples of this group are as follows:

- Quinlan\(^2\) proposed C4.5 algorithm which is well known in the classification because the algorithm is able to process quickly and the model is easily understandable. The main concept of this algorithm is the use of heuristic search to construct a decision tree and prune tree.

- Cohen\(^9\) proposed an algorithm called RIPPER (Repeated Incremental Pruning to Produce Error Reduction) developed from the IREP\(^8\) algorithm by the principle of growing and pruning techniques to select the low error rate rules.

- Holte\(^3\) proposed an algorithm named OneR or One-Attribute Rule, which is an algorithm that easy-to-understand algorithm and the model is a compact set of rules. OneR is simple because it select a single attribute with the fewest error to build the model.

Secondly, the group of associative classification uses the association mining to build the association rules, then using many techniques of data classification to make association rules appropriate for classification. Examples of this group are as follows:

- Liu and Ma\(^4\) proposed CBA algorithm, which produced is a high accuracy for the data classification. The CBA composed of 2 steps: to build the model with CBA-RG (to create the association rules), and CBA-CB (to make association rules for classification by selecting the low error rate rules).

- Chen and Zhang\(^6\) proposed an algorithm called GARC (Gain Based Association Rule Classification) with a efficacy in classifying data used the model is a compact. Because of the GARC was using the information gain threshold to create frequent item-sets and then it used redundant and conflictive techniques to build the class association rules.

- Hu and Li\(^6\) proposed an algorithm named OAC (Optimal Association Classifier) with a good efficacy in classifying data. The principle of OAC was to generate association rules by using the constraint of apriori\(^10\) algorithm, then it selected the class association rules by the method call OCARM (Optimal Class Association Rule Mining\(^11\)).

Finally, the group of fuzzy associative classification uses the fuzzy set to control the continuous value in the association rule mining processing. Examples of this group are as follows:

- Pach et al.\(^7\) proposed an algorithm called CFARC (Compact Fuzzy Association Rule-Based Classifier) to resolve the problem of how to define the minimum support and minimum confidence in association rule mining by applying fuzzy correlation threshold. In their experimental results, they showed a good accuracy and a compact set of model.

- Chen\(^8\) proposed an algorithm named CFAR (Classification with Fuzzy Association Rules). It generated rules by using the apriori algorithm, then it selected the rules that have highest confidence value and deleted remaining rules with lowest confidence value.

- Hühn and Hüllermeier\(^12\) proposed an algorithm called FURIA (An Algorithm For Unordered Fuzzy Rule) based on the Ripper algorithm. But the FURIA used unordered rule set to reduce bias of the target class and it used stretching technique to obtain the generalized rules.

### 3. Preliminaries

In this section, we introduce the basic definitions of fuzzy sets, fuzzy association rules and fuzzy associative classification rules.

#### 3.1 Fuzzy Sets

Fuzzy sets are sets that cannot explain something clearly. For example, an explanation of the people who are very tall, someone say that the people who are 180 cm. high are very tall, but others may say that the people who are 185 cm. high are very tall. So, we can see from this example that it is impossible to tell exactly regarding who is very tall. Therefore, to solve such problem Zadeh\(^13\) proposed the concepts of fuzzy sets to explain something that is not clear with the degree of membership function (rang value of the degree is [0, 1]). For instance, the people who are 180 cm. high; they are medium tall at the degree of 0.7 and very tall at the degree of 0.3. The people who are 185 cm high, they are medium tall at the degree of 0.1 and very tall at the degree of 0.9. From this example, we can see the people who are 180 cm. high and 185 cm. high would be moderately tall and very tall, respectively with the different degrees.

Therefore, this research applied the concept of fuzzy sets to explain the continuous value of numeric data. We used fuzzy partitions with fuzzy c-means (FCM\(^14\)) algorithm.
3.1 Fuzzy partitions

In this section, we present the important step of our CCFAR algorithm for transforming the continuous value (Fig. 1 a) to partitions with the FCM algorithm. The FCM is used to indicate the degree of membership in a set with the value ranging from 0 to 1 (Fig. 1 c), which is different from the k-means algorithm that indicates the degree of membership in a set with the discrete value (Fig. 1 b).

The processes of FCM are composed four main steps.

1) Initialize C and \( \mu_{ij} \)

2) Calculate the center vectors by Eq.(1)

\[
c_j^{(0)} = \frac{\sum_{i=1}^{N} \mu_{ij}^{(0)m}x_i}{\sum_{i=1}^{N} \mu_{ij}^{(0)m}}
\]

(1)

3) Compute and update membership of data by Eq.(2)

\[
\mu_{ij}^{(t+1)m} = \frac{1}{\sum_{k=1}^{C} \left(\frac{|x_i - c_k|}{|x_i - c_k|}\right)^{m-1}}
\]

(2)

4) If \( \|\mu_{ij}^{(t+1)m} - \mu_{ij}^{(0)m}\| < \varepsilon \) then STOP; otherwise repeat step 2.

Where \( x_i \) is data vector, \( m \) is fuzziness that can be any real number equals or greater than 1, \( \mu_{ij} \) is the degree of membership of \( x_i \) to be in the cluster \( j \), \( N \) is data for clustering, \( C \) is number of clusters and \( c_j \) is center of cluster \( j \).

![Fig. 1. The membership functions of k-means and FCM.](image)

3.2 Fuzzy Association Rule

Association rule is originally an analysis of customer purchases (also called Market Basket Analysis) by storing the items in the basket of the customer into transaction. Then, it analyzed these transactions to discover the association rules. The association rules are expressions of the type \( X \rightarrow Y \), where \( X \) and \( Y \) are sets of items. This means that if customers buy item \( X \), then customers buy item \( Y \) together. But the processing of the association rule mining is unable to handle the continuous value (Table 1). So, we used fuzzy set to convert the continuous value into a degree of belonging to a set. For example, the Table 2 shows the degree of attribute age associated with linguistic values low, medium and high.

In recent years, many researches have proposed methods to mining fuzzy association rules from continuous value\(^\text{15,16}\). The processing of fuzzy association rules are composed of 2 steps. First step, create frequent Item-sets by counting item-sets that have fuzzy support (Eq.3) higher than the minimum fuzzy support. Second step, create fuzzy association rules from frequent item-sets that have fuzzy confidence (Eq.4) higher than the minimum fuzzy confidence.

\[
FS(<Z:A>) = \frac{\sum_{k=1}^{N} \prod_{z_i \in <Z:A>} \mu_{k}(z_i)}{N}
\]

(3)

\[
FC(<X:A> \rightarrow <Y:B>) = \frac{FS(<X:A> \cup <Y:B>)}{FS(<X:A>)}
\]

(4)

Let \( <Z: A> \) be a fuzzy item-set denoted as \( <Z: A> = [<z_{11}: A_{i1}> \cup <z_{12}: A_{i1} > \cup ... \cup <z_{i1}: A_{ij}>] \), where \( q \leq n + 1 \), \( z_i \) is an attribute, such as Age, Income and Balance, \( A_{ij} \) is a fuzzy interval, such as Low, Medium and High, \( t_k \) is a transaction, and \( N \) is the number of transactions.

Table 1. The example of data.

<table>
<thead>
<tr>
<th>Id</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
</tbody>
</table>
Table 2. The degrees of attribute age with FCM algorithm.

<table>
<thead>
<tr>
<th>Id</th>
<th>Age = Low</th>
<th>Age = Medium</th>
<th>Age = High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9863</td>
<td>0.0127</td>
<td>0.0010</td>
</tr>
<tr>
<td>2</td>
<td>0.0119</td>
<td>0.9862</td>
<td>0.0019</td>
</tr>
<tr>
<td>3</td>
<td>0.4996</td>
<td>0.4900</td>
<td>0.0104</td>
</tr>
<tr>
<td>4</td>
<td>0.0074</td>
<td>0.0141</td>
<td>0.9785</td>
</tr>
<tr>
<td>5</td>
<td>0.0053</td>
<td>0.0090</td>
<td>0.9857</td>
</tr>
</tbody>
</table>

3.3 Fuzzy Associative Classification Rule

Fuzzy associative classification rule and Fuzzy association rule are similar techniques in terms of processing steps, but they differ at consequent part of rule. The fuzzy associative classification rule must contain only one class label at consequent part of rule ($C = \{C_1, \ldots, C_k\}$). Therefore, the fuzzy confidence of fuzzy associative classification rule can be defined as follows:

$$FS(<Z:A>) = \frac{\sum_{i=1}^{N} \prod_{k} c_i A_k C_k \cap e <Z,A> t_k(Z_i C_k)}{N}$$  (5)

$$FC(<X:A> \rightarrow <Y:C>) = \frac{FS(<X:A> \cup <Y:C>) - FS(<X:A> \times FS(<Y:C>))}{FS(<X:A>) \times (1 - FS(<Y:C>))}$$  (6)

4. Our Proposed Methodology: CCFAR

In this section, we described our algorithm that has been named data classification algorithm with compact fuzzy association rules (CCFAR) to obtain the fuzzy associative classification rules with optimum combination of accuracy and interpretability of the model. Our methodology is composed of five steps (Fig 2).

1) Data Screening: The data characteristic of CCFAR must be numeric data because of the transformation of fuzzy set that will convert numeric data only. At current stage, our algorithm assumes that there is no missing value in the data.

2) Data Partitioning: This step is transforming the numeric data to fuzzy intervals or fuzzy sets. In the CCFAR algorithm, we used FCM to make the transformation because of it easy-to-understand property and high performance. As an example, we used the data from Table 1 and we defined fuzzy intervals to be 3 partitions; the results were represented in table 2.

Fig. 2. The development of data classification algorithm with compact fuzzy association rules.

3) Frequent Fuzzy Item-Sets Searching: The fuzzy frequent item-sets are fuzzy item-sets that have fuzzy support more than minimum fuzzy support ($\gamma$). At this step, the process are the same as CFARC$^{(7)}$ algorithm (Fig 3).

4) Fuzzy classification association rule (FCARs) generation: This step is similar to the CFARC$^{(7)}$ algorithm (Fig. 4), but it differed in the computation of score values of all rules. In this research we computed score values of all rules by using our new metric as shown in Eq.9 fuzzy correlation (Eq.7), fuzzy confidence (Eq.6) and firing strength($\beta$) (Eq.8). The intuitive idea is that we need a high correlated, high confidence, as well as high support rules.

FCORR ($<X: A> \rightarrow <Y: C>$) =

$$\frac{FS(<X:A> \cup <Y:C>) - FS(<X:A>) \times FS(<Y:C>)}{\sqrt{FS(<X:A>) \times (1 - FS(<X:A>)) \times (1 - FS(<Y:C>))}}$$  (7)

$\beta$ ($<X: A> \rightarrow <Y: C>$) =

$$\sum_{k=1}^{N} \prod_{j=1}^{m} c_i A_j e <Z,A> t_k(Z_i)$$  (8)

$$\text{SCORE} = \text{FCORR} \times FC \times \text{Firing strength} (\beta)$$  (9)
**Frequent fuzzy item set searching** (an Apriori fuzzy implementation)

**Input:** DF fuzzy data

**Output:** the set of frequent fuzzy item set

**Method:**
1. Determine the supports of the classes by the distribution of classes;
2. Set the minimal fuzzy support (γ) to the half of the minimum frequency of classes;
3. Generate the 1-candidate fuzzy items;
4. Calculate FS values then select the frequent fuzzy items from the 1-candidate which has FS > γ, and n = 2;
5. While there exist some n-1 size frequent item sets: Generate the n-size candidate sets from n-1 size frequent (and 1-size frequencies);

![Fig. 3. The frequent fuzzy item set searching(7).](image)

**Fuzzy classification association rule generation**

**Input:** a set of frequent fuzzy item sets

**Output:** positive correlated FCARs separated by size

**Method:**
1. Generate association rules with class label consequent from all the frequent item sets to consider the size of item sets
2. Calculate the Score values of all the rules;
3. Select rules with positive SCORE value for all size;

![Fig. 4. The fuzzy classification association rule generation(7)](image)

5) Fuzzy Classification Association Rule Selection: This step is selecting the FCARs to be used in the data prediction. It’s composing of 4 parts as follows (Fig.5):

- First part: The rules that have the same class label and the same size will be grouped together. After that, this algorithm will select rule with the highest SCORE from each group.
- Second part: The rules with the highest SCORE in each group will be counted the frequency of attributes in antecedent of rule. For example:

  From Table 3, the frequency of attribute Age is 5 (No. 1, 2, 3, 4 and 5), attribute Inc is 3 (No. 3, 5 and 6) and attribute Bal is 1 (No. 2). Thus, the highest frequent attribute is the attribute Age. *If frequencies of attributes are equal, attributes will be randomly selected.

![Fig. 5. The fuzzy classification association rule selection.](image)

- Third part: A set of FCARs with the highest SCORE and shortest rule from each fuzzy interval is selected. For examples, the rules from Table 3 that contain attribute Age (the highest frequency) are No.1, 2, 3, 4 and 5. After that, it selects the FCARs from the rules No. 1-5 with the highest SCORE and the shortest rule, which are rule No. 1(Shortest rule of Age = high), 2 (Highest SCORE of Age = high), and 4 (highest SCORE and shortest rule of Age = medium). But this selection is incomplete became the rule that contains Age = low is missing. Therefore, this algorithm will create new rule that contains Age = low with don’t cared support and SCORE values, such as:

  Rule 1: Age = low → yes, SCORE = -0.12
  Rule 2: Age = low → no, SCORE = 0.12

  In this algorithm, the Rule 2 will be selected because it has positive SCORE value. Thus, we can conclude the results as follows:
  1. Age = high → yes
  2. Age = high and Bal = low → no
  3. Age = medium → no
  4. Age = low → no

- Fourth part: we will remove rules that are superset in antecedent part (left-hand-side) with the same class as other rules. For instance,

  Rule 1: Age = low → yes,
  Rule 2: Age = low and Bal = high → yes

  From example, we removed the Rule 2 because of it is superset in antecedent part and has same class of Rule 1.
Table 3. The FCARs with SCORE.

<table>
<thead>
<tr>
<th>No.</th>
<th>FCARs</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age = high → yes</td>
<td>1.2725</td>
</tr>
<tr>
<td>2</td>
<td>Age = high and Bal = low → no</td>
<td>1.9225</td>
</tr>
<tr>
<td>3</td>
<td>Age = high and Inc = me → no</td>
<td>1.1619</td>
</tr>
<tr>
<td>4</td>
<td>Age = medium → no</td>
<td>1.2608</td>
</tr>
<tr>
<td>5</td>
<td>Age = medium, Inc = high → no</td>
<td>0.8432</td>
</tr>
<tr>
<td>6</td>
<td>Inc = high → yes</td>
<td>1.1232</td>
</tr>
</tbody>
</table>

Table 4. Data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Size</th>
<th># Attribute</th>
<th># Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Heart</td>
<td>270</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>Pima</td>
<td>768</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

5. Experimentation and Results

In this section, we evaluated our CCFAR algorithm using three measurements accuracy: Acc (Eq. 10), number of the models (Compact Value: CV) that we normalized it to be the Normalization of Compact Value: NCV (Eq. 11), and the combination of Acc and NCV in which we call Suitable Rule: SR (Eq. 12). Our algorithm was tested on three different data sets and compared with nine different algorithms. These algorithms are grouped as: data classification, associative classification and fuzzy associative classification. The nine algorithms are namely C4.5, RIPPER, OneR, CBA, GARC, OAC, FURIA, CFAR, and CFARC. The data are taken from the UC Irvine Machine Learning Repository, namely Iris, Heart, and Pima (details shown in Table 4). The classification performances of all algorithms were measured by ten-fold cross validation.

\[
\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}
\]

Where TP is the number of true positive examples, FP is the number of false positive examples, TN is the number of true negative examples and FN is the number of false negative examples.

\[
\text{NCV}_{\text{AL}} = \frac{\text{Avg}(\text{CV})_{\text{max}} - \text{Avg}(\text{CV})_{\text{AL}}}{\text{Avg}(\text{CV})_{\text{max}} - \text{Avg}(\text{CV})_{\text{min}}} \tag{11}
\]

\[
\text{SR}_{\text{AL}} = \frac{\text{Avg}(\text{Acc})_{\text{AL}} + \text{NCV}_{\text{AL}}}{2} \tag{12}
\]

Let Avg(CV) is an average of compact value, AL is an Algorithm.

Table 5. The Accuracy (Acc) of classification of CCFAR algorithm and other algorithms.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>CLASS</th>
<th>AC</th>
<th>FCAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
<td>RIP PER</td>
<td>OneR</td>
</tr>
<tr>
<td>Iris</td>
<td>0.933</td>
<td>0.933</td>
<td>0.940</td>
</tr>
<tr>
<td>Heart</td>
<td>0.770</td>
<td>0.822</td>
<td>0.729</td>
</tr>
<tr>
<td>Pima</td>
<td>0.734</td>
<td>0.747</td>
<td>0.724</td>
</tr>
<tr>
<td>Avg(Acc)</td>
<td>0.812</td>
<td>0.834</td>
<td>0.797</td>
</tr>
<tr>
<td>Rank.</td>
<td>8</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 6. The number of classification rules (Compact Value) of CCFAR algorithm and other algorithms.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>CLASS</th>
<th>AC</th>
<th>FCAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
<td>RIP PER</td>
<td>OneR</td>
</tr>
<tr>
<td>Iris</td>
<td>4.9</td>
<td>3.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Heart</td>
<td>17.4</td>
<td>4.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Pima</td>
<td>20.3</td>
<td>4.1</td>
<td>7.8</td>
</tr>
<tr>
<td>Avg(CV)</td>
<td>14.2</td>
<td>3.9</td>
<td>4.2</td>
</tr>
<tr>
<td>NCV</td>
<td>0.86</td>
<td>0.983</td>
<td>0.980</td>
</tr>
<tr>
<td>Rank.</td>
<td>8</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
The results in Tables 5-7 were summarized and discussed as follows:

- Table 5 shows the accuracy of classification of CCFAR algorithm compared to the 9 algorithms. The symbol “-” means no available published result. GARC algorithm shows the highest accuracy (0.867) in all dataset and it has been ranked number 1 in the comparison among classifier data from 10 algorithms. Our proposed CCFAR algorithm was ranked number 4 (0.828) in classification accuracy.

- Table 6 represents the number of classification rules of CCFAR algorithm and the other 9 algorithms. The measure NCV is a normalization of the compact values computed as in Eq.8. The CFARC algorithm gives the smallest number of rules (2.6) comparing to 10 algorithms. Our CCFAR algorithm was ranked number 2 with average number of rules equals (3.3).

- From table 7, we showed the combination of accuracy and normalization of compact value which is called the SR values of CCFAR algorithm and other nine algorithms. We can see that the best SR is our presented method CCFAR (0.9104). It means our algorithm is a good classifier in terms of both accuracy and good compact of fuzzy associative classification rules considered all together.

6. Conclusions

This paper proposes the development of data classification algorithm with compact fuzzy association rules called (CCFAR) to optimize both an accuracy and interpretability of the classification model. To evaluate the performance of the proposed method, our algorithm was tested on three different data sets and compared the results with the other nine different algorithms. The results showed that our proposed CCFAR algorithm was ranked number 3 based on accuracy of classification, and it was ranked number 2 based on the smallest number of rules. CCFAR is the best algorithm when we combine accuracy and compact value together.

Acknowledgment

The first author has been funded by scholarship from the Suranaree University of Technology.

References

(3) Holte Robert C: “Very simple classification rules perform well on most commonly used datasets”, Machine learning, Vol. 11, No. 1, pp. 63-90, 1993

(11) Li Jiuyong Hong Shen, and Rodney Topor: “Mining the optimal class association rule set”, Knowledge-Based Systems. 15(7), pp. 399-405, 2002


