***I.J. Education and Management Engineering,* 2013, 2, 20-19**
Published Online February 2013 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2013.02.04

Available online at http://www.mecs-press.net/ijeme

A Weighted Relational Classification Algorithm Based on Rough Set

Fu Jinghong[a1], Zhang Chunying[a], Wang Jing[a], Tian Fang[b]

[a] *College of Science Hebei United University Tangshan, Hebei, China*
[b] *Hebei University of Engineering Han Dan, Hebei, China*

Abstract

A Weighted Relational Classification Algorithm Based on Rough Set is proposed in this paper. The relations of tables are classified in database, relational graph is converted into 0 - 1 matrix, the weight is calculated using UCINET; at the same time, different condition attributes are weighted differently by using attribute frequency of Rough Set. It is improved effectively. Experiments have proved that new classifier has good classification performance**.**

Index Terms: Multi-relational classification, 0-1 matrix, attributes frequency.

1. INTRODUCTION

Multi-relational classification is important in Multi Relational Data Mining. Its purpose is to derive a prediction model from training set. The existing classifiers are based on ILP (Inductive Logic Programming)[1] or relational database[2]. The representative of Multi-relational Classification Algorithm is CrossMine[3] algorithm and the Graph-NB[4] algorithm. Graph-NB algorithm is based on Semantic Relational Graph (SRG) that builds a semantic graph before classification, uses Pruning strategy of “cutting off” to prune the table in order to improve the Classification accuracy rate. But this method is not appropriate that directly remove the weak link table. This will allow incomplete information, thereby affecting the classification performance. Meanwhile, in the Bayesian classifier attributes are involved in the classification. In multi relation, in addition to the relationship between tables, attributes of each table are important. However, the influence of different attributes is inconsistent for classification.

In paper[5], a RS-RBC (Multi-Relational Bayesian Classification Algorithm with Rough Set) is proposed. The concept of relational graph used to dynamic choice associative table associated with the target table, and a tuple ID propagation approach is used to solve directly the association rule mining problem with multiple database relations, and the concept of Core in Rough Set is introduced, simplify the associative table.

* Corresponding author.
E-mail address: [1]Fujinghong111@126.com

Compared with the traditional algorithm, it improves the accuracy rate. This algorithm support relation Database directly. Its running rate is much higher than ILP. It makes the algorithm easier for reduce the associative table and classification attribute set, but it does not consider the different effects. In fact, the influence of different attributes is inconsistent for classification. It is not realistic without considering its impact.

So considering the influence of the associative table and attributes, a Weighted Relational Bayesian Classification Algorithm with Rough Set (RS-WRBC) is proposed. The relations of tables are classified in database, relational graph is converted into 0 - 1 matrix, the weight is calculated using UCINET; at the same time, different condition attributes are weighted differently by using attribute frequency of Rough Set. It is improved effectively. Experiments have proved that new classifier has good classification performance.

2. Related Concept

Relational database is composed by many relationship tables and association relationship of each table.

- **Definition 1: Relational graph:** In a given database D, create Relational graph through link between the table's primary and foreign key. It is a directed acyclic graph. Each table is as a node, each side is the connection between the relationship tables. Arrow points to the table where the primary key has.
- **Definition 2: Associative table:** In database D, for a classification task T, it is called association table that related by classification task and associated with the target table.
- **Definition 3: Isolation table:** In database D, for a classification task T, It is called isolation table that has nothing with target table.

As we can be seen, isolation table and target table are just in the same database. Moreover, isolation tables and relationships table, target table is relative. A table will become an isolated form, depending on the classification task.

Figure 1 is a Relational graph of seven tables.

$R_1, R_2, R_3, R_4, R_5, R_6, R_7$ compose a Relational graph. R_1 is target table. R_2, R_3, R_4, R_5, R_6 is Associative table, R_7 is Isolation table.

To show the role of associative table, using class label propagation method, a given tuple in the target table contains class label, while the topples in other tables are no class label.

- **Definition 4: Class label propagation:** it is supposed that there are two relations R_1 and R_2 where R_1 the target relation is, and they can be connected by the attributes $R_1.A$ and $R_2.A$ where A is the primary key or foreign key. Then the class labels of the tuple in R_1 can be transmitted to the tuples in R_2 by attribute A.

In fact, the method of the class label propagation is a virtual connection in contingency table. The result from the physical connection to the table is a table that contains large amounts of data, and contains a large number of redundant data that the operation is extremely troublesome. While if it is used the method of the class label propagation, there is no such problems. And as so, we can also get the class labels of the tuples in contingency table, and can easily calculate the nuclear properties in the table.

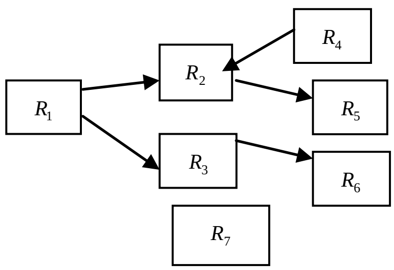


Figure 1 A Relational graph

- **Definition 5: Classification Attribute Set:** From multi-relational table, we can select all or part of properties to determine which class attribute set the given data belongs to, known as class attributes set.

The size and the properties of classification attribute set has a direct impact on the efficiency and accuracy of classification, so it is agential about how to determine the classification attribute set.

3. Weight Calculation of Associative table and Attribute

A. Weight Calculation of Associative table

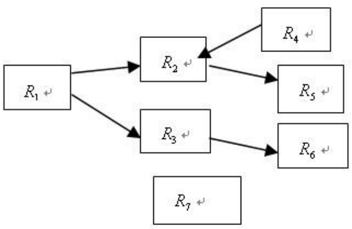
In multi relations, the relativities of tables are different. This is called as weight. In the Relational graph, each table is seen as a node in a graph, each line represents the connection of tables, arrow points to the primary key of the table. Using the Relational graph, it's converted into 0-1 relation matrix, matrix element is expressed whether there is a direct relationship between the table i and j. When the cable is directly connected, $a_{ij} = 1$; else, $a_{ij} = 0$.

Each table involved in the classification as a node, direct dependencies of tables as side, they form Network structure. UCINET is used to calculate the weight of each node.

Firstly, according to the nodes in the relational graph have relations or not, we can change the graph into a 0-1 matrix. If node A has a line pointing to node B, the element in the row A and column B is 1, otherwise it is 0.Figure 2 is the result of Figure 1 into 0-1 matrix.

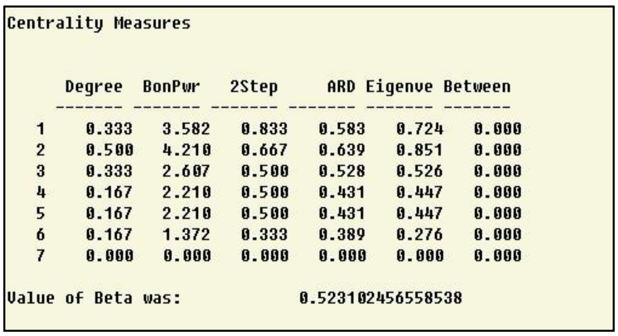
The weights of all nodes in the figure 1 are the degrees in the figure3.

Figure 1 a Relational graph ⟶ change it into 0-1 matrix.



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2 the result of Figure 1 into 0-1 matrix


Centrality Measures

	Degree	BonPwr	2Step	ARD	Eigenve	Between
1	0.333	3.582	0.833	0.583	0.724	0.000
2	0.500	4.210	0.667	0.639	0.851	0.000
3	0.333	2.607	0.500	0.528	0.526	0.000
4	0.167	2.210	0.500	0.431	0.447	0.000
5	0.167	2.210	0.500	0.431	0.447	0.000
6	0.167	1.372	0.333	0.389	0.276	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000

Value of Beta was: 0.523102456558538

Figure 3 The result of UCINET

B. Weight Calculation of Attribute

Give a decision table $S=(U,C\bigcup D,V,f)$, $U=\{x_1,x_2,\cdots x_n\}$, $D=\{d\}$.
Its identified matrix is denoted by M. $M=(m_{ij})_{n\times n}$,
And its element is as

$$m_{ij}=\{a\in C|(x_i,x_j\in U)\ \wedge\ f_a(x_i)\neq f_a(x_j)\ \wedge\ d(x_i)\neq d(x_j)\}\tag{1}$$

By the definition of discernibility matrix， the higher the frequency, the greater its importance. Therefore, when the matrix is created, at the same time recording the frequency of each attribute.

- **Definition 6: Attribute Frequency (AF):**

$$\forall a\in C,\lambda(a)=|\{m|a\in m,\forall m\in M\}|\tag{2}$$

The frequency of attribute is the number that discernibility matrix occurrences.
In the weighted Naive Bayesian classification algorithm, according to their importance and the importance of mathematical expectation, we should compared and re-allocate all the attributes.

- **Definition 7: Weight of AF:**

$$w_{a_i}=\frac{\lambda(a_i)}{\frac{1}{n}\sum_{i=1}^{n}\lambda(a_i)}\qquad(i=1,2\cdots n)\tag{3}$$

4. Weight Calculation of Associative table and Attribute

A. Weighted Naive Bayesian Model

In multi-relationship, Multi-relational Naive Bayesian Classification formula is

$$\begin{aligned}C_{MAP}&=\operatorname{argmax}_{C_j\in C}P(C_j)P(x_1,\cdots,x_n,y_{k_11},\cdots,y_{k_1r},\cdots,y_{k_p1},\cdots,y_{k_pr}|C_j)\\&=\operatorname{argmax}_{C_j\in C}P(C_j)\prod_{i=1}^{n}P(x_i|C_j)\prod_{q=k_1}^{k_p}\prod_{t=1}^{r}P(y_{qt}|C_j)\end{aligned}\tag{4}$$

Different tables and attributes are given different weights to make Naive Bayesian extends.
Then the new model is

$$\begin{aligned}C_{MAP}&=\operatorname{argmax}_{C_j\in C}P(C_j)P(x_1,\cdots,x_n,y_{k_11},\cdots,y_{k_1r},\cdots,y_{k_p1},\cdots,y_{k_pr}|C_j)\\&=\operatorname{argmax}_{C_j\in C}P(C_j)\prod_{i=1}^{n}P(x_i|C_j)\prod_{q=k_1}^{k_p}\prod_{t=1}^{r}P(y_{qt}|C_j)^{w_{AF}}\ w_{Table}\end{aligned}\tag{5}$$

w_{AF} is the weight of attributes, w_{Table} is the weight of associative table.

B. Algorithm Process

Step 1: Analyze the primary and foreign keys of the contingency tables in database, resulting diagram.
Step 2: Calculate the nuclear properties of the target table , recorded as ;

Step 3: Do the attribute reduction to the contingency table after the tuple class label transmitted, then find its nuclear property ;

Step 4: Scan diagram, using the relationship of Associative table to build 0-1 matrix, with UCINET software to calculate the weight of the associated table;

Step 5: Using rough sets and class label propagation, calculate core attributes set as Classification Attribute Set and reduce the training set;

Step 6: Scan multi-relational training samples, calculate identification matrix, while recording the frequency of each attribute, so as to derive attribute weights;

Step 7: Use Multi-relational Weighed Bayesian Model to classification.

5. Experiment

In order to verify the effectiveness of the algorithm, the algorithm is tested. The data sets are from UCI [8], continuous attribute values of all data sets are discredited. The results are shown in table 1.

From it, the ability of RS-WRBC improves obviously.

Table 1 Experiment data sets and classification results

Data Set	Attributes	Classification	Training Set	NBC％	AFWNB ％
Australian	14	2	690	68.63	85.42
Car	7	4	1880	85.78	86.29
Cleve	10	2	296	82.43	81.53
Crx	15	2	653	86.58	87.58
German	15	2	1000	75.5	75.01
Hepatitis	19	2	80	92	91.02
Iris	4	3	150	93.73	94.21
Letter	16	26	20000	74.64	73.98
average				82.41125	84.38

6. Conclusion

A Weighted Relational Classification Algorithm Based on Rough Set is proposed in this paper. The relations of tables are classified in database, relational graph is converted into 0 - 1 matrix, the weight is calculated using UCINET; at the same time, different condition attributes are weighted differently by using attribute frequency of Rough Set. It is improved effectively. Experiments have proved that new classifier has good classification performance.

It is the direction of future research that how to improve Multi-relational Bayesian Classifier performance without increase the time complexity.

References

[1] Muggleton s.Inductive Logic Programming [M]. New York, NY: Academic Press,1992

[2] Blockeel H,De Raedt L, Ramon J. Top-down induction of logical decision tree [J].Artificial Intelligence 1998 (101)1-2:285-297.

[3] Xiaoxin Yin, Jiawei Han, Jiong Yang, Efficient Multi-relational Classification by Tuple ID Propagation. [C]//OzsoyogluM, Zdonik S. Proc 2004 Int Conf on Data Engineering (ICDE’04),Boston, MA: 2004:399-410

[4] Hongyan Liu, Hailiang Chen, Graph-NB: an Efficient and Accurate Multi-relational Naïve Bayesian Classifier. China Journal of Information Systems. 2008:1-11
[5] Zhang Chunying, Wangjing. Multi-relational Bayesian Classification Algorithm with Rough Set[J]
[6] Borgatti S P, Everett M G, Freeman L C. UCINET for Windows: Software for Social Networks Analysis [M]. Har-vard: Analytic Technologies, 2002.
[7] Chunying Zhang, Wangjing. A new Weighted Naive Bayesian Classification Algorithm [J] Micro Computer Information .2010
[8] LiuJun. Lectures on Whole Network Apptoach.2008.34-75
[9] Deng Wei-Bin, Wang Guoyin, Wang Yan. Weighted Naive Bayesian Classification Algorithm Based on Rough Set [J] Computer Science. 2007.34(2):204-209
[10] Kang M G. Katsaggelos A K. General choice of the regularization functional in regularized restoration [J]. IEEE Tran on Imge Processing, 1995, 4(5): 594-602.

How to cite this paper: Fu Jinghong, Zhang Chunying, Wang Jing, Tian Fang,"A Weighted Relational Classification Algorithm Based on Rough Set", IJEME, vol.3, no.2, pp.20-19, 2013.