Efficient Rules Extraction using Rough Set Theory for Weather Data Analysis

Nyein Nyein Ei
University of Computer Studies, Mandalay
nyeinnyeinei@gmail.com

Abstract

The system extracts optimal rule from weather data set based on rough set theory. The main idea of rough set theory is to obtain as simple as rules from the given database by reducing the database while holding the original degree of consistency. In order to find the optimal rule of weather from the historical data it provide easy and accurate for the weather forecast. This system included the processes of indiscernibility, set approximation, attributes reduction, rules extraction and optimal rule selection. GDT-RS (Generalization Distribution Table for Rough Sets) are used for the rules extraction and optimal rule selection. This system analyzes the relationship between the weather condition attribute and other attributes of weather data set by calculating the dependency and accuracy between them.

1. Introduction

Rough Sets theory is a mathematical tool for data analysis [1]. Rough sets have many applications for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc.

When a dataset contains irrelevant features these can be eliminated, reducing in this way the dimension of the problem, Rough sets can be used to find subsets of relevant features. Other application of rough sets is instance selection. Rough set theory provides tools to perform the knowledge discovery task handling the instances with missing values like the instances that behave similar to them. One of main function of rough set theory is to delete unnecessary condition attributes without reducing the degree of consistency.

Weather Forecasting System with advanced technology is essential for every nation nowadays. This system aims to extract optimal rule for weather data set to easy to forecast the weather and to reduce time complexity.

This system uses the Myanmar weather forecasting data set in Mandalay city in December, 2010. In this data set, there are 9 condition attributes and one decision attribute. The goal of this system is to analyze the relationship between the condition attributes and decision attribute of weather data set.

The rest of this analysis paper is organized as follow. Section 2 describes related works. Section 3 describes Rules Extraction based on Rough Set theory, and then advantages and disadvantages of this theory. Section 4 describes Rules Extraction from weather data set. The last section 5 concludes and future works this paper.

2. Related Works

There are several approaches have been attempted for rough set theory to forecast weather and to classify weather data set. J.F.

Peters et. al reported on a rough set approach to classifying meteorological volumetric radar data used to detect storm events responsible for summer severe weather. They used a rough set approach to classify different types of meteorological storm events. They proposed the criterion for comparison the accuracy coefficient in the classification over the Radar Decision Support System database of Environment Canada [2]. Empirical Statistical Modeling of Rainfall Prediction over Myanmar described the modeling of monthly rainfall prediction over Myanmar in detail by applying the polynomial regression equation. They compared the proposed model results to the results produced by multiple linear regression model (MLR). Their experiments indicated that the prediction model based on MPR has higher accuracy than using MLR [3].

H. Nakayama suggested a few modified methods in the rough set theory, and to apply them to a medical data analysis. Their experimental results are considered to be reasonable from a practical viewpoint by medical doctors [4].

Rule selection method was developed for filtering large number of extracted rules from Coronary Artery Disease (CAD) data set. Experiment on CAD data set showed that the proposed method is able to select small number of rules while maintaining the quality of rule based classifier. The proposed method had better quality compared to previous rule selection methods [5].

3. Rule Extraction Based on Rough Set Theory

A rough set is a formal approximation of a crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set [6].

Let I=(U,A) be an information system, where U is a non-empty set of finite objects and A is a non-empty, finite set of attributes such that $a: U \to V$ for every a A. V_a is the set of values that attribute a may take. The information table assigns a value $a(x)$ from V_a to each attribute a and object x in the universe U.

3.1. Indiscernibility Relation

The equivalence relation is a binary relation which is reflexive (xRx for any object x), symmetric (if xRy then yRx), and transitive (if xRy and yRz then xRz).

The equivalence class $[x]_R$of an element $x \in X$ consists of all objects $y \in X$ such that xRy. Let IS = (U, A) be an information system, then with any $B \subseteq A$ there is an associated equivalence relation:

$IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}$ (1)

where $IND_{IS}(B)$ is called the B-indiscernibility relation. If $(x, x') \in IND_{IS}(B)$ then objects x and x' are indiscernible from each other by attributes from B. The equivalence classes of the B-indiscernibility relation are denoted by $[x]_R$.

3.2. Set Approximation

Let T = (U, A) and let $B \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained in B by constructing the B-lower and B-upper approximations of X, denoted $\underline{B}X$ and $\overline{B}X$ respectively, where

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad (2)$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

B-boundary region of X, $BN_B(X) = \overline{B}X - \underline{B}X$ consists of those objects that we cannot decisively classify into X in B. B-outside region of X $U - \overline{B}X$, consists of those objects that can be with certainty classified as not belonging to X. A set is said to be rough if its boundary region is non-empty, otherwise the set is crisp.

3.3. Accuracy of Approximation

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \quad (4)$$

where |X| denotes the cardinality of $X \neq \emptyset$. Obviously$0 \leq \alpha_B \leq 1$. If $\alpha_B(X) = 1$, X is crisp with respect to B. If $\alpha_B(X) < 1$, X is rough with respect to B.

3.4. Dependency of Attributes

Discovering dependencies between attributes is an important issue in KDD. Set of attribute D depends totally on a set of attributes C, denoted $C \Rightarrow D$ if all values of attributes from D are uniquely determined by values of attributes from C.

$$k = \gamma(B, D) = \sum_{X \in U/D} \frac{|\underline{B}(X)|}{|U|} \quad (5)$$

If k=1 we say that D depends totally on B.
If k<1 we say that D depends partially (in a degree k) on B.

3.5. Main Features of GDT-RS

Unseen instances are considered in the discovery process, and the uncertainty of a rule, including its ability to predict possible instances, can be explicitly represented in the strength of the rule.

Biases can be flexibly selected for search control, and background knowledge can be used as a bias to control the creation of a GDT and the discovery process. F(x): the possible instances (PI) and possible generalization.
G(x): the possible generalizations
(PG): the probability relationships between PI & PG.

3.6. Rule Strength

$$S(X \rightarrow Y) = s(X)(1 - r(X \rightarrow Y)) \quad (6)$$

The strength of the generalization *X* (BK is no used),

$$s(X) = s(PG_k) = \sum_l p(PI_l|PG_k) = \frac{N_{ins-rel}(PG_k)}{N_{PG_k}} \quad (7)$$

$N_{ins-rel}(PG_k)$is the number of the observed instances satisfying the *i*th generalization.
The rate of noises

$$r(X \rightarrow Y) = \frac{N_{ins-rel}(X) - N_{ins-class(X,Y)}}{N_{ins-rel}(X)} \quad (8)$$

$N_{ins-rel}(X, Y)$ is the number of instences belonging to the class Y within the instances satisfying the generalization X.

4. Rules Extraction from Weather Data Set

Weather prediction system computerized is important for every country. In this paper, extracting rules are developed by using the weather data set for Mandalay city in Myanmar in December in 2010 [7].

This data set includes 9 conditional attributes that are Temperature, Dew Point, Humidity, Sea Level Precipitation, Visibility, Wind Direction, Wind Speed, Events and Wind Direction Degree and one decisional attribute that is weather Condition.

This system cleans the data set as preprocessing step. In preprocessing step, the system uses stepwise backward elimination in the dimensionality reduction method [8]. This method reduces the data set size by removing worse attributes or dimensions remaining in the set. The output of the preprocessing step is the cleaned decision table.

And then the system computes indiscernibility from weather decision table can be seen as examples in section 4.1. By using indiscernibility relation, the system calculates set approximation can be seen in section 4.2. And the system computes accuracy of approximation and dependency between conditional attributes from weather data set can be seen in section 4.3.

The system use Quick Reduct Algorithm [Figure2] for a minimal subset without exhaustively generating all possible subsets.

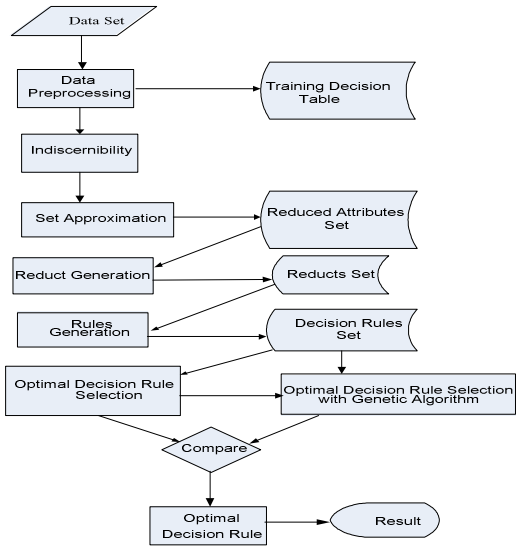

Figure1. General overview of the proposed system

For rule generation, the system uses rule discovery system that is Generalization Distribution Table-Rough Set (GDT-RS). The system measures rules strength that got from GDT-RS in the step of the optimal decision rule selection. The system uses the Optimal Set of Rules Algorithm [Figure4]. The system can decide the better rule for largest number of strength.

QuickReduct(C, D)
C: the set of all conditional features;
D: the set of decision features.
(1) R ← {}
(2) do
(3) T ← R
(4) ∀x ∈ (C − R)
(5) if $\gamma_{R\cup\{x\}}$(D) > γ_T (D)
(6) T ← R ∪ {x}
(7) R ← T
(8) until γ_R(D) == γ_C(D)
(9) return R

Figure 2. The Quick Reduct algorithm

4.1. Indiscernibility Relation of Weather Data Set

Let I=(U,A) be a weather data set, where U is a non-empty set of finite objects which include 117 instances and A is finite set of attributes such that Temperature, Dew Point, Humidity, and Condition and so on.

The Insdiscernibility class are identified the following:

IND(T)={Temperature,Condition}={{2,5,27},{4,28},{5,10,14,25,29},{20,32},{6,9,13,46,50},{66,73},{7},{11,15,46,50,76},{33},{71,86,116},{13,31,44,76},{17,34,41},{36},{37,40},{16,26,30,48,52,72,82,108},{85,89,92,96,111,113},{27},{19},{22},{35},{38},{42},{27},23,45,49},{78},{58,79,115},{39},{43},{47,51},{64,80,84,103,115,117},{87,103,106,112},{83,86,93,90,97,114},{100,104}}

IND(WD)={Wind Direction,Condition} = {{2,4,5,6,9,10,11,12,13,15,16,19,21,23,25,26,27,29,30,31,44,45,46,48,49,50,52,72,82,108},{3,47,76},{14,28},{7,22},{17,34,41,42},{18,32},{20},{24},{33,43},{35,37,40},{36,38,39},{58,63,66,67,68,69,73,74,75,78,79,83,85,86,97,89,90,92,93,96,97100,101,102,104,105,106,109,110,111,112,113,114,115,116,117}}

IND (Events) = {Events, Condition} = {{35,37,40} , {36,38,39}}

IND(WDD) = {Wind Driection Degree,Condition} = {{3,47,51,76},{14,28},{20,24},{64},{74,80,81,103,107},{84,88},{95},{99}}

where T, WD, WS and WDD are Temperature, Wind Direction, Wind Speed and Wind Direction Degree respectively and 1,2,3,… are index numbers.

4.2. Set Approximation of Weather Data Set

Let C1 = {Condition=Clear} = {7,22}

C2 = {Condition=Scattered Clouds} = {2,3,4,5,6,9,10,11,12,13,14,15,16,19,21,23,25,26,27,28,29,30,31,44,45,46,47,48,49,50,51,52,72,76,82,99,108}

C3 = {Condition=Partly Cloudy} = {17,34,41,42}
C4 = {Condition=Haze} = {18,20,24,32}
C5 = {Condition=Overcast} = {33,43}
C6 = {Condition=Light Drizzle} = {35,37,40}
C7 = {Condition=Light Rain} = {36,38,39}
C8 = {Condition=Mist} = {58,63,64,66,67,69, 70,71,73,74,75,78,79,80,81,83,84,85,86,87,88,89, 90,92,93,95,96,97,100,101,102,103,104,105,106, 107,109,110,111,112,113,114,115,116,117}

4.3. Lower approximation of weather data set

The lower approximation of decitional attribute, Condition, is identified the following:

$$\underline{C_1}(T) = \{7,22\}$$
$$\underline{C_2}(T) = \{6,9,13,46,50\}$$
$$\underline{C_3}(T) = \{42,17,34,41\}$$
$$\underline{C_4}(T) = \{18,20,24,32\}$$
$$\underline{C_5}(T) = \{33,43\}$$
$$\underline{C_6}(T) = \{35,37,40\}$$
$$\underline{C_7}(T) = \{36,38,39\}$$
$$\underline{C_8}(T) = \{110,87,102,106,112,58,59,115,$$
$$67,70,74,75,105,109,63,84,86,90,93,$$
$$97,101,114,69,78,66,73,100,104,$$
$$85,89,92,96,111,113,64,80,81,$$
$$84,103,117,95,107,71,88,116\}$$

where T is Temperature and 1,2,3,… are index numbers of weather data set.

4.4. Upper Approximation of Weather Data Set

The upper approximation of decitional attribute, Condition, is identified the following:

$$\overline{C_1}(T) = \{7,22\}$$
$$\overline{C_2}(T) = \{6,9,13,46,50\}$$
$$\overline{C_3}(T) = \{42,17,34,41\}$$
$$\overline{C_4}(T) = \{18,20,24,32\}$$
$$\overline{C_5}(T) = \{33,43\}$$
$$\overline{C_6}(T) = \{35,37,40\}$$
$$\overline{C_7}(T) = \{36,38,39\}$$
$$\overline{C_8}(T) = \{110,87,102,106,112,58,59,115,$$
$$67,70,74,75,105,109,63,84,86,90,93,97,$$
$$101,114,69,78,66,73,100,104,85,89,92,96,$$
$$111,113,64,80,81,84,103,117,95,107,71,$$
$$88,116\}$$

where T is Temperature and 1,2,3,… are index numbers of weather data set.

4.5. Attribute Accuracy and Dependency of Weather Data Set

In rough set theory, the accuracy and dependency values can be compute by using the equation (4) and (5). The accuracies of attributes from weather data set are 9%, 7%, 24%, 2%, 2%, 3%, 3%, 0% and 1% respectively. Dependencies of attributes for conditions are described in Figure 4. It can be see that dependency value for Humidity is highest. The second highest is Temperature dependency value. Therefore, Humidity attribute is the most dependence on the Condition attribute. Other attributes such as Dew Point, Sea Level Precipitation, Visibility, Wind Direction and Wind Speed are little dependence upon decision attribute. And, Events and Wind Direction Degree are not depending on it.

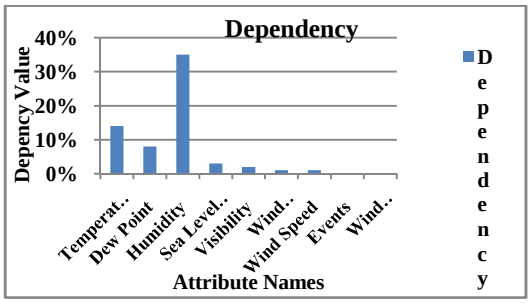

Figure 3. Condition of the dependency values between attributes

4.6. Optimal Set of Rules from Weather Data Set

This system selects the optimal rules from rule sets of weather by using Optimal Set of Rules Algorithm [1].

Optimal Set of Rules Algorithm
u: Instances
U: Compound Instances
r: rate of noise
PI: Possible instance
PG: Possible generalization
$N_{ins\text{-}rel}(X)$: number of instances belonging to the class X
$N_{ins\text{-}class}(X,Y)$: number of instances belonging to the class Y within the instances satisfying the generalization X
S: strength
DM: discernibility matrix
1. rules ←{}
2. u ←same instances ∈ U
3. Calculate $r(X \rightarrow Y)=\frac{N_{ins-rel}(X)-N_{ins-class}(X,Y)}{N_{ins-rel}(X)}$
4. Begin
5. Select u ∈ U then DM for u
6. Calculate Reduct ∈ DM
7. Begin
8. Acquire rules from Reduct ∈ u
9. Calculate $S(X)=s(PG_k)=\sum_l p(PI_l|PG_k) = \frac{N_{ins-rel}(PG_k)}{N_{PG_k}}$
10. Select better rule ←rules acquire in step 8
11. End
12. u ←U-{u}
13. If ∩ ≠ ∅ Then go to step go to step 4
14. Else go to step 15
15. Begin
16. If Rule==1 Then
17. Else Find minimal set of rule ∈ u
18. End
19. End.

Figure 4. Optimal Set of Rules Algorithm

5. Conclusion and Future Works

The system calculated the dependency and accuracy between condition attributes and decision attribute of the rough set representation. Experiment on figure 4 showed dependency values that the system is able to decide for weather forecast upon Humidity attribute at most. In this paper, the system extracted the efficient rules of weather of Mandalay in December in 2010. In future work, the system can be used in every city in Myanmar and compare with other methods. To select optimal rule; this system will calculate rule strength and will decide the better rule for largest number of strength. This system will compute the optimal rule that got from decision rule set with genetic algorithm. This system will compare the optimal rules from rough set theory and genetic algorithm. Our proposed system will give the efficient rules.

References

[1] Z. Pawlak, A. Skowron, "Rough Sets and Boolean reasoning", Information Sciences, 177, 2007, pp. 41-73

[2] J.F. Peters, Z. Suraj, S. Shn, S. Ramana, W. Pedrycz, N. Pizzi, "Classification of Meteorological Volumetric Radar Data using Rough Set Methods", Pattern Recognition Letters 24, 2003, pp. 911-920

[3] W. Thida Zaw and T. Thu Naing, "Empirical Statistical Modeling of Rainfall Prediction over Myanmar", World Academy of Science, Engineering and Technology 46, 2008

[4] H. Nakayama, Y. Hattori and R. Ishii, "Rule Extraction based on Rough Set Theory and its Application to Medical Data Analysis, IEEE, 1999

[5] N.A. Setiawan, P.A. Venkatachalam, and Ahmad Fadzil M.H, "Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set", International Journal of Recent Trends in Engineering, Volume 2, No. 5, November 2009

[6] A. Skowron, N. Zhong, "Rough Sets in KDD - Tutorial Notes", PAKDD 2000, Kyoto, Japan, Pacific-Asia Conference on Knowledge Discovery and Data Mining

[7] www.wunderground.com

[8]J. Han, M. Kamber, "Data Mining: Concepts and Techniques", ISBN 1-55860-489-8