KKU Engineering Journal

https://www.tci-thaijo.org/index.php/kkuenj/index

Classifying rubber breed based on rough set feature selection

Phanarut Srichetta*

Department of Computer Science and Information Technology, Faculty of Science,
Udon Thani Rajabhat University, Udon Thani, 41000, Thailand.

Received April 2016
Accepted June 2016

Abstract

Rubber is the economic crop that is planted widely in almost all regions of Thailand and makes a lot of income for the export of this country. Selecting a rubber breed for a particular region is one of the principal factors for the achievement of the rubber plantation. If the agriculturists get the rubber breeds unsuitable to be plant in their rubber garden, once the time to slit, the rubber water may have low quality and quantity. The objective of this work is to generate the rubber breed classifier by using the k-nearest neighbor technique based on selected set of features of rubbers. Rough set feature selection is proposed in this research to select a subset of relevant features of rubber optimally while retaining semantics. The data samples of 10 well-known breeds of rubber, 30 samples per breed, cultivated in the northeast of Thailand were used to generate the breed classifier. The accuracy rate of classifying the breed of rubber is rather good. Therefore, this generated breed classifier can assist the agriculturists classify and select the correct breed of rubber from the features of rubber in hand before cultivate in the rubber garden.

Keywords: Classification, Feature selection, K-nearest neighbor, Rubber, Rough sets

1. Introduction

Rubbers make a lot of income for the export of Thailand. They are plant widely and currently the cultivated areas of papa rubber are increasing in almost all regions of the country. In different regions, the suitable breeds of rubber are regarded as the principal factors for the accomplishment of the rubber plantation. Some breeds have features that identify themselves differentiate from others whereas some breeds are not. Sometimes, the features of some breeds have adjusted themselves to be suitable for different topographies and climates. Therefore, identifying the breed of rubber to be plant explicitly by considering its features is difficult for the less skill experts and agriculturists. This situation may affect the agriculturists in which they often get the unsuitable rubber breeds to be plant. Once the time to slitting comes, the rubber-water obtained from the unsuitable rubber breeds may have low quality and quantity. It is not worthwhile for investment.

In general, the breeds of rubbers can be classified by two ways: DNA and human eyes [1]. With DNA, the accuracy result of classification is high but the practical process is difficult, long-time, and costly. Therefore, human eye is another way but it needs the experts who know the exact feature values of each breed. Some features have indiscernible values compared with others; this may cause the difficulty for the experts to classify the breed of rubber. The Rubber Research Institute of Eastern Thailand [2] has studied and divided the features of rubber into various main features, e.g. leaf storey, leaf, petiole, etc. Each feature also has various sub-features. It can be seen obviously that there is high dimension of rubber features related to the breed classification of the rubbers. In order to lead to more compactness of the model learned, decrease the classification time and improve classification accuracy, searching for a minimal representation of rubber features by discarding redundant or least information carrying features is needed.

Several approaches have been proposed for discovering the suitable subset of features by carried out on feature selection [3-6]. Rough set theory, proposed by Zdzislay Pawlak [7], has been widely applied in machine learning, data mining and knowledge discovery. One of the applications of rough set theory is the feature selection especially for classification problems by discarding some of the redundant or irrelevant features while retaining semantics. It has become a topic of a great interest with much success in a number of real world domains, including medicine, pharmacology, control systems, social science, switching circuits, image processing, text documents, and movie reviews [8-11]. The reasons for its success are: only the facts hidden in data are analyzed; no additional information about the data (threshold or expert knowledge) is required; and a minimal knowledge representation is found [12-13]. Therefore, this research employs the rough set theory to find the optimal subset of rubber features useful for producing the desired learning results in the breed classification phase. In this paper, the breed classifier is generated with the k-Nearest Neighbor (k-NN) algorithm

*Corresponding author. Tel.: +6691 864 4715
Email address: phanarut@udru.ac.th

doi: 10.14456/kkuenj.2016.53

[14] and tested its performance. The generated classifier could be used to classify the breed of a new rubber seedling. It would assist the agriculturists identify and select the correct breed of rubber seedlings to cultivate before planting in the rubber garden.

2. Rough set-based feature selection

In rough sets theory, data is organized in a *decision table* where rows of the table correspond to objects and columns correspond to the conditional features and a decision feature. That is, the decision table is a pair $(U, A \cup \{d\})$ where U is a non-empty finite set of objects, A is a non-empty finite set of conditional features, and $d \notin A$ is the decision feature indicated the class to which each object belongs.

2.1 Indiscernibility relation

Within the decision table, it is possible that same objects may be represented several times with respect to the available features. For any non-empty finite subset of conditional features $B \subseteq A$, there is an associated equivalence relation $IND(B)=\{(x, y) \in U \text{x} U \mid \forall a \in B, a(x)=a(y)\}$. If $(x, y) \in IND(B)$, then the objects x and y are indiscernible from each other by features from B.

The indiscernibility relation induces a partition of the universe U into block of indiscernible objects. The partition of U determined by $IND(B)$ is $U/IND(B)= \otimes\{U/IND(\{a\}) \mid a \in B\}$ where $A \otimes B= \{X \cap Y \mid X \in A, Y \in B, X \cap Y \# \phi\}$. The equivalence class of an element $x \in X$ consists of all objects $y \in X$ such that symmetric. The equivalence classes of an element $x \in X$ of the B-indiscernibility relation are denoted $[x]_B$.

2.2 Set approximation

Let $X \subseteq U$, X can be approximated using only the information contained in $B \subseteq A$ by constructing the *B-lower* and *B-upper approximations* of the set X ($\underline{B}X = \{x \mid [x]_B \subseteq X\}$ and $\overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\}$), where the objects in $\underline{B}X$ can be classified as members of X on the basis of knowledge in B while the objects in $\overline{B}X$ can be only classified as possible members of X on the basis of knowledge in B. The objects that cannot be decisively classified into X on the knowledge in B are composed in the boundary set $BN_B(X) = \overline{B}X - \underline{B}X$. If this boundary region is non-empty, the set X is rough otherwise it is crisp.

2.3 Feature dependency

Discovering dependencies between features is an important issue in data analysis. Let B, $C \subset A$, it is said C depends on B in a degree k $(0 \leq k \leq 1)$, denoted $B \Rightarrow_k C$, if $k = \gamma_B(C) = |POS_B(C)|/|U|$ where $|N|$ stands for the cardinality of set N and $POS_B(C)$ is the *B-positive region of D* defined by

$$POS_B(C) = \bigcup_{X \in U/IND(C)} \underline{B}X$$

If k=1, C depends totally on B, if $0<k<1$, C depends partially (in a degree k) on B, and if k=0, then C does not depend on B. Once the dependencies for all possible subsets of B are calculated, a minimum subset of B will be chosen.

2.4 Reduct

There are usually several subsets of features and those which are minimal are called *reducts*. For an initial conditional feature set C and a given set of decision features D, $R \subset C$ is a reduct if $\gamma_R(D)= \gamma_C(D)$. Moreover, R is a minimal subset if $\gamma_{R-\{a\}}(D) \neq \gamma_R(D)$ for all $a \in R$. In order to find a minimal reduct without exhaustively generating all possible subsets, the following QuickReduct algorithm [15] was used in this research.

QuickReduct(C, D)

Input: C, the set of all conditional features; D, the set of decision features

Output: R, the feature subset

Step:

```
$R \leftarrow \{\ \}$
while $\gamma_R(D) \neq \gamma_C(D)$
    $T \leftarrow R$
    foreach $x \in (C-R)$
        if $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
            $T \leftarrow R \cup \{x\}$
    $R \leftarrow T$
return $R$
```

3. K-Nearest Neighbor algorithm

k-Nearest Neighbor (k-NN) [14] is an algorithm used for classifying a class for an unknown object in the feature space. An object is classified by a majority vote of its neighbors. First of all, the dataset which composes of the feature vectors and class label is partitioned randomly into two sets: training and test. A test point is classified by assigning the class label which is most frequent among the k training set nearest to that point in the test set. The steps for classifying the class for test data with k-NN algorithm are as follows:

1.) Specify the value of k

2.) Compute the distances between data in training set and test set

3.) Rank the computed distances in ascending order and then select the first k items of training set which have least distances

4.) Assign a class for a given test data based on the class of the selected k items according to the majority voting or weighted voting method.

At step 2, the distance computation based on the Variables of Mixed Type [16] is needed in this research in order to support different types of features in the rubber dataset. Suppose two data $x_i= (x_{i1}, x_{i2,}, \ldots, x_{if} \ldots, x_{in})$ and $x_j= (x_{j1}, x_{j2,} \ldots, x_{jf,} \ldots, x_{nf})$, their distance is

$$d(i, j) = \frac{\sum_{f=1}^{n} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{n} \delta_{ij}^{(f)}}$$

where f is the order of feature, n is the total number of features, $\delta_{ij}^{(f)}$ is the indicator between x_i and x_j if consider at the feature f and $d_{ij}^{(f)}$ is the distance between x_i and x_j if consider at the feature f.

4. Research steps

The steps of classifying the breed for a rubber seeding are presented as follows.

Table 1 Dataset of rubber to be studied

#No. of instances:	300
#No. of features:	23 (22 conditional features and 1 decision feature)
22 conditional features:	- *Leaf Storey* (shape, height, width, space) - *Leaf* (shape of the middle leaf, edge of the middle leaf, leaf color, leaf gloss, leaf base, leaf tip, leaf line color, leaf sheet, middle leaf after crosswise cut, middle leaf after lengthwise cut, size of the middle leaf, minor leaves compared with the middle leaf, edge position of the minor leaves) - *Petiole* (shape, length, base shape, property, direction of petiole compared with stem)
1 decision feature (rubber class):	1. RRIT 408 2. RRIT 251 3. RRIT 226 4. BPM 24 5. RRIM 600 6. RRIT 118 7. PB 235 8. RRIT 402 9. AVROS 2037 10. BPM 1
Class distribution	10% for each class

Table 2 Accuracy of classifying the rubber breeds according to the first ten k values of 10 training sets

k	**Accuracy (%)**										
	S_1	**S_2**	**S_3**	**S_4**	**S_5**	**S_6**	**S_7**	**S_8**	**S_9**	**S_{10}**	**Average**
1	86.67	83.33	76.67	86.67	86.67	86.67	86.67	83.33	80.00	83.33	84.00
2	90.00	80.00	80.00	83.33	76.67	86.67	90.00	80.00	76.67	80.00	82.33
3	93.33	83.33	80.00	83.33	86.67	93.33	93.33	93.33	80.00	80.00	86.67
4	96.67	86.67	83.33	86.67	90.00	93.33	96.67	93.33	86.67	86.67	90.00
5	100.00	86.67	83.33	86.67	83.33	96.67	100.00	96.67	86.67	86.67	90.67
6	100.00	90.00	80.00	90.00	90.00	93.33	96.67	86.67	80.00	83.33	89.00
7	96.67	90.00	86.67	90.00	83.33	90.00	93.33	86.67	83.33	86.67	88.67
8	100.00	83.33	86.67	86.67	80.00	93.33	96.67	80.00	83.33	86.67	87.67
9	96.67	86.67	80.00	86.67	76.67	93.33	93.33	83.33	83.33	86.67	86.67
10	93.33	93.33	80.00	93.33	83.33	93.33	93.33	83.33	80.00	76.67	87.00

4.1 Collect the rubber data

The dataset of rubber seedlings was collected from the Rubber Research Institute of Eastern Thailand according to the designed data collection form. Team of researchers selected ten breeds of rubber mostly plant in the north eastern of Thailand to be studied. Details about number of instances and features of rubber are presented in Table 1.

4.2 Find the minimal feature set with the rough set-based feature selection

In the context of rough set theory, the collected rubber dataset can be treated as a decision table of the form $T=(U, A\cup\{d\})$. Here, $U=\{x_1, x_2, \ldots, x_{300}\}$ is a set of rubber seedlings; $A=\{a_1, a_2, \ldots, a_{22}\}$ is a set of conditional features of rubbers; d is the breed feature of rubber. The indiscernibility relation, the set approximation, the positive region, the feature dependency, and the reduct are calculated from such decision table. Finally, the reduct consist of 6 features, i.e., (i) shape of leaf storey, (ii) height of leaf storey, (iii) shape between leaf storey, (iv) leaf color, (v) leaf base, and (vi) size of the middle leaf.

4.3 Generate a rubber breed classifier

To generate the breed classifier of rubber, the dataset with features in the reduct was partitioned into n mutually exclusive subsets or folds, S_1, S_2, …, S_n, each of approximately equal size. This research used 10-fold cross validation method to iterative partition the dataset into 10 learning sets where each set consists of one test set and nine training sets. Within the process of k-NN algorithm, all learning sets were tested where the k value varied from 1 to 25 using the Weighted Voting approach [17]. The accuracy results of 10 test sets with respect to k values (showed only the first ten k values) are shown in Table 2.

It can be seen that k=5 obtained the highest average accuracy result, i.e. 90.67%. Therefore, this value of k will be used in the breed classifier of rubbers using 300 samples with the selected rough set-based features. This classifier can later be used to classify the breed of a new rubber seedling.

4.4 Classify a breed for a given rubber seedling

To evaluate the effectiveness of the breed classifier on a given data, the researcher conducts experiments to compare the predictive performances with and without rough set feature selection (RSFS). Without RSFS, the Chi-Square, Spearman's Correlation, and Pearson's Correlation [18] are used. Moreover, the accuracy in breed classification performed by three experts who are assumed to have equal skillful level was compared as shown in Table 3. The new given test set consists of 10 same breeds, 5 seedlings per breed.

It can be seen that the breed classifier with RSFS can classify the breed of rubber with 86% accuracy. This rate is rather good compared with the breed classifier without RSFS and higher than classifying the breed by experts. For the accuracy obtained from the 3 experts, each expert can get 20/40/60/80/100 percentage of accuracy in classifying each breed. The last column comes from the average classification of these experts. It is low, 55.33% of accuracy. The reasons of getting low accuracy from experts' breed classification may come from (i) the features of one breed is more similar to other breeds or (ii) the mutation of rubber. The breeds which the experts and the breed classifier classify with high degree of accuracy are RRIT 402, PB 235, and BPM 1. They

Table 3 Accuracy comparison of classifying rubber breeds

Breeds of Rubber	Classifier without RSFS	Classifier with RSFS	Average Classification by Experts
1. RRIT 408	60%	60%	26.67%
2. RRIT 251	80%	100%	53.33%
3. RRIT 226	80%	80%	40.00%
4. BPM 24	60%	80%	53.33%
5. RRIM 600	80%	80%	53.33%
6. RRII 118	60%	80%	26.67%
7. PB 235	100%	100%	60.00%
8. RRIT 402	100%	100%	100.00%
9. AVROS 2037	80%	80%	46.67%
10. BPM 1	100%	100%	86.67%
Average Accuracy	**80%**	**86%**	**55.33%**

have values of leaf storey, leaf and petiole features different from other breeds explicitly.

5. Conclusion and discussion

This research employed the rough set-based feature selection to search for a subset of relevant features (termed a reduct) from the original features of rubber. The informative features within the found subset or reduct are those that are most predictive of the class feature in the rubber breed classification phase. The breed classifier was generated with the k-Nearest Neighbor (k-NN) algorithm based on the Variables of Mixed Types technique used to compute the distance between different types of features. The accuracy results of the generated rubber breed classifier are high compared with classified by the experts. It can be used to classify the breed of a new rubber seedling. Therefore, it can assist the agriculturists classify and select the correct breed of rubber seedlings to cultivate in the rubber garden.

Beyond classifying the rubber breed based on these studied features, other features of rubber could be studied in the future such as trunk of breeder or seed. In addition, the features of soil used to plant the focused rubber will be more useful to study in the future for classifying the breed of rubber. Moreover, other high performance techniques of classification can be studied on this rubber dataset in order to find the optimal breed classifier.

6. References

[1] Susevee, P. Choosing the Suitable Breeds of Rubbers for Planting. Kasikorn 2003;76(4):27-18.

[2] The Rubber Research Institute of Eastern Thailand [Internet]. [cited 2016 Apr 15]. Available from: http://www.live-rubber.com/

[3] Miller AJ. Subset Selection in Regression. London: Chapman and Hill; 1990.

[4] Langley P. Selection of Relevant Features in Machine Learning. AAAI Fall Symposium on Relevance. 1994:1-5.

[5] Dash M, Lin H. Feature Selection for Classification. Intelligent Data Analysis 1997;1(3):131-156.

[6] Lio H, Motodo H. Feature Extraction, Construction and Selection: A Data Mining Perspective. Dordrecht: Kluwer Academic; 1998.

[7] Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences 1982;11(5):341-356.

[8] Raghavan VV, Sever H. The State of Rough Sets for Database Mining Applications. 23rd Computer Science Conference Workshop on Rough Sets and Database Mining; 1995. p. 1-11.

[9] Shang C. Shen, Q. Rough Feature Selection for Neural Network based Image Classification. International Journal of Image Graph 2002;2(4):541-556.

[10] Gupta KM, Moore PG, Aha DW, Pal SK. Rough Set Feature Selection Methods for Case-Based Categorization of Text Documents. In: Pal SK, Bandyopadhyay S, Biswas S, editors. Pattern Recognition and Machine Intelligence, LNCS. Berlin Heidelberg: Springer-Verlag; 2005. p. 792-798.

[11] Agarwal B, Mittal N. Sentiment Classification using Rough Set based Hybrid Feature Selection. Proceeding of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2013 June 14; Atlanta, Georgia. USA: Association for Computational Linguistics; 2013. p. 115-119.

[12] Polkowski L. Rough Sets: Mathematical Foundations. Advances in Software Computing. Heidelberg: Physica; 2002.

[13] Shen Q, Jensen R. Rough Sets, their Extensions and Applications. International Journal of Automation and Computing 2007;4(3):217-228.

[14] Dasarathy BV. Nearest Neighbor (NN) Norms: Pattern Classification Techniques. Los Alamitos: IEEE Computer Society; 1991.

[15] Polkowski L, Skowron A. Rough Sets and Current Trends in Computing. LNAI 1424, Springer; 1998.

[16] Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd ed. Academic Press; 2006.

[17] Jensen R, Shen Q. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. New York: John-Wiley & Son; 2008.

[18] Myers JL, Well AD. Research Design and Statistical Analysis. 2nd ed. Lawrence Erlbaum; 2003.