ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

INTERNATIONAL JOURNAL OF RESEARCH – GRANTHAALAYAH
A knowledge Repository

Science

COMPARISON OF ALGORITHMS BASED ON ROUGH SET THEORY FOR A 3-CLASS CLASSIFICATION

Yonca Yazirli [1], Betül Kan-Kilinç [*2]

[*1] Institute of Graduate Programs, Department of Statistics, Eskisehir Technical University, Eskisehir, Turkey

[*2] Department of Statistics, Faculty of Science, Eskisehir Technical University, Eskisehir, Turkey

Abstract

There are various data mining techniques to handle with huge amount of data sets. Rough set based classification provides an opportunity in the efficiency of algorithms when dealing with larger datasets. The selection of eligible attributes by using an efficient rule set offers decision makers save time and cost. This paper presents the comparison of the performance of the rough set based algorithms: Johnson' s, Genetic Algorithm and Dynamic reducts. The performance of algorithms is measured based on accuracy, AUC and standard error for a 3-class classification problem on training on test data sets. Based on the test data, the results showed that genetic algorithm overperformed the others.

Keywords: Attribute Reduction; Rough Set Theory; Classification; Real Estate.

Cite This Article: Yonca Yazirli, and Betül Kan-Kilinç. (2019). "COMPARISON OF ALGORITHMS BASED ON ROUGH SET THEORY FOR A 3-CLASS CLASSIFICATION." *International Journal of Research - Granthaalayah*, 7(8), 394-401. https://doi.org/10.5281/zenodo.3401362.

1. Introduction

The rapid development of online platforms or availability of storing data is an emerging area for researchers to form and process the huge amount of data stacks. The growing volume of larger data sets has gained considerably attention among researchers. Therefore, data mining and its related approaches have become useful and valid for identifying the data patterns. One of the important criteria regarding to that is the attribute reduction. The reduction describes as an attribute set for generating efficient rule sets. In other words, relevant attributes are needed to be selected, called as attribute reduction. This is important for researchers as decision makers save time and cost by excluding attributes that do not contribute positively to the solution of the problem. Thus, creating the efficiency of the algorithms that can be the removal of negligible variables from the data set, is the following emerging area.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

Cluster and regression analysis, neural network, fuzzy sets, Bayesian methods, machine learning can be included in the field of data mining theories and techniques. Kusiak (2006) generally described data mining techniques in two classes, descriptive and predictive [1]. The first class included a model created by the training data such as in neural network and regression analysis. The second was creation of a number of models in the form of decision models such as machine learning algorithms. Rough set theory is a novel approach, proposed by Pawlak, for researchers in data mining to handle with vagueness in data patterns. Attribute reduction without losing the necessary information from the data set is one of the most capable approaches used for this purpose is offered by the Rough Set Theory [2].

The reduct generation or approximations to reduction generation in rough set theory was studied by many researchers. In this regards, Johnson (1974) provided a possible classification of optimization problems as to the behaviour of their approximation algorithms [3]. An approximate approach for reduct computation that utilized a weighting mechanism to determine the significance of an attribute to be considered in the reduct was provided by Al-Radaideh in 2005 [4]. To produce small reducts by a genetic algorithm with a greedy algorithm was offered by Wroblewski [5]. Swiniarski and Skowron [6] and Zeng [7] provided algorithms to knowledge acquisition based on rough set and principal component analysis. Srivastava et al. [8] introduced Rough Support Vector Machine approach based on the hybridization of SVM and Rough Set Exploration System. It was applied to find reducts which then used to SVM to get better classification results. Yamany et. al [9] developed an innovative use of an intelligent optimisation method, namely the flower search algorithm (FSA), with rough sets for attribute reduction. FSA has robust search capabilities and can effectively find small attribute reducts based on a suitable definition of a fitness function that combines both classification accuracy and attribute set size. Experimental results proved competitive performance for FSA-based approach showing that FSA combined with rough sets forms a useful technique for the attribute reduction problem.

In this paper, we evaluate reduction algorithms based on rough set theory for efficient classification with a minimum set of attributes for real estate in Istanbul. The paper is structured as follows. In Section 2, rough set theory preliminaries are defined. Reduction algorithms such as Johnson' s, Genetic Algorithm and Dynamic reducts are explained in Section 3. The reduction algorithms are evaluated by using the same classifier which is the voting method. Then, the comparisons of reduction methods are given in Section 4. The last section concludes the paper.

2. Rough Set Theory Preliminaries

Rough Sets developed by Pawlak is a new approach for handling vagueness and uncertainty in certain data sets [2,10,11]. Following Pawlak, the information system and indiscernibility relation, discernibility matrix and function are introduced in this section.

Definition 1: Information Systems and Decision Systems
A data set is represented as a table, where each row represents an object. Every column represents an attribute (an explanatory variable or a property) that can be measured for each object; the attribute may be also supplied by a human expert or the user. Such table is called an information system. Formally, an information system is a tuple $S = (U, A)$ where U is a non-empty finite set

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

of objects called the universe and A is a non-empty finite set of attributes such that $a : U \rightarrow V_a$ for every $a \in A$. The set V_a is called the value set of a.

Decision system is the table that includes the decision attribute with the objects and conditional attributes. The elements of *A* are called conditional attributes or simply conditions. Decision system defines $DS = (U, A \cup \{d\})$ where *d* is decision attribute $d \notin A$. The decision attribute is categorical variable. In rough set theory, decision attribute is always in the last column of the table.

Definition 2: Indiscernibility Relation
Every subset of attributes $B \subseteq A$ induces indiscernibility relation:

$$IND_S(B) = \{(x_i, x_j) \in |U| \times |U| \mid a(x_i) = a(x_j), \forall\, a \in A\}$$

For each subset of attributes $B \subseteq A$, if two objects (x_i, x_j) are same values for the set of attributes *B*, they cannot be discerned from each other on the basis of the set of attributes *B*.
For every $x \in U$, there is an equivalence class $[x]_B$ in the partition of *U* defined by $IND_S(B)$.

A reduct of a decision system is any subset $B \subseteq A \cup \{d\}$ such that $IND(B) = IND(A)$ and

$IND(B - \{a\}) \neq IND(A)$ for every $a \in B$.

While $B \subseteq A$ and $a \in B$, if the subset of conditional attributes *B* maintains the $IND(A)$ indiscernibility relation, the attributes of set *a* may be omitted. Subsets that do not contain removable attributes are called reduced attribute sets. The core set of the decision system is defined as

$$Core(B) = \cap\, Red(B)$$

Where *Red(B)* is the set of all reducts of *B*.

Definition 3: Discernibility Matrix
The discernibility knowledge of the decision system is commonly recorded in a matrix called the discernibility matrix (DM). The DM is a symmetric $|U| \times |U|$ matrix with entries $[c_{ij}]$ defined as:

$$[c_{ij}] = \{a \in B \mid if\ \ a(x_i) \neq a(x_j), \forall\, a \in A\ ;\ \emptyset\ otherwise\}$$

c_{ij} of the DM includes all the attributes that discriminate between two objects x_i and x_j.

Definition 4: Discernibility Function
Discernibility function is a Boolean function that composed of variable a^* corresponds to attribute a. It represents as f_{IS} [11,12].

$$f_{IS} = \wedge \{\vee c_{ij}^* \mid 1 \leq j \leq i \leq n, c_{ij} \neq \emptyset\}$$

where $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

3. Reduction Algorithms Based on Rough Set Theory

3.1. Johnson's Algorithm

Johnson's algorithm [3] is a heuristic algorithm using a greedy technique. The idea of Johnson's algorithm is that it always selects the attribute most frequently occurring in the clause.

The reduct *B* is generated by executing the algorithm outlined below, where $\mathcal{S}$ denotes the set of sets corresponding to the discernibility function and $w(S)$ denotes a weight for set S in $\mathcal{S}$ that automagically gets computed from the data.

The algorithm is described as follows [14]:

1) Let $B = \emptyset$.
2) Let a denote the attribute that maximizes $\sum w(S)$, where the sum is taken over all sets S in $\mathcal{S}$ that contain a. Currently, ties are resolved arbitrarily.
3) Add a to *B*.
4) Remove all sets S from $\mathcal{S}$ that contain a.
5) If $\mathcal{S} = \emptyset$; return *B*. Otherwise, go to step 2.

3.2. Genetic Algorithm

Vinterbo and Øhrn [14] described genetic algorithms for computing minimal hitting sets. The algorithm has support for both cost information and approximate solutions. The algorithm' s fitness function f is described as follows:

$$f(B) = (1-\alpha) \times \frac{cost(A) - cost(B)}{cost(A)} + \alpha \times \min\left\{\varepsilon, \frac{|[S\ in\ \mathcal{S}|\ S \cap B \neq \emptyset]|}{|\mathcal{S}|}\right\}$$

Where $\mathcal{S}$ is the set of sets corresponding to the discernibility function, the parameter α defines a weighting between subset cost and hitting fraction, while ε is relevant in the case of approximate solutions.

The subsets *B* of *A* are found by an evolutionary search measured by $f(B)$, when a subset *B* has a hitting fraction of at least ε then it is saved in a list. The size of the list is arbitrary. The function cost specifies a penalty for an attribute (some attributes may be harder to collect) but it defaults to $cost(B) = |B|$. If $\varepsilon = 1$, the minimal hitting set is returned. In this algorithm the support count is the same as in Johnson' s algorithm [15].

3.3. Dynamic Reducts

The dynamic reduction algorithm is a combination of normal reduct computation with resampling techniques [16,17].

The steps of algorithms are explained as follows:

1) Randomly sample a family of subsystems $S = \{S_1, S_2, \ldots, S_n\}$ from $S = (U, A)$, where each sub-systems $S_i = (U_i, A)$ and $U_i \subseteq U$.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

2) From each sub-systems, including $S_i \in S$, compute a reduced attribute set using reduction rules.
3) Determine the most frequently generated reduced attribute set from the reduced attribute sets obtained in the previous step.

The reducts that occur the most often across sub-tables are in some sense the most "stable" [14]. After reduction algorithms based on rough set theory, the decision rules obtained as a result of the application of these algorithms are used to determine the classification performance of the algorithms. Voting method is used for classification. It is an ad hoc technique for rule-based classification. The process of voting is that the most obtained class value for each object as a result of voting is the decision class value.

4. Application

In this paper, the advertisements of real estate from on an online platform in which people can sell or buy also car, variety of goods and services were collected for Istanbul between 9 October- 13 December 2018. The data set contains the sale prices of 250 real estate for residential purposes. Also the properties such as the number of rooms, age of building, number of floor, elevator, and bathroom were considered as explanatory variables and recorded for each real estate as well. One of the explanatory variables was the district of the real estate. The variables had 5 classes where each represented a different district of Istanbul. Hence the variable district was classified in five classes. The variables of garage and balcony were classified as 1 and 0 elsewhere. Also, the convenience point or amenities for a real estate is considered as an explanatory variable and was coded as 1 for yes, 0 elsewhere. The dependent variable is the price of real estate that was converted to a categorical variable. The determine the class intervals, housing unit prices for Turkey (₺/m2) in 2018 are used (EVDS, Data Central) [18]. According to that, if the price was larger than 2315,17TL then it was classified as 2, if smaller than 2315,17TL and larger than 2118,52TL then it was classified as 1, and 0 elsewhere.

Data set is split as 70% for training and 30% for testing. All operations are calculated in ROSETTA software which developed based on rough set theory by Øhrn in 2001 [19]. Firstly, reduction algorithms are applied. Then, the decision rules obtained by reduction algorithms are used to determine the classification performance of the algorithms. Voting method is applied for classification. The accuracy, standard error of accuracy and AUC values are compared for the performance of classification. AUC is a kind of measure of separability, also it tells how much the model is capable of separating the class.

The reduction results of attribute reduction algorithms based on rough set theory are demonstrated in *Table 1*. It shows the number of reducts, attributes in reducts, decision rules and accuracy of algorithms. With respect to the number of reducts, dynamic reducts applied maximum reduct number and the number of reducts have changed from 1 to 6 attributes. The number of decision rules obtained by reducts is 1166 for dynamic reducts. Success of Johnson's algorithm with maximum 4 attributes equals to success of genetic algorithm with maximum 5 attributes for training, however genetic algorithm performed a better performance with 81.33% among reduction algorithms for testing.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

Table 1: Overall performance of attribute reduction algorithms

Reduction Algorithm	The number of reducts	The number of attributes in reducts	The number of decision rules (training)	Training Acc.	Test Acc.
Johnson's Algorithm	30	1-4	98	0.988	0.760
Genetic Algorithm	127	1-5	484	0.988	0.813
Dynamic Reducts	190	1-6	1166	0.840	0.800

Based on the previous results given in *Table 1*, the performance of standard voting classifier for each reduction is summarized *Table 2*.

Johnson and genetic algorithms have performed well as 98.8% whereas dynamic reducts has slightly worse performed as 84% for each class with respect to training accuracy. Also, genetic algorithm has performed better with respect to accuracy in testing (81.3%) than others. However, the smallest difference in accuracy for train and test data is obtained by using Dynamic reducts algorithm.

Table 2: Classificaion performance of reduction algorithms

Reduction Algorithm	Classes	*Training*			*Test*		
		Accuracy	**AUC**	**St. Error**	**Accuracy**	**AUC**	**St. Error**
Johnson Algorithm	Expensive	0.988	0.999	0.001	0.760	0.691	0.069
	Moderate	0.988	0.999	0.009	0.760	0.500	0.208
	Cheap	0.988	0.999	0.001	0.760	0.661	0.086
Genetic Algorithm	Expensive	0.988	0.999	0.001	0.813	0.710	0.064
	Moderate	0.988	0.999	0.009	0.813	0.472	0.204
	Cheap	0.988	0.999	0.001	0.813	0.706	0.079
Dynamic Reducts	Expensive	0.840	0.928	0.018	0.800	0.734	0.061
	Moderate	0.840	0.996	0.021	0.800	0.551	0.213
	Cheap	0.840	0.937	0.027	0.800	0.734	0.077

AUC score for dynamic reducts (73.4%) means that a randomly chosen expensive instance assigned to class expensive is higher than being assigned to class moderate and cheap with probability 73.4%. Hence, this score is better than Johnson and genetic algorithms have for expensive class. AUC score of moderate class within all algorithms have performed considerably weak. The AUC score for moderate class is 0.551 that means a randomly chosen moderate instance assigned to this class is 55.1% than being assigned to class expensive and cheap class. The AUC score for moderate class in genetic algorithm indicates that the model separates the moderate class poorly than the others.

5. Conclusion

In this paper, reduction algorithms based on rough set theory for efficient classification with a minimum set of attributes for real estate in Istanbul have been examined. The reduction algorithms were evaluated by using the same classifier: the voting method. The housing unit prices of real estate for sale in different districts of Istanbul obtained from an online web source and 250 real

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

estate were investigated. In the process of determining the best reduction algorithm based on rough set theory, the classification performance of the test data was taken into consideration and the genetic algorithm was chosen as the most successful reduction algorithm.

References

[1] Kusiak A., Data Mining in Design of Products and Production Systems, Proceedings of INCOM'2006: 12th IFAC/IFIP/IFORS/IEEE Symposium on Control Problems in Manufacturing, May 2006, Saint-Etienne, France, 1, 2006, 49-53.
[2] Pawlak, Z. Rough sets, International Journal of Computer and Information Science, vol.11, no.5,1982, 341-356.
[3] Johnson, D. Approximation algorithms for combinatorial problems, Journal of Computer and System Sciences, 9, 1974, 256-278.
[4] Wroblewski, J. Finding minimal reducts using genetic algorithms, Second Annual Join Conference on Information Sciences, 1995, 186-189.
[5] Al-Radaideh, Q. A., Sulaiman, M. N., Selamat, M. H., Ibrahim, H. Approximate reduct computation by rough sets based attribute weighting, 2005 IEEE International Conference on Granular Computing, Beijing, China, 2005, 25-27 July.
[6] Swiniarski, R.W., Skowron, A. Rough set methods in feature selection and recognition, Pattern Recognition Letters, vol. 24, no. 6, 2003, 833–849.
[7] Zeng, A., Pan, D., Zheng, Q. L., Peng, H. Knowledge acquisition based on rough set theory and principal component analysis, IEEE Intelligent Systems, vol. 21, issue 2, 2006, 78-85.
[8] Srivastava, D. K., Patnaik, K. S., Bhambhu, L. Data classification: A Rough-SVM approach, Contemporary Engineering Sciences, Vol. 3, no. 2, 2010, 77 – 86.
[9] Yamany, W., Emary, E., Hassanieh, A.E., Schaefer, G. Zhu, S. Y. An Innovative Approach for Attribute Reduction using Rough Sets and Flower Pollination Optimisation, Procedia Compuer Science, 96, 2016, 403-409.
[10] Pawlak, Z. Rough Sets Theoretical Aspect of Reasoning about Data. Boston, Mass, Kluwer Academic, 1991.
[11] Pawlak, Z., Grzymala-Busse, J., Slowinski, R. and Ziarko, W. Rough sets, Communications of the ACM, vol. 38, no. 11, 1995, 89–95.
[12] Skowron, A., Rauszer, C. The discernibility matrices and functions in information systems, in Slowifiski R.(ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer, Dortrecht.1992, 331-362.
[13] Zhao, Y., Yao, Y. and Luo, F. Data analysis based on discernibility and indiscernibility, Information Sciences, 177(22), 2007, 4959–4976.
[14] Vinterbo, S., Øhrn, A. Minimal approximate hitting sets and rule templates, International Journal of Approximate Reasoning, 25, 2000, 123-143.
[15] Godinez, F., Hutter, D., Monroy, R., Attribute Reduction for Effective Instrusion Detection, Advances in Web Intelligence, Second International Atlantic Web Intelligence Conference, AWIC, Cancun, Mexico, 2004, May 16-19.
[16] Bazan J.G., Skowron A., Synak P. Dynamic reducts as a tool for extracting laws from decisions tables. In: Raś Z.W., Zemankova M. (eds) Methodologies for Intelligent Systems. ISMIS 1994. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 869. Springer, 1994, Berlin, Heidelberg.
[17] Bazan, J. G. (1998) "A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables". Rough Sets in Knowledge Discovery 1: Methodology and Applications, volume 18 of Studies in Fuzziness and Soft Computing, Heidelberg, Germany Physica-Verlag, 1998, Chapter 17, pages 321–365.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.3401362

[18] EVDS Data Central. URL: https://evds2.tcmb.gov.tr/index.php?/evds/serieMarket, Accessed Date:14.02.2019.
[19] Øhrn, A. ROSETTA Technical Reference Manual. Trondheim, Norway, 2001, Department of Computer and Information Science, Norwegian University of Science and Technology.

*Corresponding author.
E-mail address: bkan@ eskisehir.edu.tr