Pak.j.stat.oper.res. Vol. 16 No.2 2020 pp 217-224 DOI: http://dx.doi.org/10.18187/pjsor.v16i2.3069

Pakistan Journal of Statistics and Operation Research

Performance of the Hybrid Approach Using Three Machine Learning Algorithms

Betül Kan-Kılınç[1*], Yonca Yazırlı[2]

* Corresponding Author

1. Eskisehir Technical University, Faculty of Science, Department of Statistics, Turkey, bkan@eskisehir.edu.tr
2. Eskisehir Technical University, Institute of Graduate Programs, Department of Statistics, Turkey, yoncayazirli@gmail.com

Abstract

One of the essential problems in data mining is the removal of negligible variables from the data set. This paper proposes a hybrid approach that uses rough set theory based algorithms to reduct the attribute selected from the data set and utilize reducts to raise the classification success of three learning methods; multinomial logistic regression, support vector machines and random forest using 5-fold cross validation. The performance of the hybrid approach is measured by related statistics. The results show that the hybrid approach is effective as its improved accuracy by 6-12% for the three learning methods.

Key Words: Rough set; Reduction; Performance; Accuracy.

Mathematical Subject Classification: 62-07; 68T05; 68Q32

1. Introduction

Recently, the amount of data collected and stored across the world has been increasing at an exponential rate. However, the data stored in the systems does not make sense on its own; therefore, techniques to predict from the available data are essential for researchers. The process of obtaining valuable information from large amounts of data can be carried out by text mining or data mining techniques. This helps to reveal hidden information from an enormous amount of data that is valuable for recognition of important facts, relationships, trends, and patterns. The process to discover interesting knowledge from large amounts of data is known as data mining. It is a combination of statistics, machine learning and computing (by Han and Kamber, 2006).

Nowadays, due to the rapid development of technology, there are online platforms where more data can be saved. This has led to the formation of large data stacks. As data sets contain multivariate and large volumes of data, analysis has become increasingly difficult and the efficiency of algorithms decreases. Therefore, attribute reduction has become a major problem in the field of data mining, machine learning or pattern recognition for decision-makers. One important criterion in increasing the efficiency of algorithms is the removal of negligible variables from the data set.

An attribute reduction describes as an attribute set for generating an efficient rule set. Therefore, this helps decision-makers to save time and reduce costs by excluding attributes that do not contribute to the solution of the problem. Attribute reduction without losing important information from the data set is regarded as a multi-criteria decision problem. One of the most capable approaches used for this purpose is offered by the Rough Set Theory (by Pawlak, 1982).

When handling a data set in many domains, such as retail, finance, insurance, medicine, social media and marketing and others, a researcher usually wishes to work on the best attribute set for the given data set and expects the best model. This expectation leads her/him to use a large collection of learning models from many families (Bayesian, generalized linear models, discriminant analysis, decision trees, support vector machines, logistic or multi-nominal regression and other methods). Of course, the researcher may implement many learning models within her/his domain of expertise and compare the performance over a data set. In the current paper, a hybrid approach of 3-class classification problem is examined to utilize both the advantages of rough set theory and learning methods.

This paper uses a hybrid approach: 1) to apply the reduction algorithms based on rough set theory for efficient classification with a minimum set of attributes; 2) to evaluate the best classifier for the selected data set collections; 3) to determine, for each classifier, its accuracy and Kappa statistics and differences; and 4) to choose the best one with the best attribute set.

2. Reduction Algorithms Based on Rough Set Theory

Rough set theory is a mathematical approach based on the theory of sets developed by Pawlak. The theory is a methodology of database mining (by Pawlak, 1982;1991;1995).

Several heuristic and approximation approaches have been proposed for the problem of the reduct generation. Johnson (1974) indicates a possible classification of optimization problems as to the behavior of their approximation algorithms. Wroblewski (1995) integrates a genetic algorithm (GA) with a greedy algorithm to produce small reducts. Al-Radaideh et al. (2005) offer an approximate approach for reduct computation. The approach utilized a weighting mechanism to determine the significance of an attribute to be considered in the reduct. Swiniarski and Skowron (2003) and Zeng et al. (2006) developed algorithms to knowledge acquisition based on rough set and principal component analysis. Srivastava et al. (2010) introduce the Rough-Support Vector Machine approach, based on the hybridization of SVM and the Rough Set Exploration System (RSES).

By defining a set in mathematics, it is usually considered to be a set of objects that possesses similar characteristics. Two of the specific sets are fuzzy set and rough set, which is the main approach of interest in this paper. Rough set is defined by approximations and requires advanced mathematical concepts.

An indiscernibility relationship is a main expression in the rough set theory which represents the relationship between objects. In an indiscernibility relationship, all the values are identical to a subset of considered attributes (by Pawlak, 1982). This concept is the starting point of rough set. Other concepts are the approximations and reduction rules. The following subsections introduce three decision rules: Johnson's Algorithm, Genetic Algorithm and Dynamic Reduction.

2.1. Johnson' s Algorithm

The Johnson algorithm (1974) is a heuristic algorithm using a greedy technique. The idea of the Johnson algorithm is that it always selects the attribute most frequently occurring in a clause. The reduct B is generated by executing the algorithm outlined below, where § denotes the set of sets corresponding to the discernibility function f and $w(S)$ denotes a weight for set S in § that automagically gets computed from the data. The algorithm is described as follows (by Øhrn, 2001):

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

2.2. Genetic Algorithm

Vinterbo and Øhrn (2000) describe 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:

$$f(B) = (1 - \alpha) \times \mathrm{cost}(A) - \mathrm{cost}(B) + \alpha \times \min\{\varepsilon, [S in \S | S \cap B /= 0]\} (1) \mathrm{cost}(A) |\S| \quad (1)$$

where § 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 $\mathrm{cost}(B) = |B|$. If $\varepsilon = 1$ f the minimal hitting set is returned. In this algorithm the support count is the same as in Johnson's algorithm (by Godinez, 2004).

2.3. Dynamic Reduction

The dynamic reduction algorithm is a combination of normal reduct computation with resampling techniques (by Bazan, 1994;1998). The steps of the 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 subsystem $S_i = (U_i, A)$ and $U_i \subseteq U$.
(2) From each subsystem 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 most often across sub-tables are in some sense the most stable (by Øhrn,2001).

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. A 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.

3. Learning Methods

The main techniques for data mining are classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and also certain new techniques, such as social network analysis and sentiment analysis (by Han and Kamber, 2006).

One of the areas where data mining is used is the real estate sector. Hromada (2015) represents a software that is used for real estate evaluation, mapping and analyzing real estate advertisements published on the Internet in the Czech Republic from 2007 until today. The software gathers price offers concerning sale or rental of apartments, houses, business properties and building lots for each half year. The author evaluates results as a steady long-term decrease of real estate market prices since the second quarter of 2008 (by Hromada, 2015).

Chogle et al. (2017) aimed to develop a real estate web application using Microsoft ASP .NET and SQL 2008. They used Naive Bayes classification algorithms for the price prediction. It helped to satisfy customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate.

Asilkan et al. (2012) examine the applicability of Hedonic Regression (HR) and ANN models in the housing market in Tirana. Prediction of implicit prices of housing attributes determined the price of a complete house according to the HR model. The house which contains the desired attributes has a greater price than others. The house price is estimated using the ANN model based on a particular input set. The ANN model obtained a more successful result than the HR model.

Liu and Zong (2017) propose Twin Support Vector Regression, based on data mining and large data, for second-hand real estate price forecasting. It compares with the traditional support vector model; the results of their experiment show that the proposed method achieves a higher predictive performance.

3.1. Multinomial Regression Model

Logistic regression is a method where the response variable is binary. Unlike a multiple linear regression model, the dependent variable in this method is binary, nominal or ordinal. One example of logistic regression models can be a multi-nominal logistic regression model (MNL) where the output variable can have more than two choices that are coded categorically. The choices of the dependent variable may be nominal or ordinal. The nominal categories are not in order and they simply imply categories. The order is taken as the reference category. The nominal logistic regression model allows researchers to predict from a categorical response and draws conclusions about the explanatory variables on these responses (by Lawson and Montgomery, 2006).

3.2. Support Vector Machines

Support vector machines (SVM) are a class of statistical models that was originally developed by Vladimir Vapnik for classification problems (by Cortes and Vapnik, 1995). It is a flexible machine learning algorithm that can be used for classification and regression problems. According to Burges's study (1998), the theoretical basis of SVM generates from statistical learning theory. It aims to find a linearly separable hyperplane with maximum margin between classes. The hyperplane separates the groups; therefore, different points on two different sides of hyperplane are the two groups different from each other. Margin is the perpendicular distance between the decision boundary and the closest of the data points. The support vector is the data points which are closest to the hyperplane. Therefore, SVM finds hyperplanes using support vectors and margins. The SVM algorithm is based on finding a hyperplane with the largest minimum distance. It is possible to distinguish linearly separable data, but it is not easy to separate the nonlinearly separable data. Therefore, SVM uses the kernel function to transform the input data into a higher dimensional space that is larger than its size. The optimal hyperplane is constructed with respect to maximum margin. Hence, the separation of classes is more easily obtained.

3.3. Random Forest

Random Forest, developed by Leo Breiman (RF), is one widely-used method in data mining. It is a group of classification or regression trees generated from a random selection of samples of the data. RF uses an ensemble (i.e., a forest) of decision tree predictors, such that each tree has no relationship with another and with the same distribution for all the trees in the forest. The correlation between trees is reduced by selecting the features. The trees are grown to maximum size (e.g. 2000 trees). This then combines the predictions from all of the trees. In addition, unlike decision trees, there is no need to prune the trees. RF can handle a high dimensional data set with missing values, continuous, categorical or binary types.

3.4. Our Approach

As the weak point of classifiers may depend on its data set, it is intuitive to combine the reduction algorithms based on rough set theory and learning methods together.

(1) (repeat)
(2) initialize k
(3) split the entire data set into training (70%) and test (30%)
(4) evaluate the accuracy and Kappa statistics for MNL, SVM and RF
(5) until k-fold cross validation is requested by user
(6) calculate the reducts based on rough set theory
(7) evaluate the accuracy for the test set
(8) choose the best reduction with respect to its highest accuracy
    (9) (repeat)
    (10) initialize k
    (11) find the number of reducts
    (12) construct the reducts obtained from step (8) as matrices
    (13) split each reduct into training (70%) and test (30%)
    (14) evaluate the accuracy and Kappa statistics for each reduct using MNL, SVM and RF
    (15) until k-fold cross validation is requested by user
    (16) choose the best classifier with highest accuracy
    (17) end

The advantage of this approach is that the researcher does not have to select the attributes in the entire data set for a good classification. Moreover, after the hybrid approach is performed, the unnecessary variables are not included in the final model and the accuracy is improved.

4. Application

This study examines the cost of Istanbul public housing for rent and house attributes and identifies the significant determinants of the cost of renting based on the hybrid approach. A total of ten variables are used to compare and analyze the relationship between the cost of rents and the housing attributes. Table 1 reports the data set where the columns represent the attributes utilized by rough set.

Table 1. Information system

Conditional Attributes				**Continuous Attributes**					**Decision Attribute**
Rooms	**Elevator**	**Garage**	**Balcony**	**District**	**Bedroom**	**Bathroom**	**Age**	**Floor**	**Price**
Yes	Yes	Yes	Yes	1	1	1	{0,33}	{0,26}	Expensive
No	No	No	No	2	2	2			Moderate
				3	3	3			Cheap
				4	4	4			
				5	5				
					6				
					7				

Real estate in Turkey is characterized by different prices. The cost of apartments that are smaller than 2118.52 TL per square meter and larger than 2315.17 TL per square meter are recorded as 0 and 2, and 1 elsewhere. The cost of housing units in 2018 for Turkey is obtained from the Electronic Data Delivery System.
First, the data set was split into two groups: training and test data sets. MNL, SVM and RF methods were performed using a 5-fold validation technique to classify the house unit cost in Istanbul. The performance of the three methods is given in Table 2.

Table 2. Performance of MNL, SVM and RF with accuracy, Kappa statistics and reducts list obtained from genetic algorithm

	Metrics	
Methods	**Accuracy**	**Kappa**
MNL	0.757	0.190
SVM	0.797	0.366
RF	0.821	0.423

Table 2 shows the accuracy value of 82.1% for RF. Accuracy represents correctly classified observations among the positive classes. As can be seen in the table, this value for RF is larger compared to MNL and SVM. This means RF is more capable of correctly classifying the positive classes than others. Accuracy is always the easiest metric to use for comparisons, however, it does not show the type of errors the classifier does. Another method that can be used is the Kappa statistic. Similar results hold for the Kappa statistic.

It should be noted that the accuracy values in Table 2 are the best for RF. However, a hybrid approach can be performed to raise the performance of these classifiers. For this purpose, three reduction algorithms, based on rough set theory, are performed to classify the housing unit cost. Johnson's algorithm, Genetic Algorithm and Dynamic Reduction techniques are performed to select the attribute set. Then, as given in section 2, the decision rules obtained by reduction algorithms are used to determine the classification performance of the algorithms. A voting method is applied for classification. For all computations based on rough set theory ROSETTA developed by Ohrn (2000) and others R-Studio and related packages were used.

The reduction results of attribute reduction algorithms, based on rough set theory, are shown in Table 3.

Table 3. Overall performance based on attribute reduction algorithms

Reduction Algorithm	**Reducts**	**Attributes in Reducts**	**Decision Rules (Training)**	**Training Accuracy**	**Test Accuracy**
Johnson' s Algorithm	22	1-3	50	0.988	0.760
Genetic Algorithm	65	1-4	203	0.988	0.813
Dynamic Reducts	113	1-6	439	0.840	0.800

The table shows the number of reducts, attributes in reducts, decision rules and accuracy of algorithms for training and test data sets. With respect to the number of reducts, dynamic reducts applied the maximum reduct number and the number of reducts changed from 1 to 6 attributes. The number of decision rules obtained by reducts is 439 for dynamic reducts. The success of Johnson's algorithm with a maximum of three attributes equals the success of the genetic algorithm with a maximum four attributes for training. However, the genetic algorithm performed a better performance with 81.33% among the reduction algorithms for testing.

As can be seen in Table 3, Johnson and the genetic algorithms performed well at 98.8%, whereas the dynamic reducts had slightly worse performance at 84% for each class with respect to training accuracy. In addition, the genetic algorithm performed better with respect to accuracy in testing (81.3%) than others. However, the smallest difference in accuracy for train and test data was obtained using the Dynamic algorithm.

The reducts from the genetic algorithm with 65 reduct sets (different combinations of attributes) were considered in a later analysis. These reduct sets are all at 70% for training and 30% for testing. For the model validation, a 5-fold cross-validation technique was used. MNL, SVM and RF models built from the training data set were used to classify the unit cost of Istanbul housing for rent. The results after the hybrid approach was performed given in Table 4 are compared in the following table.

Table 4. Performance of MNL, SVM and RF with accuracy and Kappa statistics after hybrid approach

Methods / **Metrics**	**MNL**	**SVM**	**RF**
Accuracy	0.879	0.879	0.879
Kappa	0.619	0.619	0.615
No		**Reducts**	
1	{Bathroom, Garage}	{Bathroom, Garage}	{Age, Floor, District}
2	{Age, Floor, District}	{Garage, Floor}	-
3	{Bedroom, Garage, Balcony, Floor}	{Age, Floor, District}	-
4	-	{Bedroom, Garage, Balcony, Floor}	-

The results in Table 4 suggest that the models built performed satisfactorily. The overall accuracy performance with a high model accuracy of 87.9% is the same for MNL, SVM and RF for 4 reduct sets. This value implies that all three classifiers provide the same performance after reduction with different reducts. Similar results almost hold for the Kappa statistic. In addition, the above analysis strongly suggests that the age of the building, floor level and location of the apartment in Istanbul are highly effective on the cost of rent. The implication of these attributes could be used for future research.

5. Conclusion

In this paper, a hybrid approach is discussed to improve the model accuracy of three popular classifiers on the predictive performance of models developed from a real data set.

The dynamic reduction, genetic algorithm, and Johnson algorithms, based on rough set theory for efficient classification with a minimum set of attributes, are evaluated for unit real estate costs in Istanbul.

In the process of determining the best reduction algorithm based on rough set theory and the best classifier, a hybrid approach was used. It has been shown that the genetic algorithm was chosen as the most successful reduction algorithm among the classification performance of the test data taken into consideration. The reducts from the genetic algorithm that had 65 different rules of attributes were considered and used to classify the unit cost of real estate. A multinomial logistic regression model, support vector machines and random forest methods were used to estimate the performance of the unit cost. The performance of all methods was compared with respect to accuracy and Kappa value.

The performance results before the hybrid approach and after the hybrid approach improves accuracy by 6-12% among all the methods. It is also important to note that the best performance was obtained when the reduct set was constructed age, district, and floor.

This analysis not only underlines the choice of the best classifier associated with housing costs, but is also an alternative explanatory data analysis tool for classifying the characteristics of the cost of housing. Although the results in this study are specific to Istanbul, it would also be possible to apply the hybrid approach to other housing markets.

Finally, to utilize the hybrid approach for the housing market, researchers should acquire knowledge on rough set theory and data mining tools and methodologies. The methods used in this paper are an indicator for applying a hybrid classification based on rough set theory and may be useful for future studies.

References

1. Al-Radaideh, Q. A., Sulaiman, M. N., Selamat, M. H., & Ibrahim, H. (2005). Approximate reduct computation by rough sets based attribute weighting. IEEE International Conference on Granular Computing, Beijing, China.
2. Asilkan, O., Faqolli, A., Gerdeci, A., & Cico, B. (2012). Estimating the market values of houses in Tirana using data mining. AWER Procedia Information Technology and Computer Science, 1, 1224–1234.
3. Bazan J. G., Skowron A., & Synak P. (1994). Dynamic reducts as a tool for extracting laws from decisions tables. In: Ras Z.W., Zemankova M. (eds) Methodologies for Intelligent Systems. ISMIS 1994. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), 869. Springer, Berlin, Heidelberg.
4. 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, 18 of Studies in Fuzziness and Soft Computing. Chapter 17, pp. 321-365, Physica-Verlag, Heidelberg, Germany.
5. Breiman, L., Friedman, J., Stone, C.J., & Olshen, R.A. (1984). Classification and regression trees. Chapman and Hall/CRC Press, Florida.
6. Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121-167.
7. Chogle, A., Khaire, P., Gaud, A., & Jain, J. (2017). House price forecasting using data mining techniques. International Journal of Advanced Research in Computer and Communication Engineering, 6(12), 81–90.
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
9. Hromada, E. (2015). Mapping of real estate prices using data mining techniques. Procedia Engineering, 123, 233–240.
10. Liu, G., & Zong, X. (2017). Research of second-hand real estate price forecasting based on data mining. IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference, Chengdu, China.
11. Godinez, F., Hutter, D., & Monroy, R. (2004). Attribute reduction for effective intrusion detection. Advances in Web Intelligence, Second International Atlantic Web Intelligence Conference, AWIC 2004, Cancun, Mexico.
12. Han, J., & Kamber, M. (2006). Data mining: concepts and techniques. 2nd ed., Elsevier, San Francisco, California.
13. Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning: data mining, inference, and prediction. 2nd ed., Springer, Standford, California.
14. Johnson, D. (1974). Approximation algorithms for combinatorial problems. Journal of Computer and System Sciences, 9, 256–278.
15. Lawson, C., & Montgomery, D. (2006). Logistic regression analysis of customer satisfaction data. Quality Reliability Engineering International, 22, 971-984.
16. Ohrn, A. (2001). ROSETTA: Technical Reference Manual, Department of Computer and Information

Science, Norwegian University of Science and Technology, Trondheim, Norway.
17. Pawlak, Z. (1982). Rough sets. International Journal of Computer and Information Science, 11(5), 341–356.
18. Pawlak, Z. (1991). Rough sets theoretical aspect of reasoning about data. Kluwer Academic, Boston, Mass, USA.
19. Pawlak, Z., Grzymala-Busse, J., Slowinski, R. & Ziarko, W. (1995). Rough sets. Communications of the ACM, 38(11), 89-95.
20. Skowron, A., & Rauszer, C. (1992). 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. 331–362.
21. Srivastava, D. K., Patnaik, K. S., & Bhambhu, L. (2010). Data classification: A Rough-SVM approach. Contemporary Engineering Sciences, 3(2), 77-86.
22. Swiniarski, R.W., & Skowron, A. (2003). Rough set methods in feature selection and recognition. Pattern Recognition Letters, 24(6), 833-849.
23. Tan, S., Cheng, X., & Xu, H. (2007). An efficient global optimization approach for rough set based dimensionality reduction. International Journal of Innovative. Computing, Information and Control, 3(3), 725–736.
24. Vinterbo, S., & Ohrn, A. (2000). Minimal approximate hitting sets and rule templates. International Journal of Approximate Reasoning, 25(2), 123–143.
25. Wroblewski, J. (1995). Finding minimal reducts using genetic algorithms. Second Annual Join Conference on Information Sciences, 186–189.
26. Zeng, A., Pan, D., Zheng, Q. L., & Peng, H. (2006). Knowledge acquisition based on rough set theory and principal component analysis. IEEE Intelligent Systems, 21(2), 78– 85.
27. EVDS Data Central, Accessed Date:14.02.2019. Available at: https://evds2.tcmb.gov.tr/index.php?/evds/serieMarket