ARTIFICIAL INTELLIGENCE STUDIES
2022, 4(2): 1-8

www.aistudies.org

Gazi AKADEMİK YAYINCILIK

Çeşitli Veri Setleri için Kural Tabanlı Algoritmaları ile Analiz

Çağrı DÜKÜNLÜ*,[a], Mehmet Uğraş CUMA[a]

[a] *Çukurova Üniversitesi Elektrik Elektronik Mühendisliği Bölümü, ADANA 01330, TÜRKİYE*

MAKALE BİLGİSİ

Konferans Bildirisi

Alınma: 17.11.2021
Kabul: 27.12.2021

***Anahtar Kelimeler**:*
Veri Madenciliği
Kural Tabanlı Algoritmalar,PART,JRIP,OneR

**Sorumlu Yazar:*
e-posta:
cagridknl@gmail.com

ÖZET

Sınıflandırma, kategorileri önceden tahmin edilen verilerden yararlanan bir model hazırlayarak verilen verilerin sınıfını tahmin etme işlemidir. Veri madenciliği teknikleri, önceki verilen sınıflar arasında yeni bir verinin ait olduğu sınıfı tahmin eden bir sınıflandırıcı oluşturmak için düzenli olarak kullanılır. Bu makale, veri madenciliği uygulamalarında kullanılan kural tabanlı sınıflandırıcıların karşılaştırmalı analizini sağlamayı amaçlamaktadır. PART, JRIP, OneR gibi kural tabanlı sınıflandırıcıların performansının analizi.

Bu makalenin amacı, seçilen veri kümeleri altında sınıflandırma kuralları tekniklerinden en iyi tekniği belirlemek ve ayrıca her sınıflandırıcının bir karşılaştırma sonucunu sağlamaktır. Daha iyi sınıflandırma tekniği belirlemek amacıyla diyabet, meme kanseri ve iris veri setlerine uygulanan kural tabanlı sınıflandırıcılar. Karşılaştırma sonuçları doğruluk, kesinlik, duyarlılık ve karışıklık matrisleri ile yapılır.

DOI: 10.30855/AIS.2021.04.02.01

Comparative Analysis with Rule Based Algorithms for Various Datasets

ARTICLE INFO

Conference Paper

Received: 17.11.2021
Accepted: 27.12.2021

Keywords:
Data Mining, Rule Based Algorithms, PART, JRIP , OneR

Corresponding Authors
e-mail:
cagridknl@gmail.com

ABSTRACT

Classification is the operation of predicting class of the given data by preparing a model that makes use of data whose categories already predicted. Data mining techniques are regularly used to form a classifier that predicts belonging class of a new data among the previous given classes. This paper intends to provide comparative analysis of the rule based classifiers used in data mining applications. Analyzing the performance of rule based classifiers namely PART, JRIP, OneR.

The goal of this paper is to specify the best technique from classification rules techniques under the chosen datasets and also provide a comparison result each classifier. The rule based classifiers applied to diabetes, breast cancer and iris datasets due to the purpose of determining better technique for classification. Comparison results are made with accuracy, precision, sensitivity and confusion matrixes.

DOI: 10.30855/AIS.2021.04.02.01

Publication Ethics: This paper was presented as an oral paper at the International Conference on Engineering Technologies 2021 (ICENTE'21), Konya, Turkiye.

To cite this article: C. Dükünlü, M.U. Cuma. "Comparative Analysis with Rule Based Algorithms for Various Datasets", *Artificial Intelligence Studies,* vol.3, no.1, pp. 01-09, 2021. doi:10.30855/AIS.2021.02.01

1. INTRODUCTION *(GİRİŞ)*

Data mining is mostly known as KDD or "Knowledge Discovery in databases" and is used to turn raw data into useful information by companies [1]. The data of the world is getting bigger every day and it is getting difficult to handle the complexity of data every minute. With new collected data and different types of classification methods, mankind is getting advantage of important areas like health care, technology.

Thousands of data sets are processed every day in these fields [2]. To get healthy and organized data or to train a model, it becomes serious and Data Mining comes into play in this step. With the help of data mining techniques, the collected data can be used and organized to generate the knowledge about the data. Data mining also has tools to find out unknown patterns and validate the patterns with relationships from big data.

Some of the well-known tools of Data Mining are "Knime, Weka, Orange, R, Matlab". Every day the capacity of the tools are increased and the datasets are growing with time. With more and more data and improved models, humanity and research will be able to solve bigger and bigger problems with time and take advantage of technology with Data Mining. Due to the efficiency and models variety; Weka tool is used for model training. Weka is an open source software and implementing it to data is easy [3]. These two factors show its maintainable and easy of use to any company who wants to train their models.

Three different classification techniques are chosen for the processing the different datasets to determine in which rule based algorithm has least loss and higher accuracy. From classification techniques JRIP(RIPPER), OneR (OneRule) and PART (Projective Adaptive Resonance Theory) are selected based on literature researches.

Researchers of the study [4] has found various crime patterns. Study has obtained JRip, OneR, Decision Table methods accuracy are compared for crime patterns. JRip was most accurate classifier of all, even though it takes the maximum time to build the model which it is 21.2 sec.

In [5], the study has taken two different algorithms to compare accuracy in datasets. JRip and Hoeffding trees accuracies were similar to each other's but as for the outcomes, JRip model not only gives a prediction, but also provides insight of the problem.

In [6], researchers are studied role of PB2 in influencing influenza A virus virulence in mice. OneR, JRip and PART methods used. Accuracies ranged from 65.0% to 84.4%. PART models were better than random forest models and moreover PART models were better than other rule based learning approaches for estimating contribution to virulence.

[7] is considering performance comparison of rule based classifier: JRip and Decision Table using WEKA (Waikato Environment for Knowledge Analysis) data mining tool on car reviews. As a conclusion JRip is excellent, efficient and accurate model rather than Decision Table for test mode training set.

2. CLASSIFICATION TECHNIQUES *(SINIFLANDIRMA TEKNİKLERİ)*

Classification techniques are important in knowledge mining. Characterization is an information mining strategy that assigns things in a collection to a particular class. Classification models predict categorical class labels; and prediction models predict continuous valued functions [8]. The reason for classification is to accurately capture the target class for each case in the information.

Three classification procedures are taken as benchmarking calculations that are read for the taken wellbeing data set of Hypothyroid. The calculations of the three mentioned procedures are described below, these are the algorithms PART, JRIP, ONER algorithms.

2.1 PART: PART (Rule-based-Classification algorithm) is a different and overcome rule. The computation creates sets of rules, called "decision lists", which are ordered arrays of rules. Another information is contrasted with each standard in the rundown thusly, and the thing is appointed the class of the main coordinator with a rule [9,10].

2.2 JRIP: JRip (Rule-based-Classification algorithm) is one of the basic and most popular algorithms. Classes are examined in growing size, and using incrementally reduced error initial set of rules for the class is generated. JRip proceeds finds a set of rules that cover all members of a class with treating all examples of a particular decision in the training data as that class. After it proceeds to the next class and does the same, repeating this until all classes have been covered [11,12].

2.3. ONER: OneR (Rule-based-Classification algorithm), another way of saying "One Rule", is a basic computation that creates a one-level choice tree. OneR can regularly derive simple but accurate ordering rules from a bundle of events. OneR is likewise able to deal with missing qualities and numerical features, showing adaptability despite straightforwardness. The OneR computation creates a guideline for each quality in the preparation information and then at that point chooses the standard with the base mistake rate as its [13,14].

3. DATABASE STRUCTURES *(VERİ SINIFI YAPILARI)*

The JRIP, OneR and PART classification algorithms are applied on the databases given below following structures.

Databases:	breast_cancer	
Instances:	286	
Attributes:	10	
Sum of Weights	286	
Sr.No	Attributes	Type
1	age	Nominal
2	menopause	Nominal
3	tumor-size	Nominal
4	inv-nodes	Nominal
5	node-caps	Nominal
6	deg-malig	Nominal
7	breast	Nominal
8	breast-quad	Nominal
9	irradiat	Nominal
10	Class	Nominal
Test mode	3-fold cross-validation	

Table 1: Breast cancer dataset.

Databases:	diabetes	
Instances:	768	
Attributes:	9	
Sum of Weights	768	
Sr.No	Attributes	Type
1	preg	Nominal
2	plas	Nominal
3	pres	Nominal
4	skin	Nominal
5	insu	Nominal
6	mass	Nominal
7	pedi	Nominal
8	age	Nominal
9	class	Nominal
Test mode	3-fold cross-validation	

Table 2: Diabetes cancer dataset.

Databases:	iris	
Instances:	150	
Attributes:	5	
Sum of Weights	150	
Sr.No	Attributes	Type
1	sepal_length	Nominal
2	sepal_width	Nominal
3	petal_length	Nominal
4	petal_width	Nominal
5	class	Nominal
Test mode	3-fold cross-validation	

Table 3: Iris dataset.

3. PARAMETERS FOR MEASURING PERFORMANCE OF CLASSIFICATION TECHNIQUES *(SINIFLANDIRMA TEKNİKLERİNİN ÖLÇÜMÜ İÇİN PARAMETRELER)*

Accuracy

The simplest intuitive performance metric is accuracy, which is just the ratio of properly predicted observations to all observations. One would believe that if our model is accurate, it is the best.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total} \qquad (1)$$

Precision

The ratio of accurately predicted positive observations to total expected positive observations is known as precision.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (2)$$

Recall

The ratio of accurately predicted positive observations to all observations in the actual class is known as recall.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (3)$$

F1 Score

The weighted average of Precision and Recall is the F1 Score. As a result, this score considers both false positives and false negatives. Although it is not as intuitive as accuracy, F1 is frequently more useful than accuracy, especially if the class distribution is unequal.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (4)$$

4. RESULTS *(SONUÇLAR)*

Testing results are evaluated from the iris dataset, diabetes dataset and breast cancer dataset. Data mining tool Weka Explorer is used for testing and taking results.

Various rule-based algorithms used with three different sample datasets to determine the best classifier rule. Comparing the various algorithm techniques can be made to find out the better mining classification algorithm.

	breast cancer	diabetes	iris
Accuracy	66.4	72.5	94.6
Precision	0,653	0,730	0,948
Recall	0,664	0,725	0,947
F-Measure	0,658	0,727	0,947
ROC Area	0,599	0,781	0,953

Table 4: PART Rule

	breast cancer	**diabetes**	**iris**
Accuracy	67.1	70.9	94.0
Precision	0,645	0,697	0,940
Recall	0,671	0,710	0,940
F-Measure	0,654	0,695	0,940
ROC Area	0,566	0,646	0,955

Table 5: OneR Rule

	breast cancer	**diabetes**	**iris**
Accuracy	70.9	74.8	95.3
Precision	0,688	0,741	0,954
Recall	0,710	0,749	0,953
F-Measure	0,693	0,740	0,953
ROC Area	0,598	0,717	0,966

Table 6: JRIP Rule

	breast cancer	**diabetes**	**iris**
JRIP	70,9	74,8	95,3
OneR	67,1	70,9	94
PART	66,4	72,5	94,6

Table 7: Rule Algorithm Comparison

Visualization graph of distributions and algorithm accuracy, precision, recall, f-measure, roc area given with tables and in Fig1 and Fig 2. It gives insight about algorithm prediction is well or not. The below graphs shows the superiority of algorithms to each other.

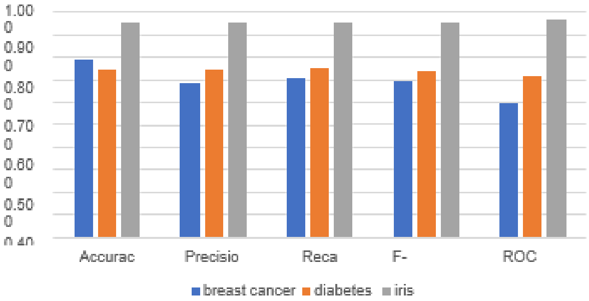

Fig 1: Dataset Result Comparison

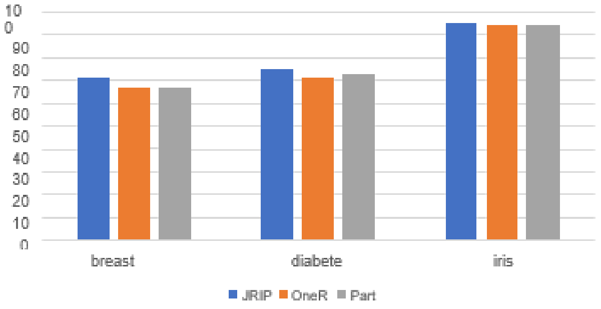

Fig 2: Algorithm Accuracy Comparison

5. CONCLUSION *(SONUÇLAR)*

As a result of taken datasets from iris, diabetes and breast cancer, the parameters values for classification techniques JRIP, OneR, PART rule can be compared. The resultants table reveals that the best results for the specified rules are seem to be JRIP. In this rule based classifier training cross validation (n=3) is used. The results in the table also show accuracy, precision, recall, f-measure, roc-area value are also seeming to be better at JRip algorithm rather than the others.

For an another vision, excluding JRIP algorithm; for low instanced datasets PART algorithm accuracy is more preferable over OneR algorithm. We may conclude sensitivity is a key for deciding the better algorithm due to having more data in dataset is more accurate for determining the true algorithm. Thus OneR algorithm can be more useful with more instanced datasets.

Publication Ethics

This paper was presented as an oral paper at the International Conference on Engineering Technologies 2021 (ICENTE'21), Konya, Turkiye.

REFERENCES *(KAYNAKLAR)*

[1] A. Twın ,"Data Mining, How Data Mining Works", Investopedia, September 17 2021.

[2] V.Parsania, N.Bhalodiya and N Jani Applying naïve bayes , bayesnet , part , hrip , andoner algoritms on hypothyroid database for comparative analysis.(2014)

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations Volume 11, Issue 1.

[4] A Comparative Study of Classification Algorithm using Data Mining: Crime and Accidents in Denver City the USA. Authors: Amit Gupta,Ali Syed, Azeem Mohammad, Malka N.Halgamuge. IJACSA) Vol. 7, No. 7,(2016)

[5] AlMuhaideb S, Alswailem O, Alsubaie N, Ferwna I, Alnajem A. Prediction of hospital no-show appointments through artificial intelligence algorithms. Ann Saudi Med 2019; 39(6): 373-381 DOI: 10.5144/0256-4947.2019.373

[6] Ivan, F.X., Kwoh, C.K. Rule-based meta-analysis reveals the major role of PB2 in influencing influenza A virus virulence in mice. *BMC Genomics* 20, 973 (2019). https://doi.org/10.1186/s12864-019-6295-8

[7] Dr. S. R. Kalmegh and Mr. S. A. Ghogare, “Performance Comparison of Rule Based Classifier: Jrip and Decisiontable Using Weka Data Mining Tool on Car Reviews”, *IEJRD - International Multidisciplinary Journal*, vol. 4, no. 5, p. 5, May 2019.

[8]Tutorials Point , Data Mining: Classification & Prediction. Retrieved from https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

[9] Aubaid, Asmaa M., and Alok Mishra. 2020 “A Rule-Based Approach to Embedding Techniques for Text Document Classification” Applied Sciences 10, no. 11: 4009.(2020) https://doi.org/10.3390/app10114009

[10] “PART Algorithm” Retrieved from : https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/PART.html, (2013)

[11] Anil RAJPUT, Ramesh Prasad Aharwal, Meghna Dubey,S.P. Saxena(2011) “J48 and JRIP Rules for E-Governance Data” IJCSS-448

[12] “JRIP Algorithm” Retrieved from: https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html, (2013)

[13] Gaya Buddhinath and Damien Derry,”A Simple Enhancement to One Rule Classification” Department of Computer Science & Software Engineering University of Melbourne,Australia,(2006)

[14] “OneR Algorithm” Retrieved from http://www.saedsayad.com/zeror.htm, (2013)

[15] Breast Cancer Wisconsin(Diagnostic) Data Set,UCI Machine Learning Repository,Creator:Dr. William H.Wolberg,University of Wisconsin Hospitals,Madison,Wisconsin,USA (1994)

[16] Diabetes Data Set,UCI Machine Learning Repository,Source:Micheal Kahn,MD,PhD,Washington University,St.Louis,MO,AIM(1994)

[17] Iris Data Set,UCI Machine Learning Repository,Creator:R.A. Fisher (1936)

[18] A probabilistic Interpretation of Precision,Recall and F-Score,with Implication for Evaluation. Authors:Cyril Goutte,Eric Gaussier (LNCS,volüme 3408).ECIR 2005:Advances in Information Retrieval pp 345-359

[19] Towards Data Science (2018) Accuracy , Precision , Recall ,F1. Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

ÖZGEÇMİŞ

Çağrı Dükünlü[*,a]

Çağrı Dükünlü was born in Adana, Turkey, on 9 August 1997. He graduated from the Electrical and Electronics Engineering Departments in Cukurova University, Adana, Turkey , in 2020. In 2021, he attended the MSc programme at the Electrical and Electronics Engineering Department in Cukurova University. His studies related with software engineering and machine learning subject since he graduated. He worked as a research and development engineer at the Koluman Automotive Industry. His current fields of research are machine learning, automotive industry, software systems and autonomous robotics.

Mehmet Uğraş CUMA[b]

He was born in 1982, He received the B.Sc., M.Sc., and Ph.D, degrees in electrical and electronics engineering from Cukurova University, Adana, Turkey, in 2004, 2006, and 2010, respectively. He is currently an Associate Professor with the Electrical and Electronics Engineering Department, Adana Science and Technology University, Adana. His current research interests include power quality, power quality devices, and electrical vehicles.