

CLASSIFICATION OF GALAXIES IN SHAPLEY CONCENTRATION REGION WITH MACHINE LEARNING

Nazlı Deniz ERGÜÇ*, Graduate School of Natural and Applied Sciences , Mugla Sıtkı Kocman University, Mugla, Turkey, ndenizerguc@gmail.com

( <https://orcid.org/0000-0002-7281-8288>)

Nida GÖKÇE NARİN, Department of Statistics, Faculty of Science, Mugla Sıtkı Kocman University, Mugla, Turkey , nidagokce@yahoo.com

( <https://orcid.org/0000-0002-4840-5408>)

Received: 08.04.2019, Accepted: 24.05.2019

*Corresponding author

Research Article

DOI: 10.22531/muglajsci.550814

Abstract

The galaxies, are the systems consisting of stars, gas, dust and dark matter combined with the gravitational force. There are billions of galaxies in the universe. Since the cost of examining each galaxy one by one is high, the classification of the galaxy is an important part of the astronomical data analysis. Galaxies are classified according to their morphologies and spectral properties. Machine learning methods aiming the revealing of hidden pattern within the data set by analyzing the available data, can be used to estimate unidentified natural groups of galaxies. This will save time and cost for both researchers and astronomers. This study has been classified five-variables (Right ascension, Declination, Magnitude, Velocity, and Sigma of Velocity) of 4215 galaxies. Galaxies whose natural groups were determined with IDL were classified by using machine learning algorithms with Weka program. Bayes classifier methods, Naive Bayes and Bayes net, Decision tree methods J48, LMT and Random Forest algorithms, Artificial Neural Networks Multilayer Perceptron and Support vector classifier methods were used. The obtained classification results were compared with the natural groups and the predictive performance of the methods were evaluated.

Keywords: Galaxies Classification, Classification Algorithms, Machine Learning, Shapley Concentration Region

MAKİNE ÖĞRENİMİ İLE SHAPLEY KONSANTRASYON BÖLGESİNDE GALAKSİLERİN SINIFLANDIRILMASI

Özet

Galaksiler, kütle çekim kuvvetiyle bir arada bulunan yıldızlar, gaz, toz ve karanlık maddeden meydana gelen sistemlerdir. Evrende milyarlarca galaksi bulunmaktadır. Her bir galaksinin tek tek incelenmesinin maliyeti yüksek olduğundan galaksi sınıflandırması astronomik veri analizinde önemli bir yer tutmaktadır. Galaksiler morfolojilerine ve spektral özelliklerine göre sınıflandırılmaktadır. Veri seti içindeki gizli örüntüyü ortaya çıkarmayı amaçlayan makine öğrenme yöntemleri mevcut veriyi analiz ederek doğal grupları henüz tespit edilmemiş olan galaksilerin hangi gruba ait olduğunu tahmin etmek amacıyla kullanılabilir. Bu da gerek araştırmacılara gerekse astronomlara zaman ve maliyet açısından kazanç sağlayacaktır. Bu çalışma da Shapley Konsantrasyon bölgesindeki 4215 galaksi, 5 değişken (enlem, boylam, parlaklık, hız ve hızdaki sapma) dikkate alınarak sınıflandırılmıştır. IDL programlama ile doğal grupları tespit edilen galaksiler Weka programı ile makine öğrenme algoritmaları kullanılarak sınıflandırılmıştır. Bayes Sınıflandırıcı yöntemlerinden Naive Bayes ve Bayes net, Karar Ağaçları yöntemlerinden J48, LMT ve Random Forest algoritmaları, Yapay Sinir Ağlarından Çok Katmanlı Algılayıcılar ve Destek Vektör sınıflandırıcı yöntemleri kullanılmıştır. Elde edilen sınıflandırma sonuçları doğal gruplarla karşılaştırılmış ve yöntemlerin tahmin performansları değerlendirilmiştir.

Anahtar Kelimeler: Galaksi Sınıflandırması, Sınıflandırma Algoritmaları, Makine Öğrenmesi, Shapley Konsantrasyon Bölgesi

Cite

Ergüç, N.D., Gökçe Narin, N. (2019). "Classification of galaxies in shapley concentration region with machine learning", *Mugla Journal of Science and Technology*, 5(1), 119-126.

1. Introduction

The galaxies are huge systems consisting of stars, gas, dust and dark matter combined with gravity. There are billions of galaxies in the universe. Since the cost of examining of each galaxy one by one is high, the classification of the galaxies has an important place in astronomical data analysis. Galaxies are classified according to their morphologies and spectral characteristics. The galaxy classification was first made by Edwin Hubble in 1926. In this study, galaxies were classified into three main groups: elliptical, spiral and irregular, which are known as the Hubble Scheme and then accepted as a common form of morphological classification [1].

Parallel to the developments in recent years in the technology of celestial bodies, stars and carried out numerous observations of galaxies and celestial bodies are newly identified. By observing the properties of the sky bodies obtained as a result of observations able them to be classified according to their similarities. The classification of a galaxy is very important in terms of astronomical data analysis. Galaxies are basically classified in two different ways: morphological and spectral classifications. Morphological classification is performed by examining the brightness and density of the galaxies. Spectral classification is based on the stellar populations and emission-line properties of galaxies. In this study, morphological classification is discussed. Large catalogs have been produced which provide more compact information to the researchers by classifying the galaxies determined according to the similarities of the identified galaxies.

There are some studies contained the morphological classification of galaxies in literature. Galaxies were first identified into three main groups, elliptical, spiral, and irregular by Edwin Hubble using the Hubble scheme which is considered as a morphological classification. [1]. The elliptical, spiral and irregular classes defined in the Hubble scheme were studied in four groups including lenticular galaxies [2]. Sandage has worked on the sub classification of previously known as Hubble classes [3]. Lotz et al. [4] investigated the clustering measurements of 148 bright Hubble galaxies. They showed that the Gini coefficient of bright-core galaxies is high, the galaxies with multiple cores and the galaxies with bright tidal tails have a second-order relative distribution $M(20)$. Dressler [5] showed that galaxies were clustered according to galaxy densities rather than distances to the cluster center. He stated that spiral and regular galaxies clustered irregularly according to density and this situation was caused by the relationship between morphology and local galaxy density.

Kasivajhula et al. [6] studied 119 astronomical images and examined the performance of classification algorithms according to the classification labels from the Zolt Free catalog. As a result, they showed that the Random Forest algorithm performance was higher.

Mariben et al. [7] examined 152 galaxies in 24 astronomical layers from the INAOE dataset, which contain astronomical strata, obtained from the Schmidt camera over a 50-year period. However, due to the insufficient numbers, they produced artificial samples with the transformations obtained from the original observations. That work has shown that the Random Forest algorithm performs better than the Naive Bayes algorithm and contributes significantly to the classification performance of artificial samples. Miller and Coe [8] made a morphological classification with 98% accuracy using Self Organizing Map to distinguish stars and galaxies. In that study, Bailin and Harris [9] examined the data of the Sloan Digital Sky Survey (SDSS), which includes a large number of sky studies, and identified three types of galaxies. They identified that the galaxies defined as an early type are red, high density and round shaped, whereas the galaxies defined as late-type are low density, disc form and blue. The results were consistent with the data in the Millennium Galaxy Catalog.

Gauci et al. [10] used decision tree learning algorithms and fuzzy inference systems to distinguish between galaxy types and galactic objects. The results were compared with the Galaxy Zoo catalog and the Random Forest method has the best match. In the study by Gauthier et al. [11], galaxies were classified as spiral, elliptical, round, disc or other and random forest method provided the best performance (67%). In addition, a regression model was used to estimate the galaxy classes. Dobrycheva et al. [12], used color indices and classified the galaxies into three groups such as elliptic, 98% for spiral, 88% for spiral and 57% for irregularity. It has been found that the Random Forest method provides the highest accuracy as a result of classification using Naive Bayes, Random Forest and Support Vector Machine methods. Remya and Mohan [13] used a convolutional neural network for galaxy classification. The network was trained with data from the Galaxy Zoo Project and the galaxy morphology was estimated directly from the raw pixel data. Selim et al. [14] in the Zolt Frei catalog [15], a set of 113 images and a test set of 20 images were used to classify galaxies and they showed that galaxies could be automatically classified with a 93% accuracy relative to the classical classification. Goderya and Lolling [16] have shown that an automatic galaxy classification based on shape properties and an artificial neural network can be developed for geometric shape classification. Abell [17] has cataloged the galaxies in the Northern Hemisphere according to the Richness, Density, Distance Galactic Latitude criteria. Abell [18] also included in this catalog the 1,361 clusters in the Southern Hemisphere. Driver et al. [19] cataloged the galaxies according to intensity and brightness.

Machine learning (ML) methods aimed at revealing the hidden pattern in the data set by analyzing the available data can be used to estimate the group of galaxies whose natural groups have not yet been determined. This will

save time and cost for both researchers and astronomers. For this purpose, 4215 galaxies in the Shapley concentration region [20] were classified according to the five variables (Right ascension, Declination, Magnitude, Velocity, and Sigma of Velocity). Galaxies with natural groups IDL were classified by using machine learning algorithms with Weka program. Bayesian classification methods, Naive Bayes and Bayes net, Decision tree methods J48, LMT, and Random Forest algorithms, Artificial Neural Networks Multilayer Perceptron and Support vector classifier methods were used. The obtained classification results were compared with the natural groups. The comparison of predictive performances of the methods were also made.

2. Material and Method

In this study, the galaxy observations of the Shapley concentration area [20] were used. The data were obtained from a 24-inch Bruce telescope in the Bloemfontein in 1930. The galaxies in the Shapley concentration region are in 8 different groups in the catalog created by Abell [17, 18]. These groups were accessed from the Vizier database [21].

2.1. Machine Learning

ML is a general name given to the approaches producing their own information by extracting patterns from raw data with the help of intelligent systems. ML makes possible to solve the real-world problem by computers. A simple ML algorithm such as Support Vector Machines and Naive Bayes, can predict whether an observation belongs to a predefined class. The performance of ML algorithms depends on the representation of the data. Each piece of information involved in the representation of data is known as a feature. Many real-world problems can be solved by designing the right set of features to reveal these problems and then using a simple ML algorithm that processes these features. It is nevertheless difficult to know what features should be extracted for many tasks [22].

In this paper, we use eight classification algorithms Bayes Nets, Naive Bayes, Multilayer Perceptron, Support Vector Machines, J48, Logistic Model Tree and Random Forest of the Weka interface. When comparing the performance of all algorithms we found the Logistic Model Tree is a better algorithm in most of the cases in terms of estimation to galaxies natural groups.

2.2. Classification

Classification is a technique to predict the class of given data points. Classes are called as labels or categories. Classification predictive modeling is an approximation function from independent variables to the categorical dependent variable. There is a lot of classification algorithms in machine learning such as neural networks, support vector machines, decision tree, Bayesian classifiers [22].

2.3. Neural Networks

Neural networks are the basis of machine learning by modeling the learning principle of the human brain. The architecture of the neural networks generally consists of one input layer, one or more hidden layer, and one output layer. Each layer may have a different number of neurons. Network architecture may vary depending on the problem of interest. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples [23]. Multilayer Perceptron (MLP) network is used to solve the classification problems. For a classification problem, the number of neurons in the input layer is equal to the number of variables, and the number of neurons in the output layer is up to the number of classes. MLP uses the multilayer feed-forward neural network approach to classify data [24, 25]. An example of a multilayer feed-forward network is shown in Figure 1.

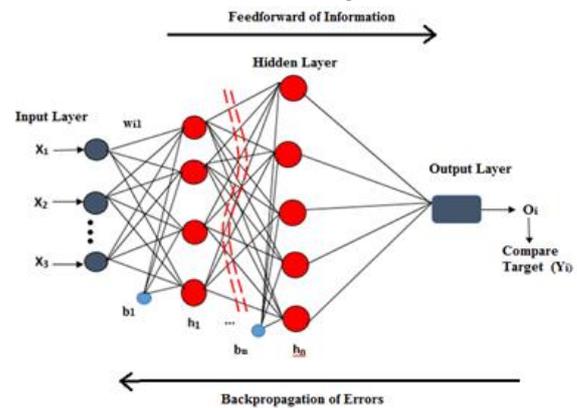


Figure 1. Multilayer Network Model

2.4. Support Vector Machines

Support Vector Machines (SVM) is a machine learning method in which data is divided into two or more classes with the help of decision planes that define decision boundaries. A decision plane separates a set of objects into their respective groups with a line or plane. Most classification tasks are needed in complex structures in order to make an optimal separation. SVM handles an iterative training algorithm, to construct an optimal hyperplane which is used to minimize an error function. SVM has kernel functions which are converted non-separable problem to separable problem such as linear, polynomial, radial basis function, sigmoid and etc. These functions are the most useful in the non-linear partition problem. Thanks to them SVM can perform the process to divide the extremely complex data based on the labels [26]. An example of a multilayer feed-forward network is shown in Figure 2.

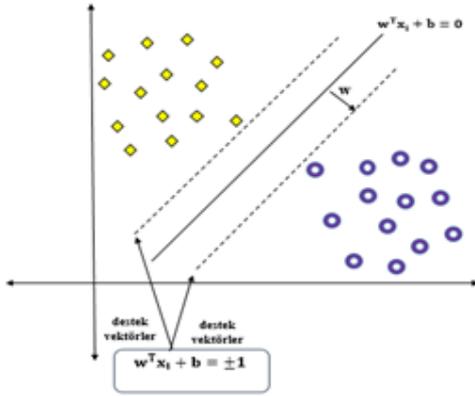


Figure 2. Support Vector Machine

2.5. Decision Tree

Decision Trees (DT) is one of the supervised learning methods and it uses to make inferences about the occurrence of consecutive events. DT aims to divide a data set containing a large number of observations into smaller sets using a set of rules. It consists of branches starting from a root and descending downwards. Both categorical and numerical data can be used in the classification. DTs compose of three basic components: root node, internal node, and leaf node. The internal nodes represent a condition based on which the tree divides into branches/edges. The leaf nodes represent a decision. In real data sets with a lot of features, DTs can produce simple and fast solutions. DT makes a variable selection or selection of features. An important advantage of DTs is that non-linear relationships between parameters do not affect tree performance. In this study, the main DT methods J48, LMT, and Random Forest algorithms were used for bone age estimation [27].

2.5.1. J48

J48 algorithm is a popular ML algorithm adapted from J.R. Quinlan C4.5 algorithm. The aim of this algorithm is to create the DT which provides the highest knowledge and makes the least number of branches. For this purpose, the algorithm calculates firstly the class entropy value [28, 29]. Then it calculates the information value of the variables for each class and the information gain of each variable. The variable with the highest information gain is determined as the root node. When the variables are continuous, the Gini index is used and the division is performed with the variable having the lowest Gini index that disrupts the continuity. For detail information about the J48 algorithm can be seen in the papers, Quinlan [27].

2.5.2. LMT

The logistic regression tree (LMT) is a supervised machine learning method which is obtained by combining the logistic regression and decision tree. A logistic model tree basically consists of a standard decision tree structure with logistic regression functions on the leaves. The logit Boost algorithm is used to

generate a logistic regression model from each node of the tree [30].

2.5.3. Random Forest

Random forest is a decision forest composed of multiple decision trees. Random Forest searches for the best feature in a random property subset, rather than searching for the most important feature when dividing a node. It usually has better modeling performance as it adds additional randomness to the model while growing trees [31].

2.6. Bayesian Classifiers

Bayes classification is a probability-based classification approach based on Bayes' theorem. A Bayesian classifier estimates the probability of class membership that a given tuple belongs to a particular class. Bayes Theorem is given as follow.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, P(B) > 0 \quad (1)$$

$P(A)$ and $P(B)$ are marginal probability of events A and B respectively. $P(A|B)$ is the conditional probability of A given B, $P(B|A)$ is the conditional probability of B given A. Bayesian networks and Naive Bayes classifiers are the most popular classifiers used in machine learning.

2.6.1. Naive Bayes (NB)

The Naive Bayes classifier calculates the odds for each variable separately and selects the most likely result from them. It assumes that the variables are independent. Even in cases where the assumption is clearly false, the Naive Bayesian classifier can give good results [32].

2.6.2. Bayesian Networks (BN)

Bayesian networks are included in network models based on conditional probability. Bayesian networks show the conditional probability relations between variables and the common probability distribution between variables on a network structure [33].

3. Performance Criteria

3.1. Correctly Classified Instances Rate (CCIR)

CCIR is a measure of the accuracy of the model used in the evaluation of methods. It is obtained by calculating the ratio of the observations classified in the model to all observations.

3.2. Kappa Statistics (K)

The Kappa statistic is the performance criterion the for used to measure compliance of two or more observations. It takes values between -1 and +1. The fact that the Kappa statistic is close to 1 shows that the model estimation and the actual class of observation are strong. Calculation of kappa can be performed according to Equation. 2.

$$K = \frac{P_T - P_e}{1 - P_e} \quad (2)$$

P_T is the total likelihood of observed compliance; P_e , denotes the probability of random of observations; K shows the Kappa statistic [34].

3.3. Confusion Matrix

The confusion matrix is a technique that summarizes the performance of a classification algorithm. If there are unequal numbers in each class or more than two classes in the dataset, the classification accuracy may be misleading. In this case, the performance measurements obtained using a confusion matrix shown in Table 1 give information about how accurate the classification model is and what kind of errors it makes [35].

Table 1. Confusion Matrix

		True Class	
		Positive	Negative
Prediction Class	Positive	TP	FP
	Negative	FN	TN

Accuracy (Acc) is a performance criterion for the correct estimation rate of classifiers. It is obtained by the ratio of correctly classified observations to all observations. *Recall (Rec)* is the ratio of the number of correctly classified positive samples to the total number of positive samples. *Precision (Pre)* is the ratio of the total number of positively classified positive samples to the total number of predicted positive samples [35].

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$Rec = \frac{TP}{TP + FN} \quad (4)$$

$$Pre = \frac{TP}{TP + FP} \quad (5)$$

4. Case Study

In this study, 4215 galaxies in the Shapley Concentration region and 5 variables (Right ascension, Declination, Magnitude, Velocity, and Sigma of Velocity) were studied. The galaxies in the Shapley concentration region are located in 8 different clusters in the Abell catalog. Using the IDL program, was determine which galaxies were found in which Abell group. Galaxies detected in natural groups Bayes Net and Naive Bayes are classified by Support Vector Machine, J48, LMT, and Random Forest methods and Artificial Neural Networks method. . A k-fold cross-validation test method is used to evaluate classification models. It separates the data set as a test set and training set according to a k number. This method provides each piece to be used for both training and testing. After the model is trained, the classifier is evaluated to verify the reliability of the model. With this method, the data source is divided into ten parts and each

section is once a set of tests, the other is used as a set of nine parts training [22]. The performances of the classification methods were evaluated according to their compliance with the Abell catalog. The performances of the classification methods used were evaluated according to their compliance with the Abell catalog. The classification method that best matches the Abell catalog has been determined.

Table 2 shows a part of the data set converted to Weka format. The 4215 galaxies in the Shapley Concentration region are distributed among 8 groups in the Abell catalog [17, 18] The distribution of galaxy numbers according to groups is given graphically in Figure 3.

Table 2. Weka format of data set

```
@relation data
@attribute R.A numeric
@attribute DEC numeric
@attribute MAG numeric
@attribute V numeric
@attribute SIGV numeric
@attribute class {1633,1648,1736,1771,1802,1816,1846,1857}
@data
193.03,-32.85,15.23,15056,81,1648
193.04,-28.54,17.22,16995,32,1633
193.04,-28.23,17.29,21211,81,1633
193.05,-28.34,18.2,29812,37,1633
193.06,-29.84,12.55,2930,38,1633
193.06,-30.33,17.62,16042,104,1633
193.06,-30.12,16.87,17264,36,1633
193.08,-27.89,16,16975,57,1633
193.08,-30.51,17.9,31572,68,1633
193.13,-28.39,17.71,14806,36,1633
193.13,-28.95,15.8,16559,118,1633
193.14,-28.45,17.9,17032,56,1633
193.14,-30.48,18.89,16618,15,1633
193.14,-30.25,17.58,27457,15,1633
193.15,-31.27,13.9,16404,157,1648
193.15,-31.89,15.18,3543,138,1648
```

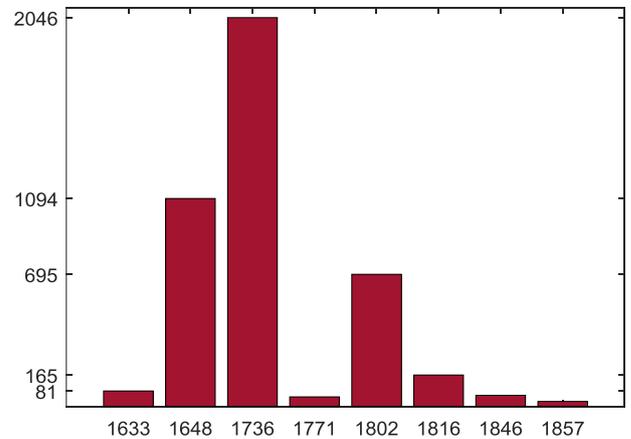


Figure 3. Number of galaxy observations in Abell groups

5. Results and Discussion

In this paper, Bayes net, Naive Bayes, Multilayer Perceptron, Support Vector, J48, LMT, and Random Forest classification methods were applied. These classification methods were applied using the Weka platform. The results of the methods according to the performance criteria are given in Table 3. In the Table, 4 different methods are compared according to 5 considered performance criteria.

Table3. Performance Comparison of Classification Method

Methods	CCIR	K	Acc	Rec	Pre
Abell	100,0	1	1	1	1
Bayes Net	94.52	0.92	0,986	0,70	0,79
Naive Bayes	93.29	0.90	0,990	0,78	0,84
MLP	96.37	0.94	0,993	0,71	0,90
SMO	92.36	0.88	0,988	0,48	0,54
J48	97.51	0.96	0,993	0,88	0,90
LMT	99.45	0.99	0,998	0,96	0,96
Random Forest	98.15	0.97	0,995	0,89	0,95

The best correct classification rate is closest to 100%. Correct classification rates are generally high for all methods. The Kappa coefficient is a statistic that measures the cohesion between classifiers in classification problems. It is expected to be close to 1. In general, because it takes into account the possibility of coincidence, it is a more stable measure than the correct classification rate. The LMT algorithm, which is one of the decision tree classification methods, has the best correctly classification percentage with rate of 99.45%.

According to the Kappa statistics, the performance of LMT classification was found to be the best, while the predictions with Support Vector Machines and Naive Bayes were found to be low. Accuracy refers to the ratio of accurately classified observations to all observations in the model. A value close to 1 indicates that the observations are mostly classified correctly. When the accuracy values of the methods are examined in Table 9, it is seen that LMT has the highest accuracy rate. Similarly, Bayes Net and SMO classification methods have the lowest accuracy rate. Similar results were obtained when Recall and Sensitivity values were taken into consideration. As a result, LMT has the highest predictive success in detecting natural groups of galaxies according to all performance criteria.

The Friedman test is one of the nonparametric tests used to determine whether more than two dependent samples have different distributions. It is the nonparametric equivalent of two-way analysis of variance [36]. Hypotheses are as follows:

H_0 : There is no significant difference between the classification methods

H_a : There are significant differences between the classification methods.

Friedman test was performed with Matlab.. Results of the Friedman test is given in Table 4. In Table 4, SS is the sum of the squares, df is degree of freedom, MS is the mean square, Chi-sq is the test statistics and p-value indicates the level of significance. The Prob (p-value) that the Friedman test returns is used to cast doubt on the null hypothesis. A sufficiently small P-value indicates that at least one approach is significantly different in sample

median than the others. To determine whether a result is "statistically significant", a critical p-value is chosen by the researcher, generally agreed to be 0.05.

Table 4. Friedman Test Results

Source	SS	df	MS	Chi-sq	Prob > Chi-sq
Columns	202.8	7	28.971	33.8	1.87e-05
Error	7.2	28	0.2571		
Total	210	39			

Since the probe value is 1.87741e-05 in our study, H_0 is rejected. That is, there is a significant difference between the classification methods. For this reason, multiple comparison tests were used to determine which methods differ significantly. The result of the multiple comparison tests is given in Figure 4.

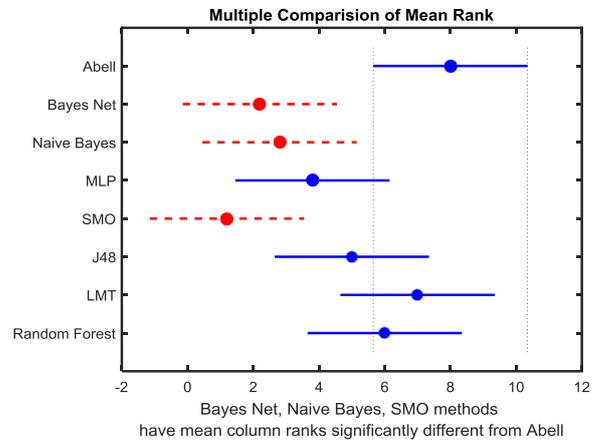


Figure 4. Multiple Comparison of Mean Ranks

In Figure 4, the comparison of the mean rank of the classification methods is given. The performances of the methods are evaluated on the basis of Abell. This means that the success of the methods close to Abell is higher. As a result, the MLP algorithm and decision tree methods used in the machine learning methods have been successful in predicting natural galaxy groups. Naive Bayes, Bayes Net, and SMO methods have shown lower predictive performance in the adaptation of galaxies to natural groups.

6. Conclusion

As a result of this study, it has been shown that the classification models based on machine learning can be used successfully to classify galaxies. Decision tree based models J48, LMT and Random Forest, and MLP method have the highest classification performance to detect natural groups of galaxies in the Abell catalog. The predictive success of Naive Bayes, Bayes Net, and SMO methods is low and there is a significant difference between Abell groups according to multiple comparison test results. In this study, the galaxies in the Shapley region with the right ascension coordinates between (193.03) - (216.03) and declination coordinates (-27.5

- (-37.65) were classified. The success of predicting the natural groups of the galaxies in the Shapley concentration region of the models created by machine learning methods is important for demonstrating that this study can be extended for different regions.

Thanks to technological advances, new galaxies or celestial bodies are discovered almost every day in the depths of the universe. The creation of regional classification models using machine learning methods especially MLP and decision tree models for currently known celestial bodies can provide great advantages to researchers in terms of time and cost in identifying natural groups of newly discovered celestial bodies. As a future work, we aim to develop classification models for galaxy systems in the whole sky catalog (SDSS) based on the findings of this study.

7. Acknowledgement

This paper has been granted by the Mugla Sitki Kocman University Research Projects Coordination Office. Project Grant Number: 18/037 and title, "Shapley Konsanstrasyon Bölgesindeki Galaksilerin İstatistiksel Öğrenme Yöntemleriyle Sınıflandırması".

In addition, we would like to acknowledge the invaluable contribution of Proff. Dr. Aysun AKYÜZ and Accos. Proff. Dr. Nazım AKSAKER, Department of Physics, Faculty of Science and Letters, Çukurova University for their helpful discussion comments

8. References

- [1] Hubble E., "Extra-galactic Nebulae", Contributions from the Mount Wilson Observatory *Astrophysical Journal*, Vol. LXIV, No.324, pp.321-369,1926.
- [2] Vaucouleurs, G. D., "Classification and Morphology of External Galaxies", *Handbuch der Physik*, Vol.11 No.53, pp. 275-310, 1959.
- [3] Sandage, A., "*Hubble Atlas of Galaxies*", Vol. 618, Carnegie Institution of Washington, Washington D.C, 1961.
- [4] Lotz J.M., Primack J, and Madau P., "A New Nonparametric Approach to Galaxy Morphological Classification", *the Astronomical Journal*, Vol.128, No.1, pp. 163-182, 2004.
- [5] Dressler, A., "A Catalog of Morphological Types In 55 Rich Clusters Of Galaxies", *The Astrophysical Journal Supplement Series*, 42, pp. 565-609, 1979.
- [6] Kasivajhula S., Raghavan N. and Shah H., "Morphological Galaxy Classification Using Machine Learning" cs229.stanford.edu, 2007.
- [7] Marin M., "A Hierarchical Model for Morphological Galaxy Classification", *Proceedings of the Twenty-Sixth International FLAIRS Conference*, 2013, Florida, USA.
- [8] Miller A. S. and Coe M. J., "Star/Galaxy Classification Using Kohonen Self-Organizing Maps", *Mon. Not. R. Astron. Soc.* Vol. 279, pp. 293-300, 1996.
- [9] Bailin J. and Harris W.E., "Inclination-Independent Galaxy Classification", *The Astrophysical Journal*, Vol. 681, pp. 225-231, 2008.
- [10] Gauci A., Kristian Zarb Adami K.Z. and Abela J., "Machine Learning for Galaxy Morphology Classification", *Mon. Not. R. Astron. Soc.*, Vol. 000, 1-8, *arXiv preprint arXiv:1005.0390*, 2010.
- [11] Gauthier A., Archa Jain A. and Emil Noordeh E., "*Galaxy Morphology Classification*", Lecture Notes, Stanford University, 16 December,2016.
- [12] D.V. Dobrycheva, I. B. Vavilova,, O. V., Melnyk, and A. A, Elyiv, "*Machine learning technique for morphological classification of galaxies at $z < 0.1$ from the SDSS*", Astronomy & Astrophysics manuscript no. Dobrycheva-EWASS-15 December 27, 2017.
- [13] Remya G. and Mohan A., "Deep Learning Approach for Classifying Galaxies", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6, Issue 4, April 2016.
- [14] Selim, I., Keshk, A. E., & El Shourbugy, B. M. Galaxy image classification using non-negative matrix factorization. *International Journal of Computer Applications*, 137(5), 4-8, 2016.
- [15] Z.Frei and J.E.Gun," A Catalog Of Digital Images Of 113 Nearby Galaxies" *Astrophysics and Space Science*, Volume 269, [Issue 0](#), pp 649-650, 1999.
- [16] Goderya N.S. and Lolling S.M., "Morphological classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational scheme", *Astrophysics and Space Science* 279: 377-387, 2002.
- [17] Abell, G. O., "The Distribution of rich clusters of galaxies", *The Astrophysical Journal Supplement Series*, 3, pp.211-288, 1957.
- [18] Abell, G. O., Corwin, H. G., Jr., Olowin, R. P., "A Catalog of Rich Clusters of Galaxies", *The Astrophysical Journal Supplement Series*, 70, pp.1-138, 1988.
- [19] Driver S.P., Liske J., Cross N. J. G., De Propriis R. and Allen P. D., "The Millennium Galaxy Catalogue: The Space Density and Surface-Brightness Distribution(S) Of Galaxies", *Mon. Not. R. Astron. Soc.* 360, 81-103, 2005.
- [20] http://astrostatistics.psu.edu/datasets/Shapley_galaxy_d_at
- [21] <http://vizier.u-strasbg.fr/viz-bin/VizieR-2>
- [22] James, Gareth, et all. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.
- [23] Haykin, Simon. *Neural networks*. Vol. 2. New York: Prentice hall, 1994.
- [24] Ilin R., Kozma R. and Werbos P. J., "Beyond feedforward models trained by backpropagation: A practical training tool for a more efficient universal approximator." *IEEE Transactions on Neural Networks* 19.6 (2008): 929-937.
- [25] Rumelhart, D. E., Hinton, G. E., & Williams, R. J., "Learning internal representations by error propagation" No. ICS-8506. *California Univ San Diego La Jolla Inst for Cognitive Science*, 1985.

- [26] Cortes C., Vapnik V., 1995, "Support-Vector Networks", *Kluwer Academic Publishers Machine Learning*, Vo. 20, No.3, pp. 273-297
- [27] Quinlan, J. R., "Simplifying decision trees." *International journal of man-machine studies* 27.3 (1987): 221-234.
- [28] Korting, T.S., "C4. 5 Algorithm and Multivariate Decision Trees." *Image Processing Division*, National Institute for Space Research-INPE Sao Jose dos Campos-SP, Brazil 2006.
- [29] Shannon C., "A Mathematical Theory Of Communication", *Bell System Tech. J.* 27: 379-423, 623-656, 1948.
- [30] Landwehr, N., Hall, M., & Frank, E., "Logistic Model Trees," *Machine Learning*, 59, pp.161-205, 2005.
- [31] Breiman L., "Random Forests", *Machine Learning*, Vol.45, No. 1, pp.5-32, 2001.
- [32] Domingos, P., and Pazzani, M., "Beyond independence: Conditions for the optimality of the simple Bayesian classifier", *Machine Learning* 29:103-130. 1997.
- [33] Friedman, N., Geiger, D., & Goldszmidt, M., "Bayesian Network Classifiers." *Machine learning* 29,2-3 131-163, 1997.
- [34] Wampold, B. E., "Kappa As A Measure Of Pattern In Sequential Data", *Quality&Quantity*, 23, 171-187, 1989,
- [35] Godbole S., Sarawagi S., and Chakrabarti S. "Scaling multi-class support vector machines using inter-class confusion." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.
- [36] Friedman M., "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." *J. Am. Stat. Assoc.* 32:675-701, 1937.