VOL.4NO.2 September 2015 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

Comparison of Different Machine Learning Algorithms for National Flags Classification

Muhammed Ali Kutlay
EmineYaman

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15,
Ilidža71210 Sarajevo, Bosnia and Herzegovina

Article Info

Article history:
Received 17 Jul.2015
Received in revised form 17 Aug. 2015

Keywords:
Flags, machine learning, data mining, decision tree, attribute selection, national flag analysis

Abstract

Each country in the world has its own combination of colors, shapes and symbols on their flags. Some of them use an animal figure such as an eagle, some use an object like a boat; some nations prefer religion figures such as a crescent, or a cross. Some questions yet remain and need an answer. What are the factors that determine the flag of a nation? What factors are affecting the color or colors of a national flag? And what are the reasons for existence of symbols on some national flags?In this paper, we worked an analysis on national flags and factors that mostly affects the design of them. In order to find out these factors, we have used feature extraction method, after that we used different machine learning algorithms to predict religion and landmass of the country. We also showed correlations of certain components that are possible to exist on a national flag such as dominant color or colors on a flag, bars or stripes, normal and sacred symbols such as sun, stars, crosses, crescents, and triangles and, finally some specific icons like a boat or an animal figure.This study shows the associations of some characteristics of countries or different nationalities. There are many affected factors and there are very close correlations between these factors. It also includes the classification of national flag data using Multilayer Perceptron, CART and C4.5 algorithms and comparison of these techniques based on accuracy and performance for classification of national flag's features.

1. INTRODUCTION

The national flag is a symbol or emblem of a country, and therefore it represents a country. Each country in the world has its own combination of colors, shapes and symbols, but conventionally almost all national flags are rectangular. Every nation in the world decides their own design of flag. Maybe, this is the reason why every national flag gives information about nation that they are belong to. Certainly, there are many factors playing role on design of a national flag. In this study, we focused on these factors and worked an analysis on different national flags.

The data set contains data that represents uniquely identifiable national flag of 194 countries [Lichman, 2013].

Each national flag has certain colors and symbol. These colors and symbols make a national flag unique and distinguished from others.While some countries use color combinations, other countries use unique symbols to represent their cultures or values. For instance, Slavic countries use the combination of blue, white, and red color in which white color represents god, blue color represents the leader and red color represents citizens. Some of the Western nations use this combination as well. When we talk about African countries, common combination of colors is red, yellow, and green and on the other handmany Arab nations have the combination of red white and black color on their flags. The flags of the Nordic countries have the same design. Thus, the flag is a single-colored background with a horizontal cross on it. The design of a national flag is also influenced by the religious status. Therefore, flags of Muslim majority countries usually have a half moon sign and Christian majority countries have a cross sign [Akhand et al., 2013].

In this study, we have used a set of data donated by Richard S. Forsyth. The data file contains details of various nations and their flags. There are 194 instance of nation in dataset.

This study aims to identify certain patterns on national flags, which is affected by social, cultural, and historical characteristics of nation using data mining tools. We firstly tried to find characteristic features with using feature extraction methods to decrease size of dataset. Then various machine learning methods have been constructed and give analysis how similar socio-cultural, historical and regional featured of countries correlate in national flags with similar colors and signs.

2. LITERATURE REVIEW

According to acquired knowledge, only one study (Akhand, M.A.H., Mahmud, A., Hossain, I., Murase, K.: 2013) have been conducted regarding national flag classification. The study investigates the correlation between national flags features with Religion, Government, and Region of countries using data mining approach and only C4.5 classification algorithm was used to classify features of the data set.

However, there have not been any study conducted regarding prediction of religion and landmass of a nation specifically. Using and comparing multiple data mining techniques such as CART and Multilayer Perceptron together with C4.5 also makes this paper distinguished from other studies.

3. MATERIAL AND METHODS

3.1. SUBJECT AND DATA ACQUISITION

This data set contains details of various nations and their flags. There are 30 different properties about 194 different countries. 10 attributes are numeric-valued and the remainder is either Boolean- or nominal-valued. We have attributes like name of country, landmass, zone, area, population, language, religion and specific information on the flag like combination of colors, symbols like bars, stripes, and position of the symbols on the flag.

3.2. FEATURE EXTRACTION FROM NATIONAL FLAG

Attribute selection (feature selection) is the process of selecting a subset of relevant features for use in model construction[Suman&Thirumagal, 2003]. It is the automatic selection of attributes in your data (such as columns in tabular data) that are most relevant to the predictive modeling problem you are working on. Attribute selection gives access to a wide variety of algorithms and evaluation criteria for identifying the most important attributes in a dataset. Due to the fact that it is possible to combine different search methods with different evaluation criteria, it is possible to configure a wide range of possible candidate techniques. Robustness of the selected attribute set can be validated via a cross-validation-based approach [Hall & Holmes, 2003].

Attribute selection techniques can be categorized according to a number of criteria. One of the most common techniques is called Information Gain Attribute Ranking. Information gain attribute ranking is one of the simplest (and fastest) attribute ranking methods and is often used in text categorization applications where the sheer dimensionality of the data precludes more sophisticated attribute selection techniques [Dumais et al., 1998].

Attribute selection is different from dimensionality reduction. Both methods intend to reduce the number of attributes in a dataset. Dimensionality reduction method do so by creating new combinations of attributes, whereas feature selection methods include and exclude attributes present in the data without changing them. Feature selection is itself useful, but it mostly acts as a filter, muting out features that are not useful in addition to your existing features. Feature selection methods aid you in your mission to create an accurate predictive model. They help you by choosing features that will give you as good or better accuracy whilst requiring less data.

In this study, we used Information gain attribute ranking technique with Ranker method to select relevant attributes and remove redundant and irrelevant attributes for all machine learning algorithms. Before attribute selection process, there were 30 different attributes in our data set. We applied some pre-processing to prune some attributes that are less than minimum support (minsup). We pointed 0.100 as our minsup rate then, we have pruned attributes that are less than our minsup. There were 13 different attributes that have values less than our minsup for religion prediction. Therefore, we have removed attributes that have values less than minsup to have a clear data in out hand. Table 1 describes the most effective features and descriptions of values in the data set:

Table 1.The most effective features of the data set for religion prediction.

Feature Name	Feature Description	Rank
Language	1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others	**1.131**

Landmass	1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania	**1.044**
Zone	Geographic quadrant, based on Greenwich and the Equator; 1=NE, 2=SE, 3=SW, 4=NW	**0.431**
Botright	Color in the bottom-left corner (moving left to decide tie-breaks)	**0.382**
Mainhue	Predominant color in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)	**0.347**
Topleft	Color in the top-left corner (moving right to decide tie-breaks)	**0.317**
Stripes	Number of horizontal stripes in the flag	**0.281**
Colors	Number of different colors in the flag	**0.263**
Population	in round millions	**0.246**
Area	in thousands of square km	**0.210**
Crosses	Number of (upright) crosses	**0.202**
Blue	same for blue	**0.153**
Green	same for green	**0.139**
Saltires	Number of diagonal crosses	**0.132**
Quarters	Number of quartered sections	**0.118**
Crescent	1 if a crescent moon symbol present, else 0	**0.105**

We implemented attribute selection technique to see the most effective factors with respect to the religion class regarding national flags. Techniqueevaluates an individual attribute by measuring the amount of information gained about the class given the attribute. Process showed that language of a nation is the biggest factor for this. Although there seem to be no logical connection between language and religion of the country, there is an indirect relation between these two attributes. For example, most of English speaking countries are catholic whereas most of Arabic speaking countries are Muslim.

In addition, process shows that landmass of the nation is as important as language but there is 0.0867-point rank difference between the most important two factors. We can easily make a connection between landmass and the religion of a country by considering that almost all of the European, North American and South American countries have majority of Christian people whereas African and Asian countries have majority of Muslim, and Buddhist people. Bars on a flag represent many cultural components in a nation which are mostly affected by language and the landmass that country is located in.

Table 2 shows the ranking features for landmass prediction:

Table 2.The most effective features of the data set for landmass prediction.

Feature Name	Feature Description	Rank
Religion	0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others	**1.044**
Language	1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others	**1.002**
Zone	Geographic quadrant, based on Greenwich and the Equator; 1=NE, 2=SE, 3=SW, 4=NW	**0.921**
Mainhue	Predominant color in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)	**0.362**
Botright	Color in the bottom-left corner (moving left to decide tie-breaks)	**0.277**
Topleft	Color in the top-left corner (moving right to decide tie-breaks)	**0.234**
Colors	Number of different colors in the flag	**0.233**
Stripes	Number of horizontal stripes in the flag	**0.202**
Green	0 if green absent, 1 if green present in the flag	**0.173**

Area	in thousands of square km	**0.159**
Population	in round millions	**0.156**
Blue	0 if blue absent, 1 if blue present in the flag	**0.121**
Bars	Number of vertical bars in the flag	**0.109**
Crosses	Number of (upright) crosses	**0.102**

For landmass prediction, 14 important features have been selected according to Information Gain Attribute Ranking method. As we see from Table 2, religion is the most important attribute with 1.044 ranking for landmass prediction.

3.3. CLASSIFICATION METHODS

3.3.1. CART

Classification and regression trees(CART) algorithm was presented by Breiman and it constructs tree's classifications and tree's regressions in 1984. It consists of the classification tree construction which relies on attributes' binary splitting. In addition to that, it also uses Hunt's model. This model is known as decision tree construction and it is applied serially [Breiman et al., 1984]. Gini index splitting measure is used for splitting attribute. A part of training data set is employed in CART in order to perform the pruning [Podgorelec et al., 2002]. Numeric attributes and categorical attributes are used by CART for the construction of decision tree. It also has extra in-built features which handle the missing attributes [Lewis, 2000]. CART is used to analyze regression with the regression trees' support unlike other algorithms that rely on Hunt's algorithm. Specified set of predictor variables over a given period of time is done by the regression analysis feature in order to predict dependent variable [Breiman et al., 1984]. To be able to decide the best splitting point, CART has a lot of single variable splitting standards such as Gini index, Symgini etc. and one multi-variable (linear combinations). In order to decide the best splitting point, each node data is categorized.

3.3.2. C4.5

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. They provide unique capabilities to supplement, complement, and substitute for traditional statistical forms of analysis (such as multiple linear regression), a variety of data mining tools and techniques (such as neural networks) and, recently developed multidimensional forms of reporting and analysis found in the field of business intelligence [Barry deVille, 2006].

There are many types of algorithms for decision trees. These algorithms usually occupy a greedy strategy that grows a decision tree by making a series of locally optimum decisions about which attribute to us for partitioning the data [Tan et al., 2006]. The most common decision three algorithms Microsoft Decision Tree Algorithm, Hunt's algorithm, ID3, Cart and, C4.5. Every algorithm has various types of implementations. In this study, we used C4.5 and CART decision tree algorithm.

C4.5 is one of the most popular algorithms for rule base classification. There are many empirical features in this algorithm such as continuous number categorization, missing value handling, etc. [Mazid et al., 2012]. C4.5 is based on ID3 algorithm that tries to find small or simple decision threes. In implementation, ID3 was not able to implement numeric type data. Despite C4.5 is based on ID3 algorithm, it is able to implement numeric type data too. In first look, it may seem hard to deal and calculate their knowledge acquisition of numeric data. But only work to do is finding appropriate threshold value between numerical value of data.

The C4.5 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained [Ritika& Paul, 2014].

3.3.3. MULTILAYER PERCEPTION

Artificial Neural Networks are the systems; learn correlation between events after given samples using the samples given before and then this system can decide about new samples by using trained data [Oztemel, 2003].

Artificial neural network is a mathematical calculation method which relies on learning and deciding according to its knowledge. It has been inspired by biological nervous system of human brain. The simplest neural network contains a single input layer and an output layer of perceptron which is called as single layer. In the second type, we have other layers between input layer and output layer and this type is called as multilayer perceptron [Yaman,E. 2015]. In the study, multilayer perceptron type was used to classify national flag data set.

Multilayer perceptron was constructed using 70 neurons in the input layer and 40 neurons in one hidden layer. Reason for using less neurons in the hidden layer is to avoid algorithm to converge before it completes learning. As rule of thumb suggests, number of neurons in the hidden layer should be as the size of;

(No of Input Neurons + No of Output Neurons) × 2 / 3

Therefore, it is always better to use half of the input neurons or anything less than the number of input neurons for hidden layer to avoid convergence.Tests which was conducted to find best construction of MLP showed that one hidden layer is sufficient for this non-linear problem. Tests conducted by using two or more hidden layer did not give any better result than as it was with one hidden layer.

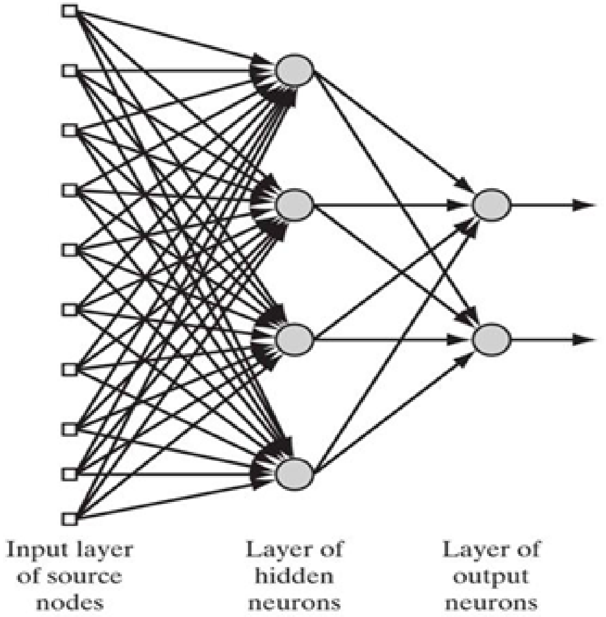

Figure 1. Multilayer feed forward network model [Haykin, 2005].

4. RESULTS AND DISCUSSION

This part presents the experimental findings from national flag data set constructing different machine learning methods. We used Cart (Classification and regression trees), C4.5 and Multilayer Perceptron algorithms to predict religion and landmass.

Table 3. Confusion Matrix of the C4.5 Decision Tree for Religion Prediction.

Classified as	A	B	C	D	E	F	G	H	Total
A=catholic	37	3	0	0	0	0	0	0	40
B=other chr.	3	56	0	0	0	0	1	0	60
C=muslim	0	1	35	0	0	0	0	0	36
D=buddhist	0	0	5	3	0	0	0	0	8
E=hindu	1	0	1	1	0	1	0	0	4
F=ethnic	0	2	2	0	0	23	0	0	27
G=marxist	2	0	0	1	0	0	12	0	15
H=others	0	0	3	0	0	0	1	0	4

It is seen from the confusion matrix table that among 40 catholic majority countries 37 are classified correctly as catholic country and only 3 are misclassified as other Christian. On the other hand, among 60 other Christian countries 56 are classified as other Christian, 3 are classified as catholic and 1 of them classified as Marxist. From total 36 instances of Muslim religion, 35 instances are classified correctly while 1 instance is classified as other Christian. The misclassification is understandable for prunes tree. Pruning reduces tree size to increase to generalize the classification ability. The misclassifications among Muslim and Christian countries indicate that they have some common characteristic principles in the national flags. After Muslim and Christian countries, Buddhist countries are 8 and only 3 are classified as Buddhist and other are classified as Muslim. From 27 instances of ethnic religion, 23 instances are classified as other Christian and 2 instances are classified as Muslim. Since number of other countries like Hindu, Marxist and others is very few in numbers most of them are also classified as Muslim or Christian.

The study which was conducted by Akhand, M.A.H., Mahmud, A., Hossain, I., Murase, K, presents muslim countries with 29 correctly classified instances, budhist countries with 2 and hindu countries with 0 correctly classified instances.

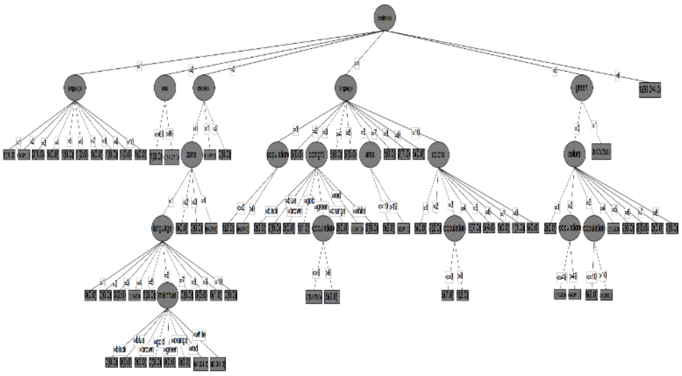

Figure 2.Pruned C4.5 decision tree from national flag data on the basis of religion condition.

Figure 2 shows us for religion prediction, landmass is the most prominent feature in national flag data set. It means if we know landmass of the country, it is easy to predict its religion using landmass attribute.

In this method, we had just 2 misclassifications in Catholics as Muslim and Marxist, 2misclassifications in other Christian as Muslim and Ethnic, 2 misclassifications in Muslim group as other Christian and others, and 1 misclassification in Hindu, Ethnic and Marxist group. It is understandable, because none of the classification method is perfect, these countries and religions have some common characteristics, because of this it is difficult to differentiate from each other.

Second algorithm that we used for religion prediction is CART algorithm. This algorithm gave us the success rate with 88%. Accuracy of this method is higher than C4.5 algorithmbut lower than Multilayer Perceptron method in the religion prediction.

Table 4. Confusion Matrix of the CART Decision Tree for Religion Prediction.

Classified as	A	B	C	D	E	F	G	H	Total

A=catholic	36	3	0	0	0	0	1	0	40
B=other chr	2	57	0	0	0	1	0	0	60
C=muslim	0	3	32	0	0	1	0	0	36
D=Buddhist	0	0	0	7	0	0	0	1	8
E=hindu	1	0	2	0	1	0	0	0	4
F=ethnic	0	1	0	0	1	25	0	0	27
G=Marxist	2	0	0	3	0	0	10	0	15
H=others	0	0	0	0	0	1	0	3	4

Also for this method, there are some misclassifications, but we can say this method is better than C4.5 to predict others religion.

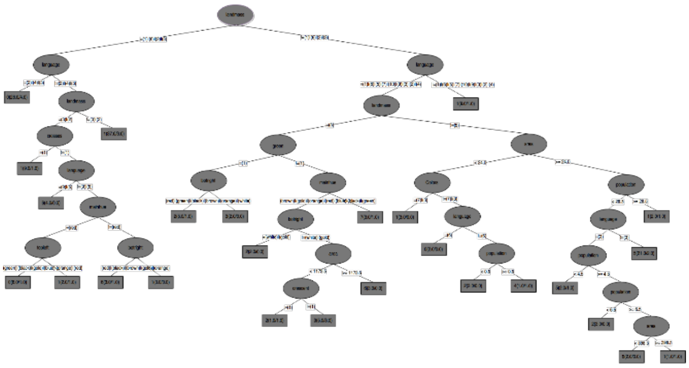

Figure 3. Pruned CART decision tree based on religion attribute.

After second method, we have analyzed our pruned dataset by using a neural network method, named as Multilayer perceptron. By using this algorithm, we got slightly better result than C4.5 and CART algorithm. With this algorithm, we have reached 95.36% classification accuracy.

Neural network model produced by MLP might be very big model. Also for religion prediction, neural network model was a very large visualization. In the figure about a part of neural network model produced by MLP for religion prediction can be seen.

Figure 4 shows just a small part of related model. Visualization of neural networks are not very clear and understandable like decision trees. Because algorithm of neural networks is very complex, it is why it gives better performance for many applications. And as a performance, neural networks are substantially slower than decision trees because of its complex algorithm.

Table 5.Confusion Matrix of the Multilayer Perceptron Algorithm for Religion Prediction.

Classified as	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**Total**
A=catholic	38	0	1	0	0	0	1	0	40
B=other chr.	0	58	1	0	0	1	0	0	60
C=muslim	0	1	34	0	0	0	0	1	36
D=buddhist	0	0	0	8	0	0	0	0	8
E=hindu	0	0	0	0	3	1	0	0	4
F=ethnic	0	0	1	0	0	26	0	0	27
G=marxist	0	0	0	1	0	0	14	0	15
H=others	0	0	0	0	0	0	0	4	4

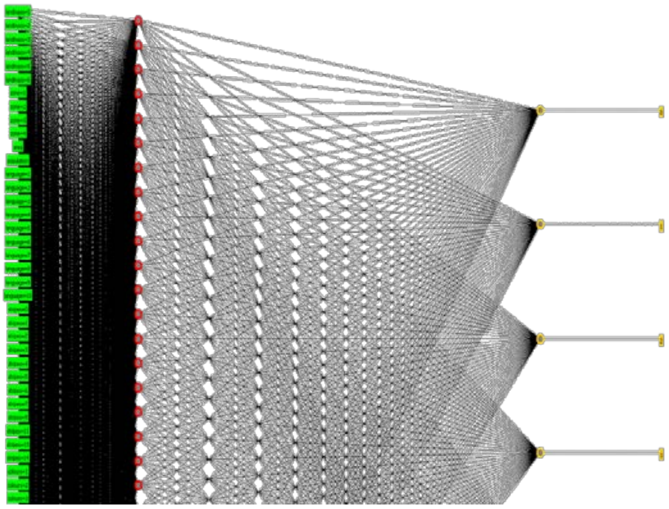

Figure 4.Neural Network Model of MLP for Religion Prediction.

As a second part of our study we tried to predict landmass of countries using same dataset. Firstly, we tried to extract significant features. We founded 14 important features, after that we used same 3 classification algorithm namely C4.5 and CART decision trees and Multilayer perceptron neural network method. Table 6, 7 and 8 shows confusion matrixes of related methods.

There are 6 different landmasses in the data set. Accuracy of C4.5 algorithm for landmass prediction is higher than religion prediction, because we have more specific features for landmass prediction. There is no misclassification in North America in 3 of our methods, and we can estimate, this landmass has more different and characteristics properties than other landmasses. For other landmasses, we took some misclassification for example we can say South America has some similar characteristics with North America, and Europe has some similarities with North America. All of other 3 landmasses have similar features with Europe. For Oceania landmass, C4.5 did not give us good results, because there are just 20 countries in this part and 6 of them are incorrectly classified.

Table 6. Confusion Matrix of the C4.5 Decision Tree for Landmass Prediction.

Classified as	1	2	3	4	5	6	Total
1=N.America	31	0	0	0	0	0	31
2=S.America	2	15	0	0	0	0	17
3=Europe	3	0	32	0	0	0	35
4=Africa	0	0	1	39	10	2	52
5=Asia	0	0	1	0	38	0	39
6=Oceania	0	0	1	5	0	14	20

There are 6 different landmasses in the data set. Accuracy of C4.5 algorithm for landmass prediction is higher than religion prediction, because we have more specific features for landmass prediction. There is no misclassification in North America in 3 of our methods, and we can estimate, this landmass has more different and characteristics properties than other landmasses. For other landmasses, we took some misclassification for example we can say South America has some similar characteristics with North America, and Europe has some similarities with North America. All of other 3 landmasses have similar features with Europe. For Oceania landmass, C4.5 did not give us good results, because there are just 20 countries in this part and 6 of them are incorrectly classified.

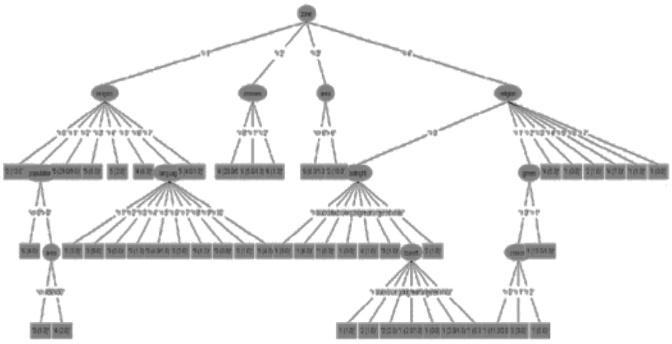

Figure 5.Pruned C4.5 decision tree from national flag data on the basis of landmass condition.

As we see from Figure 4, zone is the most important factor for landmass prediction. It is reasonable because landmass and zone are concerning attributes therefore one of them affects another one easily.

Table 7. Confusion Matrix of the Multilayer Perceptron Algorithm for Landmass Prediction

Classified as	1	2	3	4	5	6	Total
1=N.America	31	0	0	0	0	0	31
2=S.America	0	16	0	1	0	0	17
3=Europe	0	0	35	0	0	0	35
4=Africa	0	0	1	50	1	0	52
5=Asia	0	0	0	0	39	0	39
6=Oceania	0	0	0	0	0	20	20

In the second method, we took the best accuracy results with success rate of 98.96%. We have just 1 misclassification in South America and 2 misclassifications in Africa.

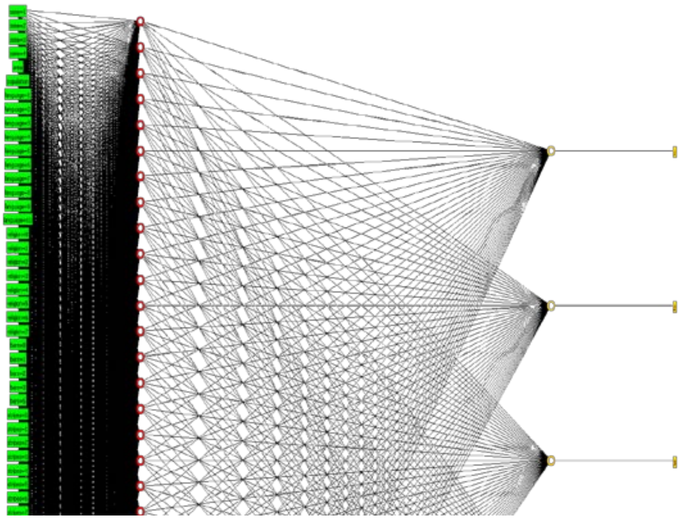

Figure 6.Neural Network Model of MLP for Landmass Prediction.

Also for landmass prediction, we could just put a small part of the related model because of its huge size. But still input layer, output layer and one hidden layer is visible.

Table 8. Confusion Matrix of the CART Decision Tree for Landmass Prediction.

Classified as	1	2	3	4	5	6	Total
1=N.America	31	0	0	0	0	0	31
2=S.America	6	8	1	1	0	1	17
3=Europe	7	0	25	0	0	3	35
4=Africa	1	0	2	33	10	6	52
5=Asia	0	0	1	0	38	0	39
6=Oceania	0	0	1	2	0	17	20

As it is seen in Table 8, CART algorithm did not give us high result, there are a lot of misclassifications in all landmasses except North America.

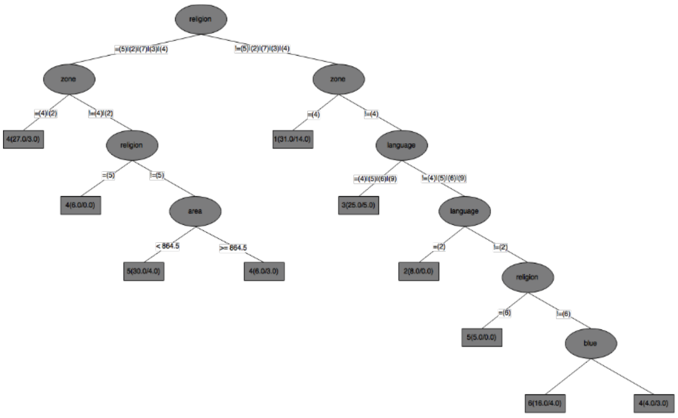

Figure 7. Pruned CART decision tree based on landmass attribute.

According to CART decision tree, religion is the most determining factor for landmass prediction. As a second important factor, we see zone.

5. FINDINGS

Table 9 and 10 includes comparison of success rates' results and performance evaluation results of various machine learning algorithms for religion and landmass prediction. Multilayer perceptron method gave us the best accuracy for these predictions. But when we compare performance of 3 methods with each other, C4.5 algorithm is the fastestclassification method and multilayer perceptron algorithm is substantially slower than other methods. Basically final result shows us that there is a tradeoff between accuracy and performance for these three algorithm. A weak performance rate was not a very big problem for our study, because our dataset was not very huge. But for enormous datasets, it might be a serious problem.

Table 9. Comparison of Success Rates of Algorithms for Religion and Landmass prediction

Algorithm Name	Religion	Landmass
C4.5	85.56%	87.11 %
CART	88.14%	81.44 %
MLP	95.36%	98.96 %

Table 10. Comparison of Performance of Algorithms for Religion and Landmass prediction

Algorithm Name	Religion	Landmass
C4.5	0.02 seconds	0.04 seconds
CART	1.09 seconds	0.74 seconds
MLP	6.3 seconds	6. 45 seconds

6. CONCLUSION

Today, data mining is a really helpful process for plenty of sectors. As helpful as it is, it may be risky too. Therefore, accuracy is really important for data mining processes. Choosing right method and choosing right algorithm for a method is very important. Day by day, number of data mining tools is increasing. This also rises another point which is choosing right software.

A National Flag not only a unique symbol of country but its colors and signs conceive the histories, socio-cultures, ideals of countries. This study investigates the correlation between national flag features with religion and landmass of countries. Based on the results of the classification, Multiple Perceptron have better accuracy than C4,5 and CART.

In the classification, total of 30 different attribute obtained from 194 countries were analyzed. C4.5 decision tree correctly classified 166 instances, 85.56%, incorrectly classified 28 instances, 14.43% for religion prediction. Multilayer Perceptron method correctly classified 185 instances, 95.36%, incorrectly classified 9 instances, 4.63% for religion prediction. CART decision tree correctly classified 171 instances, 88.14%, incorrectly classified 23 instances, 11.85% for religion prediction.

C4.5 decision tree correctly classified 169 instances, 87.11%, incorrectly classified 25 instances, 12.89% for landmass prediction. Multilayer Perceptron method correctly classified 191 instances, 98.45%, incorrectly classified 3 instances 1.54for landmass prediction. CART decision tree correctly classified 152 instances, 78.35%, incorrectly classified 42 instances, 21.64% for landmass prediction.

Total accuracy of Multilayer Perceptron algorithm is 95.36 %, total accuracy of CART is 88.14% and the total accuracy of C4.5 Decision Tree is 85.56% for religion prediction. Total accuracy of Multilayer Perceptron algorithm is 98.96%, total accuracy of CART is 81.44% and the total accuracy of C4.5 Decision Tree is 87.11% for landmass prediction.

Although the Multilayer Perceptron takes relatively more time for learning than C4.5 and CART, accuracy is considerably higher than both algorithms. It is a fact that both Decision Trees and Neural Networks has advantages and disadvantages, yet the recent researches are encouraging the construction of a Hybrid algorithm which circumscribe the advantages of both algorithms.

As a result, this study shows, there are many affected factors and there are very close correlations between these factors when we check some characteristics of countries or different nationalities.It also includes the classification of national flag data using Multilayer Perceptron, CART and C4.5 algorithms and comparison of these techniques based on accuracy and performance for classification of religion and landmass of a country using NFs. We cannot always see these relationships between various factors easily because of huge size of data. With the help of machine learning algorithms, we can clarify associations of properties for different countries. This research can throw a new light on sociological researches.

REFERENCES

AKHAND, M.A.H., MAHMUD, A., HOSSAIN, I., MURASE, K., “Knowledge Discovery from National Flag through Data Mining Approach”, International Journal of Knowledge Engineering and Research, Vol 2 Issue 4 April 2013 ISSN 2319-832X.

BARRY DEVILLE, (2006). “Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner”, SAS Institute Inc. ISBN-13:978-1-59047-567-6.

BREIMAN, L., FRIEDMAN, J.H., OLSEN, R.A., STONE, C.J., “Classification and regression trees”, Wadsworth, USA, 1984.

DUMAIS, S., PLATT, J., HECKERMAN, D., SAHAMI, M., “Inductive learning algorithms and representations for text categorization”. In Proceedings of the International Conference on Information and Knowledge Management, (P-148, 155), 1998.

HALL, M.A., HOLMES, G., “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining”, IEEE Transactions on Knowledge and Data Engineering, VOL. 15, NO.3, May/June 2003.

HAYKIN, S., “Neural Network and Learning Machines”, New Jersey: Prentice Hall, ISBN-13: 978-0131471399, 2005.

LEWIS, R.J., “An Introduction to Classification and Regression Tree (CART) Analysis”, Francisco: 2000 Annual Meeting of the Society For Academic Emergency Medicine, 2000.

LICHMAN, M., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], 2013.

MAZID, M., SHAWKAT, A., KEVIN S TICKLE, K., “Improved C4.5 Algorithm for Rule Based Classification”, Recent Advances In Artificial Intelligence, Knowledge Engineering and Data Bases, ISSN: 1790-5109.

OZTEMEL, E., “YapaySinirAğları”, Istanbul: PapatyaYayıncılık, ISBN- 9756797396, 2003.

PODGORELEC, V., KOKOL, P., STIGLIC, B., ROZMAN, I. "Decision trees: an overview and their use in medicine." Journal of Medical Systems Kluwer Academic/Plenum Press 445-463, 2002.

RITIKA, S., PAUL, A., “Prediction of Blood Donors: Population Using Data Mining Classification Technique”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014, ISSN: 2277 128X.

SUMAN K., THIRUMAGAL S., “Feature Subset Selection with Fast Algorithm Implementation” International Journal of Computer Trends and Technology (IJCTT), Volume 6, Number 1, December 2013.

TAN, P. N., STEINBACH, M., KUMAR, V., “Introduction to Data Mining”, Pearson International Edition, ISBN-13: 978-0321321367, 2006.

YAMAN, E., “Comparison of Different Feature Extraction and Machine Learning Algorithms for EMG Signal Classification”, Dissertation, Vienna University of Technology, Vienna, 2015.