Research on Income Forecasting based on Machine Learning Methods and the Importance of Features

Jinglin Wang[1, *]

[*]Corresponding author: Axon.Wang@outlook.com

[1]Foreign Language School attached to Guangxi Normal University, Guilin, Guangxi, China, 541004

Abstract: In modern society, age has a significant impact on the income distribution of employee. However, little research has focused on the precise impacts of different factors of income and their relevant applications in predicting the person's income. Using 48,842 individuals' income census data from Adult Data Set, this study aims to predict the annual income level of the individual with machine learning approaches based on 13 attributes of the person (age, workclass, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week and native-country) and determine the key factors of the prediction. For income prediction, 32,561 individuals are divided randomly for training the classification model; the Random Forest (RF), K Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR) and Naïve Bayes (NB) algorithm have been adopted. Since the accuracy of RT is greater than 0.9 in this task, Gini Importance is used to measure the relativities between each feature and the topic. Among these 5 methods, the RT and KNN models perform relatively well, with accuracies of 0.97973 and 0.8976 respectively. And the age of the employee shows the highest relativity to his or her possible income with the importance of 0.225.

Keywords: Income; Classification; Gini Importance; Random Forest; KNN

1 Introduction

In past decades, Machine Learning methods have been widely adopted in the field of prediction or classification. Some practical applications, such as precise and personalized advertising for web users, calculating the suitable credit card limit for bank clients, self-career planning, and other tasks including estimating the individual's annual income have faced the challenge of predicting the annual income into different levels with relatively high accuracy and understanding the most important factor of a person's income.

In previous studies, Kibekbaev and Duman (2016) presented income prediction by using regression algorithms based on the in-house data offered by Turkish banks. [8] And Lazar (2004, December) generated the income prediction results by employing principal component analysis and support vector machine approaches. [10] But few studies have compared the performance of various machine learning approaches or focused on the importance of each factor in the prediction task. It is hard to calculate the exact number and value of the person's income, deposit or other capital. So, in this paper, it is concentrated on the precision of classify the individual

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322745

into two levels of income, which are less (or equal) than $50K per year and more than $50K per year, in addition to determine the most relevant feature of the individual to the annual income.

This paper is one of the first attempts to compare the performances of various machine learning models including Random Forest (RF), K Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR) and Naïve Bayes (NB) on the income classification, and determine the most important feature in the prediction, by using the original 14 variables from UCI Adult Dataset. To verify the importance of the feature and confirm that the best performed model is not over-fitting, cross-validation with a new subset is used.

2 Data and pre-processing

2.1 Data

Adult Data Set was extracted from the 1994 American Census bureau database by Kohavi (1996, August). [9] The dataset was posted on the University of California Irvine (UCI) repository [6]. Each instance in the dataset consists of the following 15 attributes(variables to describe an individual): age (instance's age when be surveyed), workclass (the category of employee, like self-employed, local government, etc.), fnlwgt (final weight, the number of people the census believes the entry represents, but not standardized across states), education (instance's education degree), education-num (number of studying years), marital-status (instance's marital status), occupation (the specific job of the instance, Tech-support, Craft-repair, Sales, etc.), relationship (instance's role in his or her family), race (the race of the instance), sex (Female, Male), capital-gain (number of capital gain), capital-loss (number of capital loss), hours-per-week (weekly working hours of the instance), native-country (the native country of the instance) and income (to classify the annual income of the instance <=$50K or >$50K).

For this income prediction task, the "income" attribute of the dataset has been chosen. The origin distribution of "income" is shown in Figure.1a. The heatmap of correlations between the instance's income status and the quantitative variables is shown in Figure.1b.

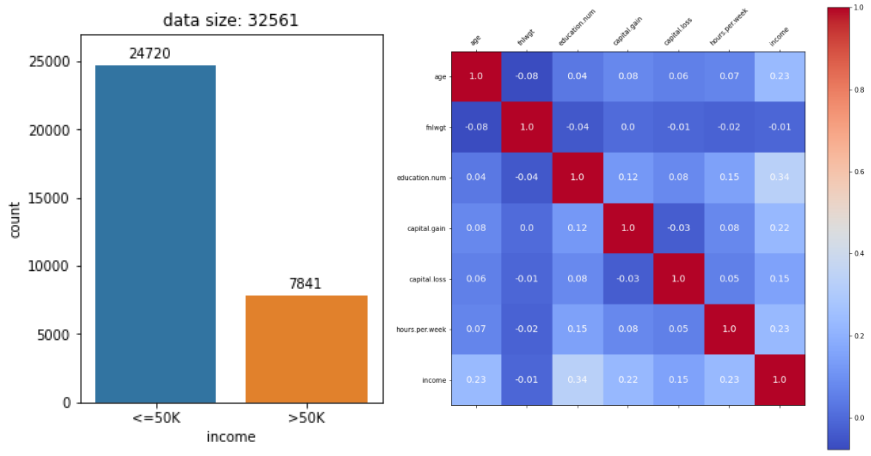

Figure 1. (a) Income distribution of respondents; (b) Pairwise correlation of numerical features and "income".

2.2 Data preprocessing

Since the task is a binary classification question, the output of income can be mapped to 0 and 1 as 0 represents “cannot earn more than $50K” and “can earn more than $50K” respectively.

From Figure.1b it can be argued that the correlations between “fnlwgt” and other numerical features are nearly zero, so “fnlwgt” can be dropped from the dataset.

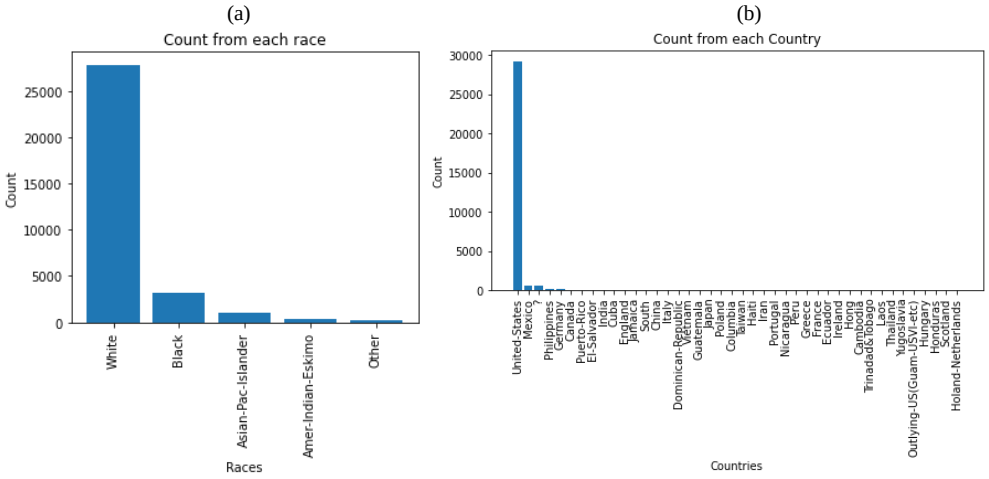

Figure 2. (a) The counts of each kind of race; (b) The counts of each kind of native-country.

As shown in Figure.2 “white” and “United-States” occupy the major part among races and countries respectively, so the rest of the races and countries can be combined together to form two new groups respectively, in order to reduce the imbalance of the dataset.

After dividing the instance’s “race” and “native-country” attributes into one major class and another class, the distributions of “race” and “native-country” attributes compared to “income” are shown in the Figure.3.

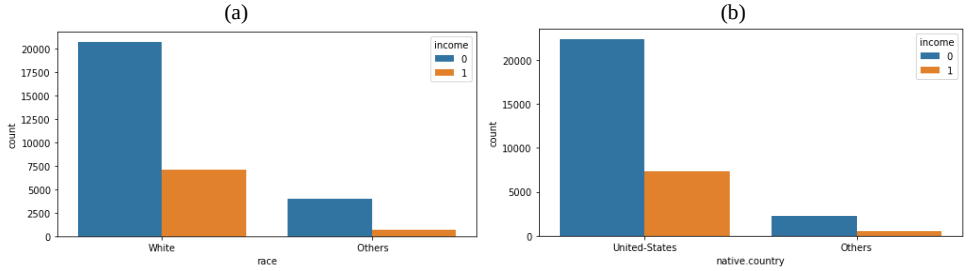

Figure 3. (a) The distribution of the majority and rest of instances’ races with different income results; (b) The distribution of the majority and rest of instances’ native countries with different income results.

For the small number of samples with unknown or missing values which will cause errors in calculation or even potential skewed results, this experiment replaced those missing value with

the mode of each attribute. As can be seen, Table.1 shows which attribute is affected by the missing values and the corresponding modes.

Table 1. The attributes including the missing value, the number of missing values, and the mode of the attribute.

Attribute	Total missing	%	Mode
occupation	1843	5.7	Prof-specialty
workclass	1836	5.6	Private

According to Figure.4(a) the samples who are only K12 educated totally occupy a small part among attribution "education" compared to other values, so these samples with "12th", "11th", "10th", "9th", "7th-8th", "5th-6th", "1st-4th" or "Preschool" of "education" attribution should be combined together to form a new group (called "K12-School"), in order to reduce the imbalance of the dataset. The result of regroup is shown in Figure.4(b).

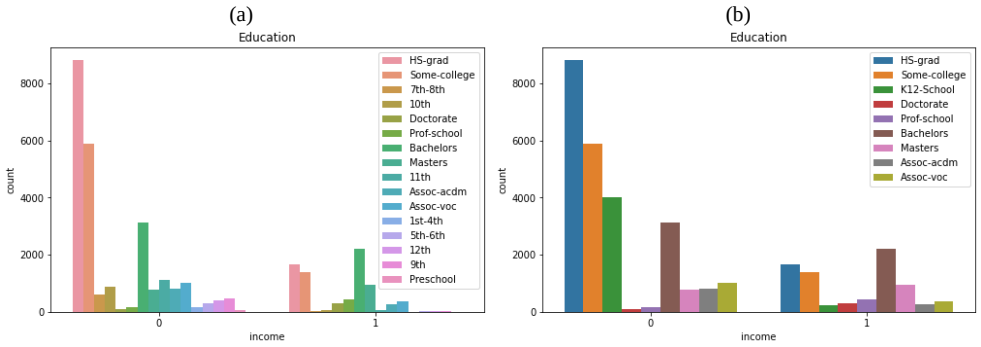

Figure 4. (a) The distribution of each value in the attribution "education"; (b) The distribution of each value in the attribution "education" after data reduction.

2.3 Split and Normalize

To evaluate the performance of classification results, the origin dataset needs to be spited into training and testing subset. In this classification, train_test_split of Scikit-learn [11] is applied to split the dataset into random training and testing subsets with 30% test subset.

Since this study planned to adopt classification algorithms including KNN and SVM that need to calculate the distance between samples (e.g., Euclidean distance), and Logistic Regression that use gradient descent method to find the optimal solution, the verification of the standard deviation of numeric attributes is needed before building those machine learning models.

If the standard deviation of an attribute is considerably larger than the standard deviation of other attributes, then it will dominate the algorithm evaluation, causing the classifier to be unable to learn other attributes as expected, which will lead to slow or even non-convergence of the final model, so it is necessary to normalize the data of such attributes as shown in Table.2.

Table 2. Four numeric attributes of the train dataset and their counts, means, standard deviations, and the interval of the values.

Attribute	age	capital.gain	capital.loss	hours.per.week
Count	22792	22792	22792	22792
Mean	38.614294	1064.423043	87.756581	40.472227
std.	7331.442737	7331.442737	403.660431	12.315027
Min	17	0	0	1
Max	90	99999	4356	99

For the standardization, the mean is removed, and the values are scaled to achieve unit variance. The StandardScaler of Scikit-learn [11] is utilized to normalize each independent column of attributes by computing the standard score on the samples in the training set as:

$$x^* = \frac{x-\mu}{\sigma} \quad (1)$$

where μ is the mean of the samples in the training set, and σ is the standard deviation of the samples in the training set.

3 Methods

To build the classification model, five supervised learning algorithms in machine learning have been applied which are Random Forest (RF), K Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression and Naïve Bayes algorithm.

3.1 Random Forest (RF)

As an emerging and highly flexible machine learning algorithm, Random Forest (RF) can perform well in many cases and has a wide range of promising applications. Random Forest is an integrated algorithm consisting of decision trees. Furthermore, Random Forest belongs to the Bagging (short for Bootstrap Aggregation) method of integrated learning [1].

According to Safavian & Landgrebe (1991), "in top-down approaches to tree design, sets of classes are successively decomposed into smaller subsets of classes" (p. 664); a Decision Tree is a simple algorithm that matches human intuitive thinking based on if-then-else rules. [12] And a Random Forest further adds random attribute selection to the training process of the decision tree based on the Decision Tree to build the classifier [7]. Specifically, the traditional Decision Tree divides attributes by selecting the optimal attribute among all potential attributes (for example, d attributes) of the current node; while in Random Forest, for each node of each composed decision tree, k attributes are first randomly selected from the potential attributes of that node to form a new subset, and then the optimal attribute is selected from that subset for further division. In this study, for the number of attributes k in each, $k = \log_2 d$.

3.2 K Nearest Neighbor (KNN)

KNN, which refers to K-Nearest Neighbors, is a typical supervised learning algorithm. Factually, the KNN algorithm places the sample to be predicted into the data set and then represents the sample to be predicted by using the K number of samples which are the closest dots around it [5].

KNN is one of the most straightforward and most popular models in the majority of classification problems because it is highly accurate and easy to implement.

3.3 Support Vector Machines (SVM)

Support vector machines (SVM) is a binary classification model that uses a learning strategy of interval maximization. Essentially, it solves for the separating hyperplane that correctly partitions the training data set and maximizes the geometric interval [4].

For a linearly divisible dataset, there are infinitely hyperplanes that can classify the data, but the geometrically spaced maximum separation hyperplane is the unique one.

3.4 Logistic Regression (LR)

Logistic regression is one common machine learning algorithms for binary classification tasks like helping doctors determine malignant tumor. It has a simple design idea, is easy to implement, and performs well in many real-life applications. Logistic regression is a type of model that uses the logistic function to describe a binary variable. Each object in the dataset will be attributed a probability between 0 and 1 by the model.

Logistic regression can be adopted to measure the relationship between the dependent variable (the predicted label) and one or more independent variables (individual's features) by estimating probabilities logistic function.

Since the probability of the prediction result is binarized, the logistic function called the Sigmoid function can be employed, which is an S-shaped curve that maps any real value to a number between 0 and 1. Then, a threshold function is chosen to transform the number between 0 and 1 to 0 or 1.

3.5 Naïve Bayes (NB)

The basis of the idea of Naïve Bayes (NB) or sometimes called plain Bayes is Bayes' theorem which applying strong independence assumptions between the features to classify a given item. That given item is classified based on the calculated probability of each category under the conditions of this item.

3.6 Gini Importance

In order to study which variable of individual has the highest relativity of the instance's possible income; it is needed to estimate the importance of each feature. In this research, Gini Importance or sometimes called Mean Decrease Impurity has been employed to measure the relativities especially in the Random Forest algorithm.

Gini Importance is the mean of a variable's total decrease in node impurity (Gini impurity is defined as: $\sum_{i=1}^{C} f_i(1 - f_i)$, where f_i is the frequency of label i at a node and C is the number

of unique labels), weighted by the proportion of samples reaching that node in each individual decision tree in the random forest [2] This is an effective measure of how important a variable is in estimating the value of the target variable across all the trees that make up the forest.

4 Results and analysis

4.1 Income prediction on test set

This study has trained the classifiers using the training subset and verified the classifiers with the test subset. The 5 machine learning algorithms all perform well in classifying income into two levels. The ACC of Random Forest model is immensely high and achieved 97.97736% for the classification.

Table 3. The results of five classifiers for the income prediction.

RT				
ACC: 0.97977				
class	precision	recall	f1-score	support
<=50K	0.880	0.924	0.902	7410
>50K	0.718	0.605	0.657	2359
KNN				
ACC: 0.89755				
class	precision	recall	f1-score	support
<=50K	0.871	0.901	0.886	7410
>50K	0.651	0.580	0.614	2359
SVM				
ACC: 0.80524				
class	precision	recall	f1-score	support
<=50K	0.810	0.971	0.883	7410
>50K	0.758	0.284	0.414	2359
LR				
ACC: 0.80353				
class	precision	recall	f1-score	support
<=50K	0.813	0.959	0.880	7410
>50K	0.704	0.307	0.428	2359

NB ACC: 0.79717 class	precision	recall	f1-score	support
<=50K	0.813	0.954	0.878	7410
>50K	0.680	0.309	0.425	2359

More specific classification results (the ACC, precision, recall, balanced F-score and occurrence number in the test subset) of the five classifiers can be observed in Table.3. As shown in Table.3, the accuracies of the RF and KNN are higher than the accuracies of the SVM, LR and NB. Among the five models, the RF classification model performs the best and the NB classification model performs the worst.

It can be observed from Table.3 that for all models, the precision rates of income <= $50K are all higher than those of income > 50K. Since there are relatively more samples of the class who earn no more than $50K to train the model, the prediction rate of income > $50K shows the higher precision rate as observed. Therefore, though the variables of the training dataset are normalized, the distribution of two classes of income is not totally balanced.

4.2 Importance of each feature

After predicting the income by existing variables from the dataset, the further study is to estimate the most relative variable to the two classes of income. For the further study, Gini Importance has been utilized to estimate the importance of each variable. Each attribution's Gini Importance can be observed in the bar chart as descending shown in Figure.5.

From Figure.5, the instance's age has the significant highest relativity with the income of that, with Gini importance of 0.228. While the instance's race, sex and native country has the lowest relativity with the income of that, with Gini importance of 0.013, 0.012 and 0.011 respectively.

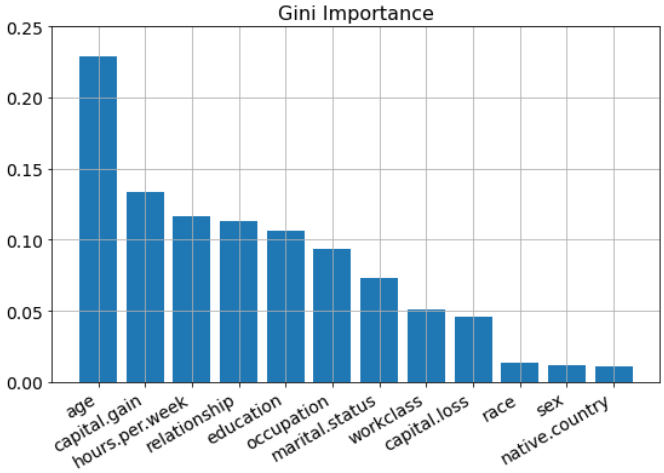

Figure 5. The Gini Importance of 12 attributions respectively based on the Random Forest classifier.

In order to research further in the considerably low importance of attributions “race”, “sex” and “native. country”, cross-validation has been utilized to test the importance. In addition, to avoid the over fitting of the best fit model, a cross-validation should also be utilized to test if the RT model has been over fitted. So, another duplicate training set without “race”, “sex”, and “native. country” was trained and tested based on the RT, which is the best fit model. The accuracy of RT with the new training dataset is 0.97539, still the highest accuracy among the five models, followed that of the KNN (accuracy: 0.89830); the accuracy of NB with the new training dataset is 0.79659, still the lowest accuracy among the five models. Moreover, the Gini Importance of the cross-validation dataset is shown in Figure.6.

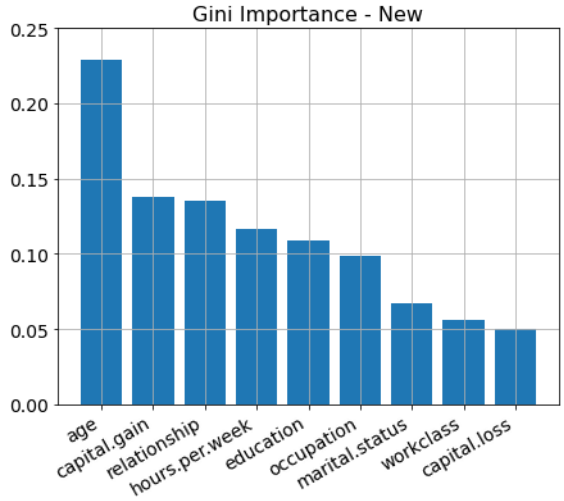

Figure 6. The Gini Importance of 9 attributions respectively based on the Random Forest classifier.

Therefore, the attributions “race”, “sex” and “native. country” can be dropped to avoid overfitting safely. And the importance of each attribution in the dataset can be proved to be authentic because the Gini importance and accuracy of prominent models nearly remain unchanged. The workclass and capital loss are the two least affected variables on the individual’s income, while age is still the most important variable of individual’s income with a large lead over the second one.

5 Conclusion

Machine learning algorithms can be employed to predict the individual’s level of income. Using Random Forest (RT) classification, the features of age, capital gain, hours per week, relationship, education, occupation, marital status, work class, and capital loss can be used to nearly precisely predict income. However, the Naïve Bayes (NB) classification predicts relatively poor performance in this task.

The accuracies of the RF and KNN are higher than the accuracies of the SVM, LR and NB in income prediction. Among the five models, the performance of RF is the best and that of NB is the poorest. And the best fit model is not over fitting.

For income <= $50K, the models all illustrate the higher prediction rates in the task. Lower prediction rate of income > $50K is calculated due to few samples in that income level.

The race, sex and native country of the person have tiny influences that can be ignored on the person's income.

Furthermore, the age of the person affects the income most while the workclass and capital loss of the person contribute least to the person's income.

Further works can be considered as following:

i Attempt to use the SMOTE (short of Synthetic Minority Oversampling Technique) method [3] to solve the imbalance problem of two classes of income, since the classification is binary in the dataset. But the risk of over-fitting caused by the SMOTE also should be considered.

ii More analysis in the reasons of why NB performed unexpected and try to improve the performance of NB.

iii More research in different parament of RT algorithm; and more experiments on different methods in data reduction based on the distribution of values in each attribution (for example, divide the work hours in light work, normal work and heavy work as 30-40-60).

References

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. Routledge.

[3] Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[5] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[6] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[7] Ho, T. K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

[8] Kibekbaev, A., & Duman, E. (2016). Benchmarking regression algorithms for income prediction modeling. Information Systems, 61, 40–52. https://doi.org/10.1016/J.IS.2016.05.001

[9] Kohavi, R. (1996). Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 202–207.

[10] Lazar, A. (2004). Income prediction via support vector machine. 2004 International Conference on Machine Learning and Applications, 2004. Proceedings., 143–149. https://doi.org/10.1109/ICMLA.2004.1383506

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel V. and Thirion, B., Grisel, O., Blondel, M., Prettenhofer P. and Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,

Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[12] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660–674. https://doi.org/10.1109/21.97458