Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 1 (2025)

A Hybrid Feature Selection through Ensemble Rank Aggregation to Improve Perdition and Classification Accuracy

Jismy Joseph[1], Dr.K.Ramesh[2]
PhD Research Scholar[1], Head cum Assistant Professor[2]
Department of Computer Science and Applications, Vivekanandha College of Arts and Science for Women (Autonomous), Elayampalayam, Thiruchengode, Tamil Nadu, India
Email: jismyjoseph2018@gmail.com[1], drjkkanimozhi@gmail.com[2]

Article History:

Received: *06-07-2024*

Revised: *23-08-2024*

Accepted: *03-09-2024*

Abstract:

Today, extending credit is a difficult operation since credit databases contain a large quantity of data as well as redundant and useless information. The surplus and unrelated data can lower the categorization and prediction accuracy. In this case, feature selection is crucial for managing massive data. Different rank aggregation (RA) methods are there to ensemble individual feature selection. However not every dataset will perform at its best when using a single RA technique, the combination of multiple RA techniques are needed to increase the perdition accuracy. This study proposes a hybrid rank aggregation model that select features that are significant across different rank aggregation methods like MC4, Kendall tau distance and Borda. This study observed that the performance of ensemble rank aggregation techniques is better than the existing individual rank aggregation methods.

Keywords: Ensemble Ensemble Rank Aggregation, CHI SQUARE.

1. Introduction

Last few decades, the size of data is exponentially increasing day by day, hence more efficient data mining techniques are necessary for credit risk analysis. If a model is trained by collecting enormous quantities of data, the model may capture insignificant pattern it the data set contains of noise, irrelevant and redundant features. The recent development familiarises new set of data types and features, hence it is necessary to build a new scalable feature selection algorithms for creating an accurate model for credit risk analysis. In this situation, frequent subset selection is a crucial strategy for dimensionality reduction and for cutting down on the training time for the data set.

The training dataset's features that have no correlation to the class labels are removed using a feature selection, also known as a variable subset selection [1]. It is a method for selecting relevant features and automatically removing noise from data. The potential to avoid over fitting, increase accuracy, shorten training times, and simplify models are all advantages of feature selection. There are three methods for selecting features Filter, wrapper, and hybrid.

In order to integrate the results of several feature selection approaches like chi-square, Information gain, Fisher score, etc., it is necessary to use the rank aggregation (RA) methodology, which combines many ranked lists into a single ranking. In this paper we propose a Hybrid Feature selection through Ensemble Rank Aggregation (HFSERA) to improve perdition and classification accuracy.

2. Literature Review

It has been demonstrated that hybrid feature selection methods increase classification accuracy. Recent works that suggest hybrid feature selection methods are scarce. A few hybrid feature selection strategies are been put out specifically for credit datasets. In the paper 'A hybrid feature selection model based on improved squirrel search algorithm and rank aggregation using fuzzy techniques for biomedical data classification' [3] Gayathri Nagarajan and L. D. Dhinesh Babu proposed a hybrid model for feature selection in biomedical data classification and their model is compared with other existing models. They claimed that their model performed better than other individual filtering methods in terms of classification accuracy and time needed for classification.

In [4] Rahi Jain, Wei Xu, presents a novel hybrid rank aggregation (HRA) strategy to carry out the rank aggregation phase in ensemble feature selection that permits feature selection based on relevance across several rank aggregation methodologies. In this model, To choose features using weighted unsupervised clustering, the HRA approach pools the feature importance from the current rank aggregation approaches. This method is adaptable since a wide range of ensemble techniques, data types, and RA methods could be used with it.

In [5] Dahiya, Shashi, Handa, S.S and Singh, Netra combines different feature rating algorithms for selecting features from credit data. They used five individual rank based feature selection methods and a rank aggregation algorithm for combining rank obtained from these individual feature selection method. This new rank aggregation algorithm uses rank order and rank score of the features. They proved that the performance of this new rank aggregation algorithm is better than other individual method.

In [6] Wanwan Zheng; Mingzhe Jin compared Feature selection methods by using rank aggregation. In this study they used 10 different feature selection techniques and Schulze (SSD) rank aggregation is used to combine ranks together. According to their analysis Mahalanobis distance is the most effective approach overall.

In the paper 'A Rank Aggregation Algorithm for Ensemble of Multiple Feature Selection Techniques in Credit Risk Evaluation' [7] Shashi Dahiya, S.S Handa and N.P Singh suggested a rank aggregation algorithm for combining different feature selection techniques. They used chi-square, information gain, gain ratio, relief-F etc. for feature selection.

In [8] the paper entitled 'The stability of different aggregation techniques in ensemble feature selection' Reem Salman, Ayman Alzaatreh and Hana Sulieman mentioned the significant differences between the score-based and rank-based aggregation procedures in the ensemble's stability and accuracy behaviour under various aggregations. In addition, it was shown that the simpler score-based techniques based on the L2-norm or Arithmetic Mean aggregation seem to be effective and convincing in most situations.

In the paper[9]' Frequent Itemsets Mining for Big Data: A Comparative Analysis' Daniele Apiletti, Elena Baralis, Tania Cerquite lli,Paolo Garza, Fabio Pulvirenti and conducted a theoretical and experimental comparative analyses of Hadoop- and Spark scalable algorithms to find frequent pattern from data.

3. Hybrid Feature Selection Using Ensemble Rank Aggregation

This section includes the information about the framework of the proposed model, frequent selection methods used, rank aggregation methods used and classification algorithms.

Ensemble Feature Ranking

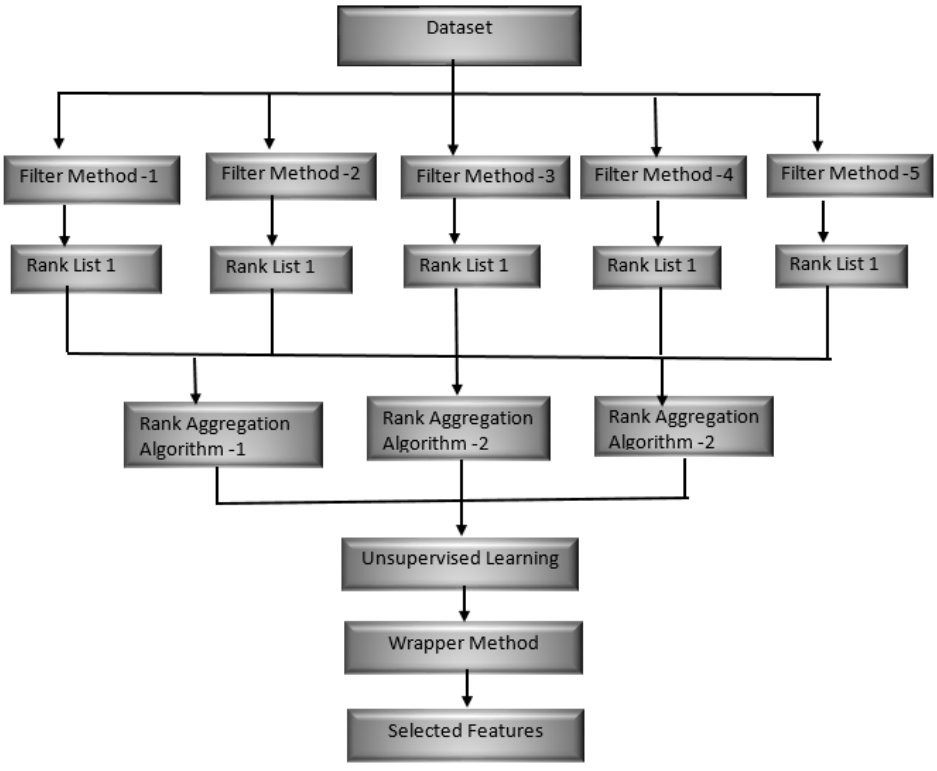
3.1 Framework Of The Proposed Model

In this study, a hybrid feature selection (HFS) method that combines the filters and wrappers methods of feature selection is applied. In the first phase, the features are ranked using different filter-based ranking algorithms like information gain, chi-square, fisher score and Brouta.

To select features and pool feature performance, a hybrid feature selection technique currently uses a single Rank Aggregation algorithm. So in the second phase of this study we used RA algorithms like MC4, Random dictator using Kendall tau distance and Borda to combine the result obtained from different feature selection methods.

However, a single Rank Aggregation approach might not necessarily provide the best results for all datasets. So in the third phase, this study uses a new Hybrid Feature selection through Ensemble Rank Aggregation (HFSERA) to improve the selection of features based on their relative importance to various RA strategies.

In the last phase, a wrapper method is applied to obtain the optimal subset for classification.



Algorithm

1. *Start with dataset D with p features.*
2. *Apply filter-based feature selection to get feature rank vector* $R = \{r_1, r_2 \ldots\ldots r_n\}$
3. *Apply N rank aggregation to get aggregated rank* $A_r = \{r_1, r_2, \ldots . r_p\}$

4. Combine aggregated rank Ar_1, Ar_2.....Ar_n by using Ensemble rank aggregation.
5. Use wrapper method to get the optimum result.

3.2 FEATURE SELECTION METHODS

Frequent selection strategies have the potential to improve classifier accuracy while simultaneously reducing computational costs by removing obtrusive and distracting characteristics. Supervised learning and unsupervised learning are the two categories used to categorise feature selection techniques. For the purpose of locating pertinent information, supervised learning makes use of labelled data while unsupervised learning makes use of unlabelled data. Filter, wrapper, and embedding methods are once more divided under the category of supervised methods.
In this work, infogain, fisherscore, chisquare, REF and Brouta are used for feature subset selection.

3.2.1 INFORMATION GAIN

To determine the reduction in entropy or surprise, Information Gain divides a dataset according to a given value of a random variable. Less surprise results from lower entropy groups of samples, which are what larger information gains suggest [13].
Mathematical formula of an entropy is
$H(X) = -\sum p_i \log_2 (p_i)$
Here 'i' is the total number of values that X can take.
In this method, the information gain (IG) for each attribute is calculated. After this, by using IG the attributes will rank in decreasing order. Mathematically, information gain is defined as,
$IG(Y/X) = H(Y) - H(Y/X)$

3.2.2 CHI SQUARE

Chi-Square is one of the commonly used method for feature selection. It is a statistical test of independence to determine the dependency of two variables. The formula for chi-square is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where

O - observed value
E - exceted value
c - degree of freedom
When two features are distinct from one another, the observed count is nearly identical to the predicted count, resulting in a lower Chi-Square value. The features with higher Chi-square value is chosen for model training because it depends more on the response.

The following are the steps used in Chi-square test:

- Define Hypothesis.
- Create a Contingency table.
- Locate the expected values.
- Determine the Chi-Square statistic.
- Accept or Reject the Null Hypothesis.

3.2.3 FISHER SCORE

Fisher Score is a supervised feature selection method and it returns the variables' ranks based on the fisher score in descending order. With the use of supervised learning for feature selection, Fisher score models have several benefits, including fewer calculations, greater accuracy, and stronger operability, which can effectively reduce time-space complexity [14].

3.2.4 BROUTA

Brouta, a random forest-based approach that removes features that are statistically less relevant. Variables that are rejected in one iteration are not taken into account in the following iteration. Initially, Brouta creates shadow features by randomly shuffling the features. Then it evaluate the importance of each features and check whether these features have higher score than the maximum score of its shadow features or not. Immediately it removes the unimportant features. Finally, the method terminates either when all features are accepted or rejected or when the number of random forest runs reaches a predetermined threshold.

3.2.4 RECURSIVE FEATURE ELIMINATION (RFE)

RFE is an effective wrapper method for feature selection method because it selects features that are most useful for predicting the target variable. Using all of the features in the training dataset as a starting point, RFE attempts to find a subset of features by successfully deleting features one at a time until the target number of features is left.

This is accomplished by first fitting the core model's machine learning algorithm, ranking the features according to relevance, eliminating the least important features, and then re-fitting the model. Up until a certain amount of traits are still present, this process is repeated [10].

3.3 RANK AGGREGATION METHODS

Rank aggregation Methods uses the rank list generated by different algorithms as inputs and it produces another rank list as output. Different rank aggregation algorithms can be used to combine the base input ranking features. They are

3.3.1 MARKOV CHAIN TYPE 4 OR MC4

A Markov chain essentially consists of a sequence of transitions that meet the Markov property and are specified by some probability distribution. In the Markov Chain methods for ranked list aggregation, the items in the various lists are represented as nodes in a graph, with the probability of transitions from node to node being determined by the relative rankings of the items in the various lists [11]. A transition matrix is used to represent the probability distribution of state transitions. This matrix is a stochastic matrix, each cell(i,j) of this matrix represents the probability of transition from state i to jMarkov chain also uses an Mx1 vector called initial state vector used to mention the probability distribution of starting at each of the **N** possible states [12]..

3.3.2 RANDOM DICTATOR THROUGH KENDALL TAU DISTANCE

The Kendall tau distance measures the number of disagreements between two ranking lists. The Kendall tau distance between two lists L_A and L_B is:

$$K(L_A, L_B) = |(i, j) : i < j, (L_A(i) < L_A(j) \wedge L_B(i) > L_B(j)) \vee (L_A(i) > L_A(j) \wedge L_B(i) < L_B(j))|$$

$K(L_A, L_B)$ will be equal to 0 if the two lists are identical and n(n-1)/2 (where n is the list size) if one list is the reverse of the other. Other formula used for Kendall tau is:

$$K(L_A, L_B) = \sum_{(i,j) \in P} \bar{K}_{i,j}(L_A, L_B)$$

Where

P is the set of unordered pairs of distinct elements in L_A and L_B.

$\bar{K}_{i,j}(L_A, L_B) = 0$ if i, and j are in the same order in L_A and L_B.

$\bar{K}_{i,j}(L_A, L_B) = 0 = 1$ if i and j are in the opposite order in L_A and L_B.

3.2.3 BORDAS RANK AGGREGATION

A set of positional voting guidelines known as the Borda count assigns each candidate a certain number of points for each ballot based on how many candidates are placed lower than them. The highest-ranked candidate receives n 1 points, where n is the number of candidates, in the original form, the lowest-ranked candidate receives 0 points, the next-lowest receives 1 point, etc. The choice or candidate with the most points after the results of all votes are tallied wins [15]. The Borda count is frequently referred to be a consensus-based voting method as opposed to a majoritarian one because it aims to elect broadly acceptable options or candidates rather than those that are favoured by the majority. [16]

3.4 CLASSIFICATION ALGORITHMS

In this study random forest is used for classification due to its high efficiency.

3.4.1 RANDOM FOREST

Random Forest is a supervised machine learning algorithm used for classification and regression. These classifiers handle the missing values and can model the **categorical values.** It creates many decision trees and merges them together to form an accurate prediction. In the method the parameters are used to increase the predictive power and speed of the model.
The advantages of random forest:

- Both classification and regression problems can be resolved using random forest.
- Handling missing values and outliers automatically.
- Random forest is considerably less influenced by noise.
- It can handle categorical and continuous variables.
- Random forest is comparatively less affected by noise.

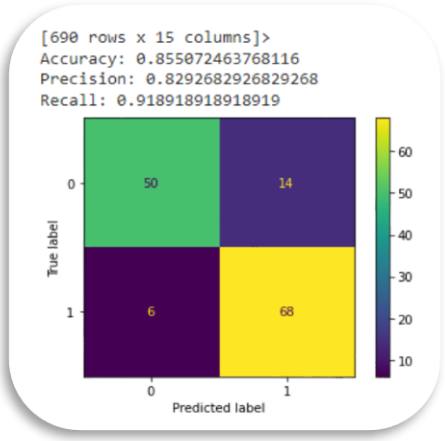
3.5 DATASETS USED

In this study, we have used two sets of credit data from the UCI repository. There are 690 instances and 15 attributes in the Australian credit data set. A German credit data set with 21 attributes and 1000 occurrences makes up the second one. Four feature selection techniques—chi-square, fisher score, information gain, and Boruta—as well as three rank aggregation techniques—Markov Chain Type 4 or MC4, Borda Count, and Random Dictator are used in this study.

4. RESULT AND DISCUSSION

Although ensemble feature selection through rank aggregation can avoid irrelevant and redundant information, hybrid feature selection through ensemble rank aggregation can make classifiers more efficient. In this study, the individual ranks lists are calculated using Fisher score, chi-square and Boruta. After that the rank aggregation algorithms MC4, Borda and Random dictator are used to combine the multiple individually ranked lists into one consensus ranking. These ranks are then combined together for getting the optimal result. Hybrid feature selection with ensemble rank aggregation method is implemented in python.

Before applying feature selection and rank aggregation algorithms, the random forest classifier is used to classify data set and the accuracy obtained for Australian and German data sets are 0.85 and 0.75 respectively.



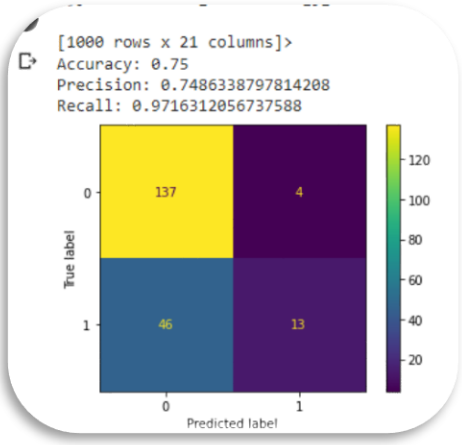


Fig1. Accuracy forAustralian Dataset **Fig2.** Accuracy for German Dataset

Table1 shows the accuracy obtained before applying different rank aggregation methods. In this study we used the random classifier for classification. The fisher score feature selection produced highest accuracy in Australian data set and chi-square produced highest accuracy for German data set.

Australian Credit Data Set				**German Credit Data Set**		
Feature selection Techniques	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Chi-Square	0.869	0.85	0.981	0.785	0.806	0.9148
FisherScore	0.884	0.862	0.932	0.74	0.743	0.964
Boruta	0.862	0.829	0.918	0.78	0.7486	0.971
InformationGain	0.869	0.841	0.932	0.76	0.7	0.9

Table 1. Acuuracy obtained after applying different feature selection techniques

Table2 shows the accuracy obtained from ensemble feature selection using rank aggregartion. The rank obtained from Chi-square, Fisher Score, Boruta and InformationGain are combined using rank aggraltion MC4, Borda count and Random Dictator.

Australian Credit Data Set				German Credit Data Set		
Rank aggregation Methods	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
MC4	0.913	0.9078	0.9324	0.775	0.8	0.96
Borda Count	0.88	0.871	0.918	0.77	0.8	0.9
Random Dictator	0.89	0.86	0.92	0.76	0.8	0.95

Table 2. Acuuracy obtained after applying different rank aggregation method

The aggregate rank obtained from MC4, Borda and random Dictator are presented in Table 3 and Table4. Table 3 show the result obtained from Australian data set and Table 4 shows the result achieved from German data set.

Borda	MC4	Random Dictator
A8 12.250	A1 13	A1 13.206910
A13 10.500	A2 9	A2 9.561778
A14 9.750	A3 6	A3 2.289081
A5 9.500	A4 11	A4 9.689620
A10 9.125	A5 4	A5 1.935323
A3 8.000	A6 12	A6 12.479884
A9 7.625	A7 8	A7 11.806383
A7 7.000	A8 1	A8 1.053010
A2 6.500	A9 7	A9 5.953927
A4 5.750	A10 5	A10 4.890913
A12 5.750	A11 14	A11 13.382848
A6 5.000	A12 10	A12 11.493168
A11 4.250		A13 1.758553
A1 4.000		A14 1.336530

Table 3. Aggregate rank obtained from Australian Data set

Borda	MC4	Random Dictator
A1 15.875	A1 : 1	A1 4.034462
A3 15.625	A2 : 3	A2 7.420083
A2 13.375	A3 : 2	A3 5.440116
A10 11.750	A4 : 8	A4 8.287495
A7 11.375	A5 : 9	A5 8.193554
A15 11.250	A6 : 11	A6 11.368984
A9 11.000	A7 : 8	A7 8.525176
A4 10.750	A8 : 12	A8 13.154061
A19 10.500	A9 : 7	A9 9.430118
A20 10.250	A10 : 4	A10 9.287205
A6 10.125	A11 : 14	A11 16.392611

A5 10.125	A12 : 13	A12 12.654030
A8 9.750	A13 : 17	A13 13.910496
A12 9.250	A14 : 18	A14 15.199035
A11 8.750	A15 : 5	A15 10.566045
A16 8.500	A16 : 16	A16 18.450728
A18 8.500	A17 : 19	A17 17.759871
A14 8.125	A18 : 15	A18 18.537267
A13 7.625	A19 : 10	A19 11.135816
A17 7.500	A20 : 6	A20 12.101246

Table 4. Aggregate rank obtained from German Data set

Table5 shows the accuracy obtained from hybrid feature selection using ensemble rank aggregartion. The aggregate rank obtained from MC4, Borda count and Random Dictator are combined together for getting more accurate result. The result shows that ensemble rank aggregation provided more accuracy and the Fig.3 and Fig.4 shows the result obtained from Australian and German data sets.

Data set	**Accuracy**	**Precision**	**Recall**
Australian Data Set	0.913	0.9078	0.9324
German Data Set	0.901	0.9	0.96

Table5 shows the accuracy obtained from hybrid feature selection

5. CONCLUSION

This paper we focus on Hybrid feature selection through ensemble rank aggregation. In this study the rank produced by fishers score, information gain, chi square and brouta are combined using different rank aggregation method MC4, Borda and Random dictator. The optimum rank for each attribute is calculated by combing the result obtained from MC4, Borda and Random dictator. Finally the wrapper method is applied for feature subset selection. The results shows that Hybrid feature selection through ensemble rank aggregation can perform better than individual rank aggregation method. In this study we used only one classification algorithm But in future, this model needs to test by using some more classification algorithms.

References

[1] Brownlee,J.,2014.Machine Learning Mastery. [Online] Available at: https://machinelearningmastery.com/an-introduction-to-feature-election/ [Accessed January 2019].

[2] H. Liu, H. M. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining.

[3] Gayathri Nagarajan, L. D. (2021). A hybrid feature selection model based on improved squirrel search. *Springer*.

[4] Rahi Jain, W. X. (2022). *Hybrid Rank Aggregation (HRA): A novel rank aggregation method for ensemble-based feature selection*. Retrieved from www.biorxiv.org: https://www.biorxiv.org/content/10.1101/2022.07.21.501057v1.full.

[5] Shashi Dahiya, S. H. (n.d.). A Rank Aggregation Algorithm for Ensemble of Multiple Feature Selection Techniques in Credit Risk Evaluation. *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*.

[6] Wanwan Zheng, M. J., 2018. *Comparing Feature Selection Methods by Using Rank Aggregation.* Bangkok, Thailand, IEEE

[7] Shashi Dahiya, S. H. ,. S., 2016. A Rank Aggregation Algorithm for Ensemble of Multiple Feature Selection Techniques in Credit Risk. *International Journal of Advanced Research in Artificial Intelligence,,* 5(9).

[8] Reem Salman, A. A. &. H. S., 2022. The stability of different aggregation techniques in ensemble feature selection. *journal of bigdata*, 9(51).

[9] Apiletti, D., Baralis, E., Cerquitelli, T., Garza, P., Pulvirenti, F., &Venturini, L. (2017). Frequent itemsets mining for big data: a comparative analysis. Big Data Research

[10] Brownlee, J. (n.d.). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Retrieved from machinelearningmastery: https://machinelearningmastery.com/rfe-feature-selection-in-python/

[11] D. Sculley, Rank Aggregation for Similar Items, Department of Computer Science Tufts University, Medford, MA, USA dsculley@cs.tufts.edu, https://www.eecs.tufts.edu/~dsculley/papers/mergeSimilarRank.pdf

[12] Devin Soni, Introduction to Markov Chains , https:// towardsdatascience.com/ introduction-to-markov-chains-50da3645a50d

[13] Brownlee, J., 2019. *https://machinelearningmastery.com/*. [Online] Available at: https://machinelearningmastery.com/information-gain-and-mutual-information/ [Accessed Feruary 2023].

[14] Lippman, D., n.d. *VotingTheory*. [Online] Available at: http://www.opentextbookstore. com/mathinsociety/current/ VotingTheory.pdf [Accessed 2023].

[15] Anon., 2018. *Devin Soni*. [Online] Available at : https:// towardsdatascience.com/ introduction-to-markov-chains-50da3645a50d
[Accessed 2023].

[16] Anon., n.d. *Borda count*. [Online] Available at: https:// en.wikipedia.org/ wiki/Borda_count [Accessed 2023].