Machine Learning Research
2024, Vol. 9, No. 2, pp. 48-53
https://doi.org/10.11648/j.mlr.20240902.13

Research Article

A Machine Learning Approach to Optimal Group Formation Based on Previous Academic Performance

Mahbub Hasan, Md. Shohel Babu*, Md. Al-Emran

Computer Science and Engineering, Southeast University, Dhaka, Bangladesh

Abstract

In today's educational institutions, student performance can vary widely due to differences in cognition, motivation, and environmental factors. These variations create challenges in achieving optimal learning outcomes. To address these challenges, Optimal Group Formation (OGF) has emerged as a promising research area. Optimal Group Formation (OGF) aims to form student groups that maximize learning efficiency based on past academic performance. Group formation problems are inherently complex and time-consuming, but their applications are extensive, spanning from manufacturing systems to educational contexts. This paper introduces a machine learning-based model designed to create optimal student groups using academic records as the primary input. The goal is to enhance overall group performance and reduce error rates by organizing students into cohesive, efficient teams. What sets this research apart is its focus on educational group formation, leveraging machine learning to improve collaborative learning outcomes. The paper also reviews prior research, emphasizing the importance of Optimal Group Formation (OGF) in various fields and its relevance in education. The model's effectiveness is demonstrated through comparative analysis, showcasing its potential to improve group dynamics in both theoretical and lab-based courses. Ultimately, the aim is to improve educational outcomes by ensuring that student groups are optimally balanced and structured.

Keywords

Optimal Group Information (OGI), Machine Learning (ML), Simulated Annealing (SA), Support Vector Machine (SVM)

1. Introduction

Currently, an enormous quantity of data is available for producing essential information across various fields such as medical, education, banking, and business. Educational institutions, in particular, can greatly benefit from this acquired information. Education is considered the backbone of any civilization. The role of an educational institution is to teach and train every student so they can achieve their desired goals in the future. Despite having a uniform learning environment in schools and colleges, student performance varies significantly. This can be attributed to differences in cognition levels, motivation, and environmental influences [13].

Optimal Group Formation (OGF) has become an evolving area of research interest among scientists and researchers. OGF aims to determine which groups yield the most significant value in various aspects. In this paper, we propose a machine learning-based group formation model that generates optimal solutions based on each student's previous academic records. It is widely believed that the performance of systems, processes, and products can be effectively improved by clustering key elements into optimal groups based on appropriate

*Corresponding author: shohel.babu@seu.edu.bd (Md. Shohel Babu)

Received: 2 September 2024; **Accepted:** 19 September 2024; **Published:** 29 September 2024

criteria. Grouping problems are inherently intricate, computationally complex, and time-consuming [20]. It is noteworthy that grouping processes are common in a wide variety of industry scenarios, including assembly line balancing [6, 7], facility location [9, 11], cell formation in manufacturing systems [17, 18, 22], advertisement allocation [10], job shop scheduling [7], order batching [19, 24], data clustering [4, 16, 28], vehicle routing problem [5], timetabling [3, 25], team formation [29], learners' grouping for cooperative learning [1], group maintenance planning [21, 27], and task assignment problem [14].

The main objective of this paper is to form groups for students in a particular theory or lab class using our proposed model. The most significant contribution of this research is that it will generate groups for future work and show the error rate.

The rest of the paper is organized as follows: In Section II, we present several background analyses of our research. In Section III, we provide the design methodology of our proposed algorithm and its pseudo-code. In Section IV, we present the results and comparative analysis. Section V contains the concluding remarks.

2. Background

Group formation is a complex and important step in designing effective collaborative learning activities [8]. In 2014, Wilmax Marreiro Cruz and Seiji Isotani used Computer-Supported Collaborative Learning (CSCL) to develop and test group formation in collaborative learning contexts using best practices and other pedagogical approaches. They used the CSCL context to address the complexity of group formation. Initially, they searched six digital libraries and collected 256 studies. After careful analysis, they verified that only 48 were related to group formation in collaborative learning contexts. They categorized the contributions of their study to present an overview of the findings produced by the community. In 2019, Anna Sapienza et al. [23] attempted to predict teams using Deep Neural Networks. They emphasized the social impact and online games. They collected data from the Dota2 game and generated a directed co-play network, whose links' weights depicted the effect of teammates on players' performance. Specifically, they proposed a measure of network influence that captures skill transfer from player to player over time. Their experimental results demonstrated that such dynamics can be predicted using deep neural networks.

In 2019, Soheila Garshasbi et al. [12] applied their algorithms to find the optimal group in the education system. They proposed a novel algorithm capable of addressing a variety of optimization problems in optimal learning group formation processes. To this end, a multi-objective version of Genetic Algorithms, i.e., Non-dominated Sorting Genetic Algorithm (NSGA-II), was successfully implemented and applied to improve the performance and accuracy of optimally formed learning groups. Their approach is applicable not only to the education system but also to other domains.

In 2019, Kaj Holmberg [15] created software that forms groups within 60 seconds. In his research, he used heuristics and metaheuristics to solve the problem. Computational tests were conducted on randomly generated instances as well as real-life instances, and some heuristics provide good solutions in a short time.

In 2010, Kalliopi Tourtoglou and Maria Virvou [26] described the differences between local search and Simulated Annealing (SA) in their book chapter. Local optimization algorithms start with an initial solution and repeatedly search for a better solution in the neighborhood with a lower cost. In contrast, SA aims to avoid getting trapped in local optima. They also emphasized the CSCL process, which improves teaching and learning with the help of modern information and communication technology.

Furthermore, Agustin-Blas et al. [2] presented a new model for team formation based on group technology. They considered different skills in staff members and set two tough constraints related to the minimum total knowledge about a resource in a team and the minimum knowledge that a given staff member must have about the resources of a team. The developed model has been shown to be well-suited for problems of team formation arising in R&D-oriented or teaching institutions.

3. Methodology

3.1. Data Acquisition

The dataset comprises 818 students from Southeast University, Dhaka, Bangladesh. Initially, the dataset included Student ID (SID), Course Code (CC), Total (Tot), Credits (Cr), Semester (Sem), Gender (Gen), Marks Round (MR), and Grades (Gr). We focused on consecutive course results and their prerequisites (Pre) to predict the best groups, each with a maximum of 3 members. The dataset covers 69 distinct courses and 36,833 rows. Additionally, we have data on prerequisite courses.

3.2. Data Preparation

First, we sorted the data according to SID, CC, Sem, and MR. Then, we dropped duplicates for SID and CC, keeping only the last value because we need only the highest grade of a student if they took the course more than once for a better grade. After that, we dropped the Tot and Gr columns because we have an alternative MR column. We also dropped rows that contain marks less than 40 to ensure the dataset does not include any failed course information. In the end, the dataset contains 23,382 rows with attributes: SID, CC, Cr, Sem, Gen, and MR. We kept only the CC and Pre columns from the prerequisite dataset. We separated 30 students (in a class or section) and grouped them into 10 different groups with arbitrary clusters based on the following table.

Table 1. Attributes Notions.

Attributes	Remarks
SID	Identity of a student
CC	Students are grouped for a specific course
Cr	For define it is theory or lab course
Sem	Semesters are also important for clustering
MR	Total marks for a course
Pr	To find out the previous record in a particular student

3.3. Algorithm

3.3.1. Simulated Annealing

This module provides a hyperparameter optimization using simulated annealing. It has a SciKit-Learn-style API and uses multiprocessing for the fitting and scoring of the cross-validation folds. The benefit of using Simulated Annealing over an exhaustive grid search is that Simulated Annealing is a heuristic search algorithm that is immune to getting stuck in local minima or maxima.

3.3.2. Simulated Annealing Algorithm

Start with some Initial T and alpha

Generate and score a random solution (score old)

Compare score old and score new:

if score new > score old: move to neighbor solution

if score new < score old: maybe move to neighbor solution

Decreases T: T*=alpha

Repeat the above steps until one of the stopping conditions is met:

T > T min

n iterations > max iterations

total runtime > max runtime

Return the score and hyper parameters of the best solution

The decision transforms into a new solution based on probability and temperature. Specifically, the comparison between the solutions is performed by computing the acceptance probability.

$$K = \frac{log(T\,min) - log(T)}{log(alpha)2a} \qquad (1)$$

3.3.3. Pseudo Code

Implementation of the algorithm was done with the help of the following tools: Python, Pandas, Numpy, Scikit-learn, matplotlib, and Google Colab.

```
Input: Student's record and prerequisite course record
Separate 30 students with the specific semester and course;
best_score = 0.0;
best_data = pd.DataFrame();
for i <- 0 to n do
    Cluster 30 students with maximum 3 members;
    Then split the data with sklearn.model selection;
    X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size = 0.2);
    Initialize linear SVM classifier;
    clf = svm.LinearSVC();
    Pass clf as a parameter of Simulated Annealing;
end
```

4. Result and Analysis

4.1. Result Analysis

We use several iterations to get the best result with success and error rates. For example, we fixed the iteration into 5 times. Each iteration we get a different success rate and error rate.

1. Iteration 1: On the 1st iteration we got 50% to 70% success rate that is not good enough. But we save the result as a current and best score.
2. Iteration 2: On the 2nd iteration it gives us a 75% to 87% success rate, now we can say it is doing better than before. Then we save the result in a current score and compare it with iteration 1. Iteration 2 gives more accuracy than before so we save it to the best score.
3. Iteration 3: On the 3rd iteration it goes the highest of all iteration that is 97% to 99%. According to our algorithm process, we save it to the current score and compare it with the previous best score that we have. We found it higher than the previous so we save it to the best score.
4. Iteration 4: On the 4th iteration, our model gives us 92% to 95% success rate. Following the same process we save it to the current score but we also don’t save it to the best score because it is not higher from the previous best score.
5. Iteration 5: On the last and 5th iteration, we can see that our proposed model gives us a success rate that belongs to 94% to 97%. In this iteration, we get a better result than iteration 4 but still now it can’t overtake the accuracy result of iteration 3.

Finally, our accuracy upon 30 students, now has been listed in the table. In this table, we can clearly see what actually happens while iteration goes on.

Table 2. Iterations Result.

Iterations	Accuracy Score
Iteration 1	66.78%
Iteration 2	86.34%
Iteration 3	98.12%

Iterations	Accuracy Score
Iteration 4	93.42%
Iteration 5	97.17%

4.2. Comparison Analysis

In this part, we will discuss our proposed model result with the actual one.

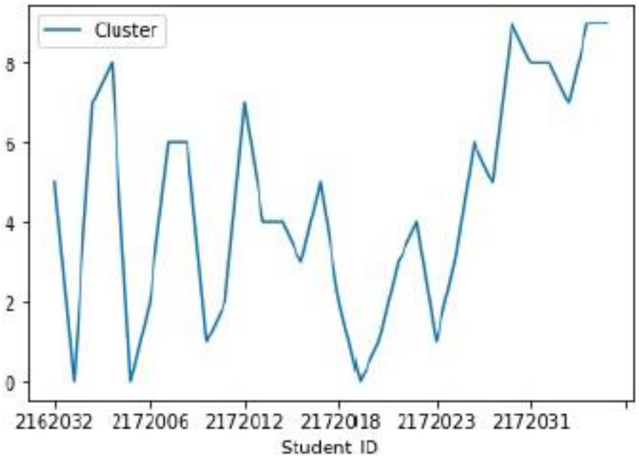

Figure 1. Cluster plot before optimization.

In Figure 1, it shows us the arbitrary cluster plot visualization where we can see that the line is in a non-linear position.

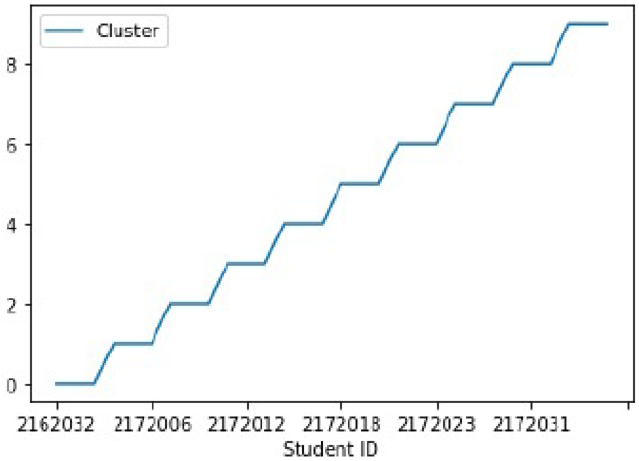

Figure 2. Cluster plot after optimization.

In Figure 2, after all iterations and optimizing with our proposed algorithm and taking the best accuracy we get the optimal group formation in a quite linear position. Where the success rate in between 97% to 99%. In our case, it gives us 98.12% among 30 students.

5. Conclusion

Our primary goal was to identify the optimal group based on the student's previous academic history. Our analysis confirms that past performance significantly impacts current performance. However, predicting the actual weight of irregular students remains challenging. Our research aimed to generate groups using Simulated Annealing (SA), which is suitable for predicting student groups in a class. With further enhancements, an algorithm similar to the one we developed could be incorporated into many academic institutions.

Abbreviations

SID	Student ID
CC	Course Code
Tot	Total
Cr	Credits
Sem	Semester
Gen	Gender
MR	Marks Round
Gr	Grades
Pre	Prerequisite
SA	Simulated Annealing

Author Contributions

Mahbub Hasan: Data curation, Formal Analysis, Methodology, Visualization

Md. Shohel Babu: Conceptualization, Formal Analysis, Investigation, Supervision, Writing – review & editing

Md. Al-Emran: Investigation, Resources

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Luis E Agust ń-Blas et al. "A hybrid grouping genetic algorithm for assigning students to preferred laboratory groups". In: Expert Syst. Appl. 36 (Apr. 2009), pp. 7234-7241. https://doi.org10.1016/j.eswa.2008.09.020.groups In: Expert Syst. Appl. 36 (Apr. 2009).

[2] Luis E Agust ń-Blas et al. "Team formation based on group technology: A hybrid grouping genetic algorithm approach". In: Computers & Operations Research 38.2 (2011), pp. 484-495.

[3] Leena N Ahmed, Ender Özcan, and Ahmed Kheiri. "Solving High School Timetabling Problems Worldwide Using Selection Hyper-heuristics". In: Expert Systems with Applications 42 (Aug. 2015), pp. 5463-5471. https://doi.org10.1016/j.eswa.2015. 02.059

[4] Mohammed Alswaitti, Mohanad Albughdadi, and Nor Ashidi Mat Isa. "Density-based Particle Swarm Optimization Algorithm for Data Clustering". In: Expert Systems with Applications 91 (Sept. 2017). https://doi.org10.1016/j.eswa.2017.08.050

[5] Mustafa Avci and Seyda Topaloglu. "A Hybrid Metaheuristic Algorithm for Heterogeneous Vehicle Routing Problem with Simultaneous Pickup and Delivery". In: Expert Systems with Applications 53 (Jan. 2016). https://doi.org10.1016/j.eswa.2016.01.038

[6] Kadir Buyukozkan et al. "Lexicographic Bottleneck Mixed-model Assembly Line Balancing Problem: Artificial Bee Colony and Tabu Search Approaches with Optimised Parameters". In: Expert Systems with Applications 50 (Dec. 2015). https://doi.org10.1016/j.eswa.2015.12.018

[7] James Chen et al. "Assembly line balancing in garment industry". In: Expert Systems with Applications 39 (Sept. 2012), pp. 10073-10081. https://doi.org10.1016/j.eswa.2012.02.055

[8] Wilmax Marreiro Cruz and Seiji Isotani. "Group Formation Algorithms in Collaborative Learning Contexts: A Systematic Mapping of the Literature". In: Collaboration and Technology. Ed. by Nelson Baloian et al. Cham: Springer International Publishing, 2014, pp. 199-214. ISBN: 978-3-319-10166-8.

[9] Sittipong Dantrakul, Chulin Likasiri, and Radom Pongvuthithum. "Applied p-median and p-center algorithms for facility location problems". In: Expert Systems with Applications 41 (June 2014), pp. 3596-3604. https://doi.org10.1016/j.eswa.2013.11.046

[10] Tuanhung Dao, Seung Ryul Jeong, and Hyunchul Ahn. "A novel recommendation model of location-based advertising: Context-Aware Collaborative Filtering using GA approach". In: Expert Syst. Appl. 39 (Feb. 2012), pp. 3731-3739. https://doi.org10.1016/j.eswa2011.09.070

[11] Ibrahim Dogan. "Analysis of facility location model using Bayesian Networks". In: Expert Systems with Applications: An International Journal 39 (Jan. 2012), pp. 1092-1104. https://doi.org10.1016/j.eswa.2011.07.109

[12] Soheila Garshasbi et al. "Optimal learning group formation: A multi-objective heuristic search strategy for enhancing inter-group homogeneity and intra-group heterogeneity". In: Expert Systems with Applications 118 (2019), pp. 506-521. ISSN: 0957-4174. https://doi.org/10.1016/j.eswa.2018.10.034

[13] Mahbub Hasan and Md. Hasan Tarque. "An Efficient Predictive Weighted Algorithm for Students Performance Prediction". In: International Journal of Emerging Technology and Advanced Engineering 8.11 (2018), pp. 94-98. ISSN: 2250-2459. URL: https://ijetae.com/files/Volume8Issuell/IJTAE_1118_15.pdf

[14] Umair ul Hassan and Edward Curry. "Efficient Task Assignment for Spatial Crowdsourcing: A Combinatorial Fractional Optimization Approach with Semi-bandit Learning". In: Expert Systems with Applications 58 (Apr. 2016). https://doi.org10.1016/j.eswa.2016.03.022

[15] Kaj Holmberg. "Formation of student groups with the help of optimisation". In: Journal of the Operational Research Society 70. 9 (2019), pp. 1538-1553. https://doi.org/10.1080/01605682.2018.1500429

[16] Saida Ishak Boushaki, Nadjet Kamel, and Omar Bendjeghaba. "A new quantum chaotic cuckoo search algorithm for data clustering". In: Expert Systems with Applications 96 (Dec. 2017). https://doi.org10.1016/j.eswa.2017.12.001

[17] Fariborz Jolai et al. "An Electromagnetism-like algorithm for cell formation and layout problem". In: Expert Syst. Appl. 39 (Feb.2012), pp. 2172-2182. https://doi.org10.1016/j.eswa.2011.07Batch

[18] Iraj Mahdavi et al. "Genetic algorithm approach for solving a cell formation problem in cellular manufacturing". In: Expert Systems with Applications 36 (Apr. 2009), pp. 6598-6604. https://doi.org10.1016/j.eswa.2008.07.054

[19] Borja Menéndez et al. "Variable Neighborhood Search strategies for the Order Batching Problem". In: Computers & Operations Research78 (Feb.2016). https://doi.org10.1016/j.cor.2016.01.020

[20] Michael Mutingi and Charles Mbohwa. Grouping Genetic Algorithms. Vol. 666. Jan. 2017. ISBN: 978- 3-319-44393-5. https://doi.org10.1007/978-3-319-44394-2

[21] Mehdi Rashidnejad, Sadoullah Ebrahimnejad, and Jalal Safari. "A bi-objective model of preventive maintenance planning in distributed systems considering vehicle routing problem". In: Computers & Industrial Engineering 120 (May 2018). https://doi.org10.1016/j.cie.2018.05.001

[22] Yeliz Buruk Sahin and Serafettin Alpay. "A metaheuristic approach for a cubic cell formation problem". In: Expert Systems with Applications 65 (Aug. 2016), pp. 40-51. https://doi.org10.1016/j.eswa.2016.08.034

[23] Anna Sapienza, Palash Goyal, and Emilio Ferrara. "Deep Neural Networks for Optimal Team Composition". In: Frontiers in Big Data 2 (2019), p. 14. ISSN: 2624-909X. URL: https://www.frontiersin.org/article/10.3389/fdata 2019.00014.

[24] André Scholz, Daniel Schubert, and Gerhard Wäscher. "Order picking with multiple pickers and due dates. Simultaneous solution of Order Batching, Batch Assignment and Sequencing, and Picker Routing Problems". In: European Journal of Operational Research 263 (Apr. 2017). https://doi.org10.1016/j.ejor.2017.04.038

[25] Vassilios Skoullis and Ioannis Tassopoulos. "Solving the high school timetabling problem using a hybrid cat swarm optimization based algorithm". In: Applied Soft Computing 52 (2017) (Nov. 2016), pp. 277-289. https://doi.org10.1016/j.asoc.2016.10.038

[26] Kalliopi Tourtoglou and Maria Virvou. "Simulated Annealing in Finding Optimum Groups of Learners of UML". In: Intelligent Interactive Multimedia Systems and Services. Ed. by George A. Tsihrintzis et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 147-156. ISBN: 978-3-642-14619-0. URL: https://doi.org/10.1007/978-3-642-14619-0_15

[27] Bo Wang and Xiaohua Xia. "A Preliminary Study on the Robustness of Grouping Based Maintenance Plan Optimization in Building Retrofitting". In: Energy Procedia 105 (May 2017), pp. 3308-3313. https://doi.org10.1016/j.egypro.2017.03.752

[28] Tanachapong Wangkhamhan, Sirapat Chiewchanwattana, and Khamron Sunat. "Efficient algorithms based on the k-means and Chaotic League Championship Algorithm for numeric, categorical, and mixed-type data clustering". In: Expert Systems with Applications 90 (Aug. 2017).
https://doi.org10.1016/j.eswa.2017.08.004

[29] Hyeongon Wi et al. "A team formation model based on knowledge and collaboration". In: Expert Systems with Applications 36 (July 2009), pp. 9121-9134.
https://doi.org10.1016/j.eswa.2008.12.031