Effect of Gender Diversity on Team performance through the Implementation of Data Mining

by Faith Grice, Dr. Rizk Nouhad, the Department of Computer Science

UNIVERSITY of HOUSTON

Background

Studies show that **collaborative learning greatly enhances students' educational and learning experiences**. Specifically, it increases retention, test scores, and engagement in the course material and it also helps students build teamwork skills, boost self-esteem, and appreciate diversity. However, **often, students get grouped together in inefficient ways** where group members have differing backgrounds, styles of learning, motivation, and more that can create an unbalanced and discouraging group experience.

This issue can be mitigated using **data mining techniques** to **create an algorithm to group students together efficiently**. Though research on the matter has been conflicting, in general, it has shown that diversity may possibly boost group performance depending on many factors**. This research is focused on determining the correlation between gender diversity in groups and group performance in a classroom setting** which will help to achieve the larger goal of efficiently grouping students together

Data and Methodology

Methodology

This study has applied 3 methods of data mining to group together students who are acquainted with each other using agglomerative hierarchical clustering which groups most similar students together first, k-means which clusters data into a specified number of groups and dbscan which clusters points which have high density within a region.

Data collection

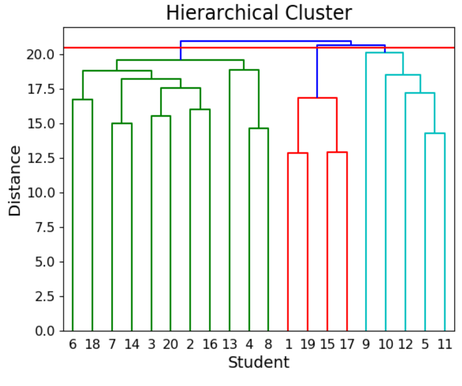
An acquaintance matrix *A*, where each cell *A[i][j]* represents student *i*'s acquaintance with student *j*. Acquaintance ranges from 1 to 10, where a low value indicates high familiarity. This matrix was randomly generated with a random male/female assignment. Using the acquaintance matrix, Euclidean distance between each student was calculated, forming a distance matrix.

Additional details about clustering methods:

Hierarchical clustering: used 'average' linkage method and optimal grouping determined by the number of clusters that had the highest silhouette score

K-Means: K = number of clusters with the best silhouette score

Dbscan: region neighborhood = 25, minimum number of points = 2



After clustering students using the methods above, statistics were computed to measure the validity of the clustering methods using **silhouette scores** and **cohesion**, and the correlation was found between diversity index in each cluster and cluster cohesion

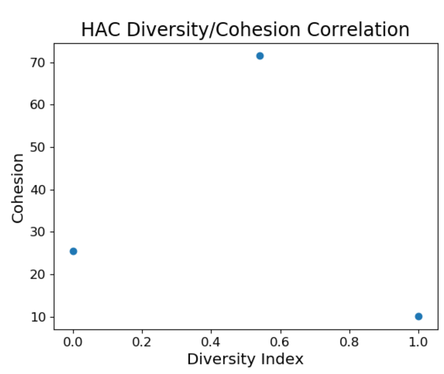
Results

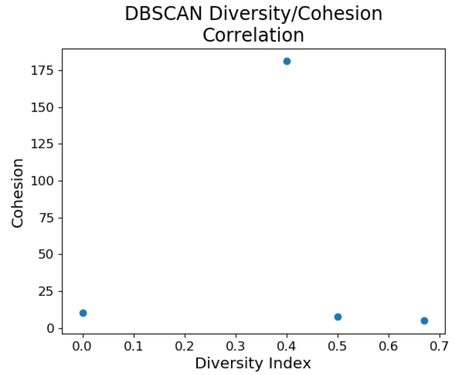
Hierarchical Clustering		
Cluster	Diversity Percentage	Average Cohesion
1	54%	71.54
2	100%	10.05
3	0%	25.52
Average silhouette score: 0.1243		

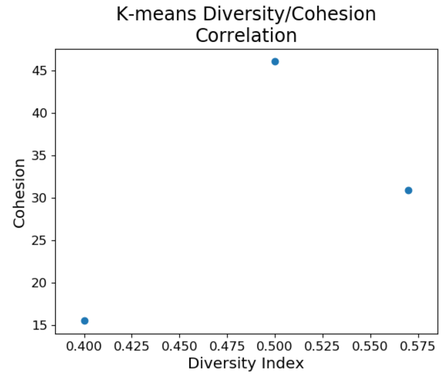
DBSCAN		
Cluster	Diversity Percentage	Average Cohesion
1	40%	180.98
2	66.67%	4.96
3	0%	10.66
4	50%	7.73
Average silhouette score: 0.0515		

K-Means		
Cluster	Diversity Percentage	Average Cohesion
1	40%	15.50
2	57%	30.89
3	50%	46.05
Average silhouette score: 0.1134		

Diversity/Cohesion correlation
HAC: -0.1968
DBSCAN: -0.00837
K-means: 0.5887







Conclusions

In the tables above,

Cohesion is the measure of how closely related observations are within a cluster (the smaller the distance the better)
Silhouette score ranges from -1 to 1 (best value is 1) and is used to evaluate how well the clustering method performed by combining the ideas of cohesion and separation.

Since a low cohesion distance means that a cluster is more cohesive, negative correlation coefficients mean there is a positive correlation between the cohesiveness of the group and diversity, therefore, the hierarchical clustering algorithm has shown the best correlation between cohesion and diversity, although not much of a correlation.

Results show that, although hierarchical clustering has the highest average silhouette score, the silhouette scores are overall low for each method, indicating that these clustering algorithms may not be great for clustering students together, or that acquaintance is not a good way to predict group efficiency. Moreover, there is not much of a correlation between diversity index and group cohesion, thus one can conclude that gender diversity may not be important to consider as a factor for clustering methods.

Acknowledgements & References

This research was made possible by an NSF grant to the University of Houston Computer Science Department (NSF CNS-1551221). I would also like to thank Dr. Rizk and Joseph Kim for providing me guidance and helping me throughout the project.