Applying Associative and Clustering Analysis to Individual Player Statistics in the 2023-2024 NBA Season.

Assignment 2

CPS844 - Data Mining

Ben Costas - 501025364

April 12th, 2024

Prepared for Professor Elodie Lugez

Background

The 2023 - 2024 NBA Player Stats, sourced from Kaggle's datasets [1] records the per-game statistics of over 400 current NBA players. Conducting association analysis and clustering analysis on the statistical measures of each NBA player can provide insightful information for basketball fanatics like myself. For instance, these analyses may be useful for placing bets on upcoming playoff games, preparing for next season's fantasy basketball league, projecting player development, and determining player valuation which is essential for contract negotiations and salary cap management. This dataset only contains individual performance statistics. For example, it doesn't contain advanced metrics such as a player's physical attributes, player efficiency rating, win shares, and other various data you may find on a website like Basketball Reference [2]. The primary objective is to explore the patterns and relationships between statistical categories and determine how to group players based on their statistical performance.

Methods

Preprocessing this dataset was a difficult task as 30 attributes needed to be examined before processing and analyzing. I filtered the dataset to players who have played in at least 30 games this season and reduced the dataset to 13 attributes. I preserved the shooting percentages of each player and removed redundant metrics such as 2 Point Attempts and 3 Point Attempts as Field Goal Attempts record the total of both Field Goal Percentage as Effective Field Goal percentage is a more suitable metric as it compensates for the value of a three-pointer. I also combined common attributes. For instance, I merged steals and blocks into "stocks" which measure the player's defensive impact. Likewise, various attributes were simplified. For instance, if a player were listed as both a PF-C, this would be simplified to PF. These attributes then had to be discretized or binarized to be analyzed through the Apriori algorithm as they are categorical and continuous attributes. To do this, I referred to the NBA League Averages [3] to determine the ranges of the bins from these metrics.

The main purpose of association analysis is to identify patterns and relationships between different statistical categories and we can use these findings to create inferences and predictions on our players. It performs this analysis by using the Apriori algorithm which identifies frequent itemsets in a dataset. I used a minimum support threshold of 0.15 which indicates that an itemset must appear

in approximately 40 NBA players to be considered frequent. The minimum confidence threshold was set at 0.9, which means that for any given X → Y association rule, at least 90% of the transactions containing X also contain Y.

The main purpose of clustering analysis is to distinguish similar items in a dataset by placing them into groups. This dataset will use cluster analysis to group players based on their statistical performance. I decided to collect a random sampling of 10% of the preprocessed data, this translates to roughly 40 NBA players being clustered into K groups. Performing the KMeans algorithm on a dataset this large would take a lot of time, thus using sampling is a suitable alternative. I used SciKitLearn's MiniBatchKMeans clustering algorithm [4] as it is more efficient at computing the algorithm with a larger dataset. The Elbow Method was used to determine the 'K' value which represents the number of clusters to be applied to the cluster algorithm. Calculating the Elbow Method involves plotting the sum of squared errors (SSE) for a range of clusters (K) and identifying where the 'elbow point' forms. I explored the hierarchical relationships between the clusters by adopting the Group Average method to generate the associated dendrogram. As the dataset specifically records the per-game average statistics of each player, the Group Average method would provide the most accurate results for this clustering algorithm.

Results

After conducting association analysis, I saved the output of all the association rules derived from the 13 attributes I kept within the dataset into the file **"association_rules_output.txt"**. The algorithm generated 1230 association rules based on the minimum support threshold of 0.15 and minimum confidence threshold of 0.9. From the 13 attributes, 2 to 6 are used in each association rule. Through most of the association rules generated, a very common pattern occurs. Most of the association rules are in the following format: X low statistic(s) → Y low statistic(s). This is most likely because more NBA players in the dataset have lower stats as per each team, there are only a couple of star-caliber players, and the remainder of their team is filled with role players, bench players, etc. This is the downside to this extensive dataset and the large number of results it produces is that it can be overwhelming and impractical to manually go through each association rule to find something unique. I would suggest implementing a filtering algorithm to reduce the redundancies and

repetitive results. Likewise, I believe that improving the threshold wouldn't be an ideal solution, as it would reduce the potential of having meaningful results from the association rules. This is because there are substantially more NBA players who have lower stats than NBA players who have all-star caliber talent.

After calculating the elbow method, I observed that the K value is 3. I processed and saved the results of the KMeans algorithm into the **"cluster_output.txt"** file. This algorithm organizes 12 players into Cluster 0, 16 players into Cluster 1, and 8 players into Cluster 2. Cluster 1 contains players with low stats, mainly due to their poor efficiency. Likewise, these players are either near retirement age and are most likely still in the NBA as a veteran presence or very young players who most likely are spending their time in the development league. Cluster 2 contains all-star and superstar-caliber NBA players. Most players on this list have received an all-star selection or an NBA reward based on their statistical performance. I would argue that some of the players do not belong in this category. They would be more suitable in cluster 2, which contains role-player and bench-player archetypes. These players complement the stars on their teams and are also capable of having a significant impact on each game. The major difference is that players in cluster 1 are more consistent performers and are heavily relied on as the first option to lead the team. Hierarchical clustering produces similar results to the output of KMeans clustering. Similarly to KMeans, it can be arguable that a couple of players in each cluster can be moved to another one (ex. Jalen Green is leading his team to a 10-game win streak, but is clustered with role players). Overall, the results of each clustering method provide insight into player valuation by grouping individuals according to similar statistical performances.

Conclusions

Overall, these two forms of analysis provide us with a breadth of NBA statistics. Association analysis allows us to discover relationships between the statistics which we can use to identify patterns between similar playstyles. Clustering analysis works by grouping players based on their statistical performances, allowing us to visualize and compare player valuation. It should also be noted that other factors contribute to the valuation of an NBA player which are immeasurable such as their physical attributes, vulnerability to injury, coachability, leadership, and basketball IQ.

Furthermore, interpreting these statistics and determining how they correlate to qualitative statistics could refine our approach to player evaluation and understanding of basketball performance.

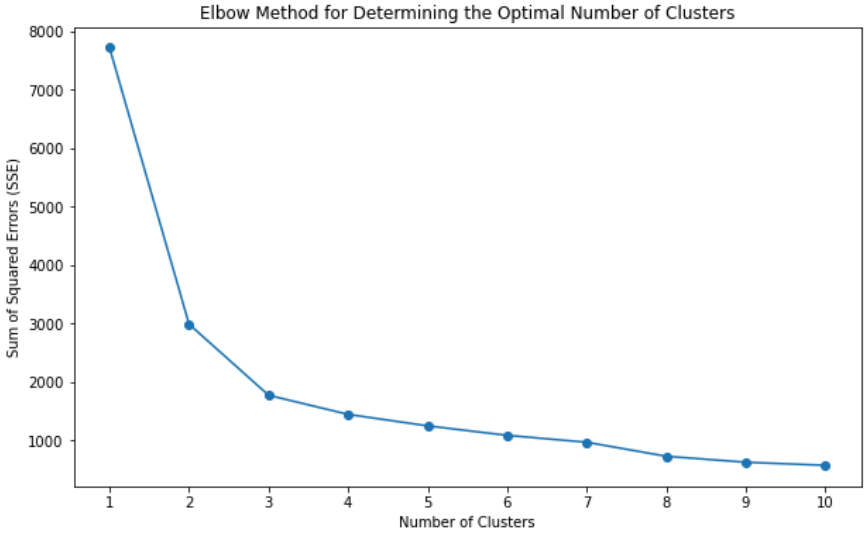

Figure 1 - Elbow Method

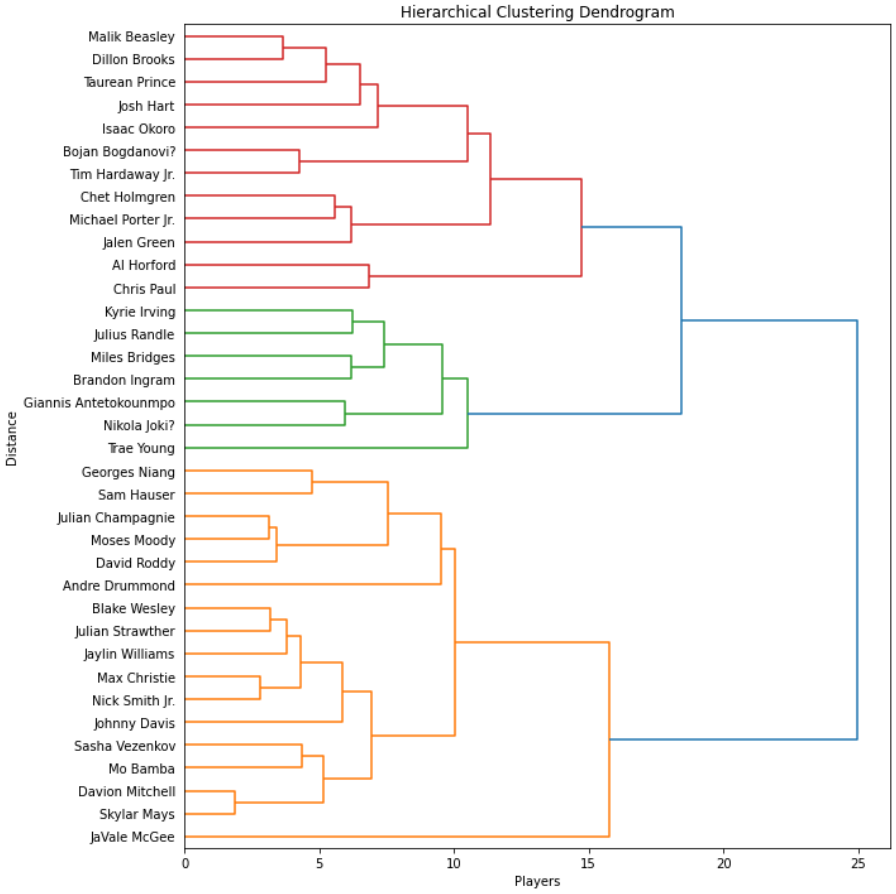

Figure 2 - Hierarchical Clustering Dendrogram

References:

1. Vivo Vinco, 2023-2024 NBA Player Stats, https://www.kaggle.com/datasets/vivovinco/2023-2024-nba-player-stats?resource=download
2. Basketball Reference, Basketball Stats and History: Statistics, scores, and history for the NBA, ABA, WNBA, and top European competition. https://www.basketball-reference.com/
3. Basketball Reference, NBA League Averages - Per Game. https://www.basketball-reference.com/leagues/NBA_stats_per_game.html
4. Scikit Learn, MiniBatchKMeans. Comparison of the K-Means and MiniBatchKMeans clustering algorithm. https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html