A Clustering Approach to Analyzing NHL Goaltenders' Performance

Ruksana Khan[2], Patrick Schena[1], Kathleen Park[1], and Eugene Pinsky[2(✉)]

[1] Administrative Sciences Department, Metropolitan College, Boston University, Boston, MA 02215, USA
{pschena,epinsky}@bu.edu

[2] Computer Science Department, Metropolitan College, Boston University, Boston, MA 02215, USA
{ruksanak,epinsky}@bu.edu

Abstract. Ice hockey is among the top 10 sports in the world by global popularity, and the National Hockey League (NHL) is one of the major professional sports leagues in United States and Canada. In the NHL there are 32 teams, 25 in the U.S. and 7 in Canada. In ice hockey, the goaltender, also known as the goalie, is one of the most important players in the game. The result of the game greatly depends on the performance of the goaltender. One of the most important statistics of the goaltender is save percentage SV% (calculated as the number of saves divided by the total number of shots attempted on the goal). In spite of the goaltender being a key player in the game, there are shortcomings in existing methods of ranking goaltenders, as these methods do not comprehensively capture the performance of the goaltender. This paper proposes the use of clustering methods from machine learning to compare performance of NHL goaltenders by using SV% and to look for patterns in their performance.

Keywords: Clustering · NHL performance comparison · Goaltender statistics

1 Introduction

One of the most important statistics of a goaltender in ice hockey is save percentage (SV%). It is a statistic that represents the percentage of attempted shots on the goal that the goaltender stops. The higher the SV% is, the better the performance of the goaltender. In this paper, we have analyzed the NHL goaltenders' SV% patterns over 5 years by applying the k-means clustering method with Manhattan distance. Inputs to the k-means model are the annual SV% for each goaltender.

We use the quantile statistics to split the SV% into clusters [3]. In this way, for each year and for each goaltender, we can describe his performance in terms

Supported by Metropolitan College, Boston University.

T. Zlateva and R. Goleva (Eds.): CSECS 2022, LNICST 450, pp. 3–10, 2022.
https://doi.org/10.1007/978-3-031-17292-2_1

of a cluster (defined by a quantile range). The performance of each goaltender over 5 years will be described by a 5-value tuple of these clusters.

We will then use k-means clustering [1,2,5–7] and divide our goaltenders into $k = 5$ groups. The outputs of the k-means are the 5 clusters defined as: World Class, Elite, Competitive, Serviceable, and Inadequate.

2 Data Analysis and Visualization

2.1 Data

We collected goaltenders performance data from two sources:

1. Worldwide ice hockey statistics (https://www.eliteprospects.com). From this source, we collected the data for each player, individually searched each player, and used a filter to view the data for the NHL goaltenders only. We obtained SV% from this dataset.
2. Instat sports performance database (https://hockey.instatscout.com). We used this dataset to obtain information about time on ice, for analyzing goaltenders who have played more than 7,000 min total in 5 years.

Apart from the minutes on ice, we have considered only goaltenders who have more than five years of playing history in the NHL. We are left with 37 goalies after applying all the above filters.

2.2 Visualization and Statistics

SV% ranges from a min of 0.750 to a max of 0.934. Inter-quartile range is defined as $Q_3 - Q_1$ and for our dataset, it is $Q_3 - Q_1 = 0.013$. Quantile skewness A is defined in Eq. (1) as the difference between the range of the lower quartile and median M (i.e. $M - Q_1$) and range of the upper quartile and median M (i.e. $Q_3 - M$) divided by inter-quartile range. For our dataset, this skewness $A = -0.61$, and the median $M = 0.915$ (Table 1).

$$A = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1} \tag{1}$$

Table 1. Save percentage (SV%) quartile statistics

Statistics	Value
Min	0.750
First quartile Q_1	0.907
Median M	0.915
Third quartile Q_3	0.920
Max	0.934
Deviation σ	0.007
Quantile skewness A	−0.61
Mean μ	0.912

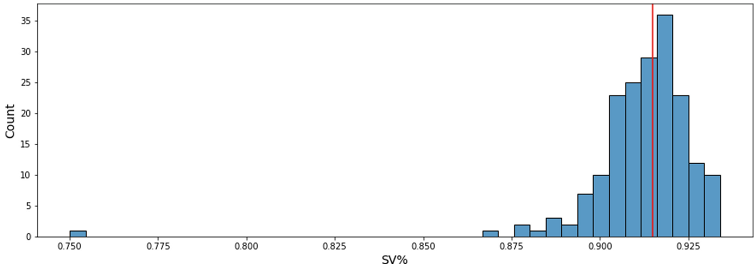

Fig. 1. Save percentage (SV%) distribution

The histogram of SV% (Fig. 1) shows that SV% is skewed to the left with a mean $\mu = 0.912$ and a standard deviation $\sigma = 0.007$.

We will use non-parametric analysis using simple quantile statistics to analyze SV% [4]. Using the Q_1, Q_2 (median M), and Q_3 we defined the patterns of SV% for every goaltender over the five years. We assign SV% values to clusters according to Table 2.

Table 2. Save percentage (SV%) pattern

Range	SV% Pattern (cluster)
$<Q_1$	1
$Q_1 \longleftrightarrow M$	2
$M \longleftrightarrow Q_3$	3
$>Q_3$	4

In Fig. 2, we plotted 5-year SV% trajectories for 37 goaltenders. Table 4 (Appendix A) contains the details of these trajectories.

As an example, consider the trajectory for the goaltender Tristan Jarry (Pittsburgh Penguins). His 5-year pattern is $[1, 2, 1, 4, 2]$. In the first year, Tristan Jarry's SV% is in cluster 1 (SV% below $Q_1 = 0.907$). In the second year, he was in cluster 2 (SV% was between Q_1 and - median M, namely in range (0.907–0.915)). In the third year, his performance was in cluster 1 again. In the fourth year he was in the top cluster 4 (SV% was above 0.920), and, finally, in the fifth year it was again between Q_1 and the median M. His trajectory shows that his performance improved over the 5 years.

By contrast, let us consider Brandon Holtby (Dallas Stars). His 5-year cluster trajectory is $[4, 4, 1, 2, 1]$. His trajectory shows that he had top performance in the first 2 years but his performance declined drastically over the next 3 years.

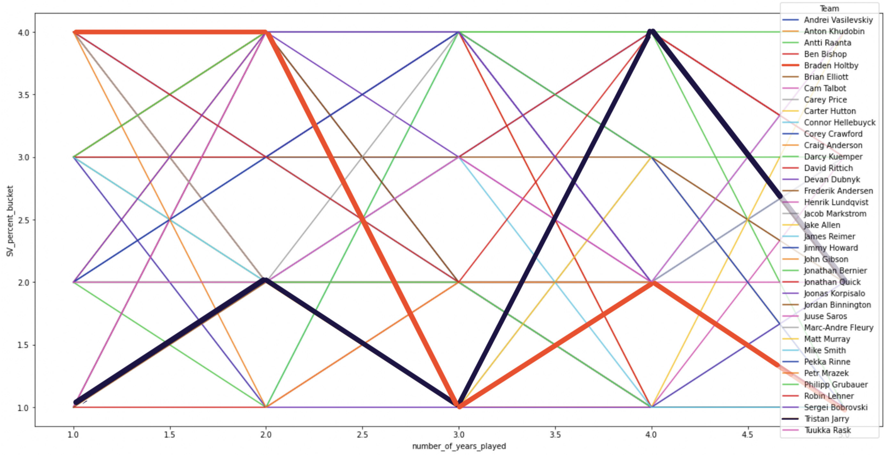

Fig. 2. Save percentage (SV%) pattern trajectories

3 *k*-Means Clustering of 5-Year SV% Trajectories

We extracted data for 37 goaltenders and their SV% pattern over 5 years. We divided these goaltenders into clusters using their 5-year performance patterns. We rank these goaltenders according to cluster trajectories and cluster signatures as explained below.

3.1 Cluster Trajectories

Clustering divides the data into groups based on certain similarities in the data. We use a standard k-means clustering to group goaltenders into k groups. Each cluster is represented by its centroid, an arithmetic mean of all the data points assigned to a cluster. In other words, a centroid represents a typical member of its cluster. In our case, a centroid is a "average" 5-year trajectory of its members. We rounded the centroid values to the nearest SV% cluster values.

The k-means algorithm works iteratively until each point has less inter-cluster distance than intra-cluster distance. Both the number of clusters k and the distance metric are the hyperparameter inputs to the model and need to be specified. Since we are working with a SV% pattern for distance, we used Manhattan distance as an input distance metric. We also decided to use $k = 5$ clusters as we believe that this number would be adequate to analyze similarities in historical goaltender performance. With this choice of hyperparameters, we divided the historical performance of 37 goaltenders into 5 clusters: *World Class, Elite, Competitive, Serviceable,* and *Inadequate.*

Table 3 shows the statistics for each of the 5 clusters. We note that there are no outlier clusters (i.e. clusters with just one member).

Table 3. k-means clusters

Cluster	Cluster name	Count	5-year SV% trajectory
0	Inadequate	10	3,2,2,1,1
1	Serviceable	4	1,4,2,2,1
2	Competitive	4	4,4,1,2,1
3	Elite	8	3,4,4,2,1
4	World class	11	3,2,3,3,3
–	Average goaltender	–	3,3,3,2,2

As can be seen from Table 3, the distribution across the 5 clusters is not equal. Instead, the *"Inadequate"* and *"World Class"* clusters each have disproportionately high representation of 10 and 11 each. In general, once past the large representation in *"Inadequate"*, the trend is toward increasing representation as the cluster rankings climb from *"Serviceable"*, to *"Competitive"* to *"Elite"* to *"World Class"*. For instance, the *"Elite"* and *"World Class"* clusters combined contain over half the players in the sample. There are implications for the longevity of play within a 5-year period being skewed toward the higher performers, with, interestingly a relatively deep bunch of backup reserve players classified as *"Inadequate"* for consistent first-rung starting play, but who have nevertheless at occasional key moments made significant enough contributions to have exceeded our baseline of 7,000 min aggregate NHL play in 5 years.

The performance of each cluster is represented by the cluster path of its centroids, as shown in Fig. 3.

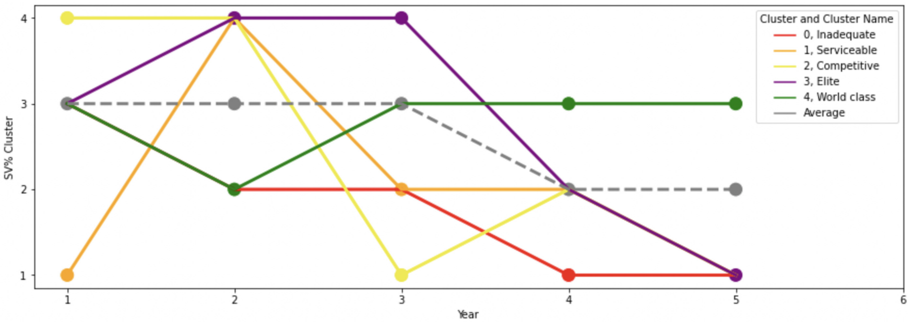

Fig. 3. Centroid trajectories

Let us examine the 5-year cluster trajectories for the centroids in more detail. We start with "world class". The goaltenders in this group have consistent and good performance over 5 years (level 3). By contrast, consider the "elite" group. Goaltenders in this group the highest performance (3–4) for at least 3 years but then much lower performance (1–2) in the last 2 years. If we examine "competitive", we see that goaltenders in this group start out very strong (highest level 4) but decline drastically in the next 3 years (1–2). Goaltenders in the "serviceable" cluster had a period of early glory (level 4 in the second year) and otherwise, show substandard performance (level 1 and 2). Finally, goaltenders in "inadequate" group never achieve a very high level (highest level is 3) and they mostly perform at the lower level (1–2).

We note that all clusters show declining performance over time. These results suggest that, with rare exceptions (e.g. Darcy Kuemper), unless goaltenders show early promise, they tend not to improve over time.

In Appendix A we show 5-year cluster trajectories for each goaltender in our dataset.

3.2 Cluster Signatures

Centroid-Signatures. From the previous discussion we have shown that the performance of every goaltender can be represented by 5 year trajectory. We now consider simplified representation which we will call signature. To compute the signature we compute two values the average cluster that best represent the first two year, and the cluster that represent the last two year. In other words, the signature is computed by taking a simple average of the first 2 years and last 2 years from 5-year cluster trajectory.

As a result we reduce 5-tuple representation to 2 values. The first value reflects the performance in the first two year whereas the last value in the signature represent the performance in the last two year.

For example, consider the performance of Tristan Jarry's in the previous section. His 5-year cluster path was $[1, 2, 1, 4, 2]$ corresponding to cluster *"World Class"*. His corresponding signature is $[2, 4]$ showing that his performance improved over the 5 years. By contrast, consider Braden Holtby. His 5-year cluster path was $[4, 4, 1, 2, 1]$ corresponding to cluster *"Competitive"*. His corresponding signature is $[2, 4]$ showing that his performance decreased over the 5 years.

Using signatures gives us a very simple summary of 5-year goaltender performance.

4 Conclusion

In this paper, we applied machine learning methods to analyze the performance of NHL goaltenders. We used the standard metric of SV% and, for every year, divided NHL goaltenders into five clusters using the k-means clustering method. We examined the 5-year performance pattern for each goaltender. We applied clustering again to these 5-year patterns using Manhattan distance and identified groups of goaltenders with similar performance over time. We use 5-year cluster trajectory and a simplified performance measure, "signature", to describe goaltenders' performance over time. This approach provides a simple and visually intuitive method to analyze and classify goaltenders. It could also be used for NHL teams to evaluate the performance of their goaltenders compared to others in the league on a game-by-game basis by entering SV% for each game played, rather than a cumulative season SV%. This method can be generalized to other sports as well. We hope to address this in our subsequent work.

Acknowledgement. The authors would like to thank Brian Daccord, the goaltending coach at Boston University, for his helpful suggestions [8]. We would also like to thank the Office of the Dean at Boston University Metropolitan College for their support.

Appendix A

Table 4. SV% trajectory patterns and signatures

Last name	First name	SV% pattern	SV% signature	Cluster
Allen	Jake	[3, 2, 1, 1, 4]	[2, 2]	Inadequate
Andersen	Frederik	[3, 3, 3, 3, 2]	[3, 2]	World class
Anderson	Craig	[3, 4, 1, 1, 1]	[4, 1]	Competitive
Bernier	Jonathan	[2, 2, 2, 1, 1]	[2, 1]	Inadequate
Binnington	Jordan	[1, 4, 2, 2, 1]	[2, 2]	Serviceable
Bishop	Ben	[4, 2, 3, 4, 3]	[3, 4]	World class
Bobrovski	Sergei	[2, 4, 4, 2, 1]	[3, 2]	Elite
Crawford	Corey	[4, 3, 4, 2, 3]	[4, 2]	World class
Dubnyk	Devan	[3, 4, 3, 2, 1]	[4, 2]	Elite
Elliott	Brian	[4, 2, 2, 1, 1]	[3, 1]	Inadequate
Fleury	Marc-Andre	[4, 2, 4, 2, 1]	[3, 2]	Elite
Gibson	John	[3, 4, 4, 3, 1]	[4, 2]	Elite
Grubauer	Philipp	[3, 4, 4, 3, 3]	[4, 3]	World class
Hellebuyck	Connor	[3, 1, 4, 2, 4]	[2, 3]	World class
Holtby	Braden	[4, 4, 1, 2, 1]	[4, 2]	Competitive
Howard	Jimmy	[1, 4, 2, 2, 1]	[2, 2]	Serviceable
Hutton	Carter	[2, 4, 2, 1, 1]	[3, 1]	Serviceable
Jarry	Tristan	[1, 2, 1, 4, 2]	[2, 3]	World class
Khudobin	Anton	[2, 4, 4, 1, 1]	[3, 1]	Elite
Korpisalo	Joonas	[3, 1, 1, 1, 2]	[2, 2]	Inadequate
Kuemper	Darcy	[2, 1, 4, 4, 4]	[2, 4]	World class
Lehner	Robin	[4, 3, 2, 4, 3]	[4, 4]	World class
Lundqvist	Henrik	[3, 2, 2, 1, 1]	[2, 1]	Inadequate
Markstrom	Jacob	[2, 2, 2, 2, 3]	[2, 2]	World class
Mrazek	Petr	[4, 1, 2, 2, 1]	[2, 2]	Inadequate
Murray	Matt	[4, 4, 1, 3, 1]	[4, 2]	Competitive
Price	Carey	[4, 4, 1, 3, 2]	[4, 2]	Competitive
Quick	Jonathan	[3, 3, 4, 1, 1]	[3, 1]	Elite
Raanta	Antti	[3, 4, 4, 4, 1]	[4, 2]	Elite
Rask	Tuukka	[2, 2, 3, 2, 4]	[2, 3]	World class
Reimer	James	[3, 3, 2, 1, 2]	[3, 2]	Inadequate
Rinne	Pekka	[2, 3, 4, 3, 1]	[2, 2]	Elite
Rittich	David	[1, 1, 2, 1, 1]	[1, 1]	Inadequate
Saros	Juuse	[1, 4, 4, 2, 2]	[2, 2]	Serviceable
Smith	Mike	[3, 2, 3, 1, 1]	[2, 1]	Inadequate
Talbot	Cam	[3, 3, 2, 1, 3]	[3, 2]	Inadequate
Vasilevskiy	Andrei	[2, 3, 3, 4, 3]	[2, 4]	World class

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, Heidelberg (2006)
2. Deisenroth, M.P.: Mathematics for Machine Learning. Cambridge University Press, Cambridge (2020)
3. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, Hoboken (2011)
4. DeGroot, M.H., Schervish, M.J.: Probability and Statistics, 4th edn. Pearson, London (2018)
5. Hastle, T.: Elements of Statistical Learning. Springer, Heidelberg (2016)
6. Kroese, D.P., Botev, Z., Taimre, T., Vaisman, R.: Data Science and Machine Learning. Chapman and Hall CRC Publishing (2019)
7. Wilmott, P.: Machine Learning: An Applied Mathematics Introduction. Panda Ohana Publishing (2019)
8. Doomany, C.: Cluster names (personal communication). Bentley University, April 2022