Analyzing NHL Goalie Stats (03-04—07-08) Using the Self-Organizing Map

Chuck Crittenden

Data Mining

05/07/2008

"In hockey, goaltending is 75 percent of the game.
Unless it's bad goaltending. Then it's 100 percent of the game,
because you're going to lose."
~ Gene Ubriaco (NHL forward)

Executive Summary

It's always been said that great goaltending will lead you to the playoffs. I'd like to know how true that statement is based on a few goaltending statistics and one team statistic. I used the Self-Organizing Map to attempt to find natural clusters of different levels (based on average standings) of NHL teams together using their average Goals Against Average (GAA), Save Percentage (SV%), and Goal Differential (DIFF) over the seasons 2003-2004 through 2007-2008 (omitting the 2004-2005 season since their was a strike for all of that season).

The team name will become my label with which to map. In short the Self-Organizing Map uses a map with a pre-defined size randomly filled with attributes. The instances of data are compared with each individual point on the map using the Euclidean distance. Whichever point on the map the instance is nearest to it is made into that map point. At this point the program also trains the instances around it using competitive learning. This process repeats for every instance. The algorithm repeats for a predetermined number of repetitions. I want to have my map as clustered as possible without having too large of a map. If I had too large of a map, the results would be so spread out that the results wouldn't be as clear.

Using NHL goaltending data (retrieved from NHL.com and Yahoo Sports) with the Self-Organizing Map, I was able to come up with a map that had a good spread for each of the levels in the standings. The high level teams mapped together, with the next level following on top of it, and so on until the low level teams mapped into the opposite corner of the map. The resulting map and the standings table follow:

Pittsburgh	x	x	x	Carolina	x	x	x	x	LosAngeles	x	x	StLouis	x	Chicago
x	x	Toronto	x	x	x	Florida	x	x	x	x	x	x	x	x
x	x	x	x	x	x	x	x	NewYorkI	x	Washington	x	x	x	x
Vancouver	x	x	TampaBay	x	x	x	x	x	x	x	x	x	Phoenix	x
x	x	x	x	x	x	x	x	x	Atlanta	x	x	x	x	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NewJersey	x	x	Calgary	x	x	x	x	x	x	x	x	x	Boston	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
x	Colorado	x	x	x	x	x	Montreal	x	x	Edmonton	x	x	x	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Buffalo	x	x	x	x	x	x	x	x	x	x	x	x	x	Philadelphia
x	x	x	x	x	x	Anaheim	x	x	x	x	x	x	x	x
x	Ottawa	x	x	x	x	x	x	x	Minnesota	x	x	x	x	x
x	x	x	SanJose	x	Dallas	x	x	x	x	x	x	x	x	x
Detroit	x	x	x	x	x	x	Nashville	x	x	x	x	x	NewYorkR	x

	overall standings	
Detroit	1.25	high
San Jose	5.75	high
New Jersey	7	high
Ottawa	7.25	high
Dallas	7.75	high
Buffalo	10.5	medhigh
Anaheim	10.75	medhigh
Nashville	10.75	medhigh
Calgary	11.75	medhigh
Colorado	12	medhigh
Montreal	12.5	medhigh
Philadelphia	14	med
Vancouver	14	med
New York R	14.5	med
Minnesota	14.75	med
Carolina	15.25	med
Tampa Bay	15.75	med
Toronto	16	med
Boston	16.25	med
Pittsburgh	18	medlow
Edmonton	18.75	medlow
Atlanta	20	medlow
New York I	20.25	medlow
Florida	22	medlow
St. Louis	23.5	low
Washington	23.5	low
Los Angeles	24.5	low
Columbus	25.5	low
Phoenix	25.5	low
Chicago	25.75	low

The findings are that the Self-Organizing Map is a very good way to separate NHL teams into their respective standing level based on GAA, SV%, and DIFF. The one major thing that must be remembered when using the Self-Organizing Map is that if the program is re-run, the results will be different, because of the random initialization. Adjusting the number of repetitions or the dimensions of the map will also result in a different map. In summary the Self-Organizing Map as shown did a very good job at grouping the different standing levels together. Proof of this is the fact that Detroit (the top team in the standings) and Chicago (the lowest team in the standings) mapped into opposite corners. This proves the long assumed statement that a good goalie leads to a playoff contending hockey team.

Problem Description

The problem that I plan to attempt involves using the Self-Organizing Map (SOM) to try to effectively cluster National Hockey League (NHL) goaltending statistics from the 2003-2004 season through the 2007-2008 season into five different levels of hockey

teams. These five levels will be decided based on each team's average standings during these particular seasons. This does not include the year of 2004-2005, because of the NHL lockout during that particular season.

Since goaltending is such an important aspect of hockey, I am expecting that these levels will appear in the map. After the goaltending data has been run through SOM, I will analyze the resulting map to decide if the algorithm was able to cluster the different levels of teams I set and how well the algorithm clustered the hockey teams.

Analysis Technique

The algorithm I am going to use to attempt to cluster the teams is the Self-Organizing Map or SOM.

The Self-Organizing Map (a type of artificial neural network) is a method of finding clusters and showing them in a 2-dimensional map. The program first randomizes instances into the points on the map (randinit). Next the program takes a specific instance (p) and tests it against each of the points on the map (q) and finds the one point where the Euclidean distance is smallest.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

(Euclidean Distance, Wikipedia) The point p is placed into that point (q) and q is trained to more closely resemble the point (p) it was being compared to. The algorithm then trains the surrounding nodes with less training the further from the original node it is. This process repeats for all the instances for 'rlen' number of times. This process along with what else is needed is described in detail in the following paragraphs. (Stepping through the algorithm, Wikipedia)

The first thing that is necessary to do is to organize the data within the spreadsheet containing the data we retrieve. We know that we want to use the teams as our labels. The format required for the executable programs is that the labels must be at the end of numerical data. In order for our map to contain the full name of each team, we must remove all spaces from the team name.

The next step is deciding what attributes are going to be used to find any clusters. I collected data for NHL goaltenders from the seasons 2003-2004 through 2007-2008 with 2004-2005 being omitted due to the strike. The data I retrieved were the average team Goals Against Average (GAA), the average team Save Percentage (SV%), and the team's goal differential (DIFF) for each year. Finally I took the average GAA, SV%, and DIFF for each team for the four seasons. This data is what I used in SOM. (Hockey Data, NHL & Yahoo Sports)

I chose to use GAA and SV% for my data, because in hockey these are the two most important statistics for goaltenders. GAA measures on average how many goals a goalie allows in a game. It is calculated by:

$$\frac{\text{Goals Allowed}}{\text{Number of Minutes Played}(1/60)}$$

SV% measures how many saves a goalie will make out of 100 shots. It is calculated by:

$$\frac{\text{Goals Allowed}}{\text{Shots Allowed}}$$

The last statistic I chose to use is DIFF. DIFF is calculated by:

Goals Scored – Goals Allowed = DIFF

The reason I am including DIFF is to attempt to remove any noise from the data. Teams that have a great goalie but a bad offense and do not make the playoffs would map higher on the map if DIFF not included. It is the same principle for teams with a very high scoring offense and a mediocre goaltender. I hypothesize that this will even my map out to cluster teams correctly.

The next step is converting it to the proper format for the executables. For each .dat file (essentially a .txt file saved as .dat), the first line should be the number of attributes (not including the label), 3 in this case. After that comes each team, with GAA, SV%, DIFF, and team name. And so on until the end with no carriage return after the last instance. This is the labeled .dat file (say nhl_label.dat). We also want an unlabeled .dat file (say nhl.dat), which is the same thing except for all of the labels are not in the file.

After this, we want to set up the proper .bat file (say nhl.bat) to execute the programs properly. The three executables referenced are randinit, vsom, and vcal. (Self-Organizing Map (SOM), Aleshunas). The .bat file should look something like this:

```
randinit -din nhl.dat -cout nhl.cod -xdim 15 -ydim 15 -topol rect -neigh bubble -rand 0
vsom -din nhl.dat -cin nhl.cod -cout nhl.cod -rlen 10000 -alpha 0.05 -radius 15
vsom -din nhl.dat -cin nhl.cod -cout nhl.cod -rlen 1000000 -alpha 0.02 -radius 5
vcal -din nhl_label.dat -cin nhl.cod -cout nhl_label.cod
```

In this example, ‘nhl.cod’ is a codebook passed between the programs, ‘xdim’ and ‘ydim’ are the dimensions of the map. This shows I am going to be using a 15x15 map. This size map allows room for a large amount of the teams to be mapped and it is not so large that it is hard to analyze. ‘rlen’ tells the program how many times to run that algorithm. ‘radius’ refers to the training radius for each instance. ‘nhl_label.cod’ is the output codebook with labels. To run the program, simply start this .bat file after making sure it is in the same folder as the ‘.exe’s and the proper .dat.

To map these results with the labels, you can either do it by hand or use som_mapper.exe. However you need to make sure of two things: there is not a carriage return at the end, and that all data points have labels even if it is just a simple ‘x’. You need to create a file named ‘control.dat’. Its contents should be:

0 nhl_label.cod nhl_output.txt

If you wanted to map the resulting attribute, you can change the 0 to a 1, 2 or 3 in this example depending on which attribute you wanted. The 1 corresponds to GAA, the 2 to SV% and, the 3 to DIFF. nhl_output.txt placed into an excel file as space delimited would give you a map that you could then color-code as you desire.

The only issue I can see arising is if the data being input into the program is not in the correct format. As I was warned against, som_mapper requires each data point to have a label. Also the user of the program must be able to make a .dat file and use an .xls file to analyze the data. If the user isn't able to, this poses an issue as to whether or not they can fully complete the problem stated.

Assumptions

Each team's goaltending statistics are accurate representations for all goaltending statistics in NHL history.
The data I retrieved was accurate information.
The algorithm performed correctly.

Results

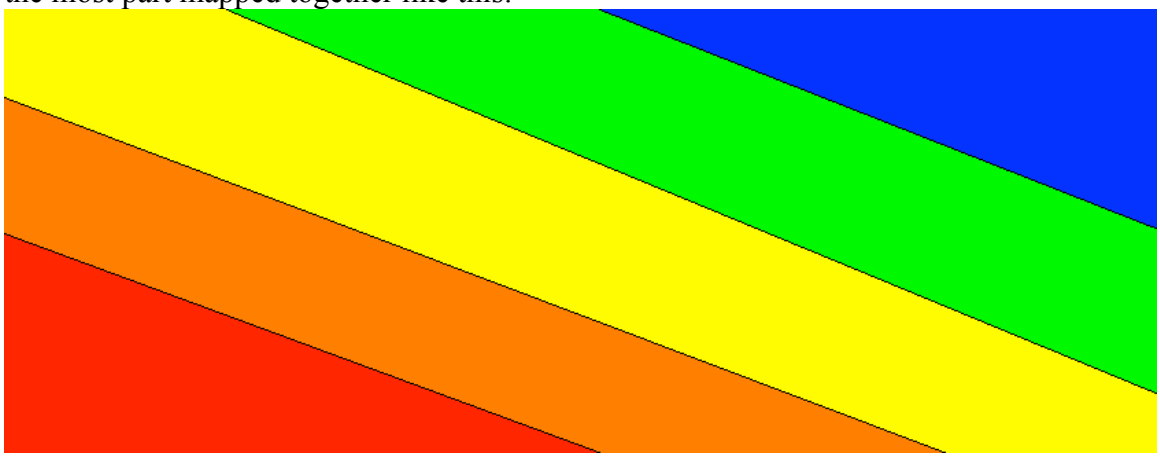
After running the three statistics through the Self-Organizing Map, the following map resulted:

Pittsburgh	x	x	x	Carolina	x	x	x	x	LosAngeles	x	x	StLouis	x	Chicago
x	x	Toronto	x	x	x	Florida	x	x	x	x	x	x	x	x
x	x	x	x	x	x	x	x	NewYorkI	x	Washington	x	x	x	x
Vancouver	x	x	TampaBay	x	x	x	x	x	x	x	x	x	Phoenix	x
x	x	x	x	x	x	x	x	x	Atlanta	x	x	x	x	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NewJersey	x	x	Calgary	x	x	x	x	x	x	x	x	x	Boston	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
x	Colorado	x	x	x	x	x	Montreal	x	x	Edmonton	x	x	x	x
x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Buffalo	x	x	x	x	x	x	x	x	x	x	x	x	x	Philadelphia
x	x	x	x	x	x	Anaheim	x	x	x	x	x	x	x	x
x	Ottawa	x	x	x	x	x	x	x	Minnesota	x	x	x	x	x
x	x	x	SanJose	x	Dallas	x	x	x	x	x	x	x	x	x
Detroit	x	x	x	x	x	x	Nashville	x	x	x	x	x	NewYorkR	x

There are five different levels of teams in this particular map. They are color-coded as follows: Red are the teams with the highest average standing, Orange, Yellow, Green, and Blue follow respectively, with Blue being the lowest set of teams. (Stats, NHL & Yahoo Sports) The list of each team, their average standing during the four seasons, and their grouping are:

	overall standings	
Detroit	1.25	high
San Jose	5.75	high
New Jersey	7	high
Ottawa	7.25	high
Dallas	7.75	high
Buffalo	10.5	medhigh
Anaheim	10.75	medhigh
Nashville	10.75	medhigh
Calgary	11.75	medhigh
Colorado	12	medhigh
Montreal	12.5	medhigh
Philadelphia	14	med
Vancouver	14	med
New York R	14.5	med
Minnesota	14.75	med
Carolina	15.25	med
Tampa Bay	15.75	med
Toronto	16	med
Boston	16.25	med
Pittsburgh	18	medlow
Edmonton	18.75	medlow
Atlanta	20	medlow
New York I	20.25	medlow
Florida	22	medlow
St. Louis	23.5	low
Washington	23.5	low
Los Angeles	24.5	low
Columbus	25.5	low
Phoenix	25.5	low
Chicago	25.75	low

The resulting map turned out to be about what was expected. All five of these levels for the most part mapped together like this:



All five of the highest level mapped to the bottom right corner: Detroit, San Jose, New Jersey, Ottawa, and Dallas. The next highest, Orange, is almost layered on top of the Red level. Yellow seems to be in between Red/Orange and Green/Blue. Green and Blue are separated but are both located in the top right of the map.

The Red level is full of teams who are expected to make the playoffs every year. All five of these teams are consistent contenders, but none of them have won the Stanley Cup in the last three seasons (not counting the 07-08 seasons). Since the Stanley Cup Playoffs

always differ from the regular season, figuring out who has the best chances of winning the Cup based on goaltending data from the season is beyond the scope of this project.

The Orange level has familiar names to Playoff hockey fans, but they do not always make it to the playoffs. They are good, but they are just not as good of a team as the Red level. The Yellow consists of teams who have been on the brink of the playoffs the past few years. These teams are usually battling for a playoff spot in the last few weeks of the season. Two of the last three Stanley Cup Champions have come out of this level of teams. Those teams are Tampa Bay and Carolina. (Stanley Cup Champions, Wikipedia)It may be that their battle at the end of the season continues throughout the duration of the playoffs, but again that is beyond the scope of this project.

The Green and Blue levels are teams that may have made the playoffs once in the past few years, but they have not been a dominant enough force to rank high enough in the standings to allow their team to appear in multiple years.

There is some overlapping within the map, but that is discussed in the Issues section, which is forthcoming.

One of the most important points to note with the map are the bottom left corner and the top right corner. Detroit, the top team in the standings, is at one extreme in the map, while Chicago, the lowest team in the standings, is at the opposite corner. Considering the fact that the standings were not included in the data mapped and the top and bottom teams were mapped as far as possible from each other is very good.

This along with the fact that the levels were mapped together is a good basis for the fact that good goaltending statistics leads the way for a team to be a playoff contender.

Issues

Overall, the only real issues that were encountered happened in the final map itself. As you can see in the map above, there are a few teams that are a bit out of place. The most apparent of these is Pittsburgh, but if you look at the overall standings list, Pittsburgh is the closest Green team to being a Yellow team. The other is Boston, and once again if you look at the overall standings, Boston is the closest Yellow team to being a Green team. The only other small issue is that Columbus didn't register on the map, however seeing how the rest of the map played out with the standings, I think it is fair to assume that it wouldn't make too large of a difference.

These small overlaps are not a problem, since the rest of the map worked out fairly well.

Appendices

For more information regarding Euclidean Distance, visit:
http://en.wikipedia.org/wiki/Euclidean_distance

For more information regarding SOM, visit:
http://en.wikipedia.org/wiki/Self-organizing_map
or
http://www.cis.hut.fi/teuvo

References

Aleshunas, John. Retrieved Apr. 17, 2008. "Self-Organizing Map (SOM)" from:
http://mercury.webster.edu/aleshunas/MATH%203210/MATH%203210%20Source%20Code%20and%20Executables.html

Goaltender's Annex. Retrieved May 5, 2008. Ubriaco Quote from:
http://www.angelfire.com/sk/goalieannex/quotes02.html

NHL.com. Retrieved Apr. 16, 2008. "Goalie Statistics and Team Standings" from:
http://www.nhl.com/nhlstats/app

Yahoo Sports. Retrieved Apr. 16, 2008. "Goalie Statistics and Team Standings" from:
http://sports.yahoo.com/nhl/teams/___/stats (Replace ___ with each team's abbreviation).

Wikipedia. Retrieved Apr. 17 2008. "Stepping through the Algorithm" from:
http://en.wikipedia.org/wiki/Self-organizing_map - Stepping_through_the_algorithm

Wikipedia. Retrieved May 1, 2008. "List of Stanley Cup Champions" from:
http://en.wikipedia.org/wiki/List_of_Stanley_Cup_champions#NHL_champion

Wikipedia. Retrieved May 6, 2008. "Euclidean Distance" from:
http://en.wikipedia.org/wiki/Euclidean_distance