SOM and MLB Stats

<MATH 3220 Final Project>
Clint Tomer

"Baseball, it is said, is only a game.
True. And the Grand Canyon is only a hole in Arizona.
Not all holes, or games, are created equal."
- George F. Will

Executive Summary

I have chosen to use the Self Organizing Map (SOM) as the algorithm for my final project. What I plan to do is use the SOM along with the hitting, fielding, and pitching statistics from each of the 2000-2006 Major League Baseball (MLB) regular seasons to see if it will correctly cluster the playoff teams together. Not only will I be looking to see if it was able to cluster the teams that made the playoffs for that particular season together, but I will also check to see if there were any other types of clustering going on. For example, if it clustered together other teams that were close to making the playoffs and teams that were at the bottom of the league. Once I run the data through SOM I will then analyze the data to see if in fact it was able to correctly cluster the data.

I am going to be using the Self Organizing Map as the algorithm for my final project. The SOM is a type of neural network that can be used to cluster data. A SOM consists of two layers of neurons; an input layer and a competition layer. The weights of the connections from the input neurons to a single neuron in the competition layer are understood to be reference vectors in the input space. This basically means that a SOM represents a set of vectors in the input space; one vector for each neuron in the competition layer (Borgelt).

The training in the SOM is utilized by competitive learning. Whenever you are training a sample, the Euclidean distance to all weight vectors is computed. The vector with the smallest Euclidean distance from the competition layer is trained so it will look more like the input vector it is closest to (Wikipedia).

The other weight vectors that are close to the one chosen with the smallest Euclidean distance are also trained, but they are not trained as much. This procedure is repeated over and over again with different input vectors and the result is a network where the closer in proximity a node is to another, the more similar the data is (Cluster Analysis).

A SOM can train data into different types of neighborhoods. Those different types are a bubble neighborhood and a Gaussian neighborhood. The first method is rectangular, where each node is connected to eight other different nodes. A second method is the hexagonal where each node is connected to only six other nodes (McKee).

In my experiment I will be using the bubble neighborhood and the rectangular method. The size of the grid that I will be using throughout this experiment will be 15x15. This was the only that I was able to get all 30 MLB teams on the grid whenever I ran the data through SOM.

I gathered my data from the Major League Baseball Website (MLB Stats). The data that I collected was from the 2000 - 2006 regular seasons. I collected statistics from three different categories: hitting, fielding, and pitching. I will be using the statistics for all 30 MLB teams for this experiment.

My experiment is going to consist of 4 parts:

Part 1: Hitting

I chose 16 different categories to run through the SOM. The different categories that I chose for the data were at-bats (AB), runs (R), hits (H), double (2B), triple (3B), homerun (HR), runs bated in (RBI), total bases (TB), walks (BB), strikeouts (SO), stolen bases (SB), on-base percentage (OBP), slugging percentage (SLG), average (AVG), hit by pitch (HBP), and on-base plus slugging percentage (OPS).

Part 2: Fielding

I chose 7 different categories to run through the SOM. The different categories that I chose for the data were total innings (INN), total chances (TC), putouts (PO), assist (A), error (E), double plays (DP), and fielding percentage (FPCT).

Part 3: Pitching

I chose 11 different categories to run through the SOM. The different categories that I chose for the data were win (W), lose (L), earned run average (ERA), complete game (CG), shut out (SHO), save (SV), hits against (H), runs against (R), walks allowed (BB), and strikeouts (SO).

Part 4: Part 1-3

For the last part I have chosen to use a total of 13 different categories to run through the SOM. I chose 4 from each of the fielding and pitching statistics and I chose 5 from the hitting statistics. I chose the following data for this part of the experiment; TC, A, E, FPCT, W, L, ERA, SV, TB, BB, SO, AVG, and OPS. I chose data that best represented each section of the data.

I will be doing this for every regular season from 2000 to the just completed 2006.

I hypothesize that my grids are going to be broken down into 3 different categories. I believe that one side will be teams that finished at the bottom of league, in the middle you are going to find the teams that were close to playoffs but just could not pull it off, and then on the other side you will have the teams that made the playoffs. I hypothesize that this is going to happen in at least 4 or 5 of the years.

The results that I received from this experiment were not exactly what I was expecting, but overall they were very impressive. I will first start with the pitching results. In 2000 and 2001, you are able to draw a box around all 8 of the playoff teams and you would only have 4 teams in 2000 and 3 teams in 2001 in each of the boxes that didn't make the playoffs. Both times the teams that played in the World Series were close to each other. In 2002, you are able to draw a box around the playoffs with 6 teams in the box that didn't make the playoffs. In 2004, you have all of the playoff teams mainly in the left corner with the exception of 5 teams. What is nice about this grid is that the farther you get away from the playoff teams, the records keep getting lower and lower for those teams. The same can go for 2006 except the playoff teams are located in the upper right and there are 4 teams that didn't make the playoffs in that corner.

The fielding results did not lead to anything really impressive. In 2001 and 2006 the teams that played in the World Series were next to each other. In 2003, the teams that were in the World Series are at extreme opposites of each other. This leads me to think that although having a strong defense doesn't necessarily mean that you are a playoff contender. You are going to have to have either great pitching or timely hitting to go along with a strong defense.

The last individual experiment I did was with the statistical data from hitting. In 2000, I was able to draw a box around all 8 playoff teams and only have 3 teams inside that didn't make the playoffs. It also grouped the teams that played in the World Series together too. In 2002 and

2004, the teams that played in the World Series were close to each other. You are also able to draw an outline of the playoff teams and only include 5 and 8 teams respectively that didn't make the playoffs. The results from 2005 almost looked like someone placed the teams where they were. First of all, on the left side of the grid are all 4 of the AL playoff teams in the same column. What else is neat about this grid is that you are able to almost draw a line between the AL and NL teams. Whenever you do this, you are only having 3 NL teams over on the AL side.

The last part of my experiment consisted of me combining all 3 parts into one grid. So I ran each year through the SOM and the results were pretty similar to what I got from the previous experiments. In 2000, I was able to draw almost an "L" shape around the playoff teams and I only had 5 other teams that in there that didn't make the playoffs. Also in the same box, were 3 teams that finished 2nd in their division. In 2001, 2002, and 2004 each had both teams that played in the World Series close to each other. Also in 2002 and 2004 you are able to separate the playoff teams with most of the teams that didn't make the playoffs. In 2003, all of the playoff teams have their closest division teams close to them.

The results that I ended up with from my experiment were very satisfying. I was able to gather a lot of information from these experiments. Out of the 28 experiments that I ran, the teams that played in the World Series were grouped together in 12 of the experiments. Another aspect that I looked at was whether or not it would group together the teams that made the playoffs. Out of the 28 experiments that I ran for this, it grouped the playoff teams together 9 times. Of those 9 times, the pitching results yielded 5 times it grouped the playoff teams together, hitting got grouped 3 times, and whenever all 3 categories were put together it grouped the playoff teams together once. As I have said earlier, pitching is one of the most important parts of the game of baseball. Hitting is also very important because if you do not score then you wont win. But you have to be able to stop the other team from scoring, so therefore you need to have great pitching and good defense.

Problem Description

I have chosen to use the Self Organizing Map (SOM) as the algorithm for my final project. What I plan to do is use the SOM along with the hitting, fielding, and pitching statistics from each of the 2000-2006 Major League Baseball (MLB) regular seasons to see if it will correctly cluster the playoff teams together. Not only will I be looking to see if it was able to cluster the teams that made the playoffs for that particular season together, but I will also check to see if there were any other types of clustering going on. For example, if it clustered together other teams that were close to making the playoffs and teams that were at the bottom of the league. Once I run the data through SOM I will then analyze the data to see if in fact it was able to correctly cluster the data.

Analysis Technique

As I mentioned above, I am going to be using the Self Organizing Map as the algorithm for my final project. The SOM is a type of neural network that can be used to cluster data. It was first described by the Finnish professor Teuvo Kohonen and it is sometimes referred to as a Kohonen map (Wikipedia). A SOM consists of two layers of neurons; an input layer and a competition layer. The weights of the connections from the input neurons to a single neuron in the competition layer are understood to be reference vectors in the input space. This basically means that a SOM represents a set of vectors in the input space; one vector for each neuron in the competition layer (Borgelt).

The training in the SOM is utilized by competitive learning. Whenever you are training a sample, the Euclidean distance to all weight vectors is computed. The vector with the smallest Euclidean distance from the competition layer is trained so it will look more like the input vector it is closest to (Wikipedia). See formula 1 for the Euclidean distance formula.

Formula 1

The other weight vectors that are close to the one chosen with the smallest Euclidean distance are also trained, but they are not trained as much. This procedure is repeated over and over again with different input vectors and the result is a network where the closer in proximity a node is to another, the more similar the data is (Cluster Analysis).

A SOM can train data into different types of neighborhoods. Those different types are a bubble neighborhood and a Gaussian neighborhood. The bubble neighborhood differs from the Gaussian neighborhood because it does not take into account the width of the neighborhood. In a bubble neighborhood, a constant training factor is applied to all nodes in the neighborhood. This means that any node within the specified distance of the selected node is trained equally (McKee).

In a Gaussian neighborhood, the training factor decreases as it gets farther from the node. The nodes that are closest to the selected node will be trained more than nodes that are further away (McKee).

There are two different methods of creating a map. The first method is rectangular, where each node is connected to eight other different nodes. A second method is the hexagonal where each node is connected to only six other nodes (McKee).

In my experiment I will be using the bubble neighborhood and the rectangular method. The size of the grid that I will be using throughout this experiment will be 15x15. This was the only that I was able to get all 30 MLB teams on the grid whenever I ran the data through SOM.

I gathered my data from the Major League Baseball Website (MLB Stats). The data that I collected was from the 2000 - 2006 regular seasons. I collected statistics from three different categories: hitting, fielding, and pitching. I will be using the statistics for all 30 MLB teams for this experiment.

My experiment is going to consist of 4 parts:

Part 1: Hitting

I chose 16 different categories to run through the SOM. The different categories that I chose for the data were at-bats (AB), runs (R), hits (H), double (2B), triple (3B), homerun (HR), runs bated in (RBI), total bases (TB), walks (BB), strikeouts (SO), stolen bases (SB), on-base percentage (OBP), slugging percentage (SLG), average (AVG), hit by pitch (HBP), and on-base plus slugging percentage (OPS). The reason why I chose all of these statistics was because I believe that the combination of these 16 truly represent the data. Once I have trained the map using the hitting data, I will then place the names of the teams on the map. I will then analyze the map and see where it places all 30 teams. I will be looking for things like the grouping of playoff teams, teams that were close to making the playoffs, teams that made the World Series, and teams that were at the bottom of the league.

Part 2: Fielding

I chose 7 different categories to run through the SOM. The different categories that I chose for the data were total innings (INN), total chances (TC), putouts (PO), assist (A), error (E), double plays (DP), and fielding percentage (FPCT). The reason why I chose all of these statistics was because I believe that the combination of these 7 truly represent the data. Once I have trained the map with the fielding data, I will then place the names of the teams on the map. I will then analyze the map and see where it places all 30 teams. I will be looking for things like the grouping of playoff teams, teams that were close to making the playoffs, teams that made the World Series, and teams that were at the bottom of the league.

Part 3: Pitching

I chose 11 different categories to run through the SOM. The different categories that I chose for the data were win (W), lose (L), earned run average (ERA), complete game (CG), shut out (SHO), save (SV), hits against (H), runs against (R), walks allowed (BB), and strikeouts (SO). The reason why I chose all of these statistics was because I believe that the combination of these 11 truly represent the data. Once I have trained the map with the pitching data, I will then place the names of the teams on the map. I will then analyze the map and see where it places all 30 teams. I will be looking for things like the grouping of playoff teams, teams that were close to making the playoffs, teams that made the World Series, and teams that were at the bottom of the league.

Part 4: Part 1-3

For the last part I have chosen to use a total of 13 different categories to run through the SOM. I chose 4 from each of the fielding and pitching statistics and I chose 5 from the hitting statistics. I chose the following data for this part of the experiment; TC, A, E, FPCT, W, L, ERA, SV, TB, BB, SO, AVG, and OPS. The reasoning behind choosing these 13 categories was because they all summarize the data from each part of the experiment the best. Once I have trained the map with all of my data, I will then place the names of the teams on the map. I will then analyze the data and see where it has placed all 30 teams based on all 3 of the categories.

I will be doing this for every regular season from 2000 to the just completed 2006. I hypothesize that my maps are going to be broken down into 3 different categories. I believe that one side will be teams that finished at the bottom of league, in the middle you are going to find the teams that were close to playoffs but just could not pull it off, and then on the other side you will have the teams that made the playoffs. I hypothesize that this is going to happen in at least 4-5 of the years.

Assumptions

- Each part of the experiment is a true representation of the whole data set.
- The algorithm preformed correctly.
- The data I entered into excel for the algorithm was entered correctly.

Results

The results that I received from this experiment were not exactly what I was expecting, but overall they were very impressive.

Pitching Results

I will first start with the pitching results. In 2000 (see chart 1) and 2001 (see chart 2), you are able to draw a box around all 8 of the playoff teams and you would only have 4 teams in 2000 and 3 teams in 2001 in each of the boxes that didn't make the playoffs. Both times the teams that played in the World Series were close to each other

Chart 1

Chart 2

In 2002 (see chart 3), you are able to draw a box around the playoffs with 6 teams in the box that didn't make the playoffs

Chart 3

In 2004 (see Chart 4), you have all of the playoff teams mainly in the left corner with the exception of 5 teams. What is nice about this grid is that the farther you get away from the playoff teams, the records keep getting lower and lower for those teams. The same can go for 2006 (see chart 5) except the playoff teams are located in the upper right and there are 4 teams that didn't make the playoffs in that corner.

Chart 4

Chart 5

The results from the pitching part of the experiment explain why pitching is such an important part of the game of baseball. As the experiment confirms this, out of the 7 experiments done it grouped the playoff teams together 5 times.

Fielding Results

The fielding results did not lead to anything really impressive. In 2001 (see chart 6) and 2006 (see chart 7) the teams that played in the World Series were next to each other.

Chart 6

Chart 7

In 2003 (see chart 8), the teams that were in the World Series are at extreme opposites of each other.

Chart 8

The results from this part of the experiment lead me to think that although having a strong defense doesn't necessarily mean that you are a playoff contender. You are going to have to have either great pitching or timely hitting to go along with a strong defense.

Hitting Results

The last individual experiment I did was with the statistical data from hitting. In 2000 (see chart 9), I was able to draw a box around all 8 playoff teams and only have 3 teams inside that didn't make the playoffs. It also grouped the teams that played in the World Series together too.

Chart 9

In 2002 (see chart 10) and 2004 (see chart 11), the teams that played in the World Series were close to each other. You are also able to draw an outline of the playoff teams and only include 5 and 8 teams respectively that didn't make the playoffs.

Chart 10

Chart 11

The results from 2005 (see chart 12) almost looked like someone placed the teams where they were. First of all, on the left side of the grid are all 4 of the AL playoff teams in the same column. What else is neat about this grid is that you are able to almost draw a line between the AL and NL teams. Whenever you do this, you are only having 3 NL teams over on the AL side.

Chart 12

The results from this experiment yielded me to believe that hitting is a very important part of the game of baseball, but it is not as important as pitching. You might be able to score 10 runs a game, but if you don't have good pitching you will give up 15 runs a game. So even though hitting is very important, I believe that pitching is just a little bit more important.

<u>All 3 Categories Combined</u>

The last part of my experiment consisted of me combining all 3 parts into one grid. So I ran each year through the SOM and the results were pretty similar to what I got from the previous experiments. In 2000 (see chart 13), I was able to draw almost an "L" shape around the playoff teams and I only had 5 other teams that in there that didn't make the playoffs. Also in the same box, were 3 teams that finished 2nd in their division.

Chart 13

In 2001(see chart 14) it grouped both teams that played in the World Series close to each other.

Chart 14

Also in 2002 (see chart 15) and 2004 (see chart 16) you are able to separate the playoff teams with most of the teams that didn't make the playoffs, it also grouped both teams that made the World Series.

Chart 15

Chart 16

In 2003 (see chart 17), the results were different than any other results that I had gotten in the previous experiments. It didn't group the playoff teams together, but it did group the divisional races together. Teams that were within 6 games of the division leader were close to each other.

Chart 17

The results that I ended up with from my experiment were very satisfying. I was able to gather a lot of information from these experiments. Out of the 28 experiments that I ran, the teams that played in the World Series were grouped together in 12 of the experiments. Another aspect that I looked at was whether or not it would group together the teams that made the playoffs. Out of the 28 experiments that I ran for this, it grouped the playoff teams together 9 times. Of those 9 times, the pitching results yielded 5 times it grouped the playoff teams together, hitting got grouped 3 times, and whenever all 3 categories were put together it grouped the playoff teams together once. As I have said earlier, pitching is one of the most important parts of the game of baseball. Hitting is also very important because if you do not score then you wont win. But you have to be able to stop the other team from scoring, so therefore you need to have great pitching and good defense.

Issues

The only thing that happened with my experiment was the fact that my data didn't cluster the way I thought it would. I was expecting SOM to cluster it very nicely for me, it did cluster the data to some extent but just not as good as I would have liked for it to be.

Appendices

- *Cluster Analysis.* (n.d.). Retrieved December 6, 2006 from http://www2.chass.ncsu.edu/garson/pa765/cluster.htm
- *Self-Organizing Map.* (n.d.). Retrieved December 6, 2006 from http://en.wikipedia.org/wiki/Self_organizing_map
- Borgelt, Christian. (n.d.). *Self-Organizing Map Training Visualization.* Retrieved December 6, 2006 from http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/somd/
- McKee, Kevin. (n.d.). *The Self-Organized Map Applied to 2005 NFL Quarterbacks.* Retrieved December 6, 2006 from http://mercury.webster.edu/aleshunas/MATH%203220/MATH%203220%20Course%20Support%20Materials.html
- *Major League Baseball Website for stats.* (n.d.). Retrieved December 6, 2006 from http://mlb.mlb.com/NASApp/mlb/stats/sortable_team_stats.jsp?c_id=mlb
- *Major League Baseball Website for Playoff Teams.* (n.d.). Retrieved December 6, 2006 from http://mlb.mlb.com/NASApp/mlb/mlb/schedule/ps_03,04,05,06.jsp
- *CBS Sportsline Website for Playoff Teams.* (n.d.). Retrieved December 6, 2006 from http://cbs.sportsline.com/mlb/postseason/pastresults/
- Heldt, S & Kreismer, J. *Baseball Almanac*. 2007. Red-Letter Press Inc. Sattle River, NJ