The Self Organized Map Applied to 2005 NFL Quarterbacks

By Kevin McKee

Executive Summary

A self organizing map is a clustering algorithm that takes data and creates a two dimensional map by training each node based on the data that it is given. Once the map is created, the self organizing map (SOM) can then assign an individual set of data to a specific node, based on which node the set of data is closest to.

I took 14 different playing statistics from NFL Quarterbacks in the 2005 season and trained the SOM with the described data. I then allowed the SOM to create a map that displays the name of each quarterback. The goal of this experiment is to find out whether or not the SOM can effectively group the players, and if it can, discover what we can learn about the players based on the SOM clustering.

I ran the data and the players through the self organizing map, and the result was the following map:

DavidCarr				JPLosman		MichaelVick			DavidGarrard			JeffGarcia	
		AlexSmith						MikeMcMahon					
KellyHolcomb		CharlieFrye				AaronBrooks					BrianGriese		
								BrooksBollinger					
		KyleBoller		JoshMcCown					AnthonyWright			JamieMartin	
DaunteCulpepper													
							JoeyHarrington						
			TrentDilfer			ChrisSimms			KyleOrton			BradJohnson	
	KurtWarner												
						MarkBrunell	GusFrerotte		ByronLeftwich		BenRoethlisberger		
MarcBulger			DrewBledsoe		JakeDelhomme								JakePlummer
	DonovanMcNabb						DrewBrees		SteveMcNair			MattHasselbeck	
KerryCollins			BrettFavre			EliManning			TrentGreen		TomBrady		CarsonPeytonManning

After analyzing the map and why players mapped to the places that they did, I came to the conclusion that the SOM did effectively group the players, and these are the groups I believe the players were placed into:

Each different section tells about a quarterback's style of play, the type of offense he is in, and/or how well he plays.

Problem Description

The problem I am attempting to tackle is to see whether or not the self organizing map can effectively cluster a group of quarterbacks from the National Football League based on playing statistics. The goal is to run the statistics and players through the algorithm, and then analyze the output map and find out how well the players were grouped. Once that is done, I will conduct a second experiment to see if the SOM can take non playing statistics and group quarterbacks in the same way.

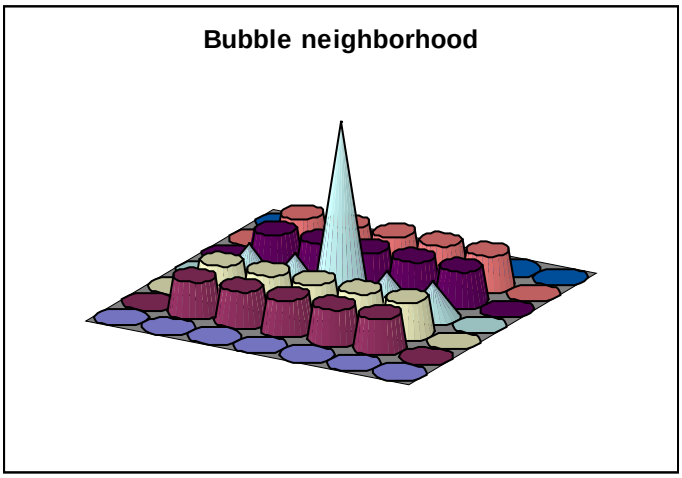
Analysis Technique

A Self-Organized Map (SOM) is a neural network used for clustering data onto a two dimensional map. The SOM consists of an input layer and a competition layer of neurons. Reference vectors are created by taking the weights of the connections from the input layer to a single neuron in the competition layer. The SOM represents a set of vectors in the input space: one vector for each neuron in the competition layer. During the SOM, the program picks the weight vector that has the smallest Euclidian distance from the competition layer and trains that vector to more closely resemble the input vector it is closest to. The Euclidian distance is defined by the following formula:

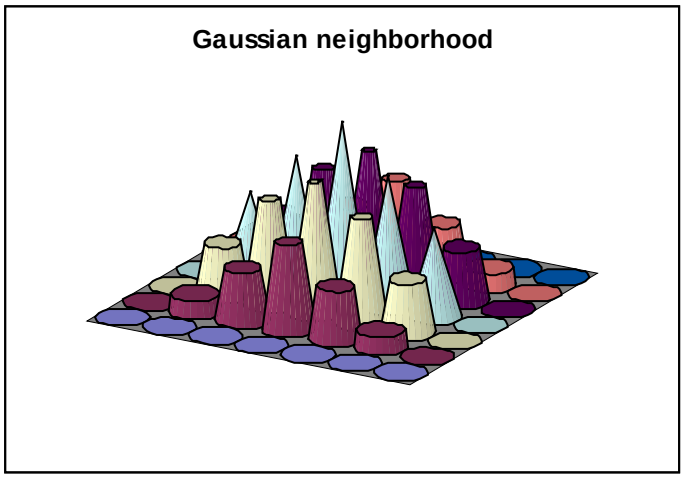
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

In this formula, n = the number of columns of data, p_i is the value of the chosen vector and q_i is the value of the vector that is being compared. The other weight vectors that are close to the one chosen with the smallest Euclidian distance are also trained, though not 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; thus a map is trained.

A SOM can train data into different types of neighborhoods: a bubble neighborhood and a Gaussian neighborhood. The bubble neighborhood differs from the Gaussian neighborhood because the bubble neighborhood 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.



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.



There are two different methods of creating a map. The first is rectangular, where each node is connected to eight other different nodes. A second topology is the hexagonal topology where each node is connected to only six other nodes. In my experiment I will use the bubble neighborhood and the rectangular topology.

I downloaded data containing statistics of NFL quarterbacks during the 2005-2006 regular season. The different categories are Passer Rating, Completions, Attempts, Yards, Passing Touchdowns, Interceptions, Rushes, Rushing Yards, Rushing Touchdowns, Sacks, Yards Lost (from sacks), Fumbles and Fumbles Lost.

I choose only those quarterbacks who had thrown at least 150 passes last season, giving me 42 players. Some of the players had played in less than 16 games, so I normalized their statistics so that it is as if all players played in 16 games that year. For example, if a player played in only 9 games, I multiplied all of their attributes by 16/9 so that their numbers would be on the same level as all the other players. Lastly, to prepare the data I took the minimum and maximum values from each column, then normalized each number by subtracting the minimum from each value, then dividing by the maximum minus the minimum. This formula

$$(X-min)/(max-min) \qquad (2)$$

normalized each value X to a corresponding number between zero and one. This way, the passing yardage column with numbers in the thousands will not dominate the rest of the data.

Once the map is trained with the quarterback data, I will have SOM place the names of the quarterbacks onto the map. I will use a 15 by 15 square map in the hopes that the map is big enough that no quarterback will map over another, while it will be small enough to easily analyze. I hypothesize that quarterbacks who play in similar offensive schemes and have comparable success will map together. I also believe that the "pocket" passers, or quarterbacks who prefer to throw on every play will not map closely to the mobile players who have a tendency to run with the football. Another hypothesis of mine is that players who have been in the league for about the same number of years will map together. There are many relations that I can test after the algorithm is run and the players are mapped to find out why players mapped where they did. However, the most important aspect of this experiment is that the players map in groups first. Once they map into groups, those groups can be analyzed later.

Assumptions

- The data from the 42 NFL quarterbacks I used is representative of all NFL quarterbacks.
- There were no flaws in the algorithm.
- The data sets I was given were contained accurate and precise data.

- Normalizing the data did not compromise the integrity of the numbers.
- When normalizing, it is fair to assume that I can normalize the data from players missed games and still have an accurate set of data
-

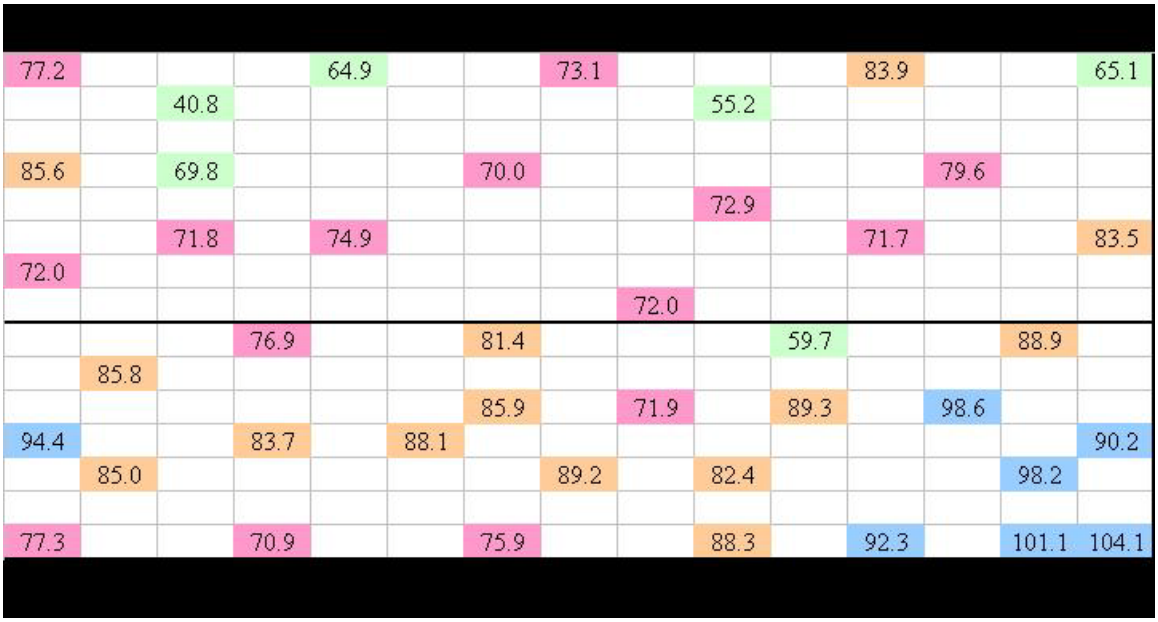
Results

I ran the data through the self organizing map, and the algorithm produced the following map:

DavidCarr				JPLosman		MichaelVick			DavidGarrard		JeffGarcia			
		AlexSmith						MikeMcMahon						
KellyHolcomb		CharlieFrye				AaronBrooks				BrianGriese				
								BrooksBollinger						
		KyleBoller		JoshMcCown					AnthonyWright		JamieMartin			
DaunteCulpepper														
							JoeyHarrington							
			TrentDilfer			ChrisSimms			KyleOrton			BradJohnson		
	KurtWarner													
						MarkBrunell		GusFrerotte	ByronLeftwich		BenRoethlisberger			
MarcBulger			DrewBledsoe		JakeDelhomme								JakePlummer	
	DonovanMcNabb						DrewBrees		SteveMcNair			MattHasselbeck		
KerryCollins			BrettFavre			EliManning			TrentGreen		TomBrady		Carson	PeytonManning

This is a 15 by 15 grid containing the names of all 42 players entered into the algorithm. It takes someone with a considerable amount of knowledge about each player, their teams and other factors to be able to break these players up into groups, as they appear to be scattered about the map almost evenly. However, to analyze these results I decided first to see if it grouped the good players together and the bad players together. I did this by replacing the names of the quarterbacks with their passer rating.

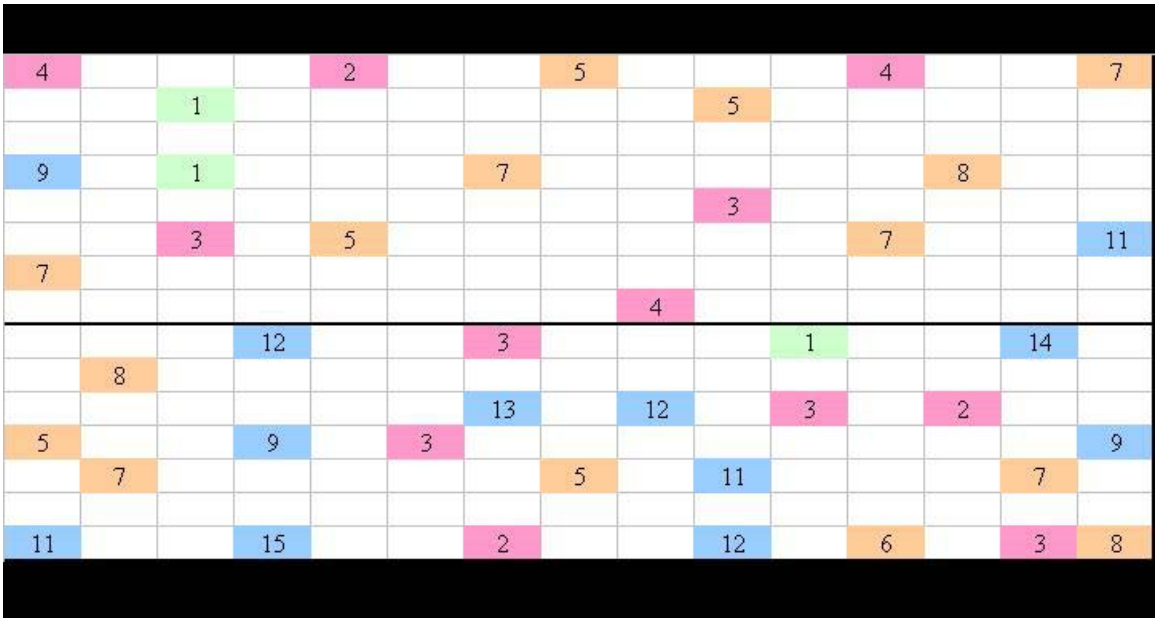
A passer rating is a mathematical formula that takes into account completion percentage, yards per attempt, touchdowns per attempt, and interceptions per attempt. Once the names were replaced with passer ratings, I colored any passer rating over 90 blue, between 80 and 90 orange, between 70 and 80 pink, and lower than 70 green. Here is the resulting map:



I added the line below the eighth row because that is where the map seems to change the most. I have kept this line consistent in all of the following graphs.

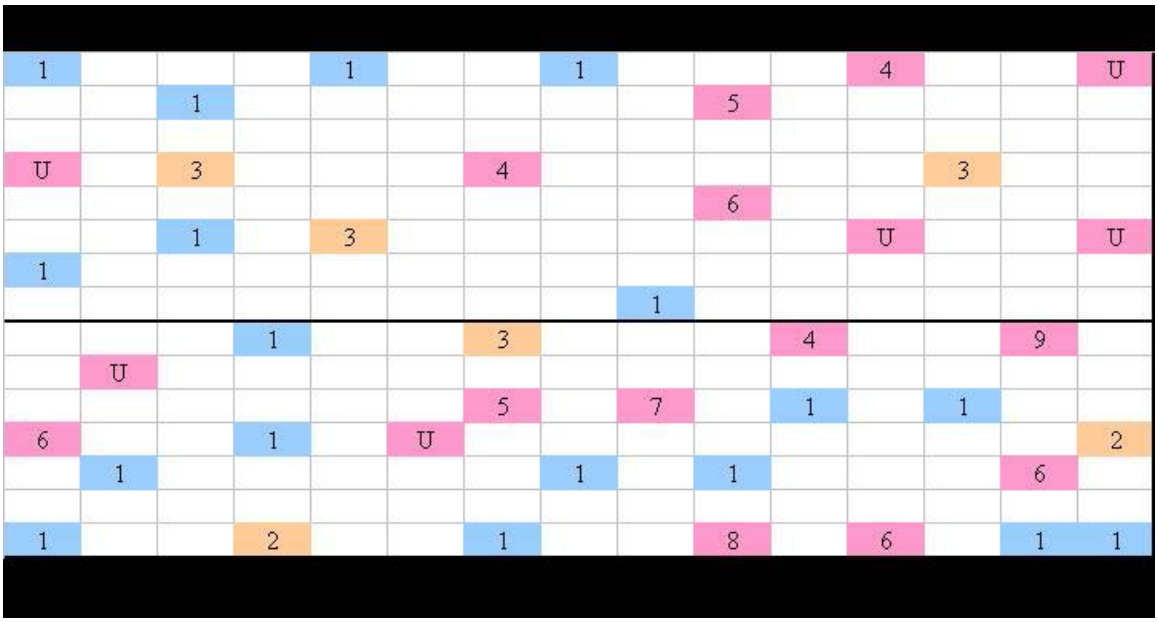
As you can see, most of the players with a passer rating above 80 grouped to the bottom half of the map. It is also apparent that the passer ratings over 90 mapped to the bottom right corner, with one exception. It appears that the SOM grouped these players well according to the passer rating. Of course there are a few values that seem to be misplaced, but overall it did a good job.

Next I wanted to see if players who have been in the league for about the same amount of time had mapped to the same places. From the original map, I replaced the names with the number of years that player had been in the league and came up with the following map:



This map is a lot less definitive than the passer rating map. There are low and mid range numbers all over the map. However, one consistent theme is that the players with nine or more years of experience were placed at the bottom of the map. Another prevalent theme is that players with four or less years of experience mapped to the top left corner. There is one exception where a player who has been in the league for nine years mapped to that spot, but further inspection will find that player to be Kelly Holcomb, who has been a career backup and has relatively little actual NFL game experience.

My next task was to see if players who were taken in the same round of the NFL draft would map together. Here are the results:



This map gives us a little more information about where the players lie. As you can see, the players who were drafted in the first round are all over the map, except for the top right corner. It also seems that the players drafted on the second day (rounds four through nine) or who went undrafted (U) mapped to the right side of the map. This gives us some information about the players, but the only real definitive group is the upper right corner. This did not give as much information as I had hoped it would.

Lastly, I wanted to find a random distribution so I replaced the player names with their height. The following is the result:

Surprisingly enough, height was the best indicator of all that mapped players into the bottom right corner. In fact, not one player who measures six feet, five inches mapped above the line. Conversely, the top right corner is full of players who are between six feet and six feet two inches. There is an obvious benefit of being tall because a taller player can see over the offensive line, which is normally made up of players above six feet, six inches tall. Nonetheless, it is surprising that height has such a strong correlation with how well players perform.

After analyzing all of these maps, I came to the conclusion that the SOM did effectively group the players. I came up with the following groups:

- The top left corner of the map consists mostly of players who don't have a lot of experience and in general are not very good. As they spend more time in the league, they should move out of that corner. This section is full of young players; however, that does not mean all young players map there, it only means older players generally did not map into that section.
- The quarterbacks who have a tendency to run with the football often grouped to the same area at the top of the map. Coincidently, quarterbacks who run often are average passers, which is why they map between the inexperienced players and the backups.
- Backup players mapped to the top right. These players are normally of average height, were drafted in the later rounds and have been in the league for a while. This is one of the most defined areas of the map as almost all players share the same characteristics.
- The best players mapped to the bottom.
- The elite players mapped in the bottom right corner. These players are great players but play on balanced teams where they are not asked to do too much.
- The high octane group differs from the elite group as they throw more passes for more yards, but also more interceptions and mistakes. They are good players, but play in different systems than the elite group.
- The very good players mapped in between the high octane and the elite groups, where they are not asked to do too much and they play well. However, they are not quite as good as the players to the right of them.
- The last section I labeled "Prone to Mistakes" because these are the players that are moving down into the sections of the good players, but aren't quite there yet. They make mistakes and keep themselves from joining the rest of the good players.

Supplemental Experiment

After I did all this work, I was curious if I could use the non playing data, such as round drafted, height, and years of experience and have the SOM create a map that would be similar to the one created from the playing statistics. I ran these through the algorithm and found the players mapped as follows, shown by their passer rating:

My hypothesis was partially correct. It appears that the players are beginning to group together. The players with low passer ratings mapped to the top left while players in the 70 to 80 range generally mapped to the bottom left. The players with high passer ratings seemed to be moving towards the right-middle of the graph and the players between the ranges of 80 to 90 mapped all over the place. I believe that if more non playing data is isolated and is found to relate to how well players perform, it is possible to predict how well a player will do using only non playing statistics. However, proving or disproving this hypothesis is beyond the scope of this project.

Issues

The players did separate into groups, but the groups were not as strictly defined as I hoped they would be. From the second hypothesis I created, it would be a very daunting task to find the categories that could be used to predict where a quarterback will fall without using actual playing data, and may not be possible at all.

Appendices

Aleshunas, J.J., Daniel C. St. Clair and W.E. Bond. "Classification Characteristics of SOM and ART2."

Borgelt, Christian. "Self-Organizing Map Training Visualization." http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/somd/

Rauber, Andreas and Dieter Merkl. "Automatic Labeling of Self-Organizing Maps: Making a Treasure-Map Reveal It's Secrets." http://www.ifs.tuwien.ac.at/ifs/research/pub_html/rau_pakdd99/rau_pakdd99.html

"Stats." 2005 NFL Statistics. Access Date: 4/21/06. http://sports.yahoo.com/nfl/stats;_ylt=AgV3dfAhsMDkpLXqjgU3s4ZDubYF