

CHAPTER ELEVEN: NEURAL NETWORKS

CONTEXT AND PERSPECTIVE

Juan is a statistical performance analyst for a major professional athletic team. His team has been steadily improving over recent seasons, and heading into the coming season management believes that by adding between two and four excellent players, the team will have an outstanding shot at achieving the league championship. They have tasked Juan with identifying their best options from among a list of 59 experienced players that will be available to them. All of these players have experience, some have played professionally before and some have many years of experience as amateurs. None are to be ruled out without being assessed for their potential ability to add star power and productivity to the existing team. The executives Juan works for are anxious to get going on contacting the most promising prospects, so Juan needs to quickly evaluate these athletes' past performance and make recommendations based on his analysis.

LEARNING OBJECTIVES

After completing the reading and exercises in this chapter, you should be able to:

- Explain what a neural network is, how it is used and the benefits of using it.
- Recognize the necessary format for data in order to perform neural network data mining.
- Develop a neural network data mining model in RapidMiner using a training data set.
- Interpret the model's outputs and apply them to a scoring data set in order to deploy the model.

ORGANIZATIONAL UNDERSTANDING

Juan faces high expectations and has a delivery deadline to meet. He is a professional, he knows his business and knows how important the intangibles are in assessing athletic talent. He also

knows that those intangibles are often manifest by athletes' past performance. He wants to mine a data set of all current players in the league in order to help find those prospects that can bring the most excitement, scoring and defense to the team in order to reach the league championship. While salary considerations are always a concern, management has indicated to Juan that their desire is to push for the championship in the upcoming season, and they are willing to do all they can financially to bring in the best two to four athletes Juan can identify. With his employers' objectives made clear to him, Juan is prepared to evaluate each of the 59 prospects' past statistical performance in order to help him formulate what his recommendations will be.

DATA UNDERSTANDING

Juan knows the business of athletic statistical analysis. He has seen how performance in one area, such as scoring, is often interconnected with other areas such as defense or fouls. The best athletes generally have strong connections between two or more performance areas, while more typical athletes may have a strength in one area but weaknesses in others. For example, good role players are often good defenders, but can't contribute much scoring to the team. Using league data and his knowledge of and experience with the players in the league, Juan prepares a training data set comprised of 263 observations and 19 attributes. The 59 prospective athletes Juan's team could acquire form the scoring data set, and he has the same attributes for each of these people. We will help Juan build a **neural network**, which is a data mining methodology that can predict categories or classifications in much the same way that decision trees do, but neural networks are better at finding the strength of connections between attributes, and it is those very connections that Juan is interested in. The attributes our neural network will evaluate are:

- **Player_Name:** This is the player's name. In our data preparation phase, we will set its role to 'id', since it is not predictive in any way, but is important to keep in our data set so that Juan can quickly make his recommendations without having to match the data back to the players' names later. (Note that the names in this chapter's data sets were created using a random name generator. They are fictitious and any similarity to real persons is unintended and purely coincidental.)
- **Position_ID:** For the sport Juan's team plays, there are 12 possible positions. Each one is represented as an integer from 0 to 11 in the data sets.
- **Shots:** This the total number of shots, or scoring opportunities each player took in their most recent season.

- **Makes:** This is the number times the athlete scored when shooting during the most recent season.
- **Personal_Points:** This is the number of points the athlete personally scored during the most recent season.
- **Total_Points:** This is the total number of points the athlete contributed to scoring in the most recent season. In the sport Juan's team plays, this statistic is recorded for each point an athlete contributes to scoring. In other words, each time an athlete scores a personal point, their total points increase by one, and every time an athlete contributes to a teammate scoring, their total points increase by one as well.
- **Assists:** This is a defensive statistic indicating the number of times the athlete helped his team get the ball away from the opposing team during the most recent season.
- **Concessions:** This is the number of times the athlete's play directly caused the opposing team to concede an offensive advantage during the most recent season.
- **Blocks:** This is the number of times the athlete directly and independently blocked the opposing team's shot during the most recent season.
- **Block_Assists:** This is the number of times an athlete collaborated with a teammate to block the opposing team's shot during the most recent season. If recorded as a block assist, two or more players must have been involved. If only one player blocked the shot, it is recorded as a block. Since the playing surface is large and the players are spread out, it is much more likely for an athlete to record a block than for two or more to record block assists.
- **Fouls:** This is the number of times, in the most recent season, that the athlete committed a foul. Since fouling the other team gives them an advantage, the lower this number, the better the athlete's performance for his own team.
- **Years_Pro:** In the training data set, this is the number of years the athlete has played at the professional level. In the scoring data set, this is the number of year experience the athlete has, including years as a professional if any, and years in organized, competitive amateur leagues.
- **Career_Shots:** This is the same as the Shots attribute, except it is cumulative for the athlete's entire career. All career attributes are an attempt to assess the person's ability to perform consistently over time.
- **Career_Makes:** This is the same as the Makes attribute, except it is cumulative for the athlete's entire career.

- **Career_PP:** This is the same as the Personal Points attribute, except it is cumulative for the athlete's entire career.
- **Career_TP:** This is the same as the Total Points attribute, except it is cumulative for the athlete's entire career.
- **Career_Assists:** This is the same as the Career Assists attribute, except it is cumulative for the athlete's entire career.
- **Career_Con:** This is the same as the Career Concessions attribute, except it is cumulative for the athlete's entire career.
- **Team_Value:** This is a categorical attribute summarizing the athlete's value to his team. It is present only in the training data, as it will serve as our label to predict a Team_Value for each observation in the scoring data set. There are four categories:
 - **Role Player:** This is an athlete who is good enough to play at the professional level, and may be really good in one area, but is not excellent overall.
 - **Contributor:** This is an athlete who contributes across several categories of defense and offense and can be counted on to regularly help the team win.
 - **Franchise Player:** This is an athlete whose skills are so broad, strong and consistent that the team will want to hang on to them for a long time. These players are of such a talent level that they can form the foundation of a really good, competitive team.
 - **Superstar:** This is that rare individual whose gifts are so superior that they make a difference in every game. Most teams in the league will have one such player, but teams with two or three always contend for the league title.

Juan's data are ready and we understand the attributes available to us. We can now proceed to...

DATA PREPARATION

Access the book's companion web site and download two files: Chapter11DataSet_Training.csv and Chapter11DataSet_Scoring.csv. These files contain the 263 current professional athletes and the 59 prospects respectively. Complete the following steps:

- 1) Import both Chapter 11 data sets into your RapidMiner repository. Be sure to designate the first row as attribute names. You can accept the defaults for data types. Save them

with descriptive names, then drag them and drop them into a new main process window. Be sure to rename the retrieve objects as Training and Scoring.

- 2) Add three Set Role operators; two to your training stream and one to your scoring stream. Use the first in the training stream to set the Player_Name attribute's role to 'id', so it will not be included in the neural network's prediction calculations. Do the same for the Set Role attribute in the scoring stream. Finally, use the second Set Role attribute in the training stream to set the Team_Value attribute as the 'label' for our model. When you are finished with steps 1 and 2, your process should look like Figure 11-1.

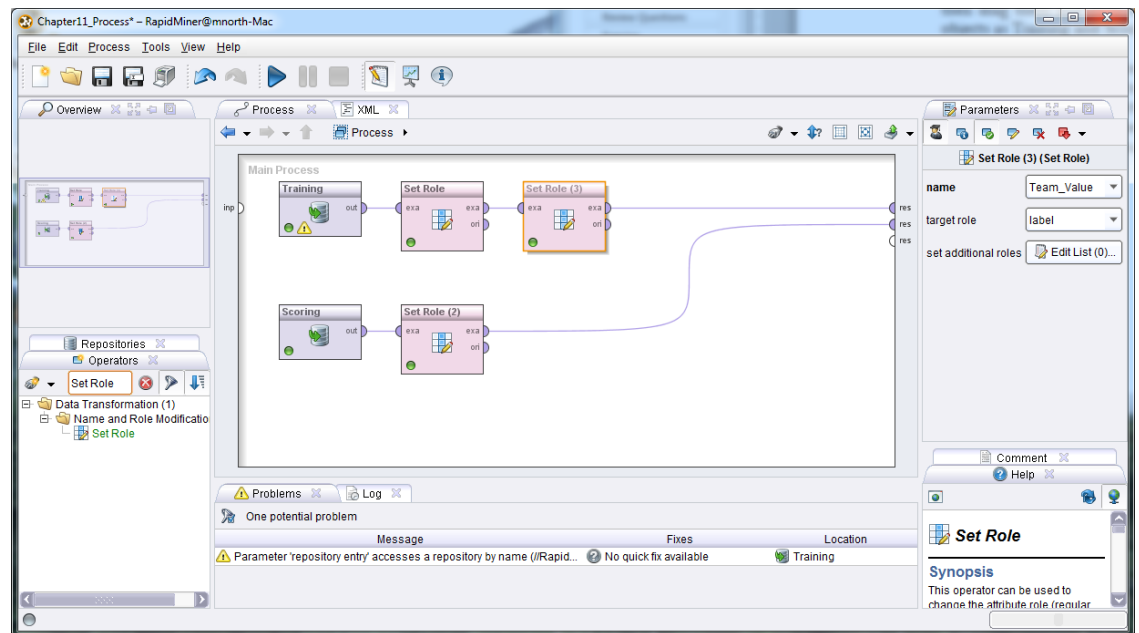


Figure 11-1. Data preparation for neural network analysis.

- 3) Go ahead and run the model. Use the meta data view for each of the two data sets to familiarize yourself with the data. Ensure that your special attributes have their roles set as they should, in accordance with the parameters you configured in step 2 (see Figures 11-2 and 11-3 which show meta data).

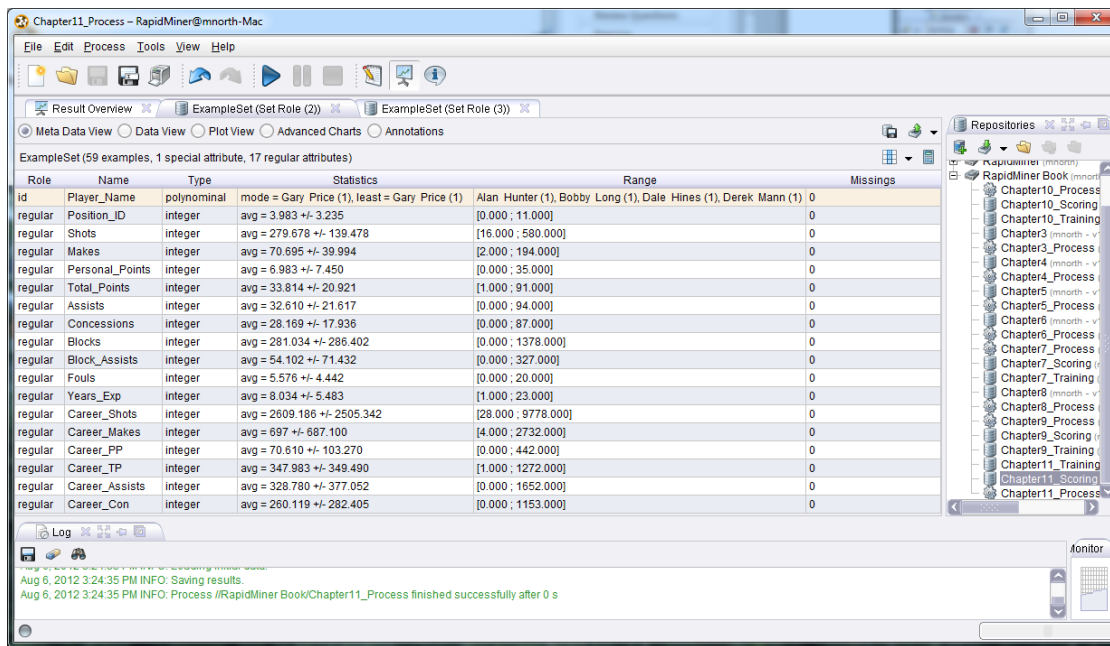


Figure 11-2. The scoring data set's meta data with special attribute Player_Name designated as an 'id'.

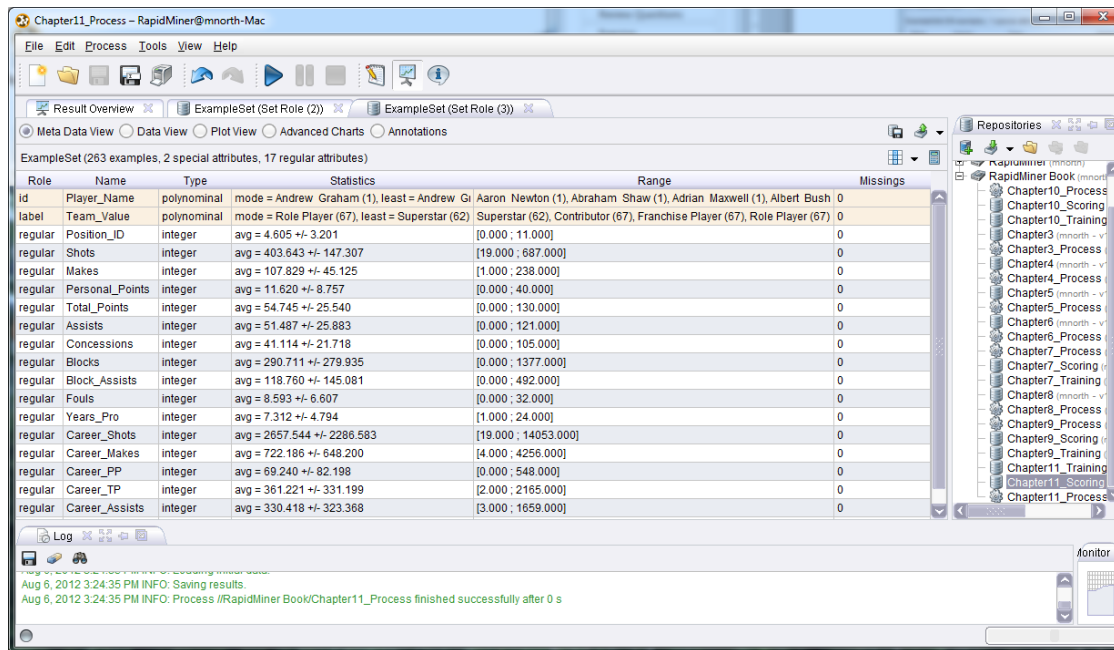


Figure 11-3. The training data set with two special attributes: Player_Name ('id') and Team_Value ('label').

- As you review the data sets, note that these two have one characteristic that is unique from prior example data sets: the ranges for the scoring data sets are not within the ranges for the training data set. Neural network algorithms, including the one used in RapidMiner,

often employ a concept called **fuzzy logic**, which is an inferential, probability-based approach to data comparisons allowing us to *infer*, based on probabilities, the strength of the relationship between attributes in our data sets. This gives us added flexibility over some of the other predictive data mining techniques previously shown in this book. Having reviewed the data sets' meta data, return to design perspective so that we can continue with...

MODELING

- 5) Using the search field on the Operators tab, locate the Neural Net operator and add it to your training stream. Use Apply Model to apply your neural network to your scoring data set. Be sure both the *mod* and *lab* ports are connected to *res* ports (Figure 11-4).

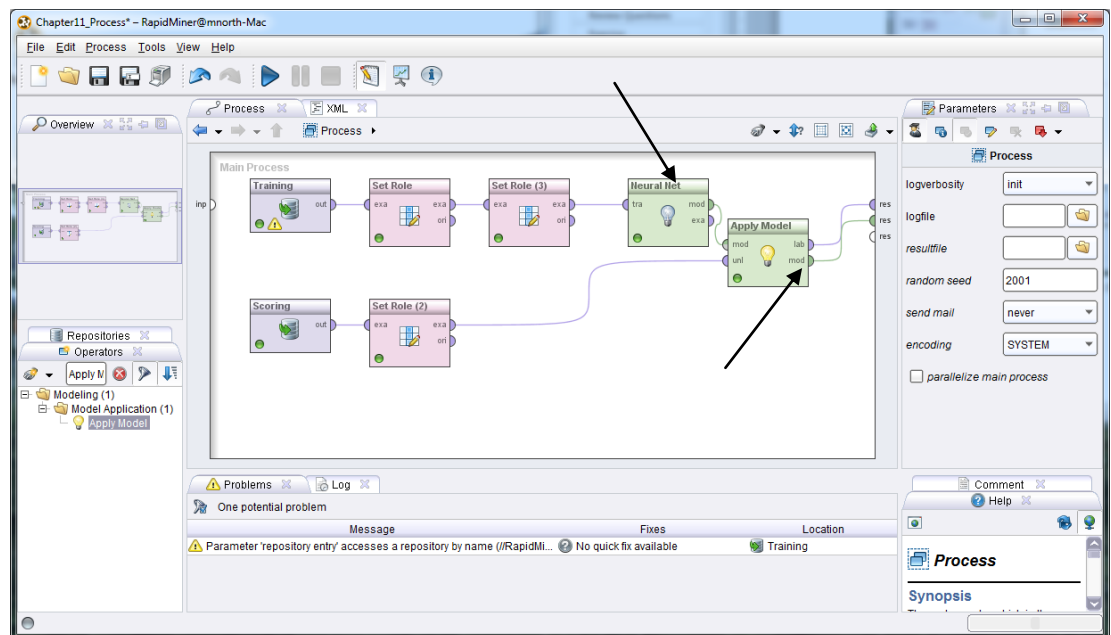


Figure 11-4. Generating a neural network model and applying it to our scoring data set.

Run the model again. In results perspective, you will find both a graphical model and our predictions. At this stage we can begin our...

EVALUATION

Neural networks use what is called a 'hidden layer' to compare all attributes in a data set to all other attributes. The circles in the neural network graph are nodes, and the lines between nodes

are called neurons. The thicker and darker the neuron is between nodes, the stronger the affinity between the nodes. The graph begins on the left, with one node for each predictor attribute. These can be clicked on to reveal the attribute name that each left-hand node represents. The hidden layer performs the comparison between all attributes, and the column of nodes on the right represent the four possible values in our predicted (label) attribute: Role_Player, Contributor, Franchise Player, or Superstar.

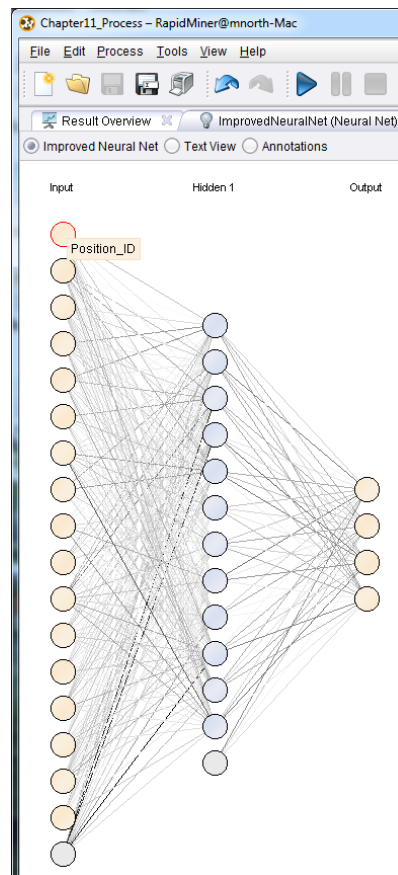


Figure 11-5. A graphical view of our neural network showing different strength neurons and the four nodes for each of the possible Team_Value categories.

Switch to the ExampleSet tab in results perspective. Again, as with past predictive models, we can see that four new special attributes have been generated by RapidMiner. Each of our 59 athlete prospects have a prediction as to their Team_Value category, with accompanying confidence percentages.

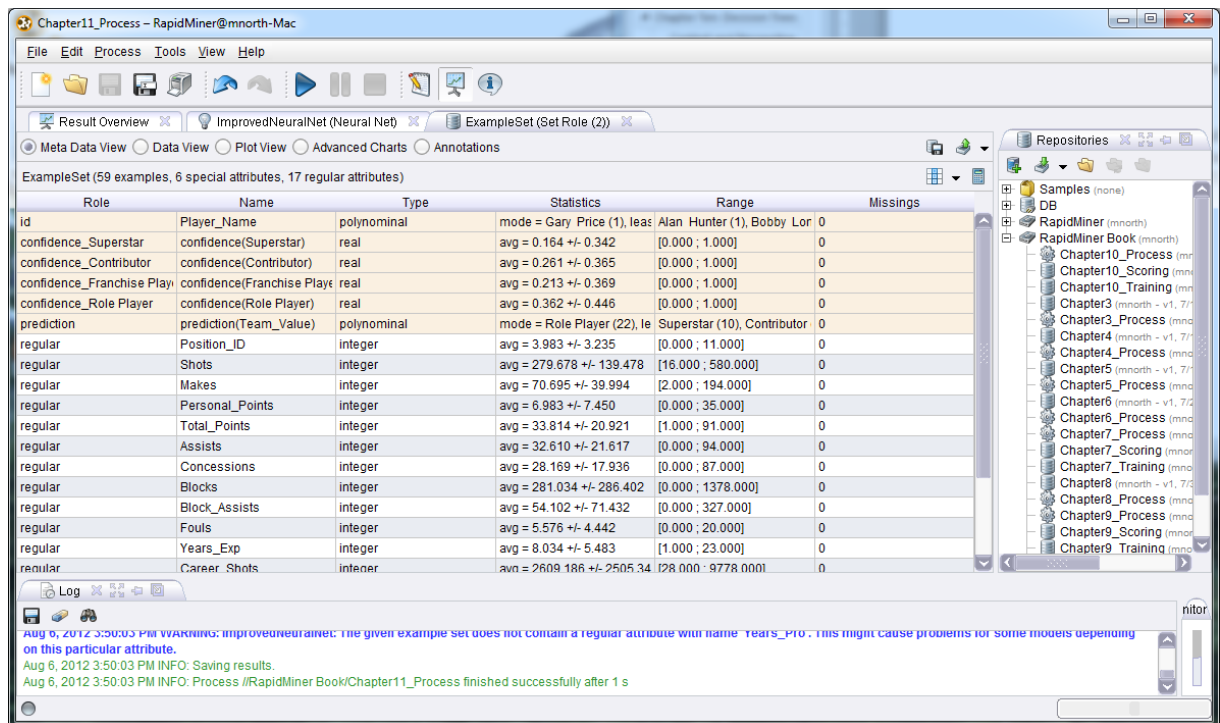


Figure 11-6. Meta data for neural network predictions in the scoring data set.

Change to Data View using the radio button. By now the results of this type of predictive data mining model should look fairly familiar to you.

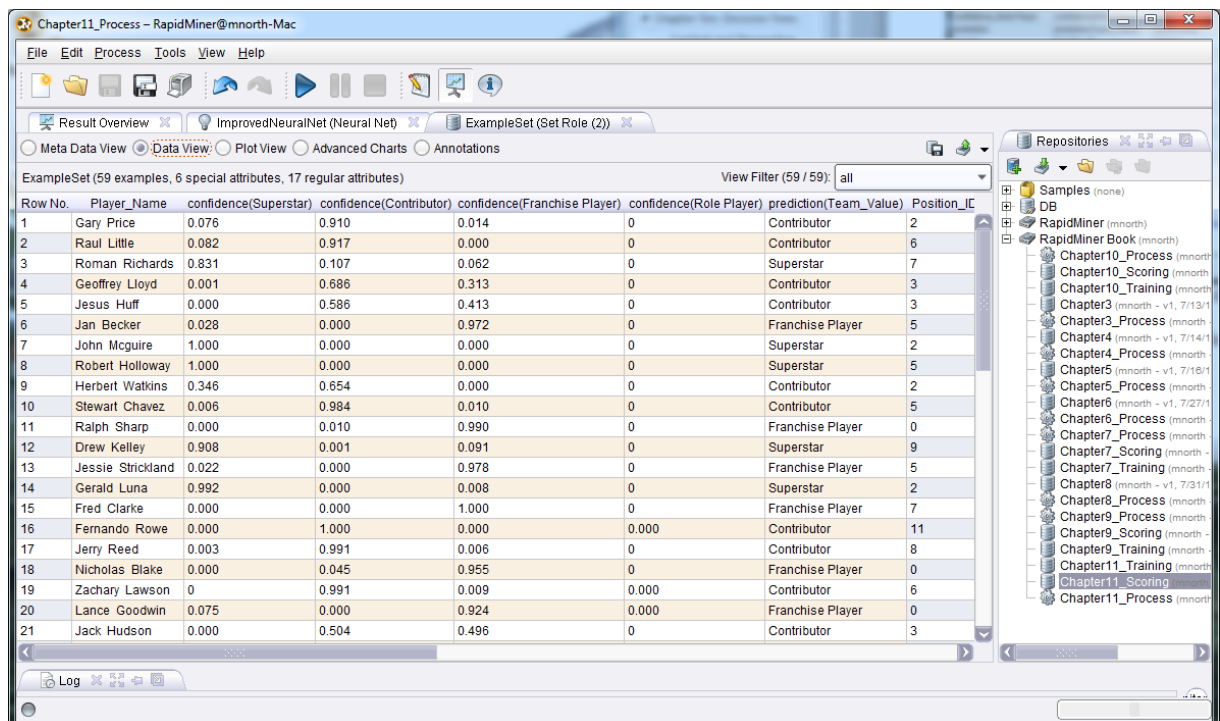
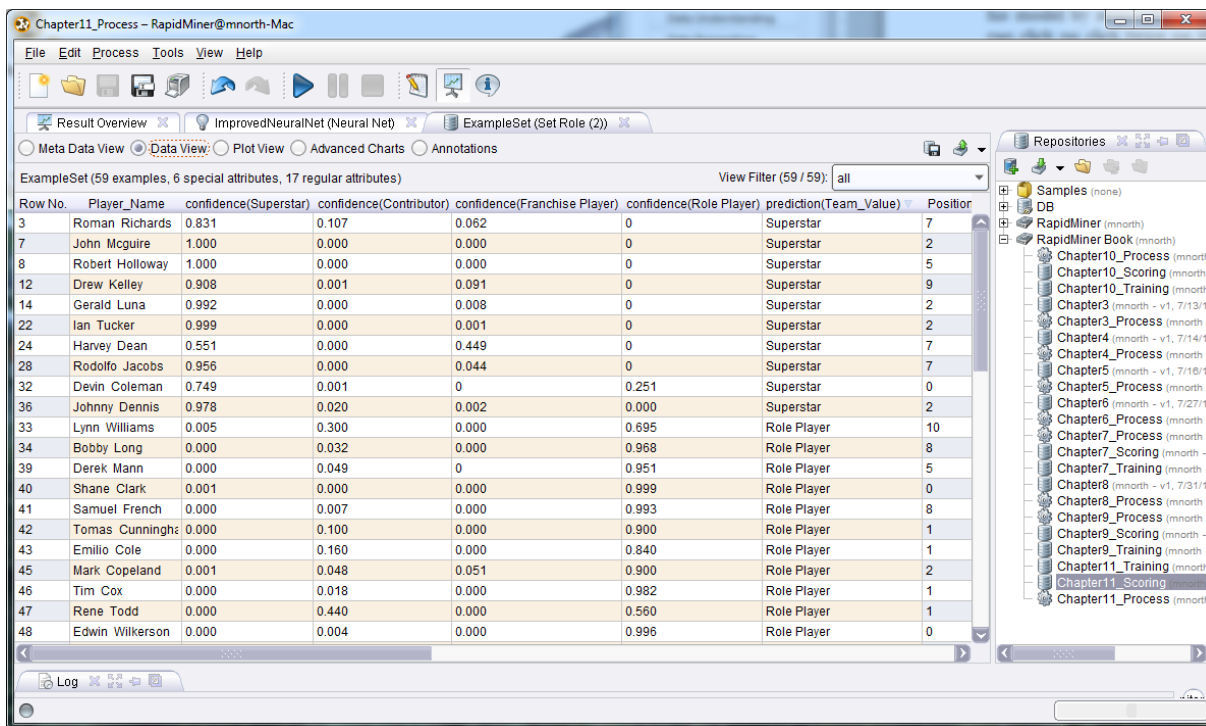


Figure 11-7. Predictions and confidences for our neural net model.

All 59 prospects are now predictively categorized. We know how confident RapidMiner is based on our training data, and Juan can now proceed to...

DEPLOYMENT

Juan wanted to quickly and easily assess these 59 prospects based on their past performance. He can deploy his model by responding to management with several different outputs from our neural network. First, he can click twice on the prediction(Team_Value) column heading to bring all of the Superstars to the top. (Superstar is the last of our values in alphabetical order, so it is first in reverse alphabetical order).



Chapter11_Process - RapidMiner@mnorth-Mac

File Edit Process Tools View Help

Result Overview ImprovedNeuralNet (Neural Net) ExampleSet (Set Role (2))

Meta Data View Data View Plot View Advanced Charts Annotations

ExampleSet (59 examples, 6 special attributes, 17 regular attributes)

View Filter (59 / 59): all

Row No.	Player Name	confidence(Superstar)	confidence(Contributor)	confidence(Franchise Player)	confidence(Role Player)	prediction(Team_Value)	Position
3	Roman Richards	0.831	0.107	0.062	0	Superstar	7
7	John McGuire	1.000	0.000	0.000	0	Superstar	2
8	Robert Holloway	1.000	0.000	0.000	0	Superstar	5
12	Drew Kelley	0.908	0.001	0.091	0	Superstar	9
14	Gerald Luna	0.992	0.000	0.008	0	Superstar	2
22	Ian Tucker	0.999	0.000	0.001	0	Superstar	2
24	Harvey Dean	0.551	0.000	0.449	0	Superstar	7
28	Rodolfo Jacobs	0.956	0.000	0.044	0	Superstar	7
32	Devin Coleman	0.749	0.001	0	0.251	Superstar	0
36	Johnny Dennis	0.978	0.020	0.002	0.000	Superstar	2
33	Lynn Williams	0.005	0.300	0.000	0.695	Role Player	10
34	Bobby Long	0.000	0.032	0.000	0.968	Role Player	8
39	Derek Mann	0.000	0.049	0	0.951	Role Player	5
40	Shane Clark	0.001	0.000	0.000	0.999	Role Player	0
41	Samuel French	0.000	0.007	0.000	0.993	Role Player	8
42	Tomas Cunningham	0.000	0.100	0.000	0.900	Role Player	1
43	Emilio Cole	0.000	0.160	0.000	0.840	Role Player	1
45	Mark Copeland	0.001	0.048	0.051	0.900	Role Player	2
46	Tim Cox	0.000	0.018	0.000	0.982	Role Player	1
47	Rene Todd	0.000	0.440	0.000	0.560	Role Player	1
48	Edwin Wilkerson	0.000	0.004	0.000	0.996	Role Player	0

Log

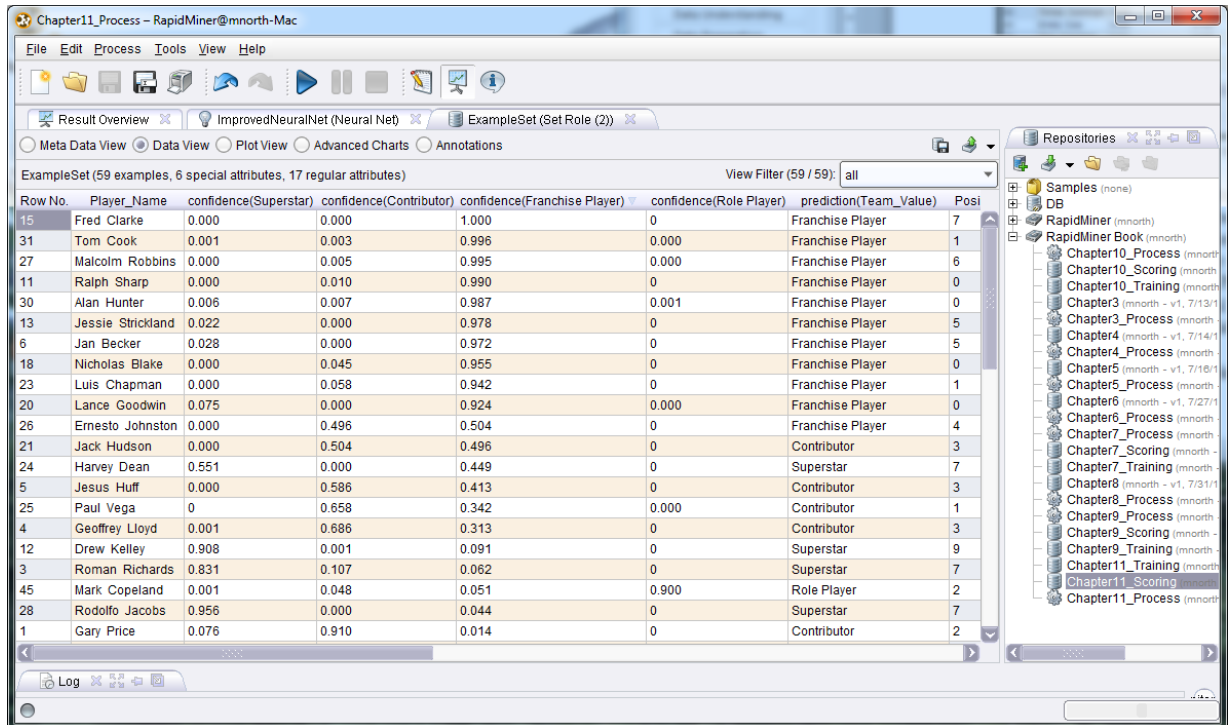
Repositories

- Samples (none)
- DB
- RapidMiner (mnorth)
 - RapidMiner Book (mnorth)
 - Chapter10_Process (mnorth)
 - Chapter10_Scoring (mnorth)
 - Chapter10_Training (mnorth)
 - Chapter3 (mnorth - v1, 7/13/1)
 - Chapter3_Process (mnorth)
 - Chapter4 (mnorth - v1, 7/14/1)
 - Chapter4_Process (mnorth)
 - Chapter5 (mnorth - v1, 7/16/1)
 - Chapter5_Process (mnorth)
 - Chapter6 (mnorth - v1, 7/27/1)
 - Chapter6_Process (mnorth)
 - Chapter7_Process (mnorth)
 - Chapter7_Scoring (mnorth)
 - Chapter7_Training (mnorth)
 - Chapter8 (mnorth - v1, 7/31/1)
 - Chapter8_Process (mnorth)
 - Chapter9_Process (mnorth)
 - Chapter9_Scoring (mnorth)
 - Chapter9_Training (mnorth)
 - Chapter11_Training (mnorth)
 - Chapter11_Scoring (mnorth)
 - Chapter11_Process (mnorth)

Figure 11-8. The scoring data set's predicted values, with Superstars sorted to the top.

The ten athletes with superstar potential are now shown at the top. Furthermore, the confidence for two of them, John McGuire and Robert Holloway have confidence(Superstar) percentages of 100%. Juan may want to go ahead and quickly recommend that management take a hard look at these two athletes. Gerald Luna and Ian Tucker are extremely close as well, with only slight probabilities of being Franchise Players instead of Superstars. Even Franchise Players are athletes with huge potential upsides, so the risk of pursuing either of these two players is minimal. There are a couple of others with predicted superstar status and confidences above 90%, so Juan has a solid list of players to work from.

But Juan knows that these players are likely already on the radar screen for many other teams in the league as well. Perhaps he should look at a few potential alternatives, that aren't quite as obvious to everyone. Juan might be able to score a real win by thinking creatively, and his savvy and experience has told him that sometimes the best player acquisitions aren't always the most obvious ones. Click on confidence(Franchise_Player) twice.



Chapter11_Process - RapidMiner@mnorth-Mac

File Edit Process Tools View Help

Result Overview ImprovedNeuralNet (Neural Net) ExampleSet (Set Role (2))

Meta Data View Data View Plot View Advanced Charts Annotations

ExampleSet (59 examples, 6 special attributes, 17 regular attributes) View Filter (59 / 59): all

Row No.	Player_Name	confidence(Superstar)	confidence(Contributor)	confidence(Franchise Player)	confidence(Role Player)	prediction(Team_Value)	Post
15	Fred Clarke	0.000	0.000	1.000	0	Franchise Player	7
31	Tom Cook	0.001	0.003	0.996	0.000	Franchise Player	1
27	Malcolm Robbins	0.000	0.005	0.995	0.000	Franchise Player	6
11	Ralph Sharp	0.000	0.010	0.990	0	Franchise Player	0
30	Alan Hunter	0.006	0.007	0.987	0.001	Franchise Player	0
13	Jessie Strickland	0.022	0.000	0.978	0	Franchise Player	5
6	Jan Becker	0.028	0.000	0.972	0	Franchise Player	5
18	Nicholas Blake	0.000	0.045	0.955	0	Franchise Player	0
23	Luis Chapman	0.000	0.058	0.942	0	Franchise Player	1
20	Lance Goodwin	0.075	0.000	0.924	0.000	Franchise Player	0
26	Ernesto Johnston	0.000	0.496	0.504	0	Franchise Player	4
21	Jack Hudson	0.000	0.504	0.496	0	Contributor	3
24	Harvey Dean	0.551	0.000	0.449	0	Superstar	7
5	Jesus Huff	0.000	0.586	0.413	0	Contributor	3
25	Paul Vega	0	0.658	0.342	0.000	Contributor	1
4	Geoffrey Lloyd	0.001	0.686	0.313	0	Contributor	3
12	Drew Kelley	0.908	0.001	0.091	0	Superstar	9
3	Roman Richards	0.831	0.107	0.062	0	Superstar	7
45	Mark Copeland	0.001	0.048	0.051	0.900	Role Player	2
28	Rodolfo Jacobs	0.956	0.000	0.044	0	Superstar	7
1	Gary Price	0.076	0.910	0.014	0	Contributor	2

Log

Repositories

- Samples (none)
- DB
- RapidMiner (mnorth)
- RapidMiner Book (mnorth)
 - Chapter10_Process (mnorth)
 - Chapter10_Scoring (mnorth)
 - Chapter10_Training (mnorth)
 - Chapter3 (mnorth - v1, 7/13/1)
 - Chapter3_Process (mnorth)
 - Chapter4 (mnorth - v1, 7/14/1)
 - Chapter4_Process (mnorth)
 - Chapter5 (mnorth - v1, 7/16/1)
 - Chapter5_Process (mnorth)
 - Chapter6 (mnorth - v1, 7/27/1)
 - Chapter6_Process (mnorth)
 - Chapter7_Process (mnorth)
 - Chapter7_Scoring (mnorth)
 - Chapter7_Training (mnorth)
 - Chapter8 (mnorth - v1, 7/31/1)
 - Chapter8_Process (mnorth)
 - Chapter9_Process (mnorth)
 - Chapter9_Scoring (mnorth)
 - Chapter9_Training (mnorth)
 - Chapter11_Training (mnorth)
 - Chapter11_Scoring (mnorth)
 - Chapter11_Process (mnorth)

Figure 11-9. The scoring data set's predicted values, with highest Franchise_Player confidences sorted to the top.

There are 11 predicted Franchise Players in the list of 59 prospects. Perhaps Juan could suggest to management that a solid, long-term building block player could be Fred Clarke. Clarke may be easier to persuade to come to the team because fewer teams may already be in contact with him, and he may be less expensive in terms of salary than most of the superstars will be. This makes sense, but there may be an even better player to pursue. Consider Lance Goodwin on Row 20. Goodwin is predicted to be a Franchise Player, so Juan knows he can play—consistently and at a high level. He would be a solid and long term acquisition for any team. But add to this Goodwin's confidence percentage in the Superstar column. Our neural network is predicting that there is almost an 8% chance that Goodwin will rise to the level of Superstar. With 10 years of experience, Goodwin may be poised to reach the pinnacle of his career within the next season or two. Although he was not the first or most obvious choice in the data set, Goodwin certainly appears to

be an athlete worth taking a hard look at. He may just be the final piece to the puzzle of bringing Juan's franchise the championship at the end of next season.

Of course Juan must continue to use his expertise, experience and evaluation of other factors not represented in the data sets, to make his final recommendations. For example, while all 59 prospects have some number of years experience, what if their performance statistics have all been amassed against inferior competition? It may not be representative of their ability to perform at the professional level. While the model and its predictions have given Juan a lot to think about, he must still use his experience to make good recommendations to management.

CHAPTER SUMMARY

Neural networks try to mimic the human brain by using artificial 'neurons' to compare attributes to one another and look for strong connections. By taking in attribute values, processing them, and generating nodes connected by neurons, this data mining model can offer predictions and confidence percentages, even amid uncertainty in some data. Neural networks are not as limited regarding value ranges as some other methodologies.

In their graphical representation, neural nets are drawn using nodes and neurons. The thicker or darker the line between nodes, the stronger the connection represented by that neuron. Stronger neurons equate to a stronger ability by that attribute to predict. Although the graphical view can be difficult to read, which can often happen when there are a larger number of attributes, the computer is able to read the network and apply the model to scoring data in order to make predictions. Confidence percentages can further inform the value of an observation's prediction, as was illustrated with our hypothetical athlete Lance Goodwin in this chapter. Between the prediction and confidence percentages, we can use neural networks to find interesting observations that may not be obvious, but still represent good opportunities to answer questions or solve problems.

REVIEW QUESTIONS

- 1) Where do neural networks get their name? What characteristics of the model make it ‘neural’?
- 2) Find another observation in this chapter’s example that is interesting but not obvious, similar to the Lance Goodwin observation. Why is the observation you found interesting? Why is it less obvious than some?
- 3) How should confidence percentages be used in conjunction with a neural network’s predictions?
- 4) Why might a data miner prefer a neural network over a decision tree?
- 5) If you want to see a node’s details in a RapidMiner graph of a neural network, what can you do?

EXERCISE

For this chapter’s exercise, you will create a neural network to predict risk levels for loan applicants at a bank. Complete the following steps.

- 1) Access the companion web site for this text. Locate and download the training data set labeled Chapter11Exercise_TrainingData.csv.
- 2) Import the training data set into your RapidMiner repository and name it descriptively. Drag and drop the data set into a new, blank main process.
- 3) Set the Credit_Risk attribute as your label. Remember that Applicant_ID is not predictive.
- 4) Add a Neural Net operator to your model.

- 5) Create your own scoring data set using the attributes in the training data set as a guide. Enter at least 20 observations. You can enter data for people that you know (you may have to estimate some of their attribute values, e.g. their credit score), or you can simply test different values for each of the attributes. For example, you might choose to enter four consecutive observations with the same values in all attributes except for the credit score, where you might increment each observation's credit score by 100 from 400 up to 800.
- 6) Import your scoring data set and apply your model to it.
- 7) Run your model and review your predictions for each of your scoring observations. Report your results, including any interesting or unexpected results.

Challenge Step!

- 8) See if you can experiment with different lower bounds for each attribute to find the point at which a person will be predicted in the 'DO NOT LEND' category. Use a combination of Declare Missing Values and Replace Missing Values operators to try different thresholds on various attributes. Report your results.