Um Modelo de Redes Neurais Artificiais para Predição do Limite Mínimo de Gols em uma Partida de Futebol

Felipe Augusto Pereira Fernandes
Departamento de Computação
Centro Federal de Educação Tecnológica de Minas Gerais
Belo Horizonte, Brasil, CEP 30510-000
Email: felipeapfernandes@gmail.com.br

Elizabeth Fialho Wanner
Flávio Vinícius Cruzeiro Martins
Laboratório de Algoritmos Metaheurísticas e Otimização
Departamento de Computação
Centro Federal de Educação Tecnológica de Minas Gerais
Belo Horizonte, Brasil, CEP 30510-000
Email: [efwanner, flaviocruzeiro]@decom.cefetmg.br

***Abstract*—Futebol é considerado o esporte mais popular do mundo, por isto, pode despertar vários interesses, como por exemplo no mercado de apostas. Hoje no Brasil as casas físicas de apostas em esporte são proibidas, no entanto, não existem restrições para Internet.Há várias formas de apostar, e uma delas é sobre a quantidade de gols que terá uma determinada partida. Este trabalho irá apresentar um modelo de predição baseado em redes neurais artificiais de múltiplas camadas (*multilayer perceptron*) para prever se em uma determinada partida ocorrerá um certo número de gols. Alguns trabalhos da literatura relatam que um dos fatores críticos ao modelar um sistema de predição é definir os fatores que influenciam na decisão, ou seja, os parâmetros de entrada. Neste artigo será mostrado um novo modelo de entrada que tenta capturar as influências da dinâmica de uma partida de futebol. Os resultados obtidos foram satisfatórios, atingindo em média 57.85% de taxa de acerto, demonstrando ser uma boa ferramenta de apoio a decisão no mercado de apostas. Este trabalho também contribui para ampliar estudos que envolvem predições em partidas de futebol, pois não existem muitos trabalhos na literatura que abordam este tema e este trabalho demonstra que é possível adquirir bons resultados neste ramo.**

***Keywords*—Predição em partidas de Futebol, Classificação de padrões, Redes Neurais Artificiais.**

I. Introdução

Futebol é o esporte mais popular do mundo e no Brasil não é diferente. Futebol é um assunto sempre presente em mesas de bares, reuniões familiares e grupos de amigos. Qual time vai ganhar a partida? Quem vai vencer o campeonato? Qual time é melhor? são indagações sempre presentes nas conversas. Justamente por isso, veículos de comunicação como jornais e programas de televisão sempre possuem especialistas para discutir e fazer projeções, empíricas, sobre quem serão os vencedores dos jogos da rodada do campeonato.

Determinar o resultado de uma partida não é uma tarefa simples, pois o futebol é um esporte coletivo sujeito a inúmeras variáveis, como padrão de um time, esquema tático, falhas individuais de jogadores, capacidade de execução e habilidade dos jogadores, condição do gramado, motivação da equipe, montagem da equipe, juízes e etc. Todos esses fatores também influenciam na quantidade de gols que um jogo pode ter.

Um estudo, apresentado em *The numbers game: why everything you know about football is wrong* [1] e feito com mais de dez mil jogos, mostra que a média de gols por partida vem mudando ao longo dos anos. Atualmente, essa média é de 2.6 gols por jogo. A Figura 1 mostra o comportamento desta média nas temporadas do campeonato inglês da primeira divisão desde 1940 a 2014.

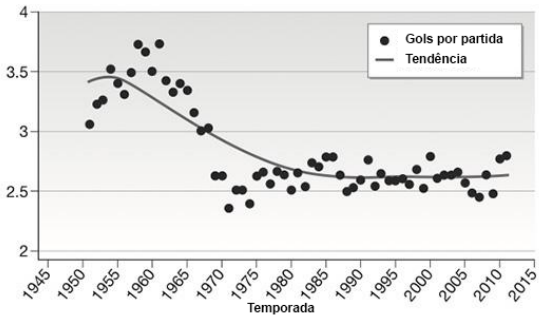

Fig. 1. Média de gols por jogo do campeonato inglês da primeira divisão de 1940 até 2014. [1]

Todas essas dificuldades para se estimar a quantidade de gols e o resultado de um jogo, criam a oportunidade de ganhar dinheiro através de apostas. Este é um mercado que vem ganhando força no cenário mundial nos últimos anos. Para exemplificar este fato, a maioria dos times ingleses da *Premier League* (primeira divisão do campeonato inglês de futebol) tem patrocínio de uma casa de aposta da Inglaterra.

O prêmio pago por cada aposta é definido por especialistas e é dado em função da probabilidade para tal evento acontecer. Quanto menor a chance de um evento acontecer, maior é o prêmio pago. Uma aposta muito utilizada pelos apostadores é relacionado a ter mais ou menos de 2 gols na partida. Essa popularidade se deve ao fato de as chances de acerto estarem sempre perto de 50% pois, como demonstrado, a média de gols em uma partida de futebol fica perto de 2.6 gols por jogo. O prêmio pago pela aposta pode ser definido como o inverso da probabilidade de certo evento acontecer, ou seja, para um evento que tem 50% de chance de acontecer o prêmio pago é 2. Este prêmio é utilizado como um multiplicador do valor apostado. Então para uma aposta de 100 reais com um prêmio

de 2, caso a aposta seja vencedora, terá um retorno de 200 reais, proporcionando assim um lucro de 100 reais.

Ao analisar os resultados dos jogos das equipes durante uma temporada é possível estabelecer um histórico de desempenho, como quantidade de gols marcado, gols sofrido, quantidade de vitórias, derrotas, empates e etc. A partir destes valores é possível perceber que diferentes times apresentam valores próximos em momentos distintos do campeonato, quando os valores comparados são divididos pela quantidade de jogos já realizados. Desta forma, é possível estabelecer o padrão de um time em um determinado jogo em uma determinada rodada em um campeonato.

Este artigo propõe um modelo com base em uma rede Perceptron de múltiplas camadas para predizer se haverá a ocorrência de mais de dois gols em um jogo de futebol. Para treinamento da rede foram utilizados dados dos jogos do campeonato inglês, *Premier League*, das temporadas 2013-2014 e 2014-2015. A taxa de acerto média atingida pela rede proposta foi igual a 57.9%, podendo esta ser uma boa ferramenta no auxílio a tomada de decisão em apostas, já que o prêmio pago para este mercado está sempre perto de 2. Assim é possível construir uma estratégia de valor esperado positivo.

II. Estado da Arte

Apesar da grande popularidade do esporte, existem poucos trabalhos científicos dedicados à predição de jogos de futebol [2]. Em [3], os autores demonstram o bom desempenho da utilização de Redes Neurais Artificiais (RNAs) para a predição de resultados de jogos de futebol. Neste trabalho, as RNAs implementadas são comparadas aos métodos estatísticos apresentando resultados superiores nos problemas tratados.

Cheng et al. [3] propõem o uso de RNAs para a predição de jogos de futebol. A Figura 2 exemplifica o modelo proposto. O modelo baseou-se em uma abordagem híbrida, criando um sistema que primeiramente avalia o nível dos times, separando os jogos em 3 conjuntos: conjunto dos jogos nos quais o time mandante é melhor; conjunto dos jogos nos quais os times apresentam níveis iguais; e jogos onde o time visitante é melhor. Para isto foi utilizado a diferença da média de pontos e a diferença da média de gols feitos pelas equipes. Para o cálculo da média de pontos, calcula-se os pontos dos times na competição até uma determinada rodada. Para cada vitória soma-se três pontos, para cada empate um ponto e zero pontos para a derrota. Ao final, divide-se a quantidade de pontos somados pela quantidade de jogos disputados pelo time. A média de gols feitos pelas equipes é calculada somando os gols feitos até determinada rodada e então divide-se esta soma pela quantidade de jogos disputados. Cada entrada da rede passa por este mecanismo seletor que define a qual RNA aquele jogo pertence. O modelo de entrada utilizado pelos autores é apresentado na figura 3. A saída dessas redes seria a predição do resultado do jogo como vitória do mandante, empate ou vitória do visitante. O modelo proposto atingiu uma taxa de acerto de 52.29%, sendo melhor que os métodos de previsão estatísticos existentes, tais como *Elo Rating* [5].

Aslan e Inceoglu [2] fizeram um comparativo de modelos de previsão de resultados de jogos de futebol. Ao fim propuseram um novo modelo de rede, baseando somente no desempenho de cada time como mandante e visitante. Para

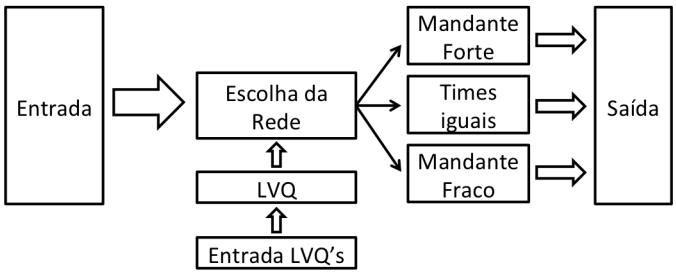

Fig. 2. Modelo de predição proposto por Cheng et al. [3]

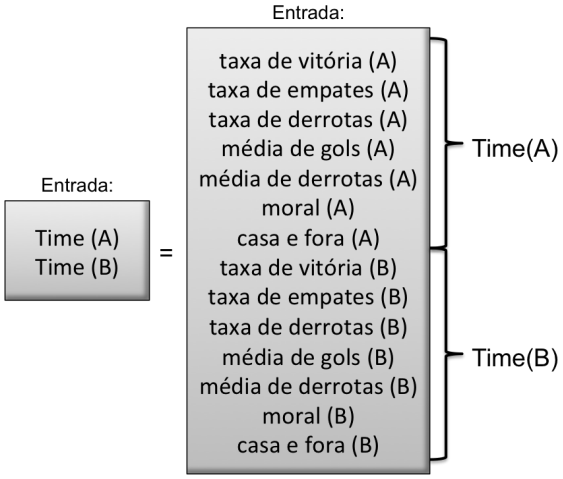

Fig. 3. Modelo de entradas proposto por Cheng et al. [3]

isso, eles simplesmente incrementam uma variável para cada vitória, decrementam para cada derrota e em caso de empate não alteram o valor da variável. Foram proposto então dois modelos de entradas para redes *Learning Vector Quantization* (LVQ): uma somente com os dados de desempenho como mandante para o time mandante e os dados de desempenho como visitante para o time visitante; e outra com os dados de mandante e visitante para ambos os times. A rede com menos entradas teve 53.25% de taxa de acerto contra 51.29% do outro modelo. Os autores concluíram que a modelagem dos dados de entrada da RNA é um problema crítico.

III. Modelagem do problema

A rede neural artificial (RNA) é uma técnica computacional amplamente utilizada. Uma RNA é composta por várias unidades de processamento. Estas unidades podem se conectar formando uma rede, onde a cada conexão é associada a um peso sinápticos. Existem diferentes arquiteturas de RNAs, e cada arquitetura se difere pela forma de conexão entre os neurônios presentes na rede. As RNAs são utilizadas comumente como técnicas de reconhecimento de padrões, aproximação de função e otimização.

Este trabalho propõe construir um modelo capaz de separar duas classes não linearmente separáveis. Para isto foi definida a arquitetura de redes Perceptron de Múltiplas Camadas (MLP - em inglês, *Multilayer Perceptron*). Este tipo de arquitetura tem sido aplicada com sucesso nas mais diversas áreas, resolvendo problemas de classificação de padrões, processamento de sinais, modelagem de funções, etc. Uma rede deste tipo é

constituída de uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada camada usa um ou mais neurônios e cada neurônio é conectado aos neurônios da camada adjacente através de pesos. A camada de saída recebe os estímulos provenientes das camadas intermediárias e é capaz de construir o padrão de resposta. As camadas intermediárias, cujos pesos representam as características presentes na entrada, permitem que a rede crie uma representação mais complexa do problema [8]. Desta forma, uma rede deste tipo é progressiva (*feedforward*), uma vez que as saídas dos neurônios em uma camada qualquer se conectam apenas aos neurônios da camada seguinte.

O projeto de uma RNA MLP necessita basicamente das seguintes especificações:

- i camada de entrada,
- ii número de camadas escondidas,
- iii número de neurônios em cada uma das camadas escondidas,
- iv camada de saída e
- v as especificações dos pesos que interconectam os neurônios nas diversas camadas.

As especificações (i) e (iv) dependem da modelagem do problema, as especificações (ii) e (iii) determinam a complexidade da rede, e por fim, a especificação (v) envolve a utilização de algoritmos de treinamento supervisionado.

O modelo aqui proposto foi baseado em uma rede de múltiplas camadas (MLP). A arquitetura da rede possui dezesseis nós de entrada, duas camadas ocultas, sendo que em cada uma das camada ocultas possuem dez neurônios, e dois neurônios de saída. A função de transferência utilizada em todos os neurônios das camadas ocultas foi a log-sigmoid. A figura 4 exemplifica a arquitetura da rede. Esta arquitetura se difere da abordagem híbrida proposta por Cheng et al, pois, faz uso de apenas uma RNA para fazer a predição. Como mostrado anteriormente os autores de [3] fazem uso de um método para identificar o potencial do mandante da partida, e assim chavear dentre as três redes construídas qual teria melhores chances de fazer a predição.

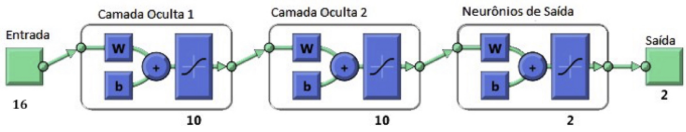

Fig. 4. Arquitetura da RNA utilizada nesta modelagem.

As entradas da rede, a princípio, se assemelham ao modelo apresentado por Cheng et al. [3]. Entretanto, para qualificar um time foi feita uma abordagem distinta. Em um jogo de futebol pode-se dividir uma equipe em três setores: defesa, meio de campo e ataque. Em geral, os especialistas afirmam que times com defesa e meio de campo fortes tendem a sofrer menos gols, enquanto times com ataque e meio de campo fortes tendem a fazer muitos gols. Portanto, indicadores de qualidade da defesa, meio de campo e ataque dos times parecem ser entradas interessante para a modelagem desse problema.

Para obter esses valores foram utilizados os indicadores de qualidade de defesa, meio de campo e ataque da franquia de jogos da EA *Sports* (*Electronic Arts Sports*), o FIFA [7]. Nele, os jogadores são avaliados de 0 a 100 em diversos atributos. O índice do time é calculado como a média dos atributos de todos os jogadores da equipe. Desta forma, o índice de defesa do time é calculado pela média dos índices defensivos de todos os jogadores da equipe, o índice de meio de campo do time pela média dos atributos de meio de campo e o índice de ataque do time pela média dos atributos de ataque dos jogadores.

Como as temporadas 2013-2014 e 2014-2015 serão utilizadas no treinamento da rede, os índices de qualidade dos times devem ser extraídos do jogo correspondente à temporada, FIFA 14 para a temporada 2013-2014 e FIFA 15 para a temporada 2014-2015. A EA *Sports* oferece uma atualização desse índice quinzenalmente, entretanto neste trabalho só foi considerado o valor inicial desses índices. Os valores indicativos de qualidade de um time não muda durante uma temporada. Os indicadores variam de 0 a 100, portanto, para a nossa rede esses indicadores serão divididos por 100, passando variar de 0 a 1.

Considerando o time mandante como A e o time visitante como B, a Figura 5 mostra o modelo proposto de entradas da rede. A Figura 6 apresenta o modelo simplificado da rede neural modelada e proposta neste trabalho.

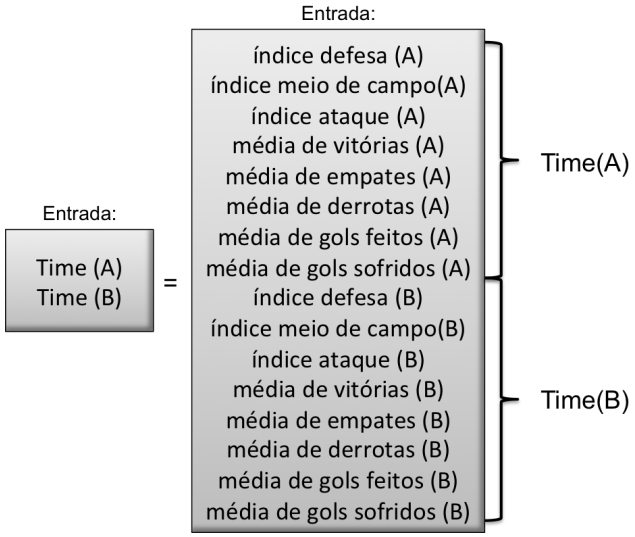

Fig. 5. Modelo de entradas proposto.

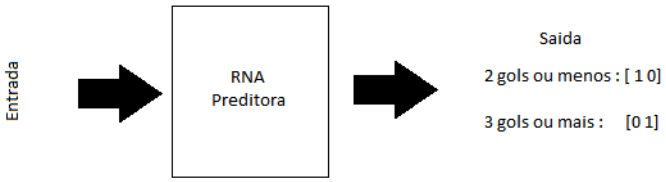

Fig. 6. Modelo da RNA proposta.

A. *Preparação dos dados*

Para o treinamento, teste e validação da rede, foram utilizados os dados dos jogos do campeonato inglês *Premier League* das temporadas 2013-2014 e 2014-2015 [6]. Os jogos das 6

primeiras rodadas não foram utilizados para o treinamento e validação da rede. O motivador para tal descarte foi o fato de acreditar que nesses jogos os times estão em processo de adaptação, tanto em relação ao campeonato, quanto aos seus novos companheiros, fato que pode tornar os resultados mais aptos a aleatoriedade dos jogos de futebol. Ao desconsiderar este período de adaptação, é esperado que a sequência dos resultados seja mais estável em relação ao padrão tático e de desempenho dos times de uma determinada partida. Vale a pena ressaltar que a escolha da quantidade de rodadas a serem descartadas foi completamente empírica, sendo um possível estudo futuro.

Após a exclusão das 6 primeiras rodadas, foram considerados 320 jogos para o treinamento em cada temporada, pois o campeonato inglês tem 20 times e 38 rodadas por temporada. Como foram utilizadas duas temporadas, foram considerados 640 jogos. Destes 640 jogos, 70% foram utilizados para o treinamento, 15% para a validação e 15% para o teste.

A cada jogo de uma rodada, são levados em consideração todos os jogos das rodadas anteriores daquela equipe. Por exemplo, em um jogo da rodada 30, os dados presentes na entrada da rede para este jogo, sofrem influências da rodada 1 à rodada 29 dos times a serem analisados.

Tomando como exemplo o jogo Liverpool vs Manchester United disputado no dia 22/03/2015, pertencente a trigésima rodada da temporada 2014-2015, a Figura 7 demonstra o desempenhos dos times até a vigésima nona rodada. Portanto, para montar a entrada da rede proposta é necessário calcular a taxa de vitórias, empates, derrotas, gols feitos e gols sofridos por jogo. A Figura 7 mostra os dados para ambos os times levando-se em consideração número de jogos, número de vitórias, empates e derrotas, o número de gols feitos e o número de gols sofridos. As Tabelas I e II apresentam os valores calculados, até a rodada 29, para o Liverpool e Manchester United, respectivamente.

	J	V	E	D	GP	GC
Manchester United	29	16	8	5	50	26
Liverpool	29	16	6	7	43	30

Fig. 7. Desempenho de Liverpool e Manchester United em vinte nove jogos. J representa o número de jogos, V o número de vitórias, E o número de empates, D o número de derrotas, GP o número de gols feitos e GC o número de gols sofridos

Tabela I. VALORES CALCULADOS DO LIVERPOOL ATÉ A RODADA 29 DA TEMPORADA 2014-2015.

Liverpool				
Vitórias/Jogo	Empate/Jogo	Derrota/Jogo	GF/Jogo	GS/Jogo
0.552	0.207	0.241	1.483	1.034

Tabela II. VALORES CALCULADOS DO MANCHESTER UNITED ATÉ A RODADA 29 DA TEMPORADA 2014-2015.

Manchester United				
Vitórias/Jogo	Empate/Jogo	Derrota/Jogo	GF/Jogo	GS/Jogo
0.552	0.276	0.172	1.724	0.896

Acrescentando os índices dos times aos dados calculados previamente, os dados de entrada da rede são determinados. Figura 8 exemplifica um dado de entrada da rede, tendo como base o jogo entre Liverpool e Manchester United mostrado no exemplo.

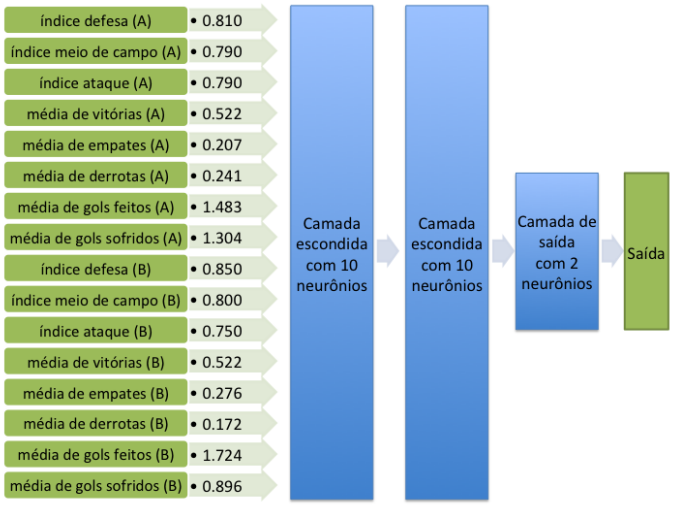

Fig. 8. Dados de entrada da rede para o jogo Liverpool (A) e Manchester United (B) da trigésima rodada da temporada 2014-2015.

IV. RESULTADOS

A proposta deste trabalho é a construção de uma RNA capaz de prever se, em uma determinada partida de futebol, haverá um certo número mínimo de gols. Com base nos indícios apresentados de que a média de gols em uma partida de futebol gira em torno de 2,6 gols por jogo, entende-se que seria interessante possuir uma rede que fosse capaz de prever se em uma determinada partida irá ocorrer mais de 2 gols ou não.

A partir desta meta, identificar se em uma determinada partida ocorrerá mais de dois gols, foi construída uma RNA de acordo com a metodologia apresentada na seção anterior. Com intuito de mostrar a robustez do modelo proposto, a base de dados escolhida, das duas temporadas do campeonato inglês (temporadas 2013-2014 e 2014-2015), foi submetida 1.000 (mil) vezes ao procedimento de treinamento, teste e validação, sendo 70% da base utilizada para o treinamento, 15% para a validação e 15% para o teste.

Após esta bateria de testes, a RNA proposta obteve uma taxa de acerto médio de 57.85%. Para melhor visualizar o resultado deste teste de robustez, a Figura 9 mostra um histograma das 1.000 redes construídas e suas respectivas taxas de acertos. Nesta figura é possível observar a baixa dispersão do comportamento das redes construídas, pois a maioria delas mantiveram a taxa de acerto entre 58% e 60%.

Outra forma de visualizar os resultados atingidos é através da matriz de confusão. A matriz de confusão de uma hipótese h oferece uma medida efetiva do modelo de classificação mostrando o número de classificações corretas versus as classificações preditas para cada classe sobre um determinado conjunto de exemplos. O número de acertos, para cada classe, se localiza na diagonal principal da matriz e os demais elementos representam os erros na classificação. Em um classificador ideal, todos os elementos fora da diagonal principal deveriam

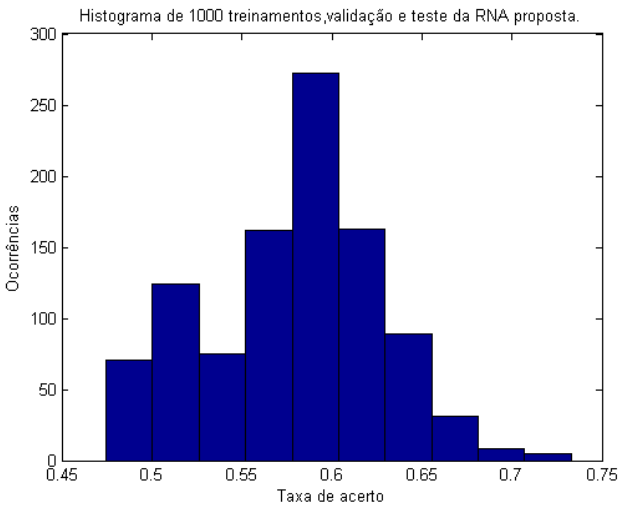

Fig. 9. Histograma das taxas de acertos nas 1000 execuções da RNA proposta.

ser iguais a zero uma vez que erros não seriam cometidos. Este tipo de matriz é uma excelente forma de expressar a quantidade de unidades amostrais e suas classificações. Cada unidade amostral é associada a uma dada categoria durante o processo de classificação efetuado, e à categoria real a que pertencem essas unidades [4].

A Figura 10 apresenta uma matriz de confusão em relação a bateria de testes realizados. A classe 1 representa partidas com 2 gols ou menos e a Classe 2 representa partidas com mais de 2 gols. Com base nesta figura, pode-se observar que em 58.2% das predições que haveriam dois gols ou menos foram corretas e 57.5% das predições que haveriam mais de dois gols foram corretas. Quando o jogo teve dois gols ou menos a rede obteve uma taxa de acerto de 62.2% e quando houve mais de dois gols a rede obteve 53.4% de acerto. Como a média dos jogos é de 2.6 gols por jogo, é interessante observar o fato da rede apresentar uma maior taxa de acerto predizendo que haverá dois gols ou menos na partida. Este fato, a princípio, não é esperado.

V. Considerações finais e trabalhos futuros

Neste trabalho foi proposto um modelo baseado em RNA para predizer o número mínimo de gols em uma partida de futebol. Uma vez que o número de gols em uma partida de futebol é na média de 2,6 gols por jogo, decidiu-se testar o modelo proposto para prever partidas que ocorrerão mais de 2 gols. O modelo proposto acrescentou elementos ainda não explorados na literatura. Foram utilizados padrões de entradas com base em elementos que puderam caracterizar melhor os padrões dos times.

Com base nos testes realizados, foi mostrado que o modelo proposto foi capaz de classificar as partidas com uma taxa de acerto média de 57.8%. Esta classificação refere-se ao acontecimento de mais de 2 gols ou não em uma partida de futebol. Como não foram encontrados na literatura trabalhos que abordaram o mesmo assunto, não é possível realizar comparações com outros trabalhos. No entanto, se esta rede fosse utilizada como ferramenta em casas de apostas, devido ao seu índice de acerto ser superior a 50%, entende-se que esta

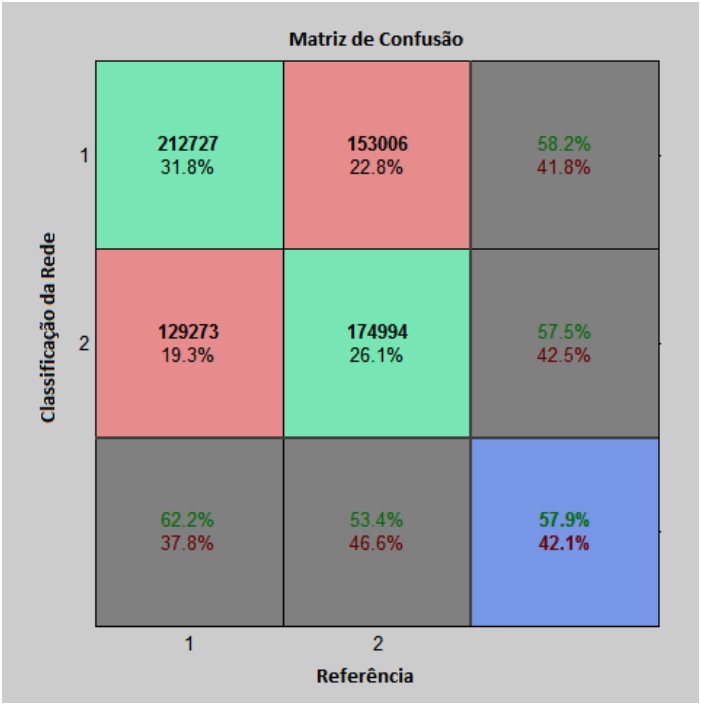

Fig. 10. Matriz de confusão das 1000 execuções. Classe 1 - Partidas com 2 gols ou menos. Classe 2 - Partidas com mais de 2 gols.

metodologia pode ser uma excelente ferramenta de auxílio na decisão.

Como visto, este trabalho visa somente a modelagem do problema. Foi utilizado uma rede MLP com *back propagation* de 2 camadas ocultas e 10 neurônios em cada camada. Como trabalho futuro será utilizado uma ferramenta de otimização para definir a arquitetura ideal para o modelo proposto.

Como este estudo apresentou resultados promissores, uma próxima fase do trabalho será a construção de uma rede neural para predizer o vencedor de um jogo de futebol. Esta nova rede terá como foco o modelo de entrada e saídas, e o modelo aqui proposto será utilizado como base. Além disso, outras variações podem ser construídas, como o uso dos dados de últimos jogos, incluindo assim o desempenho recente das equipes. Essa inclusão pode aumentar o desempenho da rede pois lesões e suspensões ou venda de jogadores podem afetar o desempenho de um time e o modelos que utilizam dados de forma global, não conseguem capturar.

Mensurar a qualidade de um time é algo complicado, portanto um estudo somente sobre este assunto pode beneficiar imensamente o desenvolvimento desta área de pesquisa. Por fim, este trabalho ainda contribui para aumentar o número de pesquisa nessa área de estudo, encorajando assim novos pesquisadores a estudar e fomentar novos trabalhos.

Agradecimentos

Os autores gostariam de agradecer o apoio das agências de fomento à pesquisa, FAPEMIG, CAPES e CNPq.

Referências

[1] ANDERSON, C. ; SALLY, D. *The numbers game: why everything you know about football is wrong*, Penguin UK, 2013.

[2] ASLAN, B. G.; INCEOGLU, M. M. A comparative study on neural network based soccer result prediction. In: IEEE. *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on.* [S.l.], 2007. p. 545–550.

[3] CHENG, T.; CUI, D.; FAN, Z.; ZHOU, J.; LU, S. A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system. In: IEEE. *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on.* [S.l.], 2003. p. 308–313.

[4] CONGALTON, R. G. *A review of assessing the accuracy classifications of remotely sensed data*. Remote Sensing of Environment, New York, v. 49, n. 12, p. 1671-1678, Dec. 1991.

[5] Elo, A.E.: *The Rating of Chess Players, Past and Present*, Arco Publishing, New York (1978).

[6] O Gol, *www.ogol.com.br :: Tudo sobre Futebol*, http://www.ogol.com.br, acessado em 14 de maio de 2015.

[7] SoFIFA, *Teams - FIFA - SoFIFA*, http://sofifa.com/teams/?hl=en-GB, acessado em 14 de maio de 2015.

[8] S. O. Haykin, *Neural Networks and Learning Machines*. Prentice-Hall, 2008.