Um modelo de regressão por vetores de suporte para previsão da diferença de gols em partidas de futebol

Mateus Ferreira, Eduardo Feitosa, Marco Cristo, Eulanda Santos
Instituto de Computação
Universidade Federal do Amazonas
Manaus, Brasil
{mateus, efeitosa, marco.cristo, emsantos}@icomp.ufam.edu.br

***Abstract*—Predict the difference of goals in football matches is a problem of great interest to the sports punters of the handicap mode. This predictive capability is also important in preparing teams for the second round of a confrontation in knockout tournaments. Furthermore, it has been demonstrated that statistical information about the past performance of football teams correlate better with the goal difference than with the score. This paper employs a machine learning model to predict the goal difference in the outcome of football matches. It was attained a RMSE of 1.267822 using the support vector regression technique. For comparison with a reference model, it was also computed the quality of the predictions of the winners of matches. In this case, the obtained accuracy of 60.53% was higher than the 55.26% accuracy achieved by the reference model.**

Keywords—machine learning; support vector regression; football; goal difference.

I. INTRODUÇÃO

Esportes exercem grande fascínio sobre a humanidade devido ao entretenimento e aos benefícios para a saúde que proporcionam. Particularmente, o futebol destaca-se atualmente como o esporte mais popular do mundo [1]. Segundo estudo realizado em 2006 pela FIFA (*Fédération Internationale de Football Association*), existem aproximadamente 270 milhões de praticantes de futebol nos 207 países associados a essa entidade [2]. No Brasil, país tradicionalmente reconhecido por sua paixão por esse esporte, cresce ainda mais o interesse pelo futebol, tendo em vista que o país se prepara para sediar a Copa das Confederações, em 2013, e a Copa do Mundo, em 2014, duas das principais competições dessa modalidade esportiva.

Além do grande número de praticantes ao redor do mundo, o futebol impacta significativamente a economia mundial. Segundo Szymanski [3], no ano de 2000, o futebol foi responsável por 3% de todo o comércio mundial. Parte dessa fatia está ligada ao mercado de apostas, que vem crescendo consideravelmente em decorrência do advento e legalização de sistemas internacionais de apostas online [4]. Dentre os diversos tipos de apostas relacionadas ao resultado de partidas de futebol, a modalidade denominada *handicap* desperta especial interesse dos apostadores, porque possibilita ampliar os ganhos dos acertadores [5]. Nessa modalidade, as casas de aposta assinalam uma vantagem para uma das equipes, através de uma compensação no placar, como forma de equilibrar as chances de vitória. Um time considerado mais fraco pode, por exemplo, receber dois gols de vantagem. Se essa equipe for derrotada por apenas um gol de diferença, os apostadores que investiram nesse competidor são premiados. Normalmente, os ganhos obtidos com a modalidade *handicap* de apostas são maiores, já que é mais difícil prever com exatidão a diferença de gols entre dois oponentes do que o vencedor do confronto.

Prever a diferença de gols é útil também para as próprias equipes, em campeonatos do tipo mata-mata. Nesses torneios, no intervalo entre a primeira e a segunda partida, os adversários já conhecem a diferença de gols que precisam alcançar para se classificarem à próxima fase da competição. Assim, a previsão correta da diferença de gols esperada para a segunda partida pode alertar os competidores para a necessidade de estabelecer medidas imediatas e aplicáveis em um curto espaço de tempo, com o intuito de aumentar a força da equipe e a consequente probabilidade de alcançar a diferença de gols necessária. Tais medidas podem incluir a escalação de um atleta importante – mesmo que fora de suas condições físicas ideais –, o treinamento de um esquema tático inesperado ou até mesmo a preparação das condições do gramado de forma a dificultar o estilo de jogo do adversário.

Tanto para a realização de apostas do tipo *handicap* quanto para a preparação das equipes em confrontos mata-mata, a capacidade de prever a diferença de gols no resultado de uma partida de futebol representa uma vantagem significativa. Trabalhos que visam prever o placar exato das partidas podem ser usados para inferir a diferença de gols. No entanto, como já foi demonstrado, ao analisar dados do passado, previsões de diferenças no placar são mais precisas do que previsões de placares exatos [6–8]. Diante desse cenário, este trabalho utiliza um modelo baseado em aprendizagem de máquina para a previsão da diferença de gols em partidas de futebol. Especificamente, será mostrado que a técnica de regressão numérica por vetores de suporte permite alcançar um RMSE (*Root Mean Square Error*) de 1,267822 para a previsão de diferença de gols e uma acurácia de 60,53% para a previsão de vencedores. Apesar de, até onde se tem conhecimento, não existirem outros modelos de aprendizagem de máquina dedicados a prever a diferença de gols em partidas de futebol, a acurácia alcançada na previsão de vencedores é ligeiramente superior à taxa de 55,26% obtida pelas previsões indicadas no modelo de referência.

Inicialmente, na Seção II deste artigo são discutidos os trabalhos relacionados. Em seguida, a Seção III apresenta o estudo de caso utilizado durante a elaboração do modelo. Os

detalhes do modelo proposto são apresentados na Seção IV. O método de aprendizagem de máquina empregado, detalhado na Subseção A, baseia-se na técnica de regressão por vetores de suporte. Cada partida é representada por um conjunto de atributos, descritos na Subseção B. A Seção IV é encerrada com a apresentação, na Subseção C, da estratégia de otimização de parâmetros empregada durante a definição do modelo. Este trabalho apresenta ainda, na Seção V, a indicação dos resultados obtidos e, na Seção VI, uma breve conclusão.

II. Trabalhos Relacionados

A tarefa que se procurou resolver neste trabalho está diretamente relacionada com a previsão do placar em jogos de futebol. Há uma vasta literatura que cobre esse problema. A maior parte das estratégias se baseia em modelos de regressão que consideram diferentes atributos, tais como: os resultados no passado [9], o impacto de se jogar em casa ou fora [10] e os aspectos particulares de torneios mata-mata [11] ou de campeonatos com muitos jogos [8], [12]. Modelos mais recentes também têm aplicado redes neurais e redes bayesianas nas previsões [7], [13]. Outros trabalhos têm investigado a relação dos resultados com sistemas de apostas (de placar), avaliando a sua eficiência [4], [14].

No entanto, nenhum dos trabalhos citados foca o problema da previsão da diferença de gols usando aprendizagem de máquina. Por outro lado, pelo menos três deles [6–8] observam que há maior correlação dos modelos estudados com as diferenças de gols do que com o placar exato, o que reforça a necessidade de um estudo para previsão da diferença de gols.

III. Estudo de caso

Para construir o modelo de regressão capaz de prever a diferença de gols em partidas de futebol foi necessário estabelecer um estudo de caso, que guiou as etapas de desenvolvimento do modelo. O estudo de caso escolhido consistiu em prever a diferença de gols em partidas da série A do campeonato brasileiro de futebol realizadas entre os anos de 2002 e 2011. Naturalmente, as estratégias empregadas no modelo utilizam somente dados que estavam disponíveis antes da realização das partidas cuja diferença de gols é estimada.

Os dados brutos foram colhidos dos sites “Chance de Gol” [15] e “RSSSF Brasil” [16], especializados em estatísticas relacionadas ao futebol. No primeiro, estão registrados os resultados de todas as partidas da série A do campeonato brasileiro de futebol, desde 1998. Além do placar, são indicados o nome da equipe mandante e da equipe visitante, a data do confronto, a chance de vitória do time da casa, a chance de empate e a chance de vitória do visitante – calculados segundo critérios estabelecidos pelo estatístico que administra o site “Chance de Gol”.

No segundo site, foram colhidas informações indicativas dos times que disputaram a Copa Libertadores da América, dos clubes que venceram a Copa do Brasil e das equipes oriundas da série B do campeonato brasileiro de futebol, tudo para o período compreendido entre os anos de 2002 e 2011. Além disso, nesse site foram verificadas também as cidades de origem de cada equipe.

IV. Protocolo Experimental

O problema de previsão da diferença de gols em uma partida de futebol pode ser interpretado como uma tarefa de regressão temporal. Nesse caso, busca-se encontrar um conjunto de características (atributos) que representem bem a qualificação das equipes envolvidas. Para isso, devem ser selecionados dados históricos relevantes que indiquem o desempenho dessas equipes no passado. Com base nesses dados, aplica-se então algum método matemático de regressão que irá estimar a diferença de gols para uma determinada partida.

Nesta seção são descritos o método de regressão aplicado aos dados de treinamento, os atributos selecionados para compor o modelo proposto e o processo de otimização de parâmetros adotado.

A. Método

A capacidade de prever uma diferença de gols não inteira possibilita a percepção do grau de proximidade entre o valor estimado e uma diferença de gols prática (valor inteiro). Para os cenários que motivaram este estudo, essa percepção pode ser mais importante do que a previsão de um valor perfeitamente compatível com o atributo alvo. Por essa razão, uma abordagem de regressão numérica apresenta previsões melhores do que as estratégias de classificação equivalentes, que fixam os resultados em valores discretos (ex: -1, 0, 1). Diante dessa constatação, optou-se por utilizar um método de regressão numérica.

Existem diversas técnicas de aprendizagem de máquina para solução de problemas de regressão numérica. Ao serem empregadas para solucionar o mesmo problema, essas estratégias podem alcançar resultados distintos. Neste trabalho, optou-se por utilizar um método de regressão por vetores de suporte (*Support Vector Regression* - SVR), tendo em vista que esse método tem apresentado bons resultados em estudos comparativos de técnicas de previsão de séries temporais, aplicados a diferentes áreas do conhecimento [17]. Para isso, utilizou-se a implementação do classificador funcional SMOreg, implementado na ferramenta Weka [18].

A abordagem SVR é uma extensão do método SVM (*Support Vector Machines*) – originalmente desenvolvido para previsão de classes – para agregar-lhe a funcionalidade de regressão numérica [17]. Esse método determina o hiperplano ótimo que separa as instâncias do atributo-alvo de um problema, por meio da análise da distância entre as instâncias posicionadas na fronteira das classes.

A família de métodos derivados do SVM utiliza funções de *kernel* para produzir transformações matemáticas nos dados, ampliando a dimensionalidade da representação de forma a torná-los linearmente separáveis. Diversos métodos de *kernel* foram propostos na literatura. Neste trabalho, optou-se por utilizar o *kernel* de funções de base radial (*Radial Basis Function* – RBF), uma vez que, quando os dados estão arbitrariamente posicionados (sem uma estrutura regular), ele apresenta boas propriedades de aproximação e é independente da dimensão onde os dados estão inseridos [19].

B. Atributos

A qualidade de uma equipe pode ser aferida por seus resultados no longo, médio e curto prazos. Pensando nisso, optou-se por representar as partidas com dados relativos ao desempenho das equipes ao longo de três períodos: nos últimos 4 jogos, no campeonato corrente e em todo o período histórico analisado (2002 a 2011). Além disso, foi discriminado o desempenho das equipes enquanto mandantes, ou seja, jogando em casa, da performance obtida como visitantes.

1) Definição e Geração

A escolha inicial dos atributos foi baseada em um processo empírico, onde procurou-se ampliar ao máximo a quantidade de variáveis no modelo. Nessa fase, a única limitação residiu na capacidade de obtenção dos dados. Quando a disponibilidade de um determinado tipo de informação era incerta, essa característica era descartada, ainda que pudesse contribuir para uma melhora nas previsões. Um exemplo de atributo dispensado foi a classificação de cada equipe nos respectivos campeonatos estaduais, visto que a coleta desses dados demandaria a pesquisa em diversas fontes, com formatações distintas e possibilidade de ausência de dados.

Inicialmente, não foram descartados atributos que pudessem apresentar um elevado nível de correlação com outra característica, visto que a etapa seguinte do modelo consistiu em selecionar o conjunto de atributos com melhor desempenho nas previsões. Assim, a lista inicial de características contemplava os seguintes aspectos:

- Saldo de gols (numérico).
- Gols marcados (numérico).
- Gols sofridos (numérico).
- Pontos ganhos (numérico).
- Número de vitórias, empates e derrotas (numérico).
- Probabilidades de vitória, empate e derrota calculadas pelo site "Chancedegol" (numérico).
- Total de jogos realizados em todos os campeonatos anteriores (numérico).
- Confronto regional (binário: S,N).
- Equipe oriunda da série B (binário: S,N).
- Equipe disputou a Copa Libertadores no ano corrente (binário: S,N).
- Equipe venceu a Copa do Brasil no ano corrente (binário: S,N).

Para cada um dos campos numéricos elencados acima (com exceção do "total de jogos" e das "probabilidades de vitória, empate e derrota"), foram calculados os valores considerando-se os fatores de tempo (últimos 4 jogos, campeonato corrente ou todos os campeonatos), de local (jogo em casa ou como visitante) e de forma de cálculo (média ou soma dos valores considerados). No total, inicialmente foram gerados 150 atributos (incluindo o atributo alvo e os campos de marcação: data, nome da equipe mandante e nome do time visitante).

Deve-se destacar a importância da inclusão de informações relacionadas ao saldo de gols e às contagens de gols marcados e sofridos – além do retrospecto de pontos ganhos, vitórias, empates e derrotas –, visto que o objetivo final da previsão é estimar a diferença de gols nas partidas e não apenas indicar o vencedor (ou empate). Isso justifica a escolha de todos os atributos numéricos listados, com exceção do campo destinado a representar o número total de jogos realizados por cada equipe. Nesse caso, a intenção é diferenciar as equipes com maior índice de participação histórica no campeonato, por não terem sido rebaixadas, daquelas que figuram recorrentemente na disputa da série B.

Por sua vez, a justificativa para a inclusão do campo "Confronto regional" no modelo reside na observação de que a maior parte dos confrontos entre equipes de uma mesma cidade são equilibrados, acarretando em um elevado índice de empates (diferença de gols igual a 0), mesmo quando o retrospecto das equipes indica uma diferença acentuada de desempenho entre elas.

As equipes oriundas da série B são destacadas porque é possível perceber que, ressalvadas algumas exceções, esses times costumam apresentar um desempenho ruim ao longo do campeonato e, muitas vezes, voltam a ser rebaixados logo no ano seguinte.

O raciocínio inverso aplica-se ao atributo destinado a destacar as equipes que disputam a Copa Libertadores da América no ano corrente, uma vez que esses clubes tendem a se reforçar para obter um bom resultado nessa competição, o que acaba refletindo em um bom desempenho no campeonato brasileiro – novamente excluindo-se algumas exceções.

Da mesma forma, a indicação do time campeão da Copa do Brasil em um determinado ano parece ser um bom indicativo de que essa equipe apresentará um bom desempenho no campeonato brasileiro do mesmo ano. Deve-se destacar que a informação do clube campeão da Copa do Brasil somente foi anotada na base de dados para as instâncias referentes a partidas realizadas após a data do jogo final da Copa do Brasil do respectivo ano. Até essa data, a equipe campeã do torneio no ano anterior permaneceu com o marcador correspondente ativo. Em outras palavras, para efeito da base de dados gerada, o campeão da Copa do Brasil permaneceu com esse status até a definição do novo campeão, no ano seguinte.

Durante os primeiros testes para seleção de atributos, surgiu a ideia de adaptar os atributos para que eles possibilitassem medir a diferença de desempenho entre as equipes, já que o atributo alvo corresponde à diferença de gols marcados. Essa hipótese levou à alteração da base inicial para que ela passasse a refletir a diferença entre as equipes para cada uma das métricas estabelecidas. Assim, os atributos "média do saldo de gols do time da casa nos últimos 4 jogos" e "média do saldo de gols do time visitante nos últimos 4 jogos", por exemplo, foram transformados no atributo "diferença da média do saldo de gols nos últimos 4 jogos entre o time da casa e o time visitante". Essa transformação foi aplicada a todas as características numéricas, reduzindo a dimensão dos dados para 81 atributos (incluindo o atributo alvo e os campos de marcação).

2) Seleção

Como o método SVR demanda um tempo de execução longo para treinamento, a seleção dos atributos foi efetuada utilizando-se o método funcional de regressão linear,

implementado no Weka [18] com o nome "LinearRegression". Portanto, foi utilizado um método de seleção de atributos do tipo *wrapper*, que permite a escolha do melhor conjunto de atributos baseando-se na análise dos resultados gerados pela execução de um método de classificação – nesse caso, a regressão linear.

Além disso, uma heurística gulosa, implementada no Weka com o nome "GreedyStepwise", foi utilizada para busca das possíveis combinações de atributos, visto que a execução dos testes para todas as combinações de atributos demandaria um tempo impraticável [20]. Assim, o conjunto de atributos que apresentou o menor RMSE quando aplicado ao método de regressão linear e que, por isso, foi utilizado no modelo proposto, foi constituído pelos seguintes campos:

- diferença das chances de vitória entre o time da casa e a equipe visitante, registradas no site "Chance de Gol".
- diferença entre a média do saldo de gols do time mandante em casa e a média do saldo de gols da equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o saldo de gols total do time mandante em casa e o saldo de gols total da equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o saldo de gols total do time mandante e o saldo de gols total da equipe visitante no campeonato corrente, independente do "fator casa".
- diferença entre a média do saldo de gols do time mandante em casa e a média do saldo de gols da equipe visitante fora de casa, em todos os jogos.
- diferença entre o total de gols marcados pelo time mandante em casa e o total de gols marcados pela equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre a média de gols marcados pelo time mandante e a média de gols marcados pela equipe visitante no campeonato corrente, independente do "fator casa".
- diferença entre o total de gols marcados pelo time mandante e o total de gols marcados pela equipe visitante no campeonato corrente, independente do "fator casa".
- diferença entre o total de gols marcados pelo time mandante em casa e o total de gols marcados pela equipe visitante fora de casa, no campeonato corrente.
- diferença entre o total de gols marcados pelo time mandante em casa e o total de gols marcados pela equipe visitante fora de casa, em todos os jogos.
- diferença entre a média de pontos ganhos pelo time mandante em casa e a média de pontos ganhos pela equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o total de pontos ganhos pelo time mandante em casa e o total de pontos ganhos pela equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o total de pontos ganhos pelo time mandante e o total de pontos ganhos pela equipe visitante em todos os jogos realizados pelas equipes, independente do "fator casa".
- diferença entre a média de pontos ganhos pelo time mandante em casa e a média de pontos ganhos pela equipe visitante fora de casa, em todos os jogos.
- diferença entre o número de vitórias do time mandante em casa e o número de vitórias da equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o número de empates do time mandante em casa e o número de empates da equipe visitante fora de casa, nos últimos 4 jogos.
- diferença entre o total de derrotas do time mandante e o total de derrotas da equipe visitante no campeonato corrente, independente do "fator casa".
- diferença entre o total de vitórias do time mandante e o total de vitórias da equipe visitante em todos os jogos realizados pelas equipes, independente do "fator casa".
- diferença entre o número de empates do time mandante em casa e o número de empates da equipe visitante fora de casa, em todos os jogos.
- indicação se o time visitante foi campeão da Copa do Brasil no ano corrente.

Os valores calculados para cada atributo foram normalizados, a fim de evitar distorções na importância que o método atribui às características quando a faixa de valores de um atributo é maior do que a de outro.

3) Ruídos

Por constatar que algumas equipes que disputaram a Copa Libertadores da América não apresentaram um bom desempenho no Campeonato Brasileiro do mesmo ano, distanciando-se do padrão esperado, foram removidas as indicações dessas equipes na base de dados, com o intuito de eliminar ruídos e melhorar o desempenho das previsões. No entanto, os experimentos comprovaram que a eliminação dessas referências não acarretou em melhoras nos resultados e, por isso, as indicações foram mantidas no modelo proposto.

Da mesma forma, a eliminação dos registros relacionados a clubes que venceram a Copa do Brasil e apresentaram um desempenho ruim no Campeonato Brasileiro do mesmo ano não resultaram em benefícios concretos para os resultados experimentais.

De fato, como os métodos baseados em vetores de suporte trabalham com uma pequena fração das instâncias, correspondentes às ocorrências que se situam na faixa de fronteira entre as classes alvo, a eliminação de ruídos tem pouca influência no resultado alcançado por esses métodos. Além disso, os parâmetros de custo e margem de erro, descritos na seção C, já apresentam mecanismos para contornar os problemas decorrentes da existência de ruídos entre os dados de treinamento.

C. Parâmetros

A precisão de um modelo SVR é largamente dependente da seleção dos parâmetros do processo. Com o *kernel* RBF, SVR requer a otimização de 3 parâmetros [17]:

- C = custo. Controla a margem para dados não separáveis linearmente, uma vez que nem sempre o *kernel* consegue separar linearmente todos os dados. Essa estratégia permite tolerar ruídos, pois considera mais pontos de treinamento – além dos que estão na fronteira –, e permite a ocorrência de erros de classificação. Diminuindo-se o valor de C, aumenta-se o custo para pontos muito distantes do padrão da classe, forçando a criação de um modelo mais ajustado às amostras e, portanto, menos generalista. Aceita valores entre 0 e ∞, usualmente potências de 10 entre 0 e 1000;
- E = margem de erro. Corresponde a uma margem dentro da qual os dados são ignorados durante o processo de regressão. Aceita valores entre 0 e ∞, usualmente muito pequenos.
- G = gama. Parâmetro usado pelo *kernel* RBF para ajustar a forma da função interpolada, controlando o seu "achatamento". Aceita valores entre 0 e ∞, usualmente "1/dimensão dos dados".

A estratégia utilizada para otimizar esses parâmetros foi baseada no modelo de pesquisa em grade (*grid search*) proposto por Hsu, Chang e Lin [21]. Na abordagem utilizada, para cada valor de C pertencente à faixa de valores entre 2^{-6} e 2^{10}, saltando em intervalos de 2^2, percorreu-se o valor de G em uma faixa de valores entre 2^{-16} e 2^3, saltando também em intervalos de 2^2. Após definir o melhor valor de C e G dentro dessas faixas, refinou-se a busca em torno do valor inicialmente encontrado, com incrementos de $2^{0.2}$ para ambos os parâmetros. Esse processo de refinamento sucessivo foi repetido até que os resultados não apresentassem melhorias.

Somente após otimizar os valores de C e G, passou-se à otimização do parâmetro E. Apesar de essa não ser a estratégia mais confiável, ela foi adotada em função das limitações de tempo para executar-se, por exemplo, a busca pelo melhor valor para o parâmetro E dentro dos laços iterativos que percorreram as faixas de valores de C e G. O parâmetro E foi otimizado em uma faixa de valores compreendidos entre 2^{-20} e 2^8, saltando em intervalos de 2^2. De maneira análoga, a busca foi refinada em torno do valor inicialmente encontrado até que os resultados não apresentassem melhorias.

Em função da reduzida quantidade de dados disponíveis para algumas classes alvo e do elevado tempo para a execução das otimizações com o SVR, optou-se por adaptar a estratégia clássica de validação cruzada [22]. Ao invés de utilizar 10 conjuntos de teste estratificados, foram empregados 10 conjuntos de teste escolhidos aleatoriamente, sem estratificação. Apesar dos benefícios garantidos pela separação estratificada dos dados, essa abordagem se tornou inviável em decorrência da pequena quantidade de exemplares pertencentes a algumas classes alvo (como as diferenças de gols maiores do que 4 ou menores que -4). Esse problema é agravado ao considerar-se a necessidade de prever os resultados de partidas realizadas no início do período trabalhado (anos de 2002 e 2003), já que os dados referentes a partidas posteriores não podem ser utilizados no conjunto de treinamento do modelo. Além disso, a amostragem sem estratificação permitiu diferenciar os parâmetros para cada ano a ser trabalhado pelo modelo.

Como as instâncias de teste podem pertencer a partidas ocorridas desde 2002 até 2011, o montante de dados disponível para treinamento é maior para os testes relativos a partidas ocorridas nos últimos anos (próximos a 2011) do que aquelas que aconteceram no início do período trabalhado (anos próximos de 2002). Além disso, as características de perfil dos campeonatos, como média de gols marcados ou percentual de empates, podem variar ao longo do tempo. Por tudo isso, a abordagem de otimizar os parâmetros do método SVR separadamente para cada ano estudado garantiu um resultado melhor do que se fossem definidos parâmetros únicos para todo o período trabalhado.

V. RESULTADOS

Foram utilizadas duas métricas para avaliar a qualidade das previsões realizadas pelo modelo proposto. Para comparar a diferença de gols prevista com a diferença real, foi escolhida a métrica da raiz quadrada do erro quadrático médio (RMSE). Essa métrica de desempenho é mais adequada para avaliar os resultados de ponto flutuante expressos pelo modelo, usados para indicar a proximidade entre a diferença de gols esperada e uma diferença de gols real (valor inteiro).

Como, até onde se tem conhecimento, não existem modelos de aprendizagem de máquina que prevejam especificamente a diferença de gols em partidas de futebol, optou-se por medir também a acurácia do modelo ao prever a equipe vitoriosa, ou o empate. Dessa forma, foi possível comparar o resultado obtido com as previsões estabelecidas pelo estatístico que administra o site "Chance de Gol" [15]. Para o modelo proposto, a acurácia é calculada considerando-se corretas as previsões (sem arredondamentos) se a diferença de gols prevista (dp) for: (a) dp > 0, em jogos com vitória do time da casa; (b) dp = 0, em jogos empatados; e (c) dp < 0 em jogos com derrota do time da casa. Em relação às previsões indicadas pelo site "Chance de Gol", o autor exprime as previsões na forma de probabilidades de vitória do mandante, empate ou vitória do visitante. Por isso, nesse caso, a acurácia é calculada considerando-se que o site escolhe como resultado previsto a opção com maior probabilidade.

Pelos motivos já discutidos na análise do processo de otimização dos parâmetros do SVR, a capacidade de generalização do modelo foi avaliada utilizando-se uma técnica de validação cruzada não estratificada. A coleção de teste foi construída sorteando-se 10 datas nas quais ocorreram rodadas do campeonato brasileiro de futebol, uma data para cada ano do período trabalhado (2002 a 2011). Representando a diferença de gols prevista para o resultado de uma partida i realizada no dia j por P_i e o valor real dessa diferença por R_i, o erro de previsão E_j para todas as N partidas realizadas no dia j foi definido pela seguinte expressão:

$$E_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - R_i\right)^2} \quad (1)$$

Dessa forma, o erro para a coleção de teste CT pode ser expresso por:

$$E_{CT} = \frac{1}{10}\sum_{j=1}^{10} E_j \quad (2)$$

O modelo apresentado neste trabalho alcançou um E_{CT} de 1,267822 para os 10 conjuntos de partidas sorteados. Esse resultado pode ser interpretado, por exemplo, assumindo-se que, em média, o modelo erra a diferença de gols de cada partida por pouco mais do que um gol. A título de exemplo, ao utilizar o modelo proposto na escolha de apostas, poderiam ser considerados vantajosos os cenários em que a disparidade entre a diferença de força dos times indicada pela casa de apostas e a diferença de gols prevista pelo modelo for maior do que 1,267822.

Para o mesmo conjunto de teste, a acurácia alcançada pelo modelo na previsão do vencedor (ou empate) das partidas foi de 60,53%, já que o modelo acertou 46 das 76 partidas sorteadas. As previsões indicadas no site especializado "Chance de Gol" atingiram uma acurácia de 55,26%, ao acertar o resultado de 42 confrontos. Deve-se destacar que o modelo utilizado como referência é fruto de um projeto de mestrado que vem sendo aperfeiçoado e amplamente testado desde 1996 [23], o que lhe confere um certo grau de maturidade.

Levando em consideração a comparação do resultado alcançado com o modelo maduro utilizado no site "Chance de Gol", acredita-se que o modelo proposto neste trabalho aponta para a viabilidade do emprego de técnicas de aprendizagem de máquina, especialmente a regressão por vetores de suporte, para a previsão da diferença de gols no resultado de partidas de futebol. Essa constatação é reforçada ao levar-se em conta que o modelo foi desenvolvido para realizar uma tarefa mais específica do que a previsão da equipe vitoriosa (ou empate) em um confronto. A tarefa de prever a diferença de gols exige uma maior especialização do modelo nos quesitos relacionados a gols marcados e sofridos, o que pode influenciar o desempenho do método em relação à previsão de vitórias, empates ou derrotas, em detrimento de um resultado mais apurado para a diferença de gols.

Além disso, os resultados alcançados pelo método empregado podem ser otimizados, especialmente através da inclusão de outros atributos não considerados no modelo. Um exemplo de atributo utilizado no site "Chance de Gol" que não foi empregado e que parece ser um bom indicativo de força das equipes, especialmente para a previsão das primeiras rodadas do campeonato brasileiro, é o desempenho das equipes nos campeonatos estaduais. Outra possibilidade é estudar o efeito da combinação de múltiplos métodos de aprendizagem de máquina na tentativa de reduzir o erro registrado pelas instâncias mais difíceis de prever.

VI. Conclusão

Neste artigo foi apresentado um modelo de aprendizagem de máquina baseado em regressão por vetores de suporte para a previsão da diferença de gols em partidas de futebol. Um estudo de caso criado para prever a diferença de gols de 76 partidas do campeonato brasileiro de futebol obteve um RMSE de 1,267822. A previsão do vencedor (ou empate) desses mesmos confrontos apresentou uma acurácia de 60,53%, resultado ligeiramente superior ao valor obtido pelo modelo de referência. Esses resultados apontam para a viabilidade de se empregar técnicas de aprendizagem de máquina no auxílio a esquemas de apostas esportivas do tipo *handicap* ou na preparação de equipes para confrontos eliminatórios.

Referências

[1] FIFA, "Activity Report 2008 / 2009," 2009.

[2] FIFA, "Big Count." [Online]. Available: http://www.fifa.com/worldfootball/bigcount/index.html.

[3] S. Szymanski, "Economics of sport: Introduction," *The Economic Journal*, vol. 111, no. 469, pp. 1–3, 2001.

[4] N. Vlastakis, G. Dotsis, and R. N. Markellos, "How efficient is the European football betting market? Evidence from arbitrage and trading strategies," *Journal of Forecasting*, vol. 28, no. 5, pp. 426–444, 2009.

[5] Online Betting, "Handicap Betting." [Online]. Available: http://www.online-betting.me.uk/articles/handicap-betting.html.

[6] R. T. Stefani, "Predicting score difference versus score total in rugby and soccer," *IMA Journal of Management Mathematics*, vol. 20, no. 2, pp. 147–158, 2009.

[7] D. Karlis and I. Ntzoufras, "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference," *IMA Journal of Management Mathematics*, vol. 20, no. 2, pp. 133–145, 2009.

[8] A. Heuer and O. Rubner, "Towards the perfect prediction of soccer matches," *ArXiv e-prints*, Jul. 2012.

[9] M. J. Maher, "Modelling association football scores," *Statistica Neerlandica*, vol. 36, no. 3, pp. 109–118, 1982.

[10] K. Courneya and A. Carron, "The Home Advantage in Sport Competitions: A Literature Review," *Journal of Sport & Exercise Psychology*, vol. 14, no. 1, pp. 13–27, 1992.

[11] D. Kuonen, "Statistical models for knock-out soccer tournaments," *Technical R eport, Department of Mathematics, Chair of Applied Statistics, Ecole P olytechniq ue Federale De Lausanne*. 1996.

[12] A. J. Lee, "Modeling scores in the Premier League: is Manchester United really the best?," *Chance*, vol. 10, no. 1, pp. 15–19, 1997.

[13] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," *Knowledge-Based Systems*, vol. 19, no. 7, pp. 544–553, 2006.

[14] M. J. Dixon and S. G. Coles, "Modelling association football scores and inefficiencies in the football betting market," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 46, no. 2, pp. 265–280, 1997.

[15] M. L. de Arruda, "Chance de Gol." [Online]. Available: http://chancedegol.uol.com.br.

[16] RSSSF, "RSSSF Brasil." [Online]. Available: http://www.rsssfbrasil.com.

[17] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Artificial Neural Networks—ICANN'97*, Springer, 1997, pp. 999–1004.

[18] The University of Waikato, "Weka 3: Data Mining Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

[19] C. M. M. Lima, "Um estudo sobre o método Mínimos Quadrados Móveis por Aproximações Iteradas," Pontifícia Universidade Católica do Rio de Janeiro, 2009.

[20] Weka, "GreedyStepwise." [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GreedyStepwise.html.

[21] C.-W. Hsu, C.-C. Chang, C.-J. Lin, and others, "A practical guide to support vector classification." 2010.

[22] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Addison Wesley, 2005.

[23] M. L. de Arruda, "Poisson, Bayes, Futebol e DeFinetti," Universidade de São Paulo, 2000.