2R-02

Jリーグの試合結果予測のための文書分類

小邦 将輝 † 奥村 紀之 †
† 香川高等専門学校 情報工学科

1 はじめに

近年，ソーシャル･ネットワーキング･サービス (SNS) の発達により，数多くのチームや選手，サポーターがTwitter やブログなどで試合や練習での情報を発信している．

本研究では，新聞記事，選手やサポーターの Twitter の投稿，ブログ記事から，チームや選手のコンディションに関わる情報を収集し，それらの解析に基づく文書分類を行うことにより，J リーグの試合結果予測を行うためのシステムを構築し，その性能の検証を行っている．

2 関連研究

関連研究として，椙本ら [1] の研究では，2009 年 J1 第 34 節のデータから，ディリクレ過程過程混合モデルを用いて各チームの攻撃力と守備力が異なる分散でランダムウォークするとして試合結果を予測している．また，青柳 [2] は，2009 年の J1 リーグ全試合のデータからチームの強さとホームアドバンテイジの関連を調査することにより，試合結果の予測を行った．以上の 2 つの研究は統計的手法を用いて試合結果を予測している．他にも，データマイニング手法を用いた研究として，Felipe[1] らが出場選手のクラブチームでの成績や，グループリーグでの成績などのデータを用いて，サッカーの試合結果を予測したものが挙げられる．

特に Felipe らは，2014 年 6 月に行われたワールドカップの決勝トーナメント 1 回戦，準々決勝の試合結果を的中させており，選手の健康状態や試合での成績は，後の試合の結果を予想する上で重要な要素となることがわかる．

3 評価実験

本研究では，2014 年度の J1 リーグに所属した 18 チームから公式アカウント 18 ユーザー，選手 138 ユーザー，サポーター 19 ユーザー，ニュースアカウント 19 ユーザーを対象として，TwitterAPI を用いて Tweet の収集を行った．Tweet の収集期間は，各チームのキャンプが始まった 2014 年 2 月 1 日からシーズンが終了した 2014 年 12 月 6 日とし，対象期間中 95136Tweet を取得した．

本研究では，取得した Tweet を試合結果予測に用いるための分類実験として，2 値ベクトルと tf・idf によって重み付けしたベクトルの 2 種類の比較実験を行った．

3.1 tf・idf，2 値ベクトルを用いた分類実験

ユーザーから取得した Tweet を形態素解析器 MeCab で形態素解析を行うことで，Tweet 中に含まれる単語の対数文書頻度 (inverse document frequency,idf) を算出した．idf は 1 式で求められる．

$$w_t^d = log\frac{N}{df(t)} + 1 \tag{1}$$

図 1 に idf の分散を示す．

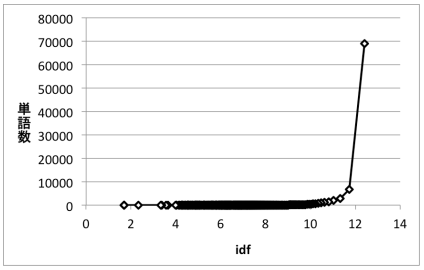

図 1: idf の分散

図 1 から idf が 10 を上回る単語，5 を下回る単語に関しては，目視での検証の結果，実験に使用するデータとして適切でないと判断し削除した．これらの単語中から，J リーグの試合結果予測を行う上で，勝利に繋がると思われる 167 単語，敗北に繋がると思われる 101 単語を本校の学生に対する調査を基に選出した．各チームの各試合前の Tweet 群から単語の出現頻度 tf を算出し，上の idf と掛け合わせて tf・idf を算出した．tf は 2 式で求められる．ただし，文書の長さによる重みを考慮して正規化を行っている．

$$w_t^d = \frac{tf(t,d)}{\sum_{S\in d} tf(s,d)} \tag{2}$$

tf・idf を用いた実験では，tf・idf を各次元ベクトルとしてサポートベクターマシン (Support Vector Machine,SVM) での分類を行った．2 値ベクトルを用い

A Document Classification for Estimating the Game Result of J-League

† Masaki Oguni(kt.guni823@gmail.com)
† Noriyuki Okumura(okumura@di.kagawa-nct.ac.jp)
Department of Information Technology, National Institute of Technology Kagawa college (†)
551 Koda, Takuma, Mitoyo, Kagawa 769-1192, Japan

[1] http://gigazine.net/news/20140704-google-predict-worldcup-2014/(閲覧:2014 年 7 月)

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

た実験では，単語が出現すれば 1 を，出現しなければ 0 を各次元ベクトルとして，SVM での分類を行った．SVM には，ソルバーとして Lib-SVM[2] を使用し，34 分割交差検定を行うことにより分類実験を行った．SVM タイプには C-SVC を，カーネル関数には RBF を指定した．勝利に関する単語を含む Tweet を正例，敗北に関する単語を含む Tweet を負例の学習データとした．

3.2 実験結果

表 1 に tf・idf を用いた分類実験，2 値ベクトルを用いた分類実験の結果を示す．ただし，表 1 中の G 大阪はガンバ大阪，C 大阪はセレッソ大阪の略である．

表 1: 各実験の分類成功率

$team$	tf・idf	2 値	$team$	tf・idf	2 値
G 大阪	55.8%	55.8%	浦和	52.9%	52.9%
鹿島	52.9%	52.9%	柏	50%	50%
鳥栖	55.8%	55.8%	川崎	47.1%	47.1%
横浜	41.2%	41.2%	広島	38.2%	38.2%
名古屋	38.2%	38.2%	東京	35.2%	35.2%
神戸	0%	0%	新潟	41.1%	41.1%
甲府	32.4%	32.4%	仙台	41.2%	41.2%
清水	52.9%	52.9%	大宮	50%	50%
C 大阪	50%	50%	徳島	76.4%	76.4%

4 考察

tf・idf を用いた分類，2 値ベクトルを用いた実験ともに分類成功率に違いはみられなかった．これは，Tweet が 140 文字という少ない文字数であるため，tf の値に大きな差が生まれず，結果的に tf・idf を用いた分類と 2 値分類で差が生じなかったと考えられる．今後，ブログ記事や Tweet に含まれるリンク先のテキストデータを分類に加えることで，tf を算出する文書空間が広がるため，tf・idf を用いた分類と 2 値分類で違った結果が得られると考えられる．

しかし，神戸の第 2 節と第 23 節の引き分けとなった試合の分類結果が，$tfidf$ を用いた分類では勝利，2 値分類では敗北となっている．これは，分離平面近くにベクトルが位置していたことで，実験ごとに結果が異なったのだと考えられる．

表 1 より 34 分割交差検定の分類成功率の平均値は 45.1%となっており，勝利，敗北，引き分けが等確率で出現するときに比べ，確率が高いことがわかる．

また，半数のチームは 50%以上の精度で分類が成功した．対して，残りの半数は 50%を下回っており，神戸の分類成功率は双方の実験で，0%となっている．この理由として，神戸は 11 勝 12 分 11 敗と勝利数が拮抗しているため，分類がうまくいかなかったと考えられる．他の分類成功率が 50%を下回ったチームについても，分類成功率が 50%を超えたチームに比べて勝敗が拮抗している．

ここで，神戸の対戦相手の分類成功率を表 2 に示す．

表 2: 神戸の対戦相手の分類成功率

分類結果	
勝　ち	負　け
45.0%(9/20)	57.1%(8/14)

表 2 から 34 試合中 17 試合の分類に成功していることがわかる．このことから，過去の傾向から分類成功率が低いチームについては，当該チームの分類結果だけでなく，対戦相手の分類結果を考慮することで，予測の成功率を高められると推察できる．

また，学習に用いたデータが，1 シーズン分のものであったため，過去数年分のデータを用いることで，より実際の試合結果に近い分類データを取得できると考えられる．

5 おわりに

本研究では，Tweet 群を，tf・idf，2 値ベクトルによって分類を行うことによって，試合結果との相関の調査を行った．実験結果より，得られた Tweet 群と試合結果には，関連性があることがわかった．

今後の課題として,「ゴール」という単語が出現した場合に，ゴールを「決めた」のか「決められた」のかによって，チーム力の評価が異なるため，文書ユニグラムを用いて文書解析を行う必要がある．文書間の相関，語間の相関など種々の観点から文書解析を行うことで，試合結果予測に役立てられるようにしたい．

テキストデータの収集や，今回行った実験内容を自動化することで，J リーグの試合結果予測システムの構築を目指す．

参考文献

[1] 椙本功弥，酒折文武：ディリクレ過程混合モデルを持つ状態空間における MCMC-based Particle Filter を用いた状態の推定，日本統計学会誌，第 41 巻第 2 号，pp.265-284，2012.

[2] 青柳領：サッカーにおけるチーム別ホームアドバンテイジの統計学的定義，体育測定評価研究，Vol.11，pp.35-47，2011.

[2] 取得先:www.csie.ntu.edu.tw/ cjlin/libsvm/

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.