Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 710 – 719

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES 2014

Sports data mining: predicting results for the college football games

Carson K. Leung*, Kyle W. Joseph

Department of Computer Science, University of Manitoba, Winnipeg, MB, R3T 2N2,Canada

Abstract

In many real-life sports games, spectators are interested in predicting the outcomes and watching the games to verify their predictions. Traditional approaches include subjective prediction, objective prediction, and simple statistical methods. However, these approaches may not be too reliable in many situations. In this paper, we present a sports data mining approach, which helps discover interesting knowledge and predict outcomes of sports games such as college football. Our approach makes predictions based on a combination of four different measures on the historical results of the games. Evaluation results on real-life college football data shows that our approach leads to relatively high accuracy in result prediction.

Peer-review under responsibility of KES International.

Keywords: American football; college football; data mining; knowledge-based and intelligent information & engineering systems; intelligent systems applications; prediction; sports data mining

1. Introduction

Data mining aims to discover implicit, previously unknown and potentially useful information or knowledge from data. These data can be ranged from traditional shopper basket market data [1,5,10] to the emerging social network data [11,12,28]. These have led to the research problems of social network mining as well as *sports data mining* [15,25,26,27]. The former helps discover useful knowledge from the linked data about, and interesting relationships among, social entities in social networks; the latter helps predict future game results for some sports (e.g., soccer, American football, basketball, hockey) by using historical game results.

* Corresponding author.
E-mail address: kleung@cs.umanitoba.ca (C.K. Leung)

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.153

Many existing sports data mining research projects focus on scheduling of games [32], visualization of games or players [2,21], sport advising [20], as well as identification and extraction of interesting moments or players from sports game video [13,16,17,30,33]. While these research projects are interesting and practical, it is also interesting to predict the outcomes of games (e.g., predict the outcomes of basketball games [22,29] or soccer games [19]).

In general, making predictions [3,31] is not an easy problem. Traditionally, some human domain experts (e.g., sports commentators, television and other media personalities, former players, coaches) make predictions on the game results based on their experience, instinct, and/or gut feeling. For example, pre-game analysis of televised sporting events often includes expert predictions. These predictions vary in accuracy as they are typically based on subjective claims and anecdotal evidence.

To reduce the subjectivity of making predications based on instinct or gut feeling, a quantitative prediction—which uses some form of objective data to predict the outcome of a game—is used. For example, some professional domain experts make their objective predictions based on historical results.

Nowadays, there is a growing contingent in the sports community that seeks to look at this research problem of predicting sports outcomes more analytically. In recent years, simple statistical algorithms (e.g., ranking algorithms) have been applied to historical results in the decision making process for the prediction of results for future games. These algorithms typically consider the two teams competing and compare their respective strengths and weaknesses in order to make a prediction. Usually, the longer the historical data, the more accurate are the results. However, predictions based on these simple statistical algorithms may not be very accurate when the two teams have not competed with each other.

The aforementioned problems become worse when trying to make predictions on college football games partially due to the lack of available data for the group of players that make up the teams. Note that the teams competing in a given college football game typically have not previously played each other in the season. Moreover, while a veteran playing in a professional sporting league may have several years of data outlining his past performance, National Collegiate Athletic Association (NCAA) athletes are only allowed to compete in their sports for four seasons. In college football, a good player in a good football program is likely not able to get significant playing time until his third or fourth year in college, further limiting the amount of relevant data. Hence, a logical question to ask is: can sports data mining help obtain more accurate predictions, especially for mining college football data? In response, we focus on mining sports data, especially mining relevant American football data (which spanned over a short period) from cfbstats.com to make accurate predictions on the outcomes of college football games.

The remainder of this paper is organized as follows. The next section provides some background information about American football with a focus on the college football. Section 3 gives related works. We then present our sports data mining approach in Section 4. Evaluation results and conclusions are shown in Sections 5 and 6.

2. Background

As this paper focuses on mining and result predictions on college football games, we provide some background information about the game of American football—especially, the games of college football. In a game of *American football* (also known as *gridiron football*), two teams compete to see who will score the most points. Each team sends out 11 players, alternating between offense and defence:

(i) The *offense* is the team with the ball, and attempts to move the ball to the end of the field (or "end zone"). This can be accomplished either by running with the football or by throwing it down the field to one's teammates downfield. Offensive players can be further divided into the following subcategories:

  1.1. The *backs* are the skill players who move the ball forward, with (i) the *quarterback* primarily doing so by throwing the football and (ii) the *running backs* doing so by running with it.

  1.2. The *offensive line* is a group of blockers whose job is to protect the backs from the defensive players coming to tackle them.

  1.3. The *receivers* run down the field to catch balls that are thrown forward.

(2) The *defence* is the team that aims to prevent the offense from moving the ball forward. Ideally, it aims to take the ball away from the offense.

In this paper, we apply sports data mining to analyze American football data from the National Collegiate Athletic Association (NCAA). In particular, we focus on data from the Football Bowl Subdivision (FBS), which is comprised of top-level college football teams in the USA. Generally, college football is highly popular in the USA. The "playoff" system seems to be fairly complicated. Teams in the Football Bowl Subdivision (FBS) are normally divided into conferences, and proceed to play a 12-game regular season (typically, 7 or 8 games against teams within a conference, and the remaining 5 or 4 against other teams). Some of the conferences then have a conference championship game between their two best teams, adding a 13th game for some teams. Traditionally, the teams are then ranked using a formula that takes into account (i) computerized rankings, (ii) the rankings by the coaches of each team, and (iii) the rankings of key members of the media. The best of these teams are placed into a final "bowl" game, with the highest ranked teams competing for the more prestigious bowls. For instance, at the end of the 2013-14 college football season, the best 70 of 124 teams competed for 35 bowls, with the highest ranked two teams (namely, Florida State University and Auburn University) playing in the Bowl Championship Series (BCS) championship—which is the most prestigious bowl game.

3. Related works

The Society for American Baseball Research (SABR) and its various counterparts in other sports have used a number of new statistics that are typically more accurate measures of an individual's or team's abilities than their traditional counterparts. For example, in baseball, *sabremetrics* are defined as the search for objective knowledge in baseball. Recently, sabremetrics have become more widely accepted amongst baseball fans and decision makers for professional baseball teams. While these statistics are primarily developed to measure the value of individuals, baseball is a sport in which the sum of individuals' performances can be used to measure the success of a team as a whole. Although baseball is a team sport, the successes or failures of individuals can often be viewed objectively and be viewed independent of the successes or failures of the rest of the team.

In American football, it is not easy to view the data collected and the resulting statistics so concretely. If a receiver drops an easily catchable ball, the quarterback's statistics are negatively affected. If an offensive lineman makes a mistake and fails to block properly on a running play, it worsens the statistics of the running back. Even if these kinds of things are taken into account, it is not always entirely clear who is responsible for successes or failures, and how responsible each individual was. Although advanced football statistics are gaining some ground in the mainstream football community, it is much harder to take individual statistics in football and effectively correlate them to wins.

Quantitative prediction models that are currently widely used can be divided into two main categories: (i) simulation based and (ii) statistics based. *Simulation based models* predict the outcome of a game by running it through a sport-specific simulation engine a single or multiple times. Modern sports video game franchises often use the newest version of their video games to simulate the outcome of a season around the time of their release, which is typically shortly before the start of its sport's respective season. The games are also sometimes used to predict the outcome of a championship game or series. As an example, the Madden NFL video game predicts the outcome of the Super Bowl—an American pro-football's championship game—a week or so before the actual game. The website WhatIfSports.com uses a college football simulation engine to simulate each of the 35 bowl match-ups a thousand times, which correctly predicted 79.8% of all 2013 FBS games. Being able to make predictions on regular season games is one of the inherent advantages to the simulation based approach. However, the lack of pre-season games and high turnover rate for its players (with new players replacing those who graduated or left college early to play professional sports) in College football would make it difficult for those algorithms that require historical statistics to make meaningful prediction. A drawback of simulation based prediction is that creating such a simulation engine can be very complicated. Moreover, in order to properly run simulations, one needs to keep track of a huge number of qualitative inputs (e.g., each player's skills and tendencies). For example, EA creates player ratings to determine the skills of the players in their games. Although some of these ratings can be created based on objective data, many of them can be considered fully or partly subjective.

In contrast, the *statistics based models* predict outcomes of games based on the statistics of the competing teams, or the players of the competing teams. For instance, a simple way to predict game outcomes is to always pick the

team with the highest *winning percentage* (*WP*), which is defined as the number of wins divided by the number of games played:

$$\begin{aligned} WP &= \#Wins / \#Games\_Played \\ &= \#Wins / (\#Wins + \#Losses). \end{aligned} \quad (1)$$

Potential problems with this approach of using the highest winning percentage include (i) two teams may have an identical winning percentage, (ii) the approach does not take into account how easy or difficult the wins were for each team to earn.

Alternatively, another common sports statistic—*strength of schedule* (*SOS*) [9], which measures the *opponents' record* (*OR*) and was originally defined as a quotient of the sum of all the opponents' wins divided by the sum of opponents' games played—is used to track the quality of a team's opponents. Moreover, there are some variants of SOS. An example is to define it as a weighted sum of the opponents' record (OR) and the *opponents' opponents' record* (*OOR*). Some statistics based prediction models take into account both the winning percentage and SOS statistics. For example, the *Ratings Percentage Index* (*RPI*) [9], which is used as part of the calculation for the NCAA's football rankings, is defined as a weighted sum of (i) winning percentage, (ii) OR, and (iii) OOR:

$$RPI = (0.25 * WP) + (0.5 * OR) + (0.25 * OOR). \quad (2)$$

Additional layers can be added to these statistics based prediction algorithms to increase their complexity, and potentially their success. For instance, *Elo ratings system* [7,8], which was originally designed to rank chess players, can be used to compare sports teams and to rank these teams. Given two teams (A and B) and their corresponding strength or ratings (denoted as Elo(A) and Elo(B)), the expected (initial) probability that Team A would win a hypothetical game can be calculated by the following equation:

$$Expected\_Prob(A) = 1 / (1 + 10^{(Elo(B) - Elo(A))/400}). \quad (3)$$

To improve the accuracy, the ratings are refined after each game has actually taken place. Specifically, if Team A has defeated Team B, then Team A's new rating becomes the sum of (i) its old rating and (ii) probability of A wins, i.e.,

$$New\_Elo(A) = Old\_Elo(A) + K * (Actual\_Prob(A) - Expected\_Prob(A)), \quad (4)$$

where K is a constant and Expected_Prob(A) is the initial probability calculated by Equation (3). On the other hand, if Team A has just lost, then its new rating becomes the difference between (i) its old rating and (ii) probability of A wins, i.e.,

$$New\_Elo(A) = Old\_Elo(A) - K * (Actual\_Prob(A) - Expected\_Prob(A)). \quad (5)$$

The Elo rating is a useful measure for sports games because it provides a reasonable expected winning probability and refines iteratively based on the team's successes and failures. However, there are a few potential problems associated with the Elo rating. First, it weighs every win equally (e.g., a 40-point victory provides the same adjustment to a team's Elo rating as a 3-point win). Hence, the Elo rating benefits those teams that just barely defeat stronger opponents over those teams that win convincingly against weaker opponents. Second, it is not easy to find a good initial ranking for teams. Although the fairest way would be to give every team in the league the same rating at the beginning of the year, this would prevent a team from archiving a substantial reward to their Elo ratings for defeating a strong opponent early in the season.

Similar to the Elo ratings (which were designed to rank chess players), *Pythagorean wins* [18,24] were proposed to measure the success of baseball teams. The intuition is that the number of points (or "runs" in baseball lingo) that a

team scored and allowed was a better way of evaluating a team's strength than the number of games that they won and lost. Specifically, the *Pythagorean wins* can be theoretically computed by the following equation:

$$\text{Original\_Pyth\_Wins} = PF^2 / (PF^2 + PA^2). \quad (6)$$

where PF ("points for") represents the number of points scored by the team and PA ("points against") represents the number of points surrendered.

Practically, the key idea of Pythagorean wins is to project how many wins a team ought to have based on how many points they are scoring and allowing. It was suggested that more accurate football prediction (i.e., the Pythagorean wins would be closest to the number of actual wins) can be obtained if the exponents in the formula were changed from 2 to 2.37:

$$\text{Pyth\_Wins} = PF^{2.37} / (PF^{2.37} + PA^{2.37}). \quad (7)$$

This statistic has the opposite effect to a team's perceived value when compared with the Elo ratings. With Pythagorean wins, teams are rewarded for the margin of victories and defeats, ignoring the number of wins and losses and the quality of opponents all together.

Example 1. To demonstrate the difference among the above formulas, let us consider Table 1.

Table 1. Statistics for Teams A and B for the prediction of the winner based on (i) WPs, (ii) RPI, (iii) Elo rating, and (iv) Pythagorean wins.

Team	Wins	Losses	OR	OOR	Elo	PF	PA
A	5	1	0.556	0.611	1850	112	57
B	0	6	0.500	0.512	1200	41	153

For the following four statistics used in this example to predict the outcome of a hypothetical game between Teams A and B, a higher value for statistics will mean that team is more likely to win:

(i) By considering only the winning percentages (WPs), Team A is superior to Team B because the WPs of Teams A and B are $5/(5+1) \approx 0.833$ and $0/(0+6) = 0$, respectively. Thus, Team A would be the predicted winner.

(ii) When considering ratings percentage index (RPI), Team A is again superior to Team B because RPIs of Teams A and B are $(0.25*0.833) + (0.5*0.556) + (0.25*0.611) = 0.639$ and $(0.25*0) + (0.5*0.500) + (0.25*0.512) = 0.378$, respectively.

(iii) Moreover, Team A with the Elo rating of 1850 is more superior to Team B with the Elo rating of 1200, i.e., a 650-point difference.

(iv) Similar comments apply when using the Pythagorean wins. Although there is a large difference between the points Team A has scored (PF for A = 112) and the number of points that Team B has scored (PF for B = 41), but there is a small difference in the number of points either team has surrendered (PA for A is 57 and PA for B is 153). As Pythagorean wins for Teams A and B are $112^{2.37}/(112^{2.37}+57^{2.37}) \approx 0.832$ and $41^{2.37}/(41^{2.37}+153^{2.37}) \approx 0.042$ respectively, Team A is gained the predicted winner (with higher Pythagorean wins than Team B).

To summarize, for all four statistics, Team A is the predicted winner over Team B. □

Example 2. Similar to Example 1, let us consider Table 2 for games between Teams C and D.

Table 2. Statistics for Teams C and D for the prediction of the winner based on (i) WPs, (ii) RPI, (iii) Elo rating, and (iv) Pythagorean wins.

Team	Wins	Losses	OR	OOR	Elo	PF	PA
C	4	2	0.444	0.412	1600	72	73
D	2	4	0.583	0.517	1450	81	75

By considering only the WPs, Team C is superior to Team D because the WPs of Teams C and D are 4/(4+2) ≈ 0.667 and 2/(2+4) ≈ 0.333, respectively. Thus, Team C would be the predicted winner. Similar comments apply to the Elo ratings (1600 for Team C vs. 1450 for Team D).

However, it is interesting to observe that a different result is predicted when using the RPI or Pythagorean wins. To elaborate, as the WPs of Teams C and D are 0.667 and 0.333, the RPIs of Teams C and D are (0.25*0.667) + (0.5*0.444) + (0.25*0.412) ≈ 0.492 and (0.25*0.333) + (0.5*0.583) + (0.25*0.517) ≈ 0.504, respectively. This means that Team D is the predicted winner. Moreover, the Pythagorean win also favours Team D (with Pyth_Wins ≈ $81^{2.37}/(81^{2.37}+75^{2.37})$ ≈ 0.545) over Team C (with Pyth_Wins ≈ $72^{2.37}/(72^{2.37}+73^{2.37})$ ≈ 0.492). Pythagorean wins suggested that, up to this point, Team D has been fortunate to have as much success as it has had, while Team C comparably unfortunate. When predicting the results over a 12-game season, Pythagorean wins suggest that (i) Team D would win 0.545 * 12 ≈ 7 games and (ii) Team C would only win 0.492 * 12 ≈ 6 games.

To summarize, unlike Example 1 (where one team is unanimously predicted to be the winner regardless which of the four statistics were applied), different prediction results are obtained here when using different statistics. For instance, (i) Team C is the predicted winner when using WPs or the Elo ratings, but (ii) Team D becomes the predicted winner when using the RPI or Pythagorean wins. □

Christodoulou [4] considered an alternative rankings-based approach and suggested that *margins of victory* should matter. He went through every game during the regular season and awards portions of 1 point to each competing team depending on who won and the margin of victory. For example, "in the case of three 7-point victories, a 42-35 win earns 0.542, a 14-7 win earns 0.630, and a 7-0 game still earns 0.769". This approach seems to give preference to good defence over good offense. Larger margins of victory also add more points for the victor. The results are adjusted by the strength of schedule (SOS) of each team, based on the rankings. Because the strength of schedule requires the strength of the teams (which, in turn, relies on the strength of schedule), this entire process is iterated until the rankings are no longer fluctuate. Christodoulou suggested that "a full regular season usually requires about 175 iterations until the system reaches stability". Again, this approach picks the teams based on their rankings in this stable system. We label this as "Christodoulou's first approach" in the evaluation results in Section 4.

Christodoulou [4] also introduced another algorithm. Instead of considering margins of victory, this algorithm goes through the list of game results, and calculates four statistics for each team: (i) points per game (PF), (ii) points against per game (PA), (iii) offensive output (Off_Output), and (iv) defensive output (Def_Output). *Offensive output* is the percentage of points that a team scores against their opponents relative to how many points these opponents typically allow. *Defensive output* is the percentage of points that a team allows to their opponents relative to how many points their opponents typically score. Moreover, Christodoulou suggested that these statistics "can be used to predict the game score". So, we could simply combine everything and take the averages. We label this as "Christodoulou's second approach" in the evaluation results in Section 4.

Example 3. Consider a game between Teams E and F, with F (before playing Team E) having PA=10.0. On the one hand, if team E scores 20 points, then its offensive output for that game would be 20/10.0 = 200%. On the other hand, if Team E only scores 3 points, then its offensive output would be 3/10.0 = 30%. □

Example 4. Let us consider another game between Teams E and F, whose statistics appear in Table 3.

Table 3. Statistics for Teams E and F for the prediction of game results based on PF, PA, Off_Output, and Def_Output.

Team	PF	PA	Off_Output	Def_Output
E	30.5	15.0	150%	50%
F	22.0	10.0	125%	40%

Based on the above table, as (i) Team E normally scores 30.5 points per game (i.e., PF(E)=30.5) and (ii) Team F only allows a team to score 40% of what it typically does, this leads to 30.5 * 40% = 12.2 points. As (i) Team F normally allows 10.0 points against per game (i.e., PA(F)=10.0) and (ii) Team E's offensive output suggests that it typically scores 150% of what opponents allow, then this leads to 10.0 * 150% = 15 points. So, the average is (12.2 + 15 points) / 2 = 13.6 points ≈ 14 points.

Similarly, as (i) Team F should be able to score to 11 points based on Team E's defensive output (because PF(F) * Def_Output(E) = 22.0 * 50% = 11 points) but (ii) F should muster 18.75 points based on Team F's offensive output (because PA(E) * Off_Output(F) = 15.0 * 125% = 18.75 points), this leads to an average of (11 + 18.75 points) / 2 = 14.875 points ≈ 15 points. Hence, Team F is predicted to defeat Team E by a score of 15-14. □

Delen et al. [6] used three prediction methods—namely, (i) neural networks, (ii) decision trees, and (iii) support vector machines (SVMs) to predict the 2010-11 bowl season based on the data collected from every bowl game between the 2002-2003 and 2009-2010 seasons (for a total of eight seasons), with "a dataset which included 36 variables, of which the first six were identifying variables ... followed by 28 input variables ... and finally, the last two [being] output variables". Their neural network consisted of a multi-layer perceptron (MLP) with a back-propagation supervised-learning algorithm. Classification and regression trees were used for the decision tree. The accuracy of predictions based on the neural network, decision tree, and support vector machine were 71.43%, 74.29%, and 82.86%, respectively.

The aforementioned statistics-based and simulation-based models are similar in the sense that they directly compare and contrast two competing teams. While this can be viewed as a benefit, there exists an unavoidable uncertainty of how two teams will matchup against one another when a future game is played. In order to avoid that direct comparison altogether, other comparisons would need to be made, introducing a different set of uncertainties.

4. Our sports data mining approach

Here, we present our *sports data mining* approach, which avoids calculating which of the two competing teams is more likely to win. The key idea is that we analyze a set of teams that are the most similar to each of the competing teams, find the results of the games between the teams in each of the two sets, and use those game results to predict the outcome of the game between the original two teams.

Our approach analyzes past game results and a number of statistics about each of the teams from each of those games (e.g., passing attempts, rushing attempts, and turnovers for and against). After scanning the statistical data, they are stored in two data structures: (i) a list storing every game played over a given time and (ii) another list storing all teams with their corresponding statistics from the season. Our approach then parses the team lists and creates a map with every point representing a team. The distance between two points on the map is proportional to their similarity. More specially, to find all of the teams that are similar to a given team, our approach represents each team as a point on a 4-dimensional space representing four different statistics: (i) RPI, (ii) Pythagorean wins, (iii) offensive strategy, and (iv) turnover differential.

Recall that RPI and Pythagorean wins, which measure the overall strength of teams, can be computed as follows:

$$RPI = (0.25 * WP) + (0.5 * OR) + (0.25 * OOR), \text{ and}$$

$$Pyth\_Wins = PF^{2.37} / (PF^{2.37} + PA^{2.37}).$$

Offensive strategy measures how the offense prefers to move the ball forward, passing or rushing:

$$Offensive\_Strategy = Passing\_Attempts / Rushing\_Attempts. \quad (8)$$

Note that, as the ideal goal of a defence is to try to take the ball away from the offense, turnover refers to an unplanned change in possession of the football. Hence, a team's *turnover differential* is defined as the number of

times they take the ball away from an opposing team (takeaway) minus the number of times that the team turns the ball over to the other team (giveaway), i.e.,

Turnover_Differential = Takeaways – Giveaways, where (9)

(9.1) Takeaways = Interceptions + Fumbles_Recovered, and

(9.2) Giveaways = Interceptions_Lost + Fumbles_Lost.

We normalize these four statistics. Specifically, to weigh the statistics, we find their maximum and minimum values. The minimum values for all four statistics are 0; the maximum values for RPI, Pythagorean wins, offensive strategy, and turnover differential are 150, 100, 50, and 50, respectively. In other words, (i) RPI $\in$ [0, 150], (ii) Pyth_Wins $\in$ [0, 100], (iii) Offensive_Strategy $\in$ [0, 50], and Turnover_Differential $\in$ [0, 50]. Then, our sports data mining approach sums the values for the four statistics. The team with the highest sum is the predicted winner.

5. Evaluation results

To evaluate our sports data mining approach, we used real-life statistical data from cfbstats.com for past college football games (e.g., regular season college football games between 2005-2006 and 2013-2014) and a number of statistics about each of the teams from each of those games (e.g., passing attempts, rushing attempts, and turnovers for and against).

To illustrate our evaluation, let us consider Example 5 showing a case drawn from the aforementioned statistical data. In the example, we analyze a future game between Teams G and H by first determining the similarity among all teams and then adding scores to the predicting teams. The team in the predicting pairs having the higher sum would be the predicted winner.

Example 5. Given from this real-life sports database that (i) Teams J and K are similar to Team G with the distance between G and J be dist(GJ)=2 and that between G and K be dist(GK)=4, and (ii) Team L is similar to Team H with the distance between H and L be dist(LH)=2. Then, as a means of accounting for closeness, we divide them by the sum of the distance between the predicted teams and the already played teams (+1 when no game has been played between either team) before adding the points.

If (i) no game has been played between G and H but (ii) J (which is similar to G) defeated L (which is similar to H) with a score of 24-17, then we add 24/5 = 4.8 points to G and add 17/5 = 3.4 points to H. Here, the denominator 5 is the sum of distances: dist(GJ) + dist(JL) + dist(LH) = 2 + 1 (already played) + 2.

Next, with additional information that (iii) K (which is similar to G) lost to L (which is similar to H) with a score of 10-35, we add another 10/7 ≈ 1.429 points to G and another 35/7 = 5 points to H. Here, the denominator 7 is the sum of distances: dist(GK) + dist(KL) + dist(LH) = 4 + 1 (already played) + 2.

Now, as H has a higher sum (of 3.4 + 5 = 8.4 points) than G (having a sum of 4.8 + 1.429 = 6.229 points), Team H is predicted to defeat G. □

Next, we test the accuracy of our sports data mining approach in predicting NCAA bowl games at the end of each of the aforementioned seasons. Although many prediction models list results over the course of the entire season, many of these games are mismatched (e.g., one participating team is clearly stronger than their opponent). Teams that are selected for bowl games are often relatively evenly matched, making predicting the outcome more challenging.

First, we compared our approach with existing models (e.g., WhatIfSports.com's simulations, Christodoulou's approaches [4], as well as Delen et al.'s three prediction models on neural network, SVM and the decision tree [6]) with data from the 2010-2011 season. The comparison results shown in Table 4 indicate that our approach—which considers four statistics—leads to relatively high accuracy of 91.43% for the 2010-2011 season (cf. accuracy ranging from 57.14% to 82.86% by other existing models) in predicting the results for the College Football games.

Table 4. Comparison between our approach and existing models for the 2010-2011 season.

The 2010-2011 season	#Wins (among 35 bowls)	Accuracy
Whatifsports.com	20	57.14%
Christodoulou's first approach [4]	23	65.71%
Christodoulou's second approach [4]	24	68.57%
Neural network [6]	25	71.43%
Support vector machine [6]	26	74.29%
Decision tree [6]	29	82.86%
Our sports data mining approach	32	91.43%

Next, we compared our approach with existing models (e.g., WhatIfSports.com's simulations, and Christodoulou's approaches [4]) with data from the 2011-2012 season. The comparison results shown in Table 5 indicate that our approach—which considers four statistics—again leads to relatively high accuracy—this time, 97.14% for the 2011-2012 season (cf. accuracy of 68.57% by other existing models) in predicting the results for the College Football games. Here, we did not compare with Delen et al.'s three prediction models on neural network, SVM, and the decision tree because they only reported their results for the 2010-2011 season [6].

Table 5. Comparison between our approach and existing models for the 2011-2012 season.

The 2011-2012 season	#Wins (among 35 bowls)	Accuracy
Whatifsports.com	24	68.57%
Christodoulou's first approach [4]	24	68.57%
Christodoulou's second approach [4]	24	68.57%
Our sports data mining approach	34	97.14%

Similarly, we also compare our approach with existing models with data from the 2012-2013 and 2013-2014 seasons. The comparison results again indicate that our approach—which considers four statistics—leads to relatively high accuracy in predicting results for College Football games. For instance, Pedersen [23] accurately predicted 19 wins. Mandel [14] and WhatIfSports both accurately predicted 20 wins. Preliminary results of our sports data mining predicted more wins (e.g., at least four more wins) for the 2013-2014 season. As ongoing work, more analyses are conducted.

6. Conclusions and future work

In this paper, we presented a sports data mining approach to predict the winners of college football bowl games. Instead of using the traditional approach of comparing the statistics of the two competing teams and projecting the outcome, our approach predicts the outcomes based on the historical results of games. The competing teams are compared to other teams, and game results are pulled from those similar teams and used in the prediction algorithm. The evaluation results show the accuracy of our sports data mining approach in predicting the outcomes of bowl games in recent seasons. As future work, we plan to analyze data for future college football bowl games seasons and/or other sports games.

Acknowledgements

This project is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Manitoba.

References

1. Budhia BP, Cuzzocrea A, Leung CK. Vertical frequent pattern mining from uncertain data. In: *Proceedings of the KES 2012*. IOS Press; 2012, p. 1273-1282.
2. Choros K. Temporal aggregation of video shots in TV sports news for detection and categorization of player scenes. In: *Proceedings of the ICCCI 2013*. Springer; 2013, p. 487-497.
3. Chowdhury NK, Leung CK. Improved travel time prediction algorithms for intelligent transportation systems. In: *Proceedings of the KES 2011, Part II*. Springer; 2011, p. 355-365.
4. Christodoulou A. Anthony Christodoulou - football algorithms. http://anthonychristodoulou.com/
5. Cuzzocrea A, Leung CK, MacKinnon RK. Mining constrained frequent itemsets from distributed uncertain data. *Future Generation Computer Systems* 2014; **37**:117-126.
6. Delen D, Cogdell D, Kasap N. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting* 2012; **28**(2):543-552.
7. Elo A. *The rating of chessplayers, past and present*. New York, NY: Arco Pub.; 1978.
8. Elo A. *The rating of chessplayers: past and present*. Bronx, NY: Ishi Press; 2008.
9. Johnson G. Baseball committee recommends changes to RPI. *NCAA News*; Aug. 03, 2011.
10. Leung CK, Jiang F. Frequent pattern mining from time-fading streams of uncertain data. In: *Proceedings of the DaWaK 2011*. Springer; 2011, p. 252-264.
11. Leung CK, Medina IJM, Tanbeer SK. Analyzing social networks to mine important friends. In: Xu G, Li L, editors. *Social media mining and social network analysis: emerging research*. IGI Global; 2013, p. 90-104.
12. Leung CK, Tanbeer SK, Cameron JJ. Interactive discovery of influential friends from social networks. *Social Network Analysis and Mining* 2014; **4**(1): art. 154.
13. Lu WL, Ting JA, Little JJ, Murphy KP. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2013; **35**(7):1704-1716.
14. Mandel S. College football pickoff: 2013-14 bowl season. *CNN Sports Illustrated*; Dec. 20, 2013.
15. Matsumotoa T, Kojirib T. Baseball coaching ability development system based on externalization of decision process. In: *Proceedings of the KES 2013*. Elsevier; 2013, p. 653-661.
16. Mentzelopoulos M, Psarrou A, Angelopoulou A, Rodríguez JG. Active foreground region extraction and tracking for sports video annotation. *Neural Processing Letters* 2013; **37**(1):33-46.
17. Messelodi S, Modena CM. Scene text recognition and tracking to identify athletes in sport videos. *Multimedia Tools and Applications* 2013; **63**(2):521-545.
18. Miller SJ. A derivation of the Pythagorean won-loss formula in baseball. *Chance Magazine* 2007; **20**(1):40-48.
19. Min B, Kim J, Choe C, Eom H, McKay RI. A compound framework for sports results prediction: a football case study. *Knowledge-Based Systems* 2008; **21**(7):551-562.
20. Mosqueira-Rey E, Prado-Gesto D, Fernández-Leal Á, Moret-Bonillo V. Prosaico: characterisation of objectives within the scope of an intelligent system for sport advising. In: *Proceedings of the KES 2012*. IOS Press, p. 238-247.
21. Mueller F, Khot RA, Chatham AD, Pijnappel S, Toprak C, Marshall J. HCI with sports. In: *Proceedings of the ACM CHI 2013, Extended Abstracts*. p. 2509-2512.
22. Odachowski K, Grekow J. Predicting the final result of sporting events based on changes in bookmaker odds. In: *Proceedings of the KES 2012*. IOS Press, p. 278-287.
23. Pedersen B. College football bowl picks 2013-14: predictions for every game. *Bleacher Report*; Dec. 08, 2013.
24. Schatz A, Benoit A, Connelly B, Farrar D, Fremeau B, Gower T, Hinton M, McCown R, McIntyre B, Stuart C, Tanier M, Tuccitto D, Verhei V, Weintraub R. *Football outsiders almanac 2013: the essential guide to the 2013 NFL and college football seasons*. Seattle, WA: CreateSpace; 2013.
25. Schumaker RP, Solieman OK, Chen H. Predictive modeling for sports and gaming. In: Schumaker RP, Solieman OK, Chen H, *Sports data mining*. Springer; 2010, p. 55-63.
26. Schumaker RP, Solieman OK, Chen H. *Sports Data Mining*. New York, NY: Springer; 2010.
27. Seif El-Nasr M, Drachen A, Canossa A, eds. *Game analytics: maximizing the value of player data*. London, UK: Springer; 2013.
28. Tanbeer SK, Leung CK, Cameron JJ. Interactive mining of strong friends from social networks and its applications in e-commerce. *Journal of Organizational Computing and Electronic Commerce* 2014; **24**(2-3):157-173.
29. Vaz de Melo POS, Almeida VAF, Loureiro AAF, Faloutsos C. Forecasting in the NBA and other team sports: network effects in action. *ACM Transactions on Knowledge Discovery from Data* 2012; **6**(3): art. 13.
30. Weber BG, John M, Mateas M, Jhala A. Modeling player retention in Madden NFL 11. In: *Proceedings of the IAAI 2011*. AAAI; 2011, p. 1701-1706.
31. Yang X, Mikamori Y, Tanaka-Yamawaki M.: Predicting the security levels of stock investment by using the RMT-test. In: *Proceedings of the KES 2013*. Elsevier; 2013, p. 1172-1181.
32. Zeng L, Mizuno S. On the generalized mirrored scheme for double round robin tournaments in sports scheduling. *Asia-Pacific Journal of Operational Research* 2013; **30**(3): art. 1340008.
33. Zhang N, Duan LY, Li L, Huang Q, Du J, Gao W, Guan L. A generic approach for systematic analysis of sports videos. *ACM Transactions on Intelligent Systems and Technology* 2012; **3**(3): art. 6.