Курсов проект по Изкуствен интелект
на тема

Предсказване на изход от футболни срещи

Михаил Николов (ФН: 80527)
студент в 4-ти курс, Компютърни науки,
ФМИ към СУ „Св. Климент Охридски“

08 февруари 2014г.

Съдържание

Кратък обзор на проблема

Предсказването на резултати от футболни срещи представлява интерес за голям процент от залагащите. Изходът от дадена среща зависи от огромен брой фактори, което прави опитите за отгатване на крайния победител изключително трудни. Поради невъзможността да се извлекат данни за всички възможни фактори влияещи на играта на двата отбора, е необходимо да се определи някакъв минимум от фактори, който има най-силно влияние върху изхода от двубоя.

Голяма част от футболните фенове, залагащи на краен изход от срещи разчитат на своите наблюдения върху играта на двата отбора и на личната си преценка за влиянието на известните им фактори. При този подход субективността на оценката е изключително голяма.

Използването на алгоритми за изкуствен интелект е един от методите за избягване на субективния момент и постигане на максимално точна преценка на базата на наличните данни. Тук от голямо значение е освен използваният алгоритъм и наличните данни и начина по който се използват.

В тази посока са правени множество опити за реализация на алгоритми за предсказване на крайния изход от двубои в най-различни отборни спортове. Най-често използваните са алгоритмите за класификация, както и Бейсовите и невронните мрежи.

Използвани технологии и извличане на данни

Проектът е реализиран изцяло на PHP, което позволява лесното му изпълнение като уеб приложение. Базата данни, използвана за съхранение на данните, е MySQL. Приложението изпълнява алгоритъма с входни данни двата отбора, изходът от двубоя между които искаме да предскажем. Данните са предварително извлечени от българският портал за футбол football24.bg.

Приложението може да извлече още данни по зададен URL адрес на страница с резултати от посочения сайт. Резултатите ще бъдат записани в базата

данни и ще бъдат асоциирани с отборите (информация, за които също се съхранява в базата). Добавянето на нови отбори става автоматично.

Реализация на алгоритъма

В конкретната реализация е използван Наивен Бейсов класификатор. Целта е от наличните данни за бъдат извлечени зависимости (знания), които да бъдат използвани за определяне принадлежността на зададения двубой към някои от трите класа – **1** (победа за домакина), **x** (равенство), **2** (победа за госта). За изчисляването вероятността за принадлежност към всеки един от трите класа се използва *формулата на Бейс*. Чрез нея се търси вероятността за изход 1, x или 2 при наличните условия, като приемаме, че те са независими едно от друго (*наивен Бейсов класификатор*).

Факторите които ще вземем предвид са следните:

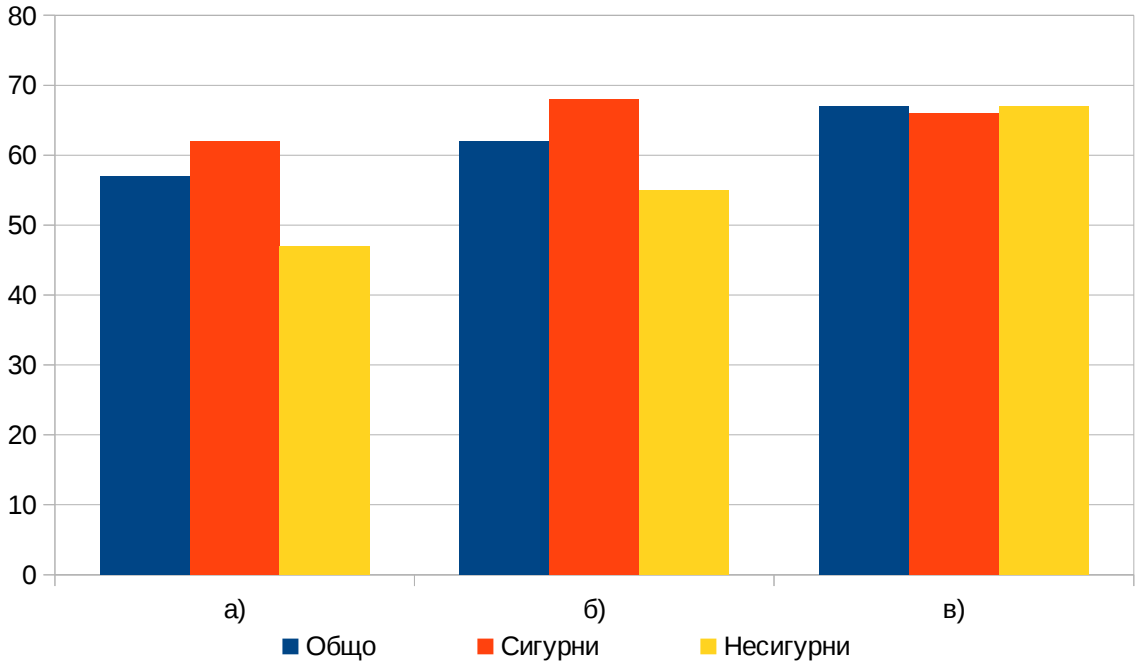
- домакинство – сред наличните данни за последните 30 мача ще се търси зависимост между това дали отборът е домакин или гост и крайния изход от двубоя.
- последни 2-3 мача - сред наличните данни ще се търси зависимост между постигнатите резултати в предходните 2-3 срещи и крайния изход от двубоя.

В стремежа да постигнем по-достоверно предположение и да повишим точността на алгоритъма ще комбинираме данните за всички изминали двубои на двата участващи в срещата отбора, както и данните от мачовете изиграни между двата отбора.

В тази реализация на алгоритъма разглеждаме отборите като едно цяло, абстрахирайки се от футболистите участващи в дадена среща. Поради това при изчисляване на вероятностите вземаме предвид само последните 30 срещи на всеки отбор (колкото се изиграват за една година). Използването на данни от по-стари срещи само би намалило точността на алгоритъма, тъй като тогава отборът е бил сформиран от други играчи, голяма част от които днес вече не са част от него.

Резултати от проведените експерименти

Първите резултати от проведените тестове са само на база последните 30 изиграни срещи и изхода от последните 3 двубоя, без да вземаме предвид двубоите между двата отбора. При тези обстоятелства получаваме малко над **57%** познаваемост (фиг. 1а).



Но нека разделим предсказанията на две категории – „сигурни“ и „несигурни“. За сигурни ще считаме резултатите, при които изчислената вероятност за даден изход от двубоя (например победа за домакина) е по-голяма от сбора на вероятностите за другите две възможности. При това забелязваме, че при посочените по-горе обстоятелства имаме почти *67%* точност за сигурните мачове, но едва *47%* за несигурните.

За да компенсираме тази разлика ще опитаме да комбинираме 1:1 получените дотук вероятности със същите изчисления, но на база само мачовете между двата отбора. Разликата е очевидна. Познаваемостта на несигурните мачове се увеличава на *55%*, на сигурните също – на *68%*, което прави **62%** общо (фиг. 1б).

Очевидно включването на статистика от двубоите между двата отбора ни осигурява по-голя точност. В такъв случай, ще опитаме да дадем предимство на този фактор, като умножим вероятностите получени от срещите между двата отбора по 0.6, а останалите – по 0.4. Резултатът е още по-добър. Вече разликата между сигурни и несигурни срещи е стопена, като и за двете категории имаме *66-67%* или общо **67%** успеваемост (фиг. 1в).

Възможности за развитие

Има възможност за постигане по-голяма точност с въвеждането на повече фактори, определящи крайния изход и извличане на повече данни за срещите.

В настоящата реализация на алгоритъма, на всеки отбор се гледа като едно неделимо цяло без да се вземат предвид играчите участващи в двубоите. Въвеждането на данни за футболистите взели участие във всеки мач и изчисляването на вероятностите на база играчи, а не отбор би дало по-достоверна информация за текущото състояние на всеки отбор и шансовете му за успех. Текущата реализация на алгоритъма не може да вземе предвид факта, че в даден

мач са взели участие едва, например, 3-ма от футболистите, които ще играят и в предстоящия двубой, и ще счита данни от този мач за също толкова достоверни, колкото и от всеки друг. Но всъщност това е друга съвкупност от футболисти със своите предимства и недостатъци, които биха се представили по различен начин на терена.

Вземането предвид на тези фактори ще увеличи точността и ще създаде по-пълна картина за предстоящите двубои.

Декларация за липса на плагиатство

- Тази курсова работа е моя работа, като всички изречения, илюстрации и програми от други хора са изрично цитирани.
- Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.
- Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.
- Трите имена и подпис на студента:

Използвани източници

- http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-8.html
- http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- http://blog.smellthedata.com/2011/02/thoughts-on-modeling-basketball.html