

А.Д. Цаконос, магістр, аспір.
Г.Д. Дуніас, доктор, лектор
Егейський університет, Хіос, Греція
С. Штовба, к.т.н., доц.
Вінницький державний технічний університет

ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ФУТБОЛЬНИХ МАТЧІВ ЗА ДОПОМОГОЮ МАШИНИ ОПОРНИХ ВЕКТОРІВ

Пропонується метод прогнозування результатів футбольних ігор, що ґрунтується на такій технології софт-комп'ютерингу, як автоматичне навчання на базі машини опорних векторів. Розроблена в статті модель прогнозування враховує такі показники команд: різниця кількості вибувливих провідних гравців; різниця ігрових динамік команд; різниця класу команд; фактор свого поля; результати особистих зустрічей команд. Тестування показує, що запропонована модель забезпечує добру збіжність прогнозованих та дійсних результатів футбольних матчів, що дозволяє рекомендувати машину опорних векторів як перспективний підхід для прогнозування результатів різних спортивних чемпіонатів.

1. Вступ

Прогнозування результатів спортивних ігор являє собою цікавий різновид сучасних реальних задач оцінки та прийняття рішень. Крім того ці задачі можуть розглядатися як добрий полігон для тестування різних технологій екстраполяції та прогнозування в складних умовах обмеженого статистичного матеріалу та визначеності впливаючих факторів.

Метою цієї статті є демонстрація застосування специфічної новітньої інтелектуальної технології SVM (Support Vector Machines – машини опорних векторів) на прикладі прогнозування переможця футбольного матчу. Під термінами [1] “інтелектуальні технології”, “софт-комп'ютеринг” та “автоматичне навчання” в статті розуміється широка гамма новітніх потужних методів інтелектуального аналізу даних, що забезпечують прийнятний шлях врахування складності, невизначеності та нечіткості реальних задач. Для синтезу та тестування моделі прогнозування результатів футбольного матчу використовується інформація про результати ігор 1–10 чемпіонатів України з футболу.

2. Постановка задачі

Задача синтезу моделі прогнозування результатів футбольного матчу може бути поставлена як пошук функціонального відображення виду [2]:

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \rightarrow y \in \{d_1, d_2, d_3\}, \quad (1)$$

де \mathbf{x} – вектор факторів впливу, таких як: рівень команди, погодні умови, місце проведення матчу, результати минулих поєдинків тощо; y – результат футбольного матчу, який оцінюється такими термами: d_1 – «перемога команди-господаря поля», d_2 – «нічия» й d_3 – «перемога гостьової команди».

Відповідно до потреб методу SVM постановка задачі (1) може бути модифікована як:

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \rightarrow y \in \{-1, 1\}, \quad (2)$$

де \mathbf{x} – як і раніше, вектор факторів впливу; y – результат футбольного матчу, що оцінюється як: -1 – “команда-господарь поля не переможе” та 1 – “гостьова команда не переможе”.

3. Вибір факторів впливу

При будь-якому прогнозуванні різні особи, що приймають рішення, можуть обрати як впливаючі різні фактори. Множина факторів впливу обирається особою, що приймає рішення, як правило, суб'єктивно, хоча і існують деякі загальні правила вибору впливаючих факторів. Згідно з [2] в роботі будуть враховуватися такі фактори впливу:

- x_1 – різниця втрат провідних гравців, що розраховується як кількість травмованих або дискваліфікованих гравців команди-господаря поля мінус кількість таких гравців гостьової команди;

- x_2 – різниця ігрових динамік команд, що розраховується як кількість здобутих очок командою-господарем поля за останні п'ять ігор мінус кількість здобутих очок гостьовою командою за останні п'ять ігор;

- x_3 – різниця в класі, що розраховується як різниця місць в турнірній таблиці, які посідають команда-господар поля та гостьова команда;

- x_4 – фактор свого поля, що розраховується як $HP/IG - GP/GG$, де HP – кількість очок, що здобула команда-господар поля вдома в даному чемпіонаті; IG – кількість зіграних домашніх ігор командою-господарем поля в даному чемпіонаті; GP – кількість очок, що здобула гостьова команда в даному чемпіонаті на виїзді; GG – кількість зіграних ігор на виїзді гостьовою командою в даному чемпіонаті;

- x_5 – показник персональних зустрічей команд, що розраховується як різниця голів, забитих командою-господарем поля та гостьовою командою в персональних зустрічах за останні 10 чемпіонатів України.

Зауважимо, що наведені фактори не містять конфіденційної інформації, і для особи, що приймає рішення, не є важкою справою визначити значення цих факторів перед початком будь-якої зустрічі.

4. Машина опорних векторів

Машина опорних векторів (SVM) [3] є відносно новою інтелектуальною технологією, що ґрунтується на концепції автоматичного навчання. SVMs використовуються для моделювання як об'єктів з дискретним виходом (задачі розпізнавання образів), так і для об'єктів з неперервним виходом (задачі регресивного аналізу). SVMs мають цікаві особливості, що принципово відрізняють їх від інших інтелектуальних технологій, наприклад, від нейронних мереж. SVMs завжди знаходять глобальний розв'язок задачі навчання, причому цей розв'язок має просту геометричну інтерпретацію. SVMs також забезпечують швидкість навчання на великих вибірках багатовимірних даних на рівні зі швидкістю настройки нейронних мереж.

При виборі класифікаційної функції виду

$$f: \mathbf{x} \rightarrow \{\pm 1\}, \quad (3)$$

найбільш важливим показником є так звана *ємність (capacity)* машини, що навчається. Мала *ємність* може не забезпечити високий рівень апроксимації складної функції, тоді як велика *ємність* може призвести до втрати властивостей узагальнення, тобто призвести до ефекту перенавчання. На відміну від нейронних мереж, які для уникнення ефекту перенавчання використовують метод "ранньої зупинки алгоритму навчання", в SVMs перенавчання є обмеженням відповідно до статистичної теорії навчання на малих вибірках даних [3].

Найпростішою вирішувальною функцією є лінійна. Застосування лінійної вирішувальної функції в SVM зводить процес навчання до пошуку гіперплощини з найширшою межею, що поділяє елементи вибірки на два класи. Ширина цієї межі визначається як мінімальна відстань від елементів навчальної вибірки до гіперплощини розтину. Процедура знаходження гіперплощини розтину з найширшою межею ("нейтральною смугою") є типовою задачею опуклого квадратичного програмування [4]. У випадку, коли дані неможливо коректно розділити гіперплощиною, за кожну помилкову класифікацію накладається деякий штраф. Якщо поверхня розтину буде нелінійною, то використання відображення

$$\phi: \mathbf{x} \rightarrow F \quad (4)$$

дозволяє трансформувати початкові дані в деякий новий факторний простір F такий, що лінійна машина, яка навчається, перетворюється у нелінійну. Таке нелінійне відображення виконується в SVMs неявно – необхідно лише забезпечити виконання операції множення пари точок з навчальної вибірки $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \in F$ в новому факторному просторі. Для розрахунку такого множення необхідно знати так звану зерняткову функцію (kernel function), що визначає новий факторний простір через

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \quad (5)$$

Вибір відповідної даним зерняткової функції може значно покращити якість процесу навчання. Але, ще раз підкреслимо, немає потреби знати ϕ явно, тому що відображення здійснюється опосередковано.

5. SVM-алгоритм

Як передбачалося в розділі 3, наявними початковими даними є навчальна вибірка $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, в якій кожна точка $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ належить до \mathbf{R}^n , а $y_i \in \{-1, 1\}$ вказує на клас, якому відповідає точка \mathbf{x}_i . Метою є визначення функції

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b, \quad (6)$$

де $\mathbf{w} = (w_1, w_2, \dots, w_n)$, b – параметри розділяючої гіперплощини;

$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_m(\mathbf{x}))$ – відображення з \mathbf{R}^n в новий факторний простір \mathbf{R}^m . Це типове відображення, що застосовується в машинах, які навчаються, з використанням зерняткові функції [5].

Відповідно до статистичної теорії навчання [3], управління генералізуючими властивостями вирішувальної функції необхідно здійснювати через розмірність Вапніка-Червопекіса шляхом мінімізації структурного ризику. Практичне застосування цієї ідеї в SVMs призводить до наступної постановки задачі квадратичного програмування [5].

Задача P1:

$$\text{Мінімізувати } \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \cdot \sum_{i=1}^N \xi_i,$$

$$\text{при обмеженнях: } y_i (\mathbf{w} \cdot \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, N,$$

де C – деякий додатний штраф за помилкову класифікацію.

Розв'язком \mathbf{w}^* цієї задачі є:

$$\mathbf{w}^* = \sum_{i=1}^N a_i^* y_i \phi(\mathbf{x}_i), \quad (7)$$

де $\mathbf{a}^* = (a_1^*, a_2^*, \dots, a_N^*)$ є розв'язком такої *дуальної задачі*:

Задача P2:

$$\text{Мінімізувати } -\frac{1}{2} \mathbf{a}^T \mathbf{D} \mathbf{a} + \sum_{i=1}^N a_i,$$

$$\text{при обмеженнях: } \sum_{i=1}^N y_i a_i = 0; \quad 0 \leq a_i \leq C, \quad i = 1, 2, \dots, N,$$

де \mathbf{D} – матриця розміром $N \times N$ така, що

$$D_{ij} = y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j). \quad (8)$$

Використовуючи рівняння (6) та (7), розв'язок задачі **P1** запишемо так:

$$\sum_{i=1}^N y_i a_i^* \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b^*. \quad (9)$$

Точки, для яких $a_i^* > 0$, називаються опорними векторами (SVs – Support Vectors). Це точки, що неправильно класифіковані вирішувальною функцією, або точки з “нейтральної смуги”, відстань від яких до лінії розтину менше за мінімальну [5]. В багатьох практичних задачах такими точками є лише невелика підмножина навчальної вибірки.

Відомо [3], що складність задачі **P2** пропорційна N^2 та не залежить від розмірності факторного простору. Цей факт дозволяє узагальнити задачу навчання на факторний простір нескінченної розмірності. Але через скінченний об'єм оперативної пам'яті та швидкодію комп'ютера обсяг навчальної вибірки для задачі **P2**, яку можна розв'язати на практиці, є обмеженим [6].

6. Результати

На сьогодні відсутні вагомі наукові результати, які дозволяють побудувати одноітераційну SVM-модель для прийняття рішень у випадку багатьох класів [7]. В зв'язку з цим початкова задача прогнозування переможців футбольного матчу з трьома можливими класами рішень (виграш команди-господаря поля, нічия та виграш гостьової команди) була модифікована в типову регресійну задачу, для якої SVM-алгоритм було використано для мінімізації середньої квадратичної помилки. Перехід від неперервного значення виходу моделі до дискретного здійснюється за такими правилами:

- якщо значення прогнозованого показника не є від'ємним, тоді результат матчу буде "гостьова команда не переможе";
- якщо значення прогнозованого показника є від'ємним, тоді результат матчу буде "команда-господарь поля не переможе".

Тобто результат футбольного матчу прогнозується за знаком вихідного показника.

Для навчання та тестування була використана вибірка даних за матчами, що були зіграні в перших десяти чемпіонатах України з футболу [2]. Для навчання використовувалась вибірка з 105 матчів, а для тестування моделі – з 70. Навчання моделі передувала нормалізація даних на діапазон $[-1, 1]$. Комп'ютерні експерименти проведені за допомогою програмного забезпечення *svmSVM* [8].

Як зрештову було обрано функцію множення, що обумовлено відсутністю у авторів будь-яких підстав для вибору інших, більш складних функцій. Навчання моделі було проведено при $C = 1000$. Інший важливий параметер навчання – коефіцієнт нечутливості (*epsilon*), який визначає коридор виходу за межі розділяючої гіперплощини без накладання штрафу. При навчанні значення коефіцієнта нечутливості було встановлено однаковим для обох класів: $epsilon+ = epsilon- = 0.01$.

Параметри SVM-моделі після навчання наведені в таблиці 1. В таблиці використовуються такі позначення:

- *опорні вектори* – кількість опорних векторів, знайдених в результаті навчання моделі;
- *граничні SVs* – кількість опорних векторів на верхній межі;
- *min SV (max SV)* – мінімальне (максимальне) значення розв'язку дуальної задачі;
- $|w|$ – абсолютне значення вектора гіперплощини;
- *Vcdim* – оцінка розмірності Ванніка-Червоненкіса;
- (w_1, \dots, w_5) і *b* – параметри розділяючої гіперплощини.

Таблиця 1

Параметри SVM-моделі

Параметри	Значення
<i>Опорні вектори</i>	97
<i>Граничні SVs</i>	90
<i>min SV</i>	-9.7087379
<i>max SV</i>	9.7087379
$ w $	0.12128035
<i>Vcdim</i>	≤ 1.3774434
w_1	0.2527201
w_2	-0.010411425
w_3	0.28175218
w_4	0.18387293
w_5	0.099184523
<i>b</i>	0.06384628

Після 1377 ітерації SVM-алгоритму були отримані такі результати:

- середня квадратична помилка на навчальній вибірці – 0.052297589;
- середня квадратична помилка на тестувальній вибірці – 0.053676842.

За правилами класифікації, які наведені на початку розділу, точність прогнозування на тестувальній вибірці складає 61,4 %, тобто з 70 футбольних матчів правильно спрогнозованими були результати 43 поєдинків. В таблиці 2 наведено порівняння результатів прогнозування, які отримані за допомогою запропонованої SVM-моделі та інших моделей, побудованих на базі різних інтелектуальних технологій [2]. Зауважимо, що в попередній роботі [2] прогнозування здійснювалось для випадку трьох можливих класів – виграш, нічия, програш. Крім того нечітка модель та нейронна мережа тестувалися на усіх 175 матчах. Але наведені в таблиці 2 дані дозволяють зробити попередні висновки щодо ефективності різних інтелектуальних технологій для прогнозування результатів футбольних матчів чемпіонату України.

Таблиця 2

Порівняння SVM-моделі з іншими підходами

Модель прогнозування	Рівень безпомилковості
Модель на основі нечіткого логічного висновку	64 % (навчальна та тестувальна вибірки)
Штучні нейронні мережі	64 % (навчальна та тестувальна вибірки)
Модель на основі генетичного програмування	64.28 % (тестувальна вибірка)
Машини опорних векторів	61.4% (тестувальна вибірка)

7. Висновки та майбутні дослідження

В статті продемонстровано застосування нового статистичного підходу до інтелектуального аналізу даних – машини опорних векторів для прогнозування результатів футбольного матчу. Відповідно до загальних принципів SVM-теорії результати прогнозування представлялися в бінарній формі: +1 – гостьова команда не виграє та -1 – команда-господар поля не виграє. Проведене тестування запропонованої SVM-моделі показує конкурентоспроможність машини опорних векторів для прогнозування результатів футбольних ігор в порівнянні з іншими інтелектуальними технологіями. На нашу думку, перспективним шляхом в області прогнозування результатів спортивних змагань є застосування гібридного підходу, в якому поєднуються позитивні сторони різних інтелектуальних методів аналізу даних. Застосування гібридного підходу для багатьох задач, пов'язаних з статистичними або хаотичними закономірностями, дозволяє отримати моделі з високими рівнями адекватності [9].

ЛІТЕРАТУРА:

1. Zadeh L. Applied Soft Computing –Foreword // Applied Soft Computing. – 2000. – Vol. 1. – P. 1–2.
2. Tsakonas A., Dounias G., Shtovba S., Viodyuk V. Soft Computing-Based Result Prediction of Football Games // Proceedings of the First International Conference on Inductive Modeling, Lviv. – 2002. – Vol. 3. – P. 15–21.
3. Vapnik V.N. Statistical learning theory. – Wiley-Interscience, 1998. – 736 p.
4. Vapnik V.N. The nature of statistical learning theory. – Springer-Verlag, 1999. – 304 p.
5. Cortes C., Vapnik V. Support Vector Networks // Machine Learning. – 1995. – Vol. 20. – P. 1–25.
6. Evgeniou T., Pontil M. Support Vector Machines with Clustering for Training with Very Large Datasets. – Springer: Lecture Notes in Computer Science. – Vol. 2308. – 2002. – P. 346–354.
7. Boser B., Guyon I., Vapnik V.P. A training algorithm for optimal margin classifiers // Computational Learning Theory. – 1992. – Vol. 5. – P. 144–52.
8. Riping S. mySVM-Manual. Technical Report.– University of Dortmund, Computer Science Department, 2000.
9. Tsakonas A., Dounias G. Hybrid Computational Intelligence Schemes in Complex Domains: An Extended Review. – Springer: Lecture Notes in Computer Science. – Vol. 2308. – 2002. – P. 494–512.

ЦАКОНОС Ацанасіос – магістр, аспірант Егейського університету, Хіос, Греція.

Наукові інтереси:

- штучний інтелект;
- прийняття рішень;
- вейвлет-аналіз;
- теорія хаосу і катастроф.

Тел.: (30937) 89-13-99.

E-mail: tsakonas@stt.aegean.gr

ДУНІАС Гіоргос – доктор філософії, лектор Егейського університету, Хіос, Греція.

Наукові інтереси:

- софт-комп'ютинг;
- прийняття рішень;
- вейвлет-аналіз;
- застосування штучного інтелекту в медицині.

Тел.: (30271) 9-44-08.

E-mail: g.dounias@aegen.gr

ШТОВБА Сергій – кандидат технічних наук, доцент Вінницького державного технічного університету.

Наукові інтереси:

- печітка логіка;
- генетичні алгоритми;
- прийняття рішень, надійність, управління якістю.

Тел.: (0432) 44-04-30.

E-mail: serg@faksu.vstu.vinnica.ua

Подано 10.01.2003