УДК 004.8

Каліберда С.С., магістрант, гр. ПІ-50м,
Морозов А.В., канд. техн. наук, доц., доцент кафедри,
Марчук Г.В., старш. викладач
Державний університет «Житомирська політехніка»

ПРОГНОЗУВАННЯ ХРОНІЧНИХ ЗАХВОРЮВАНЬ

Станом на сьогодні дуже велика увага приділяється рекомендаційним системам, які допомагають приймати рішення в умовах швидко змінюваних умов. Сучасні дослідження по машинному навчанню показують, що точність прогнозів може бути великою.

Одним з найпопулярніших алгоритмів машинного навчання в медичній сфері є метод опорних векторів (support vector machine). Цей метод в основному використовується для класифікації. Основна ідея полягає в тому, що між класами будується гіперплощина, яка розділяє об'єкти двох класів (рис.1). Положення гіперплощини обирається таким чином, щоб вона була розташована якомога далі від векторів кожного класу. Але не завжди вибірку можливо розділити лінійно, тому допускають деякий відсоток помилок класифікації. Опорними векторами в даному випадку є вектори, які розташовані біля поділяючої гіперплощини.

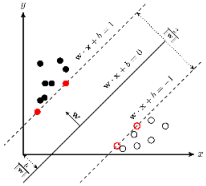

Рис. 1. Метод опорних векторів

Метод опорних векторів також можна використовувати для регресійного аналізу. Так само, як і з класифікаційним підходом, існує мотивація до пошуку та оптимізації меж узагальнення. Робота алгоритму спирається на функцію визначення втрат, яка ігнорує помилки, які знаходяться на певній відстані від істинного значення.

Слід зауважити що метод опорних векторів може вимагати велику кількість обчислень на великих вибірках, тому що складність алгоритму складає $O(n^2)$, таким чином для розвідувального аналізу слід використовувати більш прості методи регресійного аналізу, такі як лінійна або поліноміальна регресія. Але при цьому стійкість метода опорних векторів до викидів надає можливість виконувати більш якісне навчання моделі на вибірках невеликого розміру.

Для проведення експериментів було обрано дані по хронічним хворобам 500 міст США американської організації Centers for Disease

Control and Prevention[1]. Ці дані самі по собі унікальні, тому що вони охоплюють 103 млн. осіб в віці від 18 років, мають в своєму складі 27211 тисяч записів по різним територіям статистичної звітності, населення котрих складає від 50 чоловік до 26980 чоловік. Усі показники представлені у співвідношенні відсотку населення та діапазону похибки. Дані представлені у форматі csv файлу. Таким чином можна зробити висновки, що обрані статистичні дані підходять для проведення різних наукових досліджень. В результаті використання методів інтелектуального аналізу було визначено що більшість відносин були логічними при використані лінійної регресії, але і на її основі можна побачити наступні цікаві залежності між показниками (Таблиця 1).

Таблиця 1. - Результати точності прогнозування деяких показників

Прогноз на основі показника	Прогнозований показник	LinearRegression, %	Polynomial Regression, %	SVM, %
споживання алкоголю	артрит	74,90	75,64	75,12
споживання алкоголю	цукровий діабет	80,04	80,23	78,87
споживання алкоголю	інсульт	82,83	84,93	84,13
ішемічна хвороба	хвороба нирок	81,00	81,13	80,65
цукровий діабет	інсульт	88,84	90,04	89,16
хвороба нирок	інсульт	93,23	94,03	93,23
куріння	поширеність регулярних візитів до стоматологів	64.82	64.19	64.34

Стосовно використаних алгоритмів можна зробити наступні висновки: для більшості випадків прогнозування зв'язку між величинами достатньо більш простих моделей, наприклад лінійної регресії, яка дозволила швидко провести розвідувальний аналіз. Серед використаних алгоритмів, найбільш ефективним виявився метод опорних векторів, тому що від може прогнозувати нелінійні залежності, але цей метод також вимагає витрат часу на проведення аналізу. Альтернативним методом виявився метод поліноміальної регресії.

Посилання:

1. 500 Cities: Local Data for Better Health [Електронний ресурс] // Centers for disease and control prevention. – 2018. – Режим доступу до ресурсу: https://www.cdc.gov/500cities/about.htm