УДК 004.8

Левківський В. Л., старший викладач
Державний університет «Житомирська політехніка»

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ СТАТИСТИЧНИХ ДАНИХ МЕДИЧНОГО СПРЯМУВАННЯ

Інтелектуальний аналіз даних це процес визначення нових, коректних і потенційно корисних знань на основі даних, які представлені великими об'ємами. Крім того, аналіз включає в себе безліч різних підходів і методів для дослідження і перетворення даних.

Основна мета інтелектуального аналізу даних полягає в тому, щоб створити модель, що дозволяє ефективно інтерпретувати і використовувати ті дані, якими володіє дослідник на даний час, і ті дані, які отримає в майбутньому. Оскільки аналіз даних включає в себе безліч методів, то основний етап створення моделі даних - це вибір методу аналізу, що буде використаний в цій моделі. Для правильного вибору методу потрібен практичний досвід. Далі модель потрібно доопрацювати, щоб зробити її більш ефективною.

Існує безліч алгоритмів для інтелектуального аналізу даних. Ціль яких знаходити закономірності в даних. Знання, що видобуваються методами Data mining, прийнято представляти у вигляді закономірностей. В даній роботі буде розглянуто алгоритми класифікації та кореляційно-регресійні алгоритми, такі як:

- метод лінійної регресії;
- метод поліноміальної регресії;
- метод опорних векторів.

Метою дослідження є застосування алгоритмів інтелектуального аналізу для опрацювання статистичних даних медичного спрямування.

Нажаль поки що медична статистика в Україні не може надати необхідного об'єму статистичної інформації. Для проведення наукових досліджень і подальшого впровадження результатів в Українську медичну сферу будемо використовувати статистичні дані по хронічним хворобам 500 міст США. Одним з джерел цієї інформації є американсь-ка організація Centers for Disease Control and Prevention (скорочено CDC) - Центри з контролю та профілактики захворювань.

Деякі варіанти результатів прогнозування різних показників наведені у таблиці 1.

На основі тестових даних можна зробити висновки, що найбільш точним методом прогнозування є метод опорних векторів. Але слід зауважити що складність методу опорних векторів складає O(n2), де n – кількість записів в вибірці, що може бути недоліком у випадку наявнос-ті великої кількості навчальної вибірки.

Таблиця 1.

Результати прогнозування різних показників

PredictBy	PredictTo	LinearRegression, %	Polynomial Regression, %	SVM, %
артрит	ішемічна хвороба	80,92	81,37	80,72
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	хвороба нирок	91,37	91,37	90,92
цукровий діабет	інсульт	88,84	90,04	89,16

Менш якісні результати демонструє метод поліноміальної регресії, але цей вид регресії здатен знаходити залежності в даних, які не опи-суються лінійним рівнянням. При цьому використання цього методу вимагає попередньої обробки даних, на основі яких буде відбуватися прогнозування.

Найменш точні результати демонструє метод лінійної регресії, але цей метод є найбільш простим та швидким, тому використання лінійної регресії для розвідувального аналізу можна вважати обґрунтованим.

Стосовно використаних алгоритмів можна зробити наступні вис-новки: для більшості випадків прогнозування зв'язків між величинами достатньо більш простих моделей, наприклад лінійної регресії, яка дозволила швидко провести розвідувальний аналіз. Серед використаних алгоритмів, найбільш ефективним виявився метод опорних векторів, але цей метод також вимагає витрат часу на проведення аналізу. Альтернативним методом виявився метод поліноміальної регресії.

В результаті проведених досліджень зазначених алгоритмів інте-лектуального аналізу даних було розроблено моделі та методи для встановлення впливу одних хронічних захворювань на інші. Для підвищення достовірності результатів із використанням запропонованих моделей та методів, необхідно збільшувати обсяги даних.