УДК 621.391.82 **В.С. Наконечний**

АНАЛІЗ МЕТОДІВ ВІДБОРУ ІНФОРМАТИВНИХ ОЗНАК ОБ'ЄКТІВ СУПРОВОДЖЕННЯ В РАДІОЛОКАЦІЙНИХ СИСТЕМАХ РОЗПІЗНАВАННЯ

Природне прагнення обліку якомога більшої кількості потенційно корисної інформації для систем розпізнавання призводить до появи надлишкових (шумових) ознак, включення яких в модель розпізнавання може тільки погіршити її якість. Відбір ознак розпізнавання з незначною інформативністю скоротить вартість їх збору та підвищить ефективність роботи системи розпізнавання в цілому. Тому в статті проведено порівняльний аналіз методів відбору і виключення менш інформативних ознак для розпізнавання класифікаційних систем. Вперше, для пошуку максимально інформативних ознак розпізнавання об'єктів радіолокаційного моніторингу запропонована ідея застосування популяційних алгоритмів, а саме методу бджолиної колонії.

Ключові слова: інформативність, ознака розпізнавання, метод бджолиної колонії.

Вступ. У практичних задачах розпізнавання витрати на вимірювання або обчислення окремих ознак можуть бути співставленніз вартістю втрат від помилкових прогнозів. Відбір менш інформативних ознак скорочує вартість збору інформації та підвищує ефективність роботи системи розпізнавання в цілому. Проблема відбору ознак часто виникає із-за того, що на етапах постановки задачі формування даних ще не ясно, які ознаки менш ефективні або дублюють одна одну. Природне прагнення врахувати якомога більше потенційно корисної інформації призводить до появи надлишкових (шумових) ознак і їх включення в модель розпізнавання може тільки погіршити її якість. Методи навчання мають відрізняти шумові ознаки і відкидати їх. Крім того, у міру збільшення числа використовуваних ознак середня помилка на навчальній вибірці, як правило, монотонно убуває. При цьому середня помилка на незалежних контрольних даних спочатку зменшується, потім проходить через точку мінімуму і далі зростає. Це явище носить назву перенавчання, воно настає тоді, коли метод навчання уточнює вирішальні правила, зважаючи на все більш дрібні особливості розподілу класів, аж до окремих об'єктів («викидів») з різко виділеними характеристиками. Аналітичні методи цензурування таких об'єктів, що використовують знання про закони розподілу навчальної вибірки, у більшості сучасних задач розпізнавання не застосовують, так як ці закони, як правило, невідомі. Кількість ознак N часто на порядки перевищує кількість об'єктів M.

Для вирішення даної проблеми необхідний пошук і порівняльна оцінка методів, що дозволяють проводити відбір ознак за ступенем їх інформативності.

Метою даної роботи є порівняльна оцінка різних методів відбору ознак, врахування їх переваг і недоліків з метою визначення найбільш ефективного алгоритму, здатного функціонувати в умовах дефіциту часу.

На сьогоднішній день проблема аналізу та відбору найбільш інформативних ознак є досить актуальною. Аналіз джерел [1-10] показує, що існує велика кількість різноманітних методів, які у тій чи іншій мірі намагаються вирішувати поставлене перед ними завдання, проте єдиного алгоритму її рішення поки ще не знайдено.

Розвиток методів відбору ознак має багату історію. У шістдесяті роки минулого століття почали активно розвиватися крокові методи (Stepwise Regression) [1]. Головна ідея таких методів полягає у відборі ознак, що вносять найбільший вклад у залежну змінну. Вводиться критерій, на підставі якого алгоритм додає або видаляє ознаки. Широке застосування отримали методи крокової регресії – алгоритми LARS (Least Angle Regression) [2] і LASSO(Least Absolute Shrinkage and Selection Operator) [3]. Робота алгоритму LARS полягає в послідовному додаванні ознак. На кожному кроці відповідні ваги ознак змінюються таким чином, щоб принести найбільшу кореляцію з вектором регресійних залишків. Алгоритм дозволяє скоротити кількість вільних змінних і уникнути проблеми нестійкої оцінки ваг. Метод LASSO вводить обмеження на норму вектора коефіцієнтів моделі, що призводить до обігу в нуль деяких коефіцієнтів моделі і до підвищення її стійкості, дозволяє відбирати

ознаки, що мають найбільший вплив на вектор відповідей.Однією з причин виникнення завдання відбору ознак є їх мультиколінеарність, тобто існування міцної лінійної залежності або сильної кореляції між двома чи більше незалежними змінними.

У 1963р. над цією проблемою почав працювати А. В. Тихонов, який ввів поняття регуляризації – додаткового обмеження на завдання. В роботі [4] введено поняття регуляризації і описаний загальний метод розв'язання задач.

У 1970р. Hoerl і Kennard запропонували метод гребеневої регресії [5]. До мінімізованої функції вводився додатковий доданок, що підвищувало стійкість рішення, однак не дозволяло проводити відбір ознак. Пізніше стали з'являтися методи, що використовували якісно інший підхід для вирішення проблеми мультиколінеарності. Наприклад, Belsley запропонував метод для видалення ознак [6], що використовує сингулярне розкладання матриці плану. Алгоритм знаходить коефіцієнт, що характеризує ступінь залежності ознак один від одного. Пізніше з'явився метод фактора інфляції дисперсії (Variance Inflation Factor) [7], який оцінює збільшення дисперсії заданого коефіцієнта регресії, що свідчить про високу кореляцію даних.

Проведений теоретичний аналіз методів відбору ознак показав, що універсальний спосіб вирішення цієї задачі це повний перебір варіантів, оскільки він найбільш простий для реалізації і гарантує, що буде знайдений найкращий набір. Однак його практичне застосування обмежено задачами з числом ознак не більше 20-25 і проблема «комбінаторного вибуху» в принципі не вирішується простим нарощуванням обчислювальних потужностей.

Встановлено, що в тих випадках, коли повний перебір неможливий, вдаються до евристичних процедур, які істотно скорочують обсяг обчислень. Такі методи послідовного перебору варіантів, засновані на додавання (Add) та/або вилучення (Del) ознак. До переваг методу додавання та вилучення (Add-Del) слід віднести те, що на практиці йому набагато частіше вдається знайти краще рішення, ніж методам Add або Del, однак Add-Del працює довше, ніж Add і Del окремо, і також не гарантує оптимальність. До недоліків розглянутих методів слід віднести певну складність реалізації.

Аналіз методів випадкового пошуку і випадкового пошуку з адаптацією (ВПА) показав, що їх робота заснована на генерації випадкових наборів ознак. Кожен набір оцінюється за зовнішнім критерієм, і якщо він виявляється досить вагомим, то всі вхідні в нього ознаки «заохочуються» збільшенням імовірності їх появи в таких наборах. Відповідно, ознаки, що утворюють гірші набори, «наказуються» зменшенням ймовірності. Перевагою методу є простота реалізації і порівняно невелике число параметрів. Для вибору параметрів методу ВПА є досить чіткі рекомендації. На одних і тих самих прикладах алгоритм ВПА показує кращі результати, ніж описані алгоритми Del та Add. Проте недоліком методу ВПА є його повільна збіжність.

Аналіз метода спрямованого таксономічного пошуку ознак показав, що якщо в один таксон ознаки групуються за принципом максимальної подібності, то між таксонами забезпечується максимальна відмінність. Така підсистема найкращим чином відповідає вирішальному правилу, що орієнтується на використання незалежних ознак. Недоліком даного методу є значний час, що витрачається на перебір комбінацій ознак $C_\ell^{\ell'}$.

Визначено, що методи кластеризації дозволяють розбити вибірку об'єктів на кластери, які складаються із схожих об'єктів, і виділити в кожній групі по одному, найбільш типовому представнику. Те ж саме можна виконати і з ознаками, якщо визначити функцію відстані між ними, наприклад, через коефіцієнт кореляції.

До основних недоліків методу кластеризації відносять: По-перше, можуть існувати кластери, які цілком складаються із неінформативних ознак. Взявши по одному типовому представнику від кожного кластера, все одно доведеться вирішувати задачу відбору ознак, хоча обсяг перебору при цьому сильно скоротиться. По-друге, інформації про попарні подібності між ознаками в загальному випадку не достатньо для виділення оптимального

набору ознак. Зокрема, кластеризація не вирішує проблему мультиколінеарності, оскільки набір попарно некорельованих ознак цілком може виявитися лінійно залежним.

Попередню кластеризацію ознак має сенс застосовувати для скорочення перебору в інших методах відбору ознак. Наприклад, щоб заборонити надто схожим ознакам входити в один і той же набір.

Аналіз відбору ознак методами математичного програмування виявив, що вони відрізняються від попередніх методів тим, що в них немає явного перебору ознак, однак при більш глибокому розгляді відміна виявляється поверхневою – перебір здійснюється всередині стандартних процедур математичного програмування при пошуку активних обмежень, що задовольняють умовам Куна-Таккера. Застосування стандартних методів квадратичного програмування для рішення даної задачі автоматично призводить до обнуління деяких коефіцієнтів, а отже, і до відбору ознак.

На ряду з відбором часто застосовують синтез. Встановлено, що синтез ознак, вирішує завдання скорочення розмірності простору. При відборі частина ознак повністю ігнорується. Синтез має дві переваги перед відбором. По-перше, відсутній комбінаторний перебір варіантів. По-друге, вся вихідна інформація враховується в повному обсязі. Втім, 88астосування з стає недоліком в тих задачах, де присутні завідомо неінформативні шумові ознаки. Недоліком синтезу є те, що нові ознаки можуть виявитися не інтерпретованими, і отже, використовувати відбір або синтез – визначається особливостями завдання. У деяких випадках доводиться користуватися і тим, і іншим.

Аналіз існуючих підходів до вирішення комбінаторних завдань показав, що вдалими є підходи, засновані на методах еволюційного моделювання і штучного інтелекту. Відомими представниками такого підходу є генетичні алгоритми (ГА). При роботі ГА оперують з популяцією рішень. З одного боку це дозволяє швидше знаходити покращене рішення, але з іншого боку потрібний великий об’єм пам’яті для зберігання інформації про популяції рішень. Тим не менш, останні дослідження, пов’язані з використанням генетичних методів оптимізації в різних областях показали їх високу ефективність.

Перевага генетичного алгоритму – в його очевидності і багатих можливостях для введення різних евристик. Недоліком є відносно повільна збіжність. Незважаючи на безліч успішних практичних застосувань, збіжність генетичного алгоритму досі залишається відкритою теоретичною проблемою. Крім того, хороший генетичний алгоритм поряд з параметрами розміру популяціїB, максимального числа поколінь T і ймовірності мутації P_m має ще з десяток-інший параметрів, підбір яких є мистецтвом і залежить від особливостей задачі.

Загальний аналіз наведених вище методів показав, що жоден з них практично не задовольняє основному критерію –швидкому відбору ознак за ступенем їх важливості. Однак, для систем радіолокаційного розпізнавання, що функціонують в масштабі реального часу, конче потрібен метод, який би давав високі результати за мінімально короткий проміжок часу. В силу останньої обставини, для розпізнавання об’єктів супроводження в радіотехнічних системах радіолокаційного моніторингу, вперше запропонована ідея пошуку глобальних екстремумів (максимально-інформативних ознак) на основі застосування методу штучної бджолиної колонії.

Метод штучної бджолиної колонії (Artificial Bee Colony method) є досить молодим алгоритмом для знаходження глобальних екстремумів складних багатовимірних функцій. Ідея парадигми методу бджолиної колонії взята з поведінки бджіл при пошуку місць, де можна роздобути якомога більше нектару [8], і полягає у використанні дворівневої стратегії пошуку на кожній ітерації. На першому рівні за допомогою бджіл-розвідників формується безліч перспективних областей (джерел), на другому рівні за допомогою бджіл-фуражирів здійснюється дослідження околиць даних областей (джерел). Мета бджолиної колонії знайти джерело, що містить максимальну кількість нектару (глобальний екстремум).

Метод бджолиної колонії відноситься до класу популяційних, ітераційних алгоритмів. Суть ітераційної процедури полягає у виконанні повторюваних дій на кожній ітерації і полягає в тому, що рішення на кроці t формується шляхом змін рішень, отриманих на кроці t-1. При цьому зміни, що вносяться, як правило, незначні.

Найважливішим поняттям популяційних алгоритмів є поняття фітнес функції. Часто цю функцію називають функцією придатності, функцією корисності, функцією пристосованості і т. д. Важливість функції зумовлена тією обставиною, що за її допомогою оцінюють «якість» агентів популяції. Стратегічно, в процесі міграції агенти рухаються таким чином, щоб наблизитися до глобального екстремуму фітнес – функції.

Метод бджолиної колонії має наступні особливості:

1. Усі агенти діляться на різні типи у відповідність з діями, які вони виконують в процесі рішення задачі.

1.1 Зайняті фуражири забезпечують використання вже знайдених джерел нектару, тобто трохи змінюють вже знайдені раніше рішення задачі.

1.2 Незайняті фуражири забезпечують продовження пошуку нових джерел нектару, тобто агенти такого типу виконують пошук нових допустимих рішень задачі. Незайняті фуражири у свою чергу бувають двох типів.

1.2.1 Спостерігачі – чекають у вулику інших агентів. Вони не виконують ніяких дій, вони фактично чекають моменту, коли їм треба буде також почати пошук рішень;

1.2.2 Розвідники – забезпечують пошук нових джерел нектару. При цьому пошук здійснюється випадковим чином, тобто вони випадково вибирають в просторі пошуку можливе рішення.

2. Зв’язок між рішеннями агентів здійснюється шляхом моделювання виконання бджолами 89астосуван «танцю», що забезпечує утворення двох типів зворотного зв’язку: позитивного і негативного.Позитивний зворотний зв’язок полягає в тому, що агенти, ґрунтуючись на інформації про рішення інших агентів, можуть почати досліджувати рішення, отримане іншим агентом.Негативний зворотний зв’язок полягає в тому, що агенти, отримавши інформацію про знайдені рішення іншими агентами, можуть прийняти рішення про припинення розгляду свого рішення у зв’язку з гіршими характеристиками в порівнянні з іншими отриманими рішеннями.

3. Процес пошуку рішення забезпечується двома процедурами.

3.1. Пошук нових джерел нектару в усьому просторі пошуку, який досягається за допомогою агентів-розвідників. Таким чином, забезпечується дослідження усього простору пошуку.

3.2. Поглиблене використання областей, в яких знаходяться вже знайдені джерела нектару (досягається за допомогою зайнятих фуражирів). Тобто знаходяться рішення, що знаходяться в просторі пошуку поблизу від даного рішення.

Перше завдання при розробці алгоритму на основі парадигми бджолиної колонії полягає у формуванні простору пошуку. Позиція a_sпростору пошуку представляється у вигляді наборів різного роду параметрів і залежностей між ними.

Ключовою операцією методу бджолиної колонії є дослідження перспективних позицій і їх околиць в просторі пошуку. Сенс, початково вкладений у поняття околиці, полягає в тому, що рішення, які лежать в межах деякої позиції, мають високий ступінь подібності і, як правило, незначно відрізняються один від одного. Основними параметрами методу бджолиної колонії є: кількість агентів n_b, максимальна кількість ітерацій L, початкова кількість агентів-розвідників n, обмеження максимальної кількості агентів-розвідників, граничне значення розміру околиці λіт. Д.

На початку процесу пошуку всі агенти розташовані у вулику, тобто поза простору пошуку.На першій ітерації (l=1) агенти-розвідники у кількості n випадковим чином розміщуються в просторі пошуку. Іншими словами випадковим чином формуються n рішень. З них вибирається n_b кращих рішень, сукупність яких складає безліч базових позицій.

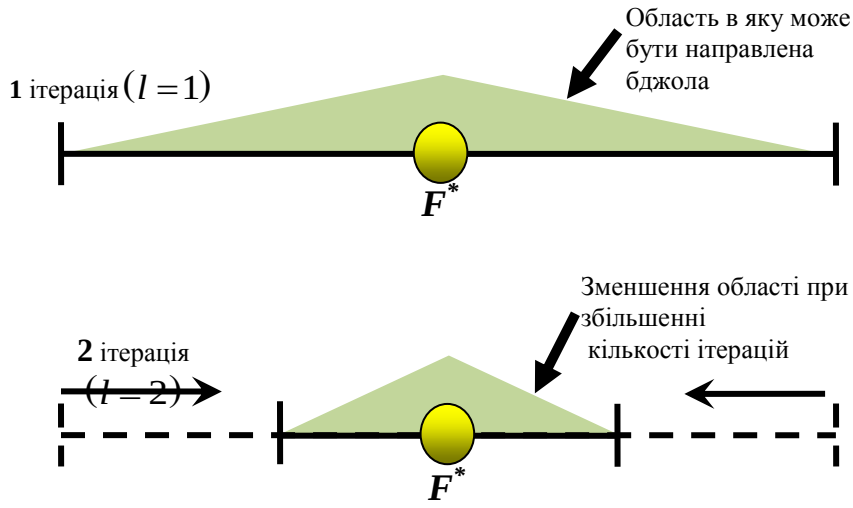

Рис. 1 Принцип одержання кращого рішення в просторі пошуку екстремуму при збільшенні кількості ітерацій

Пропонуються три підходи до визначення числа агентів-фуражирів, які направляються в околиці кожної базової позиції. При першому підході агенти-фуражири розподіляються по базових позиціях рівномірно. При другому підході агенти-фуражири розподіляються по базових позиціях пропорційно значенню цільової функції позиції. При третьому підході реалізується ймовірнісний вибір. Ймовірність вибору агентом-фуражиром базової позиції пропорційна значенню цільової функції в цій позиції.

При першому і другому підходах число рішень в околицях розраховується, при третьому підході – визначається випадково.

Після вибору агентом-фуражиром b_z базової позиції, реалізується імовірнісний вибір позиції a_z, розташованої в околиці базової позиції a_s^b.Позначимо безліч позицій, вибраних агентами-фуражирами в околиці позиції a_s^b як O_s^b. Назвемо безліч позицій $O_s^b \bigcup a_s^b$ областю D_s^b. У кожній області D_s^b вибирається краща позиція a_s^* з кращою оцінкою F_s^*. Приймемо F_s^* оцінкою області D_s^b. Серед F_s^* вибирається краща оцінка F^* і рішення, що відповідає їй, знайдене на цій ітерації спільно роєм розвідників і роєм фуражирів. Краще рішення з оцінкою F^* зберігається, а потім відбувається перехід до наступної ітерації рис. 1.

На рис. 1 показано, як із зростанням числа ітерацій відбувається скорочення площі пошуку екстремуму. Відмітимо, що в даній парадигмі бджолиної колонії не важливе, знати, яким агентом вибрана позиція в просторі пошуку. Важливо знати число агентів-розвідників і число агентів-фуражирів, а також, які саме позиції вибрані агентами-розвідниками і які агентами-фуражирами.

На другій і подальших ітераціях безліч базових позицій формується з двох частин. У першу частину включаються кращі позиції, знайдені агентами в кожній з областей, сформованих на попередній ітерації. Друга частина формується бджолами-розвідниками таким же чином, як і на першій ітерації. Далі виконуються дії, аналогічні діям, розглянутим на першій ітерації.

Порівняльний аналіз ефективності методу бджолиної колонії з детермінованим симплекс-методом, стохастичним методом імітації відпалу, генетичним алгоритмом і методом мурашиної колонії показав [10], що запропонований метод має високу надійність, забезпечуючи 100% успішних результатів розрахунків в усіх випадках. Метод сходиться до максимуму або мінімуму, не «застряючи» в локальному оптимумі із якісними показниками, перевершує методи, з якими був порівняний на предмет швидкості оптимізації і точності отриманих результатів.

Висновки.Для зменшення часу на ідентифікаціюоб'єктів радіолокаційного моніторингу, за рахунок скорочення кількості 91астосування запроп, 91астосування запропо ідея відбору максимально інформативних ознак на основі застосування методу бджолиної колонії. Враховуючи різні області 91астосування запропонованого методу можна виділити наступні його переваги: не схильний до зациклення в локальному оптимумі; пошук кращого рішення ґрунтується на рішеннях агентів усієї колонії бджіл; може використовуватися в динамічних системах, оскільки здатний адаптуватися до змін довкілля; мультиагентність реалізації; може застосовуватися для вирішення як дискретних, так і безперервних завдань оптимізації.

Література

1. Efroymson M.A. Multiple regression analysis. New York: Ralston, Wiley, 1960

2. Efron B.. Hastie T., Johnstone J., Tibshirani R. Least Angle Regression // Annals of Statistics, 2004. vol. 32, no. 3, pp. 407-499

3. Tibshirani R. Regression shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society, 1996, vol. 32. no. 1, pp. 267-288.

4. Тихонов Л. И. Решение некорректно поставленных задач и метод регуляризации. М.: ДАН. 1963, № 151, с. 501-504.

5. Hoerl A.E., Kennard R.W. Ridge regression: Biased estimation for nonortthogonal problems // Technometrics. 1970. vol. 3. no. 12. pp. 55-67.

6. Belsley D.A. Conditioning Diagnostics: Collinearity and Weak Data, in Regression. New York: John Wiley and Sons. 1991.

7. Marquardt D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation // Technometrics, 1996, vol. 12, no. 3, pp. 605-607.

8. Лебедев Б.К., Лебедев В.Б. Размещение на основе метода пчелиной колонии // Известия ЮФУ. – 2010. – № 12. – С. 12-18.

9. Курейчик В.М., Лебедев Б.К., Лебедев О.Б. Поисковая адаптация: Теория и практика. − М.: Физматлит, 2006.

10. Mathur M., Karale S.B., Priye S., Jayaraman V.K. and Kulkarni B.D. Ant Colony Approach to Continuous Function Optimization. Ind. Eng. Chem. Res. 39(10) (2000), pp. 3814 - 3822.

Надійшла 29.01.2015 р. Рецензент: д.т.н., проф. Толюпа С.В.