УДК 519.237.8
DOI https://doi.org/10.24144/2616-7700.2021.38(1).143-148

Н. Е. Кондрук

ДВНЗ «Ужгородський національний університет»,
доцент кафедри кібернетики і прикладної математики,
кандидат технічних наук
natalia.kondruk@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0002-9277-5131

ВИКОРИСТАННЯ МІР ПОДІБНОСТІ В МЕТОДАХ КЛАСИФІКАЦІЇ

Дане дослідження є розвитком напрямку застосування різних видів мір подібності в задачах інтелектуального аналізу даних. Майнінг даних – це процес видобутку неявної інформації з бази даних, яка характеризує приховані зв'язки та структури. Прогнозується, що цей вид аналізу стане надзвичайно затребуваним протягом наступного десятиліття. В роботі наведено огляд сучасних напрямків контрольованої класифікації. Найпопулярнішим прийомом класифікації об'єктів із числовими атрибутами вважається метод K-найближчих сусідів (KNN). Встановлено, що прогнозне значення мітки класу можна покращити, якщо використовувати зважений вплив кожного сусіда на результат. Таким чином, доцільно модифікувати метод KNN. При цьому, запропоновано ввести функцію, що характеризує схожість неміченого об'єкта із його найближчими сусідами у вигляді міри подібності. На її основі введено індикатори зваженого підрахунку голосів «сусідів» за певну мітку класу. Розроблено програмне забезпечення, що реалізує описаний підхід. Проведення практичних експериментів показало його ефективність при розв'язанні певних класів прикладних задач.

Ключові слова: класифікація, алгоритм k-найближчих сусідів, KNN, міра подібності, контрольоване машинне навчання.

1. Вступ. Інтелектуальний аналіз даних – це розділ науки про дані, який використовує обчислювальні прийоми статистики, машинного навчання та розпізнавання шаблонів для аналізу баз даних. Цей тип аналізу має дві мети: прогнозуючу та описову. Прогнозуючі моделі (регресія, класифікація) використовують для передбачення поведінки певних процесів та явищ. Дескриптивні моделі (кластеризація, асоціативні методи, послідовне виявлення шаблонів) визначають закономірності, що описують дані.

Методика класифікації, як метод прогнозування, – це техніка контрольованого машинного навчання, що передбачає існування групи мічених взірців (прототипів, навчаючої вибірки) для кожного класу об'єктів. Класифікація – це завдання присвоєння міток класу об'єктам даних без міток. Процес класифікації згідно [1]:

- вхід: набір атрибутів прототипів, включаючи атрибут мітки класу;

- класифікатор – модель класифікації, що буде використовуватись для прогнозування мітки класу нових об'єктів;

- вихід: класифікований шаблон, що визначив мітку класу об'єкта на основі його атрибутів.

Системний підхід до вивчення класифікаційної моделі з урахуванням навчальної вибірки називають алгоритмом навчання. Процес використання алгоритму навчання для побудови класифікаційної моделі з навчальних даних відомий як індукція. Цей процес також часто описується як "вивчення моделі"або

"побудова моделі". Застосування моделі класифікації на тестових об'єктах для прогнозування їх класу відомий як дедукція. Таким чином, процес класифікації включає два етапи: застосування алгоритму навчання до навчальних даних для вивчення моделі, а потім застосування моделі для визначення міток неміченим об'єктам.

Найпоширенішими є такі методи класифікації: дерева рішень; Байєсівські класифікатори; штучні нейронні мережі; класифікатор k-найближчих сусідів.

2. Дерева рішень (Decision Tree, DT). Структура дерева рішень організована таким чином, що воно містить корінь, гілки, які є внутрішніми вузлами та листи, які далі не класифікуються. Внутрішні вузли представляють атрибути; гілки, що з'єднують вузли, визначають значення атрибутів; листи містять мітки класів. Вони широко використовується в процесі прийняття рішень.

Існують різні алгоритми дерева рішень – ID3, C4.5, C5.0, CART, SPRINT. Вони використовують спеціальні функції, що найкраще розділяють набір даних за певним атрибутом. Серед них – приріст інформації (Information Gain) та індекс Джині (Index Gini). При цьому, будують дерево рішень як схематичну деревоподібну діаграму [2].

В процесі побудови дерева рішень можуть бути виявлені помилки. Їх можна виправити та оцінити за допомогою підходу, який називається усіченням дерева. Використовують два методи – попереднє та «після» усічення. При попередньому усіченні на початковому етапі перевіряється наявність аномалій, які зупиняють побудову дерева. При цьому вузол стає листом. У методі «після» усічення будується все дерево, потім з кореня починається видалення піддерев, якщо в них виявлено будь-яку аномалію.

Використання даного підходу виявилось успішним в багатьох прикладних задачах діагностики захворювань та прогнозування статусу студентів після навчання та передбачення рівня успішності здобувачів освіти [3-4].

3. Байєсівські класифікатори. Баєсовський класифікатор (Naive Bayes, BayesNe, NBC) – це проста, але важлива ймовірнісна модель. В основі даного методу лежить теорема Байєса про умовну ймовірність. При його використанні для баз даних великої потужності, алгоритм демонструє високу точність і швидкість [5]. Апріорі припускається, що атрибути є незалежними. Це припущення називається умовно-класовою незалежністю. Наївні байєсовські моделі прості в реалізації та популярні в додатках машинного навчання, вони забезпечують можливість кожному атрибуту однаково впливати на прийняте рішення незалежно від інших атрибутів. Крім того, NBC дуже швидко будує та оцінює моделі.

4. Штучні нейронні мережі. Штучні нейронні мережі (Single-Layer Perceptron, Multy-Layer Perceptron, SVM) виражаються в термінах біологічної нейронної системи. Вони складаються з кількості окремих одиниць (нейронів), що комунікують між собою, надсилаючи сигнали та організовані у вигляді напрямленого графу, який містить вузли та ребра, що з'єднують кожен вузол. Ребра – це взаємозв'язки між кожним вузлом. Кожне ребро, що з'єднує вузол, містить вагу. Розраховується сума ваг. Порогове значення присвоюється кожному нейрону. Якщо зважена сума перевищує порогове значення, вона видає вихід 1, інакше 0.

Класичним застосуванням нейронних мереж є розпізнавання зображень у

сфері забезпечення якості, медичній діагностиці та засобів аналізу відбитків пальців.

5. Класифікатори найближчих сусідів. Класифікатори найближчих сусідів також відомі як класифікатори на основі відстані (KNN, PEBLS). Вони засновані на вивченні аналогій. Невідомому об'єкту ставиться у відповідність мітка найпоширенішого класу серед його найближчих сусідів. Якщо $k = 1$, то невідомому екземпляру присвоюється мітка класу прототипу, який є найближчим до нього в просторі шаблонів. Вважається [6-8], що це один з найпростіших і найбільш часто використовуваних алгоритмів класифікації.

Недоліком є те, що при голосуванні об'єктів-«сусідів» за певний клас може виникати невизначеність через рівномірний розподіл їх голосів. Тому доцільно використовувати модифікований метод найближчих сусідів, де підрахунок голосів проводиться за зваженою сумою із так званими ваговими коефіцієнтами, що відповідають за подібність сусіда до нового об'єкта. Обґрунтованість підбору вагових коефіцієнтів є відкритою задачею, що потребує додаткових досліджень.

6. Постановка задачі. Нехай задано навчаючу вибірку об'єктів та новий немічений об'єкт O. Визначимо параметр k і за деякою метрикою відстані d (Евклідовою, манхетенською, Хемінга) k найближчих сусідів до O. Ставиться задача реалізувати зважене голосування за прогнозну мітку класу для нового об'єкта.

7. Матеріали і методи. Нехай k найближчих сусідів до O серед об'єктів навчаючої вибірки будуть $\widetilde{O} = \left\{O_i,\ i = \overline{1,k}\right\}$. Кожен O_i характеризується атрибутами $\overline{a_i}\,(a_1^i, a_2^i, \ldots, a_n^i)$ та міткою класу $c^i \in C$. Тоді, поставимо у відповідність об'єкту O та деякій із можливих міток класу c^* бінарний вектор $\overline{y^*} = \begin{cases} 0,\ \textit{якщо } c^i \neq c^*; \\ 1,\ \textit{якщо } c^i = c^*. \end{cases}$

Введемо нечітке бінарне відношення, що визначає схожість об'єктів серед найближчих сусідів із функцією належності виду:

$$\mu^i_{R^{KNN}} = e^{-\frac{d(O_i,O)}{d_{\max}}}. \tag{1}$$

$d_{\max}$ – визначає найбільшу відстань від неміченого об'єкту O до k найближчих точок із $\widetilde{O}$. Таким чином, $\mu_{R^{KNN}} : \widetilde{O}^2 \to \left[\frac{1}{e},\ 1\right]$. Чим більше значення величини $\mu_{R^{KNN}}$ близьке до 1, тим в більшому ступені об'єкти O_i та O будуть подібними за своїми атрибутами. Найменше значення міри подібності (1) не буде досягати нуля, тому при визначенні мітки класу голоси всіх сусідів будуть враховані.

Голоси за клас пропонується розраховувати за наступними індикаторами:

$$votes\,(class = c^*) = \sum_{i=1}^{k} \mu^i_{R^{KNN}} \cdot y_i^* \tag{2}$$

або

$$votes\,(class = c^*) = \frac{\sum\limits_{i=1}^{k} \mu^i_{R^{KNN}} \cdot y_i^*}{\sum\limits_{i=1}^{k} \mu^i_{R^{KNN}}} \cdot 100, \tag{3}$$

де y_i^* – i-ва координата вектора $\overline{y^*}$.

Індикатори (2) та (3) накопичують зважені голоси за клас: чим далі від O знаходиться сусід, тим в меншому ступені враховується його голос. Формула (3) є нормованою та визначає голоси за клас у відсотках. Підраховуємо голоси для всіх можливих класів і новому об'єкту O прогнозуємо ту мітку c^* значення індикатора (2) або (3) для якої є максимальним.

8. Експерименти. Для проведення практичних експериментів була розроблена комп'ютерна програма, що реалізує модифікований метод найближчих сусідів і використовує індикатори (2) та (3) при прогнозуванні мітки класу новому об'єкту.

Вхідною інформацією для проведення класифікації є навчаюча вибірка із міченими об'єктами та кортеж атрибутів нового об'єкту O. Кількість сусідів обчислюється за формулою $k = [\sqrt{m}]$, де m – потужність вибірки. Вихідною інформацією є мітка класу для O.

9. Висновки та перспективи подальших досліджень. Дана робота є розвитком напрямку застосування різних видів мір подібності в задачах інтелектуального аналізу даних [9-12].

Наведено огляд сучасних методів контрольованої класифікації; модифіковано метод KNN використанням зваженого голосування за прогнозовану мітку класу, при цьому описано функцію $\mu_{R^{KNN}}$, що характеризує схожість нового неміченого об'єкту із його найближчими сусідами. Розроблено програмне забезпечення, що реалізує описаний підхід. Проведення практичних експериментів показало його ефективність при розв'язанні певних класів прикладних задач.

Перспективні дослідження полягають у розвитку підходу використання різних видів мір подібності заснованих на нечітких бінарних відношеннях, які б визначали схожість об'єктів за категоріальними атрибутами та їх застосування при розв'язанні різних видів прикладних задач.

Список використаної літератури

1. Oprea C. Performance evaluation of the data mining classification methods. *Information society and sustainable development*. 2014. Vol. 1. P. 249-253. DOI: https://doi.org/10.9790/ 0661-1060106
2. Jain N., Vishal S. Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology*. 2013. Vol. 2, Iss. 11. P. 116-119. DOI: https://doi.org/10.15623/ijret.2013.0211019
3. Kumari M., Godara S. Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. *IJCST*. 2011. Vol. 2. P. 304-308.
4. Jantawan B., Tsai C. The Application of Data Mining to Build Classification Model for Predicting Graduate Employment. *International Journal of Computer Science and Information Security*. 2013. Vol. 11, N 10. P. 1-7.
5. Xhemali D., Hinde C., Stone R. Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *IJSCI: International Journal of Computer Science*. 2009. Vol. 4, N 1. P. 16-23.
6. Zhang C., Liu C., Zhang X., Almpanidis G.. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*. 2017. Vol. 82. P. 128-150. DOI: https://doi.org/10.1016/j.eswa.2017.04.003
7. Hamid P., Hoseinali A., Behrouz M. A Modification on K-Nearest Neighbor Classifier. *Global Journal of Computer Science and Technology*. 2010. Vol. 10, Iss. 14. P. 37-41.
8. Singh A., Patel S. Applying Modified KNearest Neighborn to Detect Threat in Collaborative Information Information Systems. *International Journal of Innovative Research in Science, Engineering and Technology*. 2014. Vol. 3, N 6. P. 14141-14151
9. Маляр М. М., Кондрук Н. Е. Структурування критеріального простору за кутовою мірою

подібності. *Науковий вісник Ужгородського університету. Серія «Математика і інформатика».* 2020. Вип. 1(36). С. 85-91. DOI: https://doi.org/10.24144/2616-7700.2020.1(36).85-91

10. Kondruk N. E. Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control.* 2018. N 3 (46). P. 98-105. DOI: https://doi.org/10.15588/1607-3274-2018-3-11
11. Kondruk N. E. A comparative study of cluster validity indices. *Radio Electronics. Computer Science. Control.* 2019. N 4. P. 59 – 67. DOI: https://doi.org/10.15588/1607-3274-2019-4-6
12. Kondruk N. Clustering method based on fuzzy binary relation. *Eastern-European Journal of Enterprise Technologies.* 2017. Vol. 2, N 4 (86). P. 10–16. DOI: https://doi.org/10.15587/1729–4061.2017.94961

Kondruk N. E. Use of similarity measures in classification methods.

This study is a development of the application of different types of similarity measures in data mining problems. Data mining is the process of extracting implicit information from a database, which characterizes hidden connections and structures. This type of analysis is projected to be extremely popular over the next decade. One of the methods of data extraction is classification. The paper provides an overview of modern areas of supervised classification. The most popular method of classifying objects with numerical attributes is the method of K-nearest neighbors (KNN). It has been found that the predictive value of a class label can be improved by using the weighted influence of each neighbor on the result. Thus, it is advisable to modify the KNN method. In this case, it is proposed to introduce a function that characterizes the similarity of the unlabeled object with its nearest neighbors in the form of a measure of similarity. Based on it, indicators of weighted counting of votes of neighbors for a certain class mark are introduced. Software has been developed that implements the described approach. Practical experiments have shown its effectiveness in solving certain classes of applied problems.

Keywords: classification, algorithm k-nearest neighbors, KNN, similarity measures, supervised machine learning.

Список використаної літератури

1. Oprea, C. (2014). Performance evaluation of the data mining classification methods. *Information society and sustainable development,* 1, 249-253. https://doi.org/10.9790/0661-1060106
2. Jain, N., & Vishal, S. (2013). Data mining techniques: a survey paper. *IJRET: International Journal of Research in Engineering and Technology*, 2(11), 116-119. https://doi.org/10.15623/ijret.2013.0211019
3. Kumari, M., & Godara, S. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. *IJCST*, 2, 304-308.
4. Jantawan, B., & Tsai, C. (2013). The Application of Data Mining to Build Classification Model for Predicting Graduate Employment. *International Journal of Computer Science and Information Security, 11(10).* 1-7.
5. Xhemali, D., Hinde, C., & Stone, R. (2009). Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *IJSCI: International Journal of Computer Science, 4(1)*, 16-23.
6. Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications, 82*, 128-150. https://doi.org/10.1016/j.eswa.2017.04.003
7. Hamid, P., Hoseinali, A., & Behrouz, M. (2010). A Modification on K-Nearest Neighbor Classifier. *Global Journal of Computer Science and Technology, 10(14)*, 37-41.
8. Singh, A., & Patel, S. (2014). Applying Modified KNearest Neighborn to Detect Threat in Collaborative Information Information Systems. *International Journal of Innovative Research in Science, Engineering and Technology, 3(6).* 14141-14151
9. Malyar, M. M., & Kondruk, N. E. (2020). Structuring of the ctriterional space by an anglesimilarity measure. *Scientific Bulletin of Uzhhorod University. SIeries of Mathematics and Informatics, 1(36)*, 85-91. https://doi.org/10.24144/2616-7700.2020.1(36).85-91

10. Kondruk, N. E. (2018). Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control, 3(46),* 98-105. https://doi.org/10.15588/1607-3274-2018-3-11
11. Kondruk, N. E. (2019). A comparative study of cluster validity indices. *Radio Electronics. Computer Science. Contro, 4*, 59-67. https://doi.org/10.15588/1607-3274-2019-4-6
12. Kondruk, N. (2017). Clustering method based on fuzzy binary relation. *Eastern-European Journal of Enterprise Technologies, 2(4(86))*, 10-16. https://doi.org/10.15587/1729-4061.2017.94961

Одержано 13.04.2021