

ДОСЛІДЖЕННЯ ВПЛИВУ ШУМУ В ТЕОРІЇ ПРИБЛИЗНИХ МНОЖИН

Чернишов Дмитро Владиславович

здобувач вищої освіти другого (II) ступеня факультету комп'ютерних наук
Харківський національний університет радіоелектроніки, Україна

Науковий керівник: Ситніков Дмитро Едуардович

ORCID ID: 0000-0003-1240-7900

Професор кафедри системної інженерії, канд. техн. наук, доцент
Харківський національний університет радіоелектроніки, Україна

У складному світі аналізу даних класична теорія приблизних множин (rough set theory) займає провідну позицію, особливо цінується за свою ефективність у відборі ознак. Однак, дані, з якими ми стикаємося у реальному світі, часто містять шум - змішання помилкових та неважливих елементів, що можуть спотворити аналіз. Це дослідження занурюється у цю проблему, фокусуючись на тому, як шум впливає на функціональність та ефективність Класичної теорії приблизних множин. Наша мета - дослідити нюанси того, як шум впливає на розрізнення даних та надійність методів відбору ознак, що базуються на RST.

Класична теорія приблизних множин, розроблена Здзіславом Павлаком у 1980-х роках, пропонує структурований підхід до обробки даних з внутрішньою невизначеністю та непрецизністю. Ця теорія знайшла застосування у різноманітних галузях, таких як машинне навчання, аналіз даних та відкриття знань, головним чином через використання просторів апроксимації для класифікації даних. Однак, нещодавні дослідження пролили світло на значний вплив шуму на функціональність RST. Зокрема, шум має тенденцію розширювати межу області, що розділяє нижні та верхні апроксимації, внаслідок чого зменшується точність відбору ознак. Це дослідження прагне поглибити розуміння того, як встановлені алгоритми RST, включаючи Quickreduct, Relative Reduct та Entropy-based Reduct, піддаються впливу шумних умов даних. Незважаючи на ці наукові прогреси, залишається нерозв'язана область щодо всебічного впливу шуму на RST. Ця робота має на меті заповнити цей пробіл, поєднуючи існуючу літературу та теоретичні рамки, щоб надати більш ясне розуміння викликів та потенційних рішень для управління шумом у межах класичної теорії приблизних множин.

Основою RST є принцип нерозрізнення, який групує об'єкти, які неможливо відрізнити на основі їхніх атрибутів. Введення шуму порушує ці відносини нерозрізнення, тим самим впливаючи на нижні та верхні апроксимації, що є невід'ємною частиною RST. Наш аналіз починається з детального дослідження математичних конструкцій RST, зокрема зосереджується на тому, як шум змінює матрицю розрізнення та формування редуктів - мінімальних наборів атрибутів, які зберігають класифікаційну силу повного набору даних. Наші результати вказують на те, що атрибути, що містять шум, можуть затемнити справжні взаємодії між атрибутами та класами рішень, часто призводячи до включення невідповідних або оманливих атрибутів у редукти. Практичний експеримент становить основу нашого емпіричного дослідження, використовуючи бінарну версію набору даних "Титанік" з 13 конкретними бінарними ознаками. Ми послідовно вводили шум у набір даних, спостерігаючи за наступними змінами в відносній межевій області, зображеної через

(1) для ясності. Результати графіка (рис. 1) демонструють ефект додавання N шумових ознак на межу області. Крім того, ми провели аналіз найгіршого сценарію, використовуючи бінарний характер наших ознак, як окреслено в (2).

$$\text{Relative Boundary} = \frac{\text{Boundary Size}}{\text{Total Rows}} \quad (1)$$

$$\text{WorstBoundaryDecay}(x) = \frac{\text{WorstBoundaryDecay}(x-1)}{2} \quad (2)$$

Наші спостереження виявили послідовне зниження середньої граничної області, що відповідає нашим теоретичним очікуванням. Ми також аналізували випадково згенерований набір даних з 512 зразків (рис. 2), спостерігаючи за початково великими граничними областями, які значно зменшилися після додавання певних особливостей. У класичній теорії приблизних множин (RST) залежність визначає ступінь залежності класифікації одного набору від атрибутів іншого набору. Шум може спотворювати ці виміри залежності, що може призвести до потенційно помилкових висновків щодо значущості атрибутів.

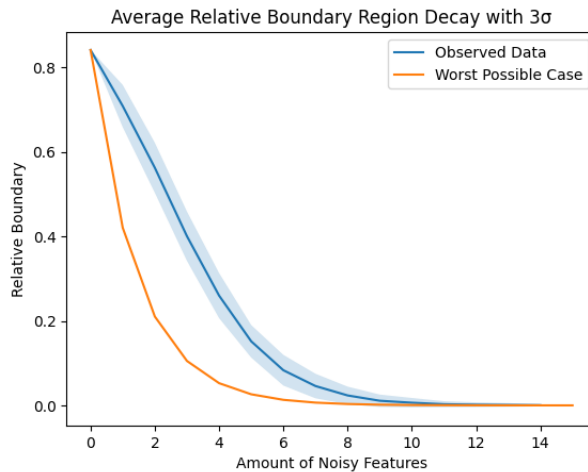


Рис. 1. Зменшення середнього відносного граничного регіону з 3σ

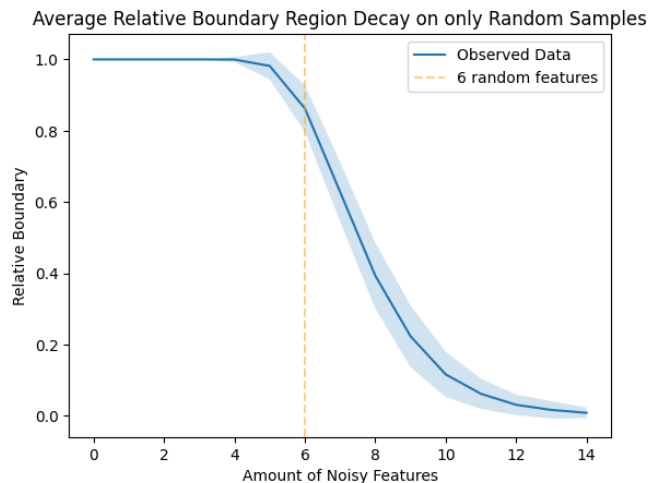


Рис. 2. Варіації граничної області у випадковому наборі даних

Цей шаблон вказує на зменшення граничної області у нашому випадковому наборі даних у діапазоні від $(\log_2 M) - 3$ до $(\log_2 M) + 3$ випадкових

особливостей, де вказує на розмір набору даних. Ця гіпотеза була додатково підтверджена шляхом повторення експерименту з наборами даних різних розмірів. Цей теоретичний та емпіричний аналіз має на меті кількісно визначити вплив шуму на RST та надати уявлення про стійкість методів вибору особливостей на основі RST у шумових середовищах.

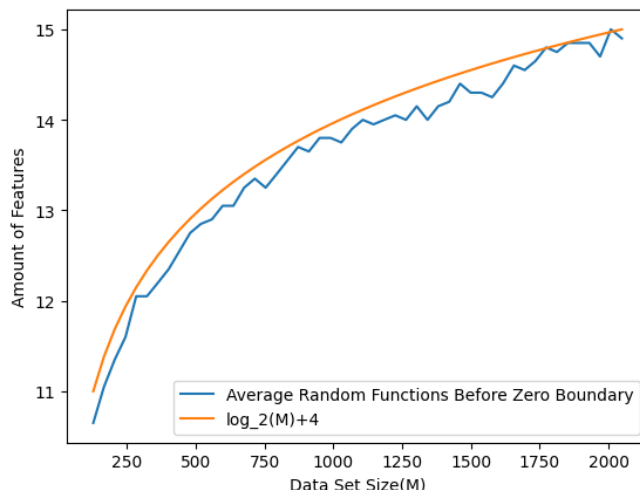


Рис. 3. Кореляція нульового кордонного регіону з розміром набору даних

Висновки. Це дослідження впливу шуму в теорії приблизних множин підкреслює важливий аспект аналізу даних: вразливість методів вибору ознак. Дослідження показує, що шум глибоко впливає на відносини розпізнавання та залежності в RST, ставлячи серйозні виклики в шумових середовищах. Наші експериментальні результати підкреслюють вплив шуму на кордонний регіон у методах вибору особливостей на основі RST, особливо у двійкових наборах даних. Рекомендується вибирати оптимальну кількість основних особливостей у діапазоні від $(\log_2 M) - 3$ до $(\log_2 M) + 3$ випадкових особливостей, де позначає розмір набору даних, щоб ефективно збалансувати представлення особливостей проти невизначеності, внесеної шумом. Ці уявлення є критично важливими для науковців та дослідників, які використовують RST, підкреслюючи важливість попередньої обробки даних для зменшення шуму та вибору відповідних методологій RST, адаптованих до конкретних характеристик шуму. Ця дисертація сприяє глибшому розумінню обмежень класичної RST та спрямовує на більш точний та надійний аналіз даних у шумових умовах.

Список використаних джерел:

1. Pawlak, Z. (1982). "Rough sets." *International Journal of Computer & Information Sciences*, 11(5), 341-356.
2. Skowron, A., & Rauszer, C. (1992). "The discernibility matrices and functions in information systems." *In Intelligent Decision Support*. Springer, Dordrecht.
3. Swiniarski, R. W., & Skowron, A. (2003). "Rough set methods in feature selection and recognition." *Pattern Recognition Letters*, 24(6), 833-849.
4. Bania, Rubul Kumar. (2015). "Comparative Review on Classical Rough Set Theory based Feature Selection Methods." *International Journal of Computer Applications*, 114, 31-35.