УДК 004.89

**А. А. Папа**
**А. А. Яровий**
**Ю. М. Паночишин**

# ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ НА ОСНОВІ АЛГОРИТМІВ ВИБОРУ ПІДМНОЖИНИ ОЗНАК

Вінницький національний технічний університет

***Анотація***

*Розглянуто актуальність задачі прогнозування відтоку клієнтів. Здійснено аналіз алгоритму вибору підмножини ознак для вирішення задачі прогнозування відтоку клієнтів, за результатами якого була підтверджена доцільність та перспективність застосування даного алгоритму у реальному програмному продукті.*

**Ключові слова:** прогнозування відтоку клієнтів, алгоритм обирання підмножини змінних.

***Abstract***

*The relevance of the problem of prediction the outflow of customers is considered. The analysis of the algorithm for selecting a subset of features for solving the problem of predicting the outflow of customers is carried out. The results confirmed the feasibility and prospects of using this algorithm in a real software product.*

**Keywords**: predicting customer churn, variable subset selection algorithm.

## Вступ

Організації, які надають послуги великій кількості користувачів, стикаються із серйозною проблемою утримання наявних та залучення нових клієнтів. Незважаючи на те, що питання утримання та залучення клієнтів викликало інтерес дослідників і аналітиків, жодне дослідження не представило реалізованих математичних алгоритмів або програмних рішень, які могли б повністю вирішити цю особливу проблему для певної послуги та регіону. Велику кількість аналогічних завдань доводиться вирішувати в умовах, коли дані містять деякі ознаки, що є або надлишковими, або недоречними, і в такому випадку їх може бути усунено без спричинення значної втрати інформації за допомогою алгоритму вибору підмножини ознак. Таким чином, для вирішення задачі збереження існуючих клієнтів доцільно проаналізувати алгоритм вибору підмножини ознак, який може бути корисним при прогнозуванні відтоку клієнтів [1].

Метою роботи є аналіз алгоритму вибору підмножини ознак, що в подальших дослідженнях дозволить підвищити точність прогнозування відтоку клієнтів.

## Результати дослідження

Алгоритм вибору підмножини ознак можливо розглядати як поєднання методики пошуку для пропонування нових підмножин ознак разом із мірою оцінки, яка встановлює бали різним підмножинам ознак. Найпростішим алгоритмом є перевірка кожної можливої підмножини ознак, шукаючи таку, яка

мінімізує рівень похибки. Це є вичерпним пошуком у просторі ознак, проте є обчислювально громіздким та неефективним для множин ознак, крім найменших. Вибір оцінювальної метрики сильно впливає на алгоритм, і саме ці оцінювальні метрики вирізняють три основні категорії алгоритмів вибору підмножини ознак на основі методів: обгортання, фільтрів та вкладення [2].

Методи обгортання – використовують модель апріорної оцінки результату для ранжування піднабору ознак. Оскільки вказані методи тренують нову модель для кожної підмножини ознак, вони є обчислювально затратними, проте зазвичай досить якісно пропонують множину ознак для окремого типу моделі. До головних недоліків цієї групи методів можна віднести: підвищення ризику перенавчання за недостатньої кількості спостережень, значний час обрахунку за великої кількості змінних [2].

Методи фільтрів – використовують опосередкований показник замість показника помилки з метою оцінювання піднабору ознак. Цей показник обирається так, щоб його можна було легко обчислити при збереженні показника корисності набору ознак. Зазвичай використовують такі показники, як взаємна інформація, поточкова взаємна інформація, коефіцієнт кореляції Пірсона, алгоритми на основі Relief, внутрішнокласова/міжкласова відстань або ж результат критеріїв значущості для кожної комбінації клас/ознака. Варто зазначити, що методи фільтрів зазвичай є менш обчислювально затратним, ніж методи обгортання, проте їх результатами є набори ознак, які не налаштовані на специфічний тип прогностичної моделі. Методи фільтрів також можна застосовувати як початковий етап для методів обгортання, уможливлюючи застосування обгорток до більшої кількості задач. Одним з інших популярних підходів є алгоритм рекурсивного усунення ознак, який зазвичай застосовується із методом опорних векторів для повторної побудови моделі та усунення ознак з низькими ваговими коефіцієнтами [3].

Методи вкладення – узагальнена група методів і методик, що виконують відбір ознак як частину процесу побудови моделі. Прикладом цього підходу є метод оцінювання коефіцієнтів лінійної регресійної моделі (Least absolute shrinkage and selection operator – LASSO). Вказані методи з погляду складності обрахунку можна наближено віднести посередині між методами фільтрів та методами обгортання [4, 5].

У традиційному регресійному аналізі найпопулярнішим видом вибору ознак є поетапна регресія, яка є методикою обгортання. Вона представляє собою жадібний алгоритм, що під час кожного раунду додає найкращу ознаку або видаляє найгіршу, і має свої недоліки [5-7].

Розглянуті вище групи методів є основою алгоритмів вибору підмножини ознак. Як підсумок варто зазначити, що вибір оцінювальної метрики сильно впливає на алгоритм вибору підмножини ознак, і саме ці оцінювальні метрики відрізняють три основні категорії методів, що є в основі алгоритмів вибору підмножини ознак.

Для розробки програмного продукту для аналізу відтоку клієнтів компанії, необхідно визначитись зі структурними елементами, які реалізовуватимуть функції, що необхідні для вирішення поставлених задач. Для забезпечення захисту вмісту потрібно розробити модуль входу, що передбачатиме авторизацію та реєстрацію користувачів. Дані користувачів, їхні звіти, файли, правила, що застосовуються для аналізу відтоку клієнтів, необхідно зберігати в базі даних. Перед прогнозуванням подій та завантаженості необхідно ввести дані в систему (анкету), яка буде оброблятись в модулі аналізу даних. Модуль прийняття рішень надає рекомендації щодо обрання однієї із двох подій. Модуль налаштування передбачає внесення змін у дані і в систему.

На рис. 1 зображено IDEF0 діаграму першого рівня декомпозиції програмного засобу, що буде аналізувати відтік клієнтів компанії. Вхідними даними є дані від користувача, характеристики клієнта, оточення.

Purpose: Цей процес має бути промодельований для демонстрації роботи поточних (AS-IS) процесів у проектованій інформаційній технології аналізу відтоку клієнтів компанії у вигляді моделі, що представляє собою ієрархічно впорядковані та взаємопов’язані діаграми, що ілюструють зв'язок системи з зовнішніми сутностями та взаємозв’язки внутрішніх процесів системи. Користувач при цьому отримує представлення про організацію системи як в цілому, так і її окремих функціональних блоків.

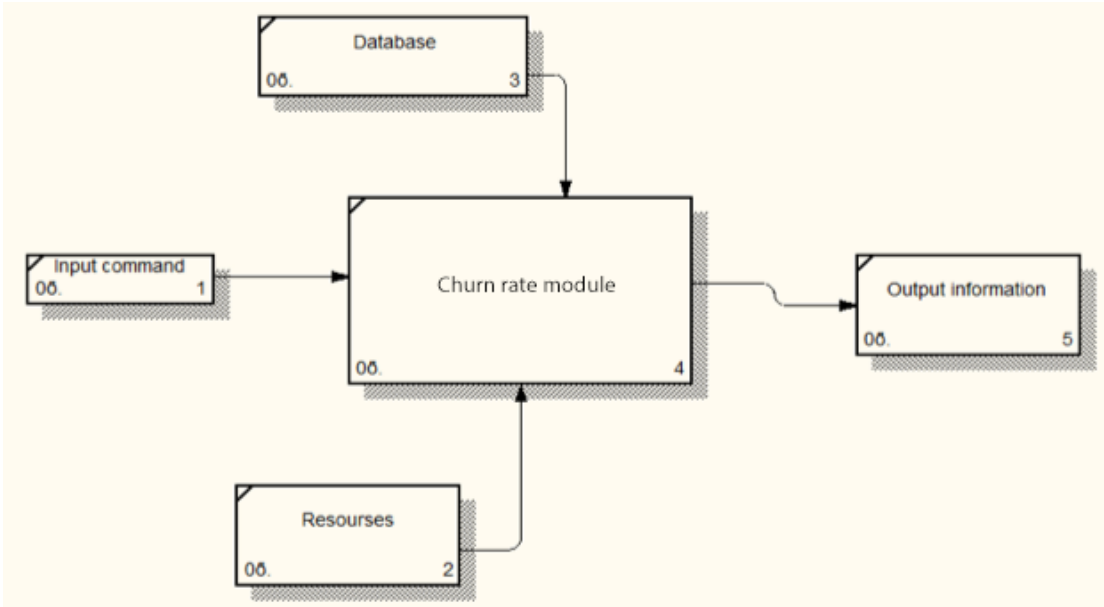

Рисунок 1 – IDEF0 діаграма першого рівня декомпозиції системи

Viewpoint: При побудові моделі система була розглянута з точки зору її користувачів.
Definition: Модель створюється для ілюстрації роботи системи на різних рівнях декомпозиції.
Scope: Загальне управління даними системи: їх отримання, обробка та збереження.

**Висновки**

Здійснено аналіз алгоритму вибору підмножини ознак для задачі прогнозування відтоку клієнтів, за результатами якого була підтверджена доцільність та перспективність застосування підходу, який побудований на основі методів машинного навчання для вирішення вказаної задачі. Досліджено модель прогнозування відтоку клієнтів телекомунікаційної компанії, що відрізняється від відомих застосуванням алгоритму вибору підмножини ознак

Проведено тестування системи аналізу відтоку клієнтів. Тестування показало повну відповідність системи поставленим задачам, а саме виконання аналізу менше ніж за 2 секунди, точність формування прогнозу відтоку клієнта більше 85%. При порівнянні розробленого продукту з аналогами, отримано кращі результати за рахунок використання комбінації методів машинного навчання.

Отримані результати планується використати в подальших дослідженнях з метою точності процесу прогнозування відтоку клієнтів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Andrii Papa, Yevhen Shemet, Andrii Yarovyi, Lyubov Vahovska “Development of information technology for analyzing the customer churn of a telecommunication company”. – Information and control systems. – Vol. 2, No. 2(64), 2022. – p. 11-15. – DOI: https://doi.org/10.15587/2706-5448.2022.255861

2. Guyon and A. Elisseeff, "Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

3. Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, USA, 1997, pp. 412-420.

4. Bach, Francis R (2008). Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of the 25th International Conference on Machine Learning. с. 33–40.

5. Phuong TM, Lin Z, Altman RB. Choosing SNPs using feature selection. Proc IEEE Comput Syst Bioinform Conf. 2005:301-9. doi: 10.1109/csb.2005.22. PMID: 16447987.

6. G. Forman and I. Guyon, "An extensive empirical study of feature selection metrics for text classification," Journal of Machine Learning Research, vol. 3, pp. 1289-1305, Mar. 2003.

7. Saghapour, E., Kermani, S., & Sehhati, M. (2017). A novel feature ranking method for prediction of cancer stages using proteomics data. PLOS ONE, 12(9), e0184203. DOI: https://doi.org/10.1371/journal.pone.0184203.

**Папа Андрій Андрійович –** асистент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: papa.andriy@gmail.com.

**Яровий Андрій Анатолійович –** д.т.н., професор, завідувач кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

**Паночишин Юрій Миколайович –** к.т.н., доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

**Andrii A. Papa –** Lecturer of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske Shose, 95, e-mail: papa.andriy@gmail.com.

**Andrii A. Yarovyi –** Doctor of Science (Eng.), Professor, Head of the Computer Science Department, Vinnytsia National Technical University, Vinnytsia.

**Yurii M. Panochyshyn** – Ph. D. (Eng), Associate Professor of the Department for Computer Science, Vinnytsia National Technical University, Vinnytsia.