Бакун С.А., Терентьєв О.М.
КПІ ім. Ігоря Сікорського, ІПСА, Київ, Україна

Методи перетворення категоріальних змінних в числові

Вступ. На сьогоднішній день все більшого масштабу набуває практичне застосування методів знаходження закономірностей у великих обсягах даних та методів прогнозування, зокрема, побудова регресійних моделей. Переважна більшість таких алгоритмів дозволяють працювати лише з числовими ознаками для опису об'єктів, або процесів, що спостерігаються. Однак на практиці часто зустрічаються задачі з категоріальними змінними, значення яких позначають приналежність об'єктів до якоїсь категорії. Приклади таких ознак: національність, громадянство, професія тощо. Нагальною необхідністю є коректне перетворення текстових значень з фіксованого набору, тобто категорій, у числовий формат, для забезпечення можливості роботи з категоріальними змінними.

Метод фіктивних змінних. Найвідомішим та простим методом перетворення категоріальних змінних у числові є метод фіктивних змінних. Суть методу полягає у розбитті кожної категоріальної змінної на множину фіктивних бінарних змінних (табл. 1). При цьому одна з категорій зазвичай не кодується з метою забезпечення функціональної незалежності множини створених змінних. Таким чином, категоріальна змінна перетворюється в набір $k-1$ бінарних змінних, де k – кількість взаємовиключних категорій даної категоріальної змінної [1].

Табл. 1. Приклад перетворення змінної "Освітньо-кваліфікаційний рівень"

Назва категорії	A1	A2	A3
Немає	0	0	0
Молодший спеціаліст	1	0	0
Бакалавр	0	1	0
Магістр	0	0	1

Недоліками даного методу є те, що метод не враховує ні характеристики розподілу категорій, ні будь-який взаємозв'язок з цільовою змінною. Ще одним недоліком такого підходу є суттєве збільшення кількості змінних при заміщенні кожної категоріальної змінної на множину фіктивних змінних. Більшість з алгоритмів не зможуть обробити отриману кількість даних на реальних задачах.

Метод використання ваг категорій. Альтернативою методу фіктивних змінних є метод на основі використання ваг категорій, або WOE (Weight of evidence). WOE вимірює статистичну значущість кожного класу змінної [2]. Якщо змінна, що спостерігається, є бінарною змінною, тобто приймає тільки два значення (наприклад, 1 — якщо певна ознака присутня; 0 — якщо вона відсутня), то WOE розраховується за формулою:

$$WOE = ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \tag{1}$$

де $d_i^{(1)}$ – відносна доля першого спостереження в i-й категорії;
$d_i^{(2)}$ – відносна доля другого спостереження в i-й категорії.

Даний метод дозволяє ставити у відповідність категоріальним значенням числові значення ваг категорій змінної (табл. 2).

Порівняльний аналіз. В рамках порівняльного аналізу методів перетворення категоріальних змінних у числові було побудовано дві регресійні моделі, для прогнозування кредитоспроможності фізичних осіб. Для побудови моделей було використано вибірку даних з німецького банку, що надає кредити фізичним особам. Набір даних містить 3000 записів по клієнтах, у яких вже закінчився строк кредитування, та включає в себе інформацію щодо 17 показників

Табл. 2. Приклад перетворення змінної "Ціль кредиту"

Назва категорії	Значення WOE
Автомобіль	-0.21
Побутова техніка	-0.11
Відпочинок	0
Покупки в магазинах	0.05
Меблі	0.19

(анкетних даних) по кожній особі, з яких 7 є категоріальними:

1. TITLE – стать особи;
2. STATUS – сімейний стан;
3. REGN – регіон;
4. PRODUCT – тип бізнесу;
5. NAT – національність особи;
6. CARDS – тип банківської карти;
7. RESID – тип проживання.

Як можна побачити з таблиць 3 та 4, побудована модель з використанням коефіцієнту WOE для перетворення категоріальних змінних в числові має кращі значення індексу GINI та загальної точності моделі ніж з використанням методу фіктивних змінних на навчальній та тестовій вибірках.

Табл. 3. Порівняльна таблиця характеристик якості моделей на навчальній вибірці

Модель	Загальна точність (CA)	Індекс GINI
Модель з використанням фіктивних змінних	0.85	0.65
Модель з використанням ваг категорій	0.87	0.67

Табл. 4. Порівняльна таблиця характеристик якості моделей на тестовій вибірці

Модель	Загальна точність (CA)	Індекс GINI
Модель з використанням фіктивних змінних	0.82	0.57
Модель з використанням ваг категорій	0.84	0.6

Висновки. Перетворення категоріальних змінних в числові є досить трудомісткою задачею. Застосування методу фіктивних змінних передбачає додавання великою кількості нових змінних у модель, що ускладнює модель і погіршує оцінку її якості. Було запропоновано альтернативний метод перетворення категоріальних змінних – на основі використання ваг категорій. В ході роботи було проведено порівняльний аналіз даних методів. В результаті отримано підтвердження, що метод на основі використання ваг категорій може застосовуватися як метод, альтернативний методу фіктивних змінних.

Література. **1.** Hosmer D.W. Applied logistic regression / David W. Hosmer, Jr., Stanley Lemeshow. — 2nd ed. — Hoboken: John Wiley & Sons, Inc., 2000. — 375 p. **2.** Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring / Naeem Siddiqi. — Hoboken: John Wiley & Sons, Inc., 2006. — 196 p.