МОДЕЛЮВАННЯ ШВИДКОСТІ ПРОДАЖ ТА РЕКОМЕНДОВАНОЇ СТРУКТУРИ ТОВАРУ НА ОСНОВІ МЕТОДУ СКОРОЧЕННЯ ДАНИХ

Р. Пасічник, Б. Маслияк, В. Віцентій

Економічний успіх підприємства прямо залежить від того, наскільки його продукція задовольняє потреби споживачів. Тому відповідність продукту вимогам ринку можна визначити безпосередньо з економічних показників діяльності підприємства. Індикаторами тут можуть служити об'єм збуту, прибуток, покриття постійних затрат. Однак існують ситуації коли неможливо оцінити продукт з точки зору економічних показників, наприклад, коли рішення про іновацію приймаються раніше, ніж з'являються дані про реакцію ринку. В цьому випадку потрібно перевірити адекватність продукту його вимогам [1].

Степінь відповідності продукту суб'єктивним уявленням споживачів можна встановити за допомогою опитувань, що, незважаючи на свою розповсюдженість, має ряд суттєвих недоліків. Сюди можна віднести низьку об'єктивність респондентів а також затрати, необхідні для організації масштабних опитувань. Інший підхід оцінки адекватності продукту полягає в диференціальній оцінці окремих його елементів та властивостей. Однією із найбільш розповсюджених моделей цього типу є компесаційна модель Розенберга:

$$A_j = \sum_i V_i I_{ij}$$

де A_j – суб'єктивна придатність продукту;

V_j – важливість мотиву для споживача;

I_{ij} – суб'єктивна оцінка придатності продукту j для задоволення мотиву i.

Недоліком цієї моделі є складність оцінок мотивів, що є важливими для продукту. Більший прикладний характер має модель, в якій значення окремих мотивів визначаються опосередковано через конкретні характеристики продукту:

$$Q_j = \sum_k X_k Y_{jk}$$

де Q_j – оцінка споживачами марки j;

X_k – важливість характеристики k;

Y_{jk} – оцінка характеристики k марки j з точки зору користувачів.

Попередні моделі грунтувалися на допущенні, що кожна характеристика бажана і що найвища оцінка дає найкращу характеристику. Цей недолік усувають моделі з ідеальною точкою

$$Q_j = \sum_k X_k \mid Y_{jk} - I_k \mid^r$$

де I_k – ідеальне значення характеристики k з точки зору споживачів, r – параметр, який визначає при r=1 постійну, а при r=2 спадну корисність.

Однак всі згадані методи опираються на величини, які встановлюються на основі суб'єктивних оцінок споживачів або експертів і тому мають невисоку достовірність. Для її підвищення необхідно будувати методи, які грунтуються на обліковій інформації товарного дистриб'ютера, зокрема на обробці об'ємів його щоденних продаж. Вони можуть фіксуватися в комп'ютеризованій системі автоматично на основі даних, що поступають від касових апаратів. Таким чином на протязі певного періоду спостережень та узагальнення інформації можна встановити усереднену денну швидкість продаж кожного товару. Величини швидкостей приймають значення в широкому діапазоні. Вони точно характеризують динаміку продаж конкретного товару, але малоінформативні в оцінці загального становища по продажу товарів даної групи. Тому потрібно згрупувати значення швидкостей по окремих категоріях. Для такої операції зручно використати метод одновимірної кластеризації.

Кластеризацію проведемо за допомогою розділення сумішей нормально розподілених величин.

При цьому віддаль між окремими товарами T_i, T_j обчислюються як абсолютні величини різниць між їхніми швидкостями продаж V_i, V_j:

$$d_V(T_i, T_j) = \mid V_i - V_j \mid$$

На першому кроці вибираємо два товари, віддаль між якими мінімальна. Їх об'єднуєм в початковий кластер. Далі послідовно вибираємо товари, що мають мінімальні віддалі до

вибраного кластеру. Якщо значення швидкості продажу вибраного товару можна віднести до однієї вибірки із швидкостями продажу кластеру, то товар включається в кластер. В іншому випадку формування кластеру припиняється

Об'єкти в кластері нумеруються в порядку зростання значень їх швидкостей продажу. До цього кластеру віднесемо об'єкти, значення швидкостей продажу яких утворюватимуть вибірку деякої нормально розподіленої випадкової величини. Тому на наступних етапах послідовно вибираємо товари T_i, що мають мінімальні віддалі до найближчого елементу T_B кластеру і попередньо включаємо їх в нього. Далі перевіряємо обгрунтованість такого включення за формулами [2]:

$$\left|\frac{V_i - V_{k(n)}}{V_i - V_{l(n)}}\right| < M_{n,\alpha} \quad при \quad n \le 20$$

$$\left|\frac{V_i - Vm}{Vs}\right| < N_{n,\alpha} \quad при \quad n > 20$$

де n – кількість об'єктів включених в підкластер $kl(ap,m)_l$,

V_i – значення швидкості продажу для об'єкту кластеру, що включений в нього останнім,

V_k – значення швидкості продажу для к-го об'єкту кластеру,

$$k(n) = \begin{cases} B(i,1) & при \quad n \le 10 \\ B(i,2) & при \quad n > 10 \end{cases}$$

$$l(n) = \begin{cases} D(i,0) & при \quad n \le 7 \\ D(i,1) & при \quad 7 < n \le 13 \\ D(i,2) & при \quad 13 < n \le 20 \end{cases}$$

B(i,s) – номер s-го сусіда в кластері до об'єкту, що включений в кластер останнім,

D(i,s) – номер s-го сусіда в підкластері до об'єкту, який протилежний до об'єкту, що включений в кластер останнім,

$M_{n,a}$ – межа значимості при перевірці екстремальних значень для одностороннього критерію,

Vm, Vs – середнє та середньоквадратичне значення вхідного параметра по кластеру без елемента, що був включений лише попередньо,

$N_{n,a}$ – верхня межа значимості стандартизованого екстремального відхилення.

Якщо приналежність об'єкта до кластеру підтверджується, то проводиться попереднє включення наступного об'єкта з наступною перевіркою. В іншому випадку формування кластеру припиняється. Далі розглядаються підмножини значень швидкостей які утворилися в результаті виділених на попередніх етапах кластерів. Якщо в деякій підмножині кількість елементів не перевищує двох, то вони автоматично утворюють кластер, межами якого служить весь даний незаповнений попередніми кластерами інтервал. Інакше на цьому проміжку проводиться спроба виділення наступних кластерів. Процедура завершується коли кожен елемент множини буде віднесений до деякого кластера. Після віднесення товарів до відповідних кластерів будується оцінка їхньої швидкості продаж. Ця оцінка формується як середнє значення швидкості продаж по даному кластеру.

Для характеристики особливостей товарів певної групи вводиться множина показників, перелік яких встановлюють фахівці даного ринку. Показники до цього переліку добираються таким чином, щоби для кожного товару групи значення показника однозначно встановлювалися, а також, щоби сукупність показників достатньо повно описувала товари даної групи. По кожному із показників проводимо кластеризацію за тим же алгоритмом, що й для швидкості продажу товарів. Для кожного кластеру фіксується узагальнене значення показника, яке обчислюється як його середнє значення для елементів кластеру.

Таким чином по сукупності товарів можна побудувати матрицю взаємозв'язку узагальнених показників та узагальненої швидкості реалізації. Ця матриця служить основою для виділення найсуттєвіших ознак товарів, які обумовлюють значення їхньої швидкості продаж. З цією метою використаємо метод скорочення даних, керований даними (Data Driven Data Reduction – DDDR) [3]. Згідно цього методу для ідентифікації залежності узагальнених швидкостей продажу товару від узагальнених показників товару будується відповідне дерево рішень. Алгоритм формування дерева побудований таким чином, щоби використати мінімальну кількість показників та розгалужень. Такий підхід узгоджується із положеннями фудаментальної теореми розпізнавання образів, доведеної В.Н.Вапником та А.Я.Червоненкісом [4]. У ній встановлюється зв'язок між імовірністю похибки розпізнавання, розмірністю простору та довжиною навчальної вибірки. Зміст теореми полягає в тому, що якщо із множини вирішальних правил вибирається таке, що на навчальній вибірці певної довжини не здій-

снюється жодної помилки, то імовірність похибки розпізнавання на нових даних тим більша, чим більше число таких правил і чим коротша навчальна вибірка, яка розділяється безпомилково.

Згідно методу DDDR стверджується, що для ідентифікації кластеру результуючого показника необхідна інформація, яка виражається ентропією

$$Info(M) = -\sum_{s=1}^{n}(\frac{|Kv_s|}{|M|}\log_2\frac{|Kv_s|}{|M|})$$

Якщо ж спочатку розбити множину стрічок матриці узагальнених показників на базі значень нецільового показника X на кластери MX_i, то інформація необхідна для ідентифікації кластеру результуючого показника на кластері MX_i відповідно дорівнює

$$Info(MX_i) = -\sum_{s=1}^{n}(\frac{|Kv_s|}{|MX_i|}\log_2\frac{|Kv_s|}{|MX_i|})$$

Математичне сподівання інформації для ідентифікації результуючого показника на множині M, яка була попередньо розбита на підмножини MX_i, буде виражатися зваженою сумою

$$Info(X,M) = \sum_{i=1}^{m}\frac{|MX_i|}{|M|}Info(MX_i)$$

Цінність попереднього розбиття по атрибуту T буде виражатися різницею висхідної ентропії розбиття на кластери по результуючому показнику та залишкової інформації для ідентифікації після розбиття по показнику X.

$$Gain(X,M) = Info(M) - Info(X,M)$$

Для побудови дерева рішень використовується алгоритм, запропонований Quinlan. Основна ідея цього алгоритму полягає в попередньому розбитті матриці по атрибуту із максимальною цінністю на кластери, кожному з яких відповідатиме один вузол першого рівня в створюваному дереві рішень. Якщо $Info(MX_i)$ для і-го кластеру дорівнюватиме 0, то відповідний кластер далі не деталізується. В іншому випадку проводиться деталізація по аналогії із висхідною матрицею та добудовою наступного рівня дерева рішень. Цим самим будується дерево, яке містить мінімальну кількість рівнів.

Таким чином в статті запропоновано алгоритм виділення найбільщ інформативних ознак товарів, по яких може бути спрогнозована швидкість його продаж. Метод грунтується на обробці статистичної інформації про об'єми продаж за допомогою процедур добування знань (data mining). Побудоване дерево рішень дозволяє віднести товар, що аналізується по значеннях його атрибутів до певного кластеру швидкості продаж. При розробці нових товарів структура дерева рішень дає інформацію про те, які атрибути товару найсуттєвіше впливають на рівень його продаж, що дозволить спрямувати пошук проектувальників в потрібне русло.

ЛІТЕРАТУРА

[1] Х. Хершген Маркетинг: основы профессионального успеха, М.: ИНФРА–М, 2000– 334с.

[2] Л. Закс Статистическое оценивание, М.: Статистика, 1976. 598с.

[3] P. Reush, A. Flemming A New Generation of Data Mining, NITE Minsk,2000.

[4] В. Вапник, А. Червоненкис. Теория распознавания образов (статистические проблемы обучения), М.: Наука, 1974. 416с.

[5] J. R. Quinlan Introduction of Decision Trees, Machine Learning, 1 (1), p. 81-106.

Богдан Маслияк 1958 року народження. В 1983 році закінчив Львівський політехнічний інститут. 1994 року при Київському політехнічному інституті захистив кандидатську дисертацію. Автор близько 40 наукових праць. Коло наукових інтересів – моделювання економічних систем.

Віталій Віцентій 1975 року народження. Закінчив Тернопільську академію народного господарства – у 1998 році отримав диплом із відзнакою інженера-економіста за спеціальністю "Інформаційні системи в менеджменті", у 1999 році отримав диплом з відзнакою магістра за спеціальністю "Економічна кібернетика". З 1999 року навчається в аспірантурі. Коло наукових інтересів – інформаційно-пошукові системи, людино-машинна взаємодія, бази знань на основі онтологій.