МЕТОДЫ ОБРАБОТКИ ПОЛЬЗОВАТЕЛЬСКИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ БИЗНЕС-ПРОЦЕССОВ

Дрозд П. С., Козадаев К. В.
Факультет радиофизики и компьютерных технологий Белорусского государственного университета
Минск, Республика Беларусь
E-mail: drozdps@gmail.com, kozadaeff@mail.ru

В работе представлен новый способ решения задачи сегментации рынка путём кластеризации транзакционных данных клиентов. Данный способ основан на алгоритме локально-чувствительного хеширования, который используется как для снижения размерности исходного пространства признаков, так и для уменьшения количества обрабатываемых образов. Для непосредственной кластеризации значений хеш-функций использовался высокоэффективный алгоритм Bisecting K-mean. Метод показал высокую точность и оказался более устойчивым по сравнению с другими алгоритмами кластеризации. Реализована система, позволяющая в автоматическом режиме сегментировать целевой рынок предприятия.

Введение

Задача сегментации рынка может быть сформулирована как выделение определённых групп потребителей, для каждой из которых могут потребоваться различные подходы в бизнес-стратегии предприятия. В условиях конкуренции компаниям необходимо определять приоритеты и таргетировать своё предложение на целевые сегменты наиболее перспективных клиентов. В работе [1] решалась задача сегментации рынка с помощью самоорганизующихся карт Кохонена и иерархической кластериазции. Метод показал высокую точность, однако при использовании на практике оказался неустойчивым. Была поставлена цель разработать подход к кластеризации клиентских транзакционных данных, который должен иметь высокую точность и быть более устойчивым.

I. Анализируемый набор данных

В качестве тестовой выборки для проверки предложенного метода нами был использован набор данных «Ta-Feng», который выложен в свободный доступ компанией ACM RecSys. Он содержит информацию о покупках различных товаров, совершённых более чем 32 тысячами уникальных клиентов. Всего в наборе содержится 817741 запись, каждая из которых описывает совершённую покупателем транзацию с помощью 9 характеристик (дата проведения платежа, тип товара, сумма транзакции и т.д.). После анализа и предварительной обработки этих данных мы получили $p = 29$ уникальных характеристик для каждой транзакции (этап feature creation).

II. Метод локально-чувствительного хеширования

Алгоритм локально-чувствительного хеширования (LSH - от англ. Locality-Sensitive Hashing) традиционно применятся для решения задачи поиска ближайшего соседа[2], однако нами было принято решение использовать его для кластеризации. Особенность этого алгоритма заключается в том, что он использует специальный набор «плохих» хеш-функций, которые, в отличии от обычных хеш-функций, должны генерировать коллизии на схожих образах. Таким образом, соседние точки скорее всего попадут в одну хеш-корзину. Пусть $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ - точка в исходном p-мерном пространстве признаков, которая в рассматриваемом случае описывает одну транзакцию, совершённую одним клиентом (при этом p получилось равным 29 на этапе feature creation). Ограничим значения всех координат этой точки: $C = sup\{x_1, x_2, ..., x_p\} + \epsilon, \epsilon \geq 0$ - это условие не накладывает никаких ограничений на специфику решаемой задачи. Преобразуем $\mathbf{x}$ в новый вектор размерности Cp по следующему правилу:

$$\begin{cases} \mathbf{v}(\mathbf{x}) = \Psi_C(x_1)\Psi_C(x_2)...\Psi_C(x_p) \\ \Psi_C(x_i) = \underbrace{11...11}_{x_i}\underbrace{00...00}_{C-x_i}, \forall i \leq p \end{cases}$$

Функция $\Psi_C(x_i)$ переводит значение i-ой компоненты $\mathbf{x}$ в последовательность из x_i единиц, за которыми следуют $C - x_i$ нолей. Например, если $\mathbf{x} = (3, 4)^T$ и $C = 5$, то $\mathbf{v}(\mathbf{x}) = (1110011110)^T$. Хеш-функция в алгоритме LSH вычисляет своё значение для вектора $\mathbf{x}$ путём конкатенации k битов (параметр локально-чувствительного хеширования) из $\mathbf{v}(\mathbf{x})$, порядковые индексы которых содержатся в предварительно сгенерированном множестве Υ, содержащем k случайных целочисленных элементов из $\{1, 2, 3, ..., Cp\}$. На практике генерируется l множеств $\{\Upsilon_1, \Upsilon_2, ...\Upsilon_l\}$ и соотвественно l хеш-функций, каждая из которых вычисляет своё значение для каждого вектора исходного набора.

III. Алгоритм Bisecting K-means

Мы подобрали хеш-функции таким образом, чтобы их значения существовали в метрическом пространстве и между ними можно было вычислять расстояние Хэмминга. После этапа LSH все данные оказались разбиты на доме-

ны, объединяющие схожие транзакции, и каждому домену соответсвовало какое-то значение хеш-функции. Для непосредственного кластерного анализа этих значений был применён алгоритм Bisecting K-means - высокоэффективная иерархическая версия K-means. Этот алгоритм является дивизионным, так как предполагает, что все точки изначально принадлежат одному глобальному кластеру[3]. На каждой итерации он делит текущий кластер на два дочерних с помощью обычного K-means с фиксированным параметром $k = 2$. В некоторых случаях [3], Bisecting K-means на порядки производительнее K-means.

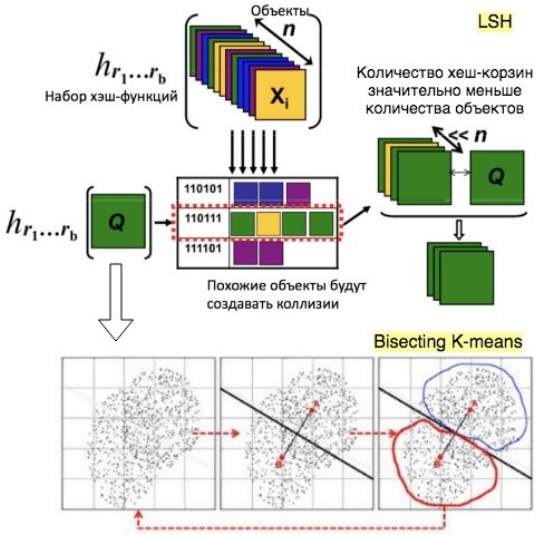

Рис. 1 – Схема предложенного метода

IV. Система для сегментации рынка

Рис. 2 – Функциональная схема системы

Предложенный подход был использован при разработке системы для автоматической сегментации рынка. Система основана на стеке технологий Apache® Spark и архитектуре **DASE** (**D**ata source and preparator, **A**lgorithm, **S**erving, **E**valutaion). В качестве клиентского приложения выступает платформа Salesforce® Sales Cloud, где сохраняются транзакционные данные клиентов после совершения ими каких-либо действий и покупок во внешнем приложении (например, интернет-магазине).

V. Результаты

Результат обработки системой набора данных Ta-Feng представлен ниже.

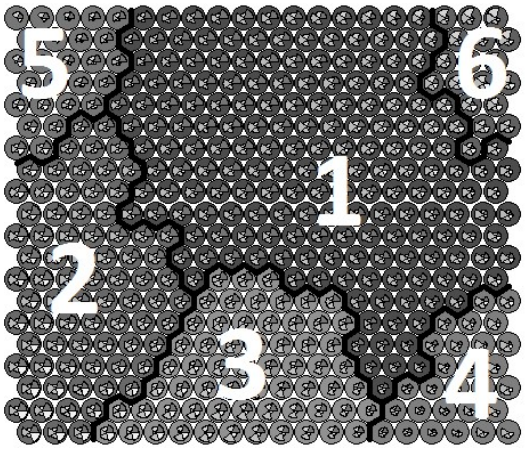

Рис. 3 – Результат сегментации Ta-Feng

Полученная информация может быть использована при планировании маркетинговой стратегии. Например, в кластере 6 набора Ta-Feng находятся VIP-покупатели, совершающие мало очень дорогих покупок. Анализ показал, что выручка от этого сегмента минимальна. Следовательно, бизнесу стоит таргетировать своё предложение на иные группы клиентов.

Заключение

Предложенный способ сегментации транзакционных данных покал высокую точность и оказался гораздо более устойчивым, чем классические алгоритмы кластеризации и алгоритмы, основанные на нейронных сетях[1]. Метод имеет высокую производительность, так как используются быстрые приёмы вычисления хеш-функций. Реализована система, автоматизирующая процесс сегментации рынка на основе локально-чувствительного хеширования и алгоритма Bisecting K-means. Система была протестирована на тестовом наборе данных Ta-Feng и в настоящее время проходит апробацию в крупном европейском ритейлере.

1. Drozd P. Kohonen's neural networks for Customer Segmention / P. Drozd // Open Readings 2018 abstract book / ed. E. Skliutas. – Vilnius, 2018. – P.103.
2. Koga, H. Hierarchial Clustering Algorithm Using Locality-Sensitive Hashing / H. Koga, T.Ishibashi, T. Watanabe // Discovery Science, 7 th International Conference, Padova, 2-5 oct. 2004 / ed. E. Suzuki. – Padova, 2014. – P.114-128
3. Fern, X.Z., Clustering ensembles for high dimensional data clustering / X.Z. Fern, C.E. Brodley // In Proc. International Conference on Machine Learning / ed. T. Fawcett. – Washington DC, 2003. – P.178-185