МЕТОДЫ КЛАСТЕРИЗАЦИИ ПОЛЬЗОВАТЕЛЬСКИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ БИЗНЕС-ПРОЦЕССОВ

Дрозд П. С., Адуцкевич И. А.
Факультет радиофизики и компьютерных технологий Белорусского государственного университета
Минск, Республика Беларусь
E-mail: {drozdps, dutsik}@gmail.com

В работе приведен способ решения задачи сегментации рынка, основанный на применении самоорганизующихся карт Кохонена и иерархического кластерного анализа. Для предварительной обработки данных были использованы методы Z-оценки и главных компонент. Были получены хорошие результаты в ходе проверки рассматриваемого подхода на реальных данных. Данный подход устраняет необходимость изменения архитектуры искусственной нейронной сети при изменении числа кластеров, что положительно сказывается на скорости работы при разведочном анализе данных в рамках определения маркетинговой стратегии предприятия. На языке R был реализован инструментарий, позволяющий нетехническим специалистам проводить кластеризацию данных клиентов, а также имеющий возможность интеграции с CRM-системами.

Введение

В рамках экономической теории выделяют понятия целевого рынка и целевого сегмента. Целевой рынок – это определённая группа потребителей, на которую таргетируется предложение товара или услуги. Целевой сегмент – это однородная группа потребителей целевого рынка предприятия. Под сегментированием рынка будем понимать выделение определённых групп потребителей, для каждой из которых могут потребоваться различные подходы в маркетинговой стратегии предприятия. Решать задачу сегментирования рынка будем с помощью кластеризации потребительских данных. На сегодняшний день существует тенденция того, что компании размещают свою ИТ-инфраструктуру в облачных системах (например, CRM Salesforce, Demandware). Поэтому выбираемый алгоритм должен быть масштабируемым и как можно более эффективно использовать возможность параллельных вычислений. Этому условию удовлетворяют самоорганизующиеся карты Кохонена. Тем не менее, карты Кохонена обладают рядом недостатков (необходимость перепроектирования архитектуры при изменении числа кластеров, возможность работы только с вещественными значениями и т.д.) [1], которые могут быть устранены с помощью описанного в работе метода.

Архитектура самоорганизующейся карты Кохонена состоит из двух слоёв – распределительного входного и выходного слоя Кохонена. Нейроны находятся в узлах двумерной сетки. Спецификой применения данного подхода является упорядоченность нейронов в этой сетки. Это свойство позволит поддерживать топологическое распределение объектов входного множества по группам. В связи с этим, использование карты Кохонена позволяет одновременно с кластеризацией данных решить задачу визуализации.

I. Анализируемый набор данных

Были выделены следующие этапы решения поставленной задачи интеллектуального анализа данных:

1. Предварительная визуализация и исследование данных с помощью гистограмм, диаграмм рассеяния;
2. создание новых переменных (feature creation) и очистка данных (data clean);
3. снижение размерности (метод главных компонент);
4. проектирование архитектуры карты Кохонена;
5. кластеризация главных компонент картой Кохонена, получение векторов весов нейронов;
6. кластеризация весов нейронов иерархическим способом
7. анализ полученных результатов.

Метод был апробирован на наборе 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.).

II. Предварительная обработка данных

Так как решалась задача кластеризации покупателей, была необходима структура данных в виде матрицы размера n на c, где n – количество покупателей, c – число характеристик [2]. «Ta-Feng» содержит данные о 32366 уникальных пользователей, однако впоследствии это значение оказалось несколько меньше. Это связано с удалением данных, содержащих сильные выбросы. На этапе создания новых переменных было

получено 29-мерное пространство признаков (см. рис. 1).

```
> str(features)
Classes 'data.table' and 'data.frame':  32266 obs. of  29 variables:
 $ customer_id              : int  1069 1113 1250 1359 1823 2189 3667 4282 4381 4947 ...
 $ age_group                : num  11 11 4 11 11 11 11 10 4 11 ...
 $ address                  : num  1 2 3 7 3 5 7 1 7 3 ...
 $ transactionNumber        : num  11 18 14 3 14 62 13 9 11 36 ...
 $ itemNumber               : int  16 23 18 4 25 141 26 9 13 40 ...
 $ totalSpend               : int  1944 2230 1583 364 2607 14056 11509 967 701 3363 ...
 $ productNumber            : num  10 17 14 3 13 58 10 9 11 35 ...
 $ basketNumber             : num  11 18 14 3 14 62 13 9 11 36 ...
 $ subclassNumber           : num  9 15 9 3 9 42 9 8 10 33 ...
 $ maxItemNumberPerDay      : int  9 9 13 4 16 89 11 7 13 24 ...
 $ meadianItemNumberPerDay  : num  2.5 5.5 9 4 6 70.5 6 4.5 13 20 ...
 $ minItemNumberPerDay      : int  2 3 5 4 3 52 3 2 13 16 ...
 $ maxTotalSpendPerDay      : int  971 628 849 364 1256 9078 8960 796 701 1875 ...
 $ medianTotalSpendPerDay   : num  393 591 792 364 918 ...
 $ minTotalSpendPerDay      : int  187 420 734 364 433 4978 329 171 701 1488 ...
 $ maxSubclassNumberPerDay  : num  5 6 6 3 6 30 6 6 10 20 ...
 $ medianSubclassNumberPerDay: num  2 4.5 4.5 3 3 23 2.5 4 10 17 ...
 $ minSubclassNumberPerDay  : num  1 2 3 3 1 16 1 2 10 14 ...
 $ subsclassPopularityA     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ subsclassPopularityB     : num  0 0 0 0 0 0.0714 ...
 $ subsclassPopularityC     : num  0 0 0.143 0 0 ...
 $ subsclassPopularityD     : num  0 0.111 0.143 0 0.143 ...
 $ subsclassPopularityE     : num  1 0.889 0.714 1 0.786 ...
 $ meanItemCost             : num  121.5 97 87.9 91 104.3 ...
 $ maxItemCost              : num  425 268 395 119 245 399 2590 155 115 269 ...
 $ totalProfit              : int  15 241 354 104 498 3289 1107 179 165 479 ...
 $ profitMargin             : num  0.772 10.807 22.363 28.571 19.102 ...
 $ profitPerBasket          : num  3.75 60.25 177 104 166 ...
 $ weekendPercent           : num  0.75 0.75 1 0 0 0.5 0 0.5 1 1 ...
 - attr(*, "sorted")= chr "customer_id"
 - attr(*, ".internal.selfref")=<externalptr>
```

Рис. 1 – Характеристики, полученные на этапе «feature creation»

Метод Z-оценки позволил удалить сильные выбросы в исходных данных. Эта оценка рассчитывается следующим образом:

$$z = \frac{x - \mu}{SE} = \frac{x - \mu}{\sigma/\sqrt{n}}$$

где x – значение случайной величины, μ – математическое ожидание, SE – стандартная ошибка, σ – среднеквадратичное отклонение генеральной совокупности, n – объём выборки [2]. При выборе максимального стандартного отклонения $\delta_{max} = 6$ было потеряно всего 3.9% исходных данных - около 30 тысяч записей.

Для уменьшения размерности пространства признаков нами был применён метод главных компонент (PCA)[3], который с помощью линейного преобразования, задаваемого матрицей $\mathbf{A}$, представляет исходные данные (матрица $\mathbf{X}$) в виде нового набора $\mathbf{Z} = (Z_1, Z_2, ..., Zp)$. Столбцы $Z_1, Z_2, ..., Zp$ называются главными компонентами, причем можно выделить $m \ll p$ первых главных компонент, которые обеспечивают требуемую долю дисперсии γ. В результате применения метода было получено 8 первых главных компонент, которые вносят 93.4% дисперсии в исходные данные, что можно считать хорошим результатом. Таким образом, размерность данных была снижена с 29 до 8.

III. Проектирование и обучение карты Кохонена

В качестве архитектуры карты Кохонена была выбрана сетка размером 20 на 20 нейронов с шестиугольной формой ячеек. В результате обучения были получены весовые векторы всех 20x20=400 нейронов карты. Далее мы выполнили иерархическую кластеризацию полученных весовых векторов нейронов, а предполагаемое количество кластеров выбрали равным 6. Существенным достоинством такого подхода является то, что при изменении количества кластеров не требуется повторение этапов проектирования и обучения карты Кохонена. Для этого нужно только выбрать определённый уровень на дендрограмме, полученной на этапе иерархической кластеризации.

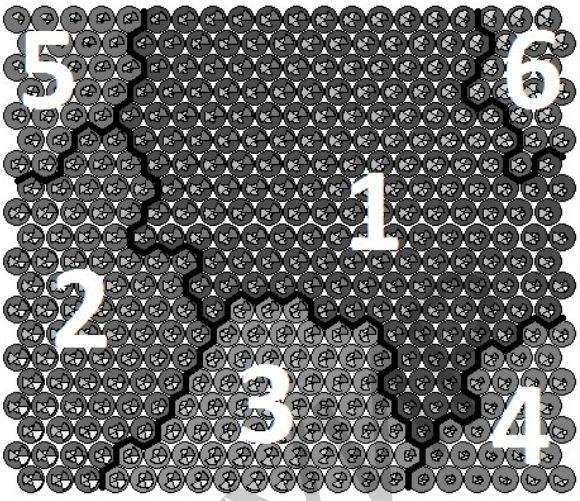

Рис. 2 – Результат кластеризации 32 тысяч клиентов из набора данных «Ta-Feng»

Приведём пример использования полученной информации маркетинговыми стратегами предприятия. В кластер 6 (см. рис. 2) попали преимущественно VIP-покупатели, которые совершают мало очень дорогих покупок, причем выручка компании от этого кластера минимальна. В свою очередь, большой доход приносят покупатели из кластера 5, который составляют преимущественно пожилые люди. Рост заинтересованности в товарах верхней ценовой категории для таких людей наблюдается по выходным дням. Следовательно, бизнесу стоит задуматься о расширении маркетинговой деятельности среди пожилых людей и распродажах по выходным дням.

Заключение

Описанный подход показал хорошие результаты на реальных данных. Методы Z-оценки и главных компонент делают возможным автоматизировать процесс подготовки информации и выбора значимых характеристик. Иерархический кластерный анализ в сочетании с картой Кохонена позволяют изменять число кластеров без необходимости повторять этапы проектирования и обучения модели, что хорошо сказывается на общем времени решения задачи интеллектуального анализа данных. На практике этот подход будет применяться при разработке модуля анализа данных для платформы Salesforce с возможностью его использования нетехническими специалистами.

1. Kohonen, T. Self-Organizing Maps (Third Extended Edition) / T. Kohonen // New York – 2001.-502 p.
2. Kaufman, L. Finding Groups in Data: An Introduction to Cluster Analysis / L. Kaufman, P. J. Rousseeuw // John Wiley Sons, Inc – 2005.-342 p.
3. Яцков, Н. Н. Интеллектуальныи анализ данных: метод. указания к лабораторным работам / Н. Н. Яцков, И. П. Шингарев // Минск:БГУ. – 2012.-51 с.