УДК 519.248:[33+301+159.9]

ЗАДАЧА КЛАСТЕРИЗАЦИИ РАЗНОТИПНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Моисейченко Е. С.,
научный руководитель канд. физ.-мат. наук Баранова И. В.
Сибирский федеральный университет

1. Введение

В работе приводится решение задачи кластеризации марок растворимого кофе на основе предпочтений потребителей. Задача решается с помощью двух методов: самоорганизующихся карт Кохонена (*SOM – Self-Organizing Map*) и разработанного аналога алгоритма самоорганизующихся карт, который использует аппарат эвентологии и метод двудольных множеств событий.

Определение 1: Кластерный анализ – обобщенное название целого ряда методов, используемых для разбиения объектов, событий или индивидов на однородные в соответствующем понимании группы (называемые кластерами), так чтобы каждый кластер состоял из схожих объектов, а объекты из разных классов существенно отличались.

Большое достоинство кластерного анализа состоит в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-теоретических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы.

Основной недостаток методов анализа многомерных данных заключается в том, что при понижении размерности данных происходит потеря информации. Кроме того, общие методы кластеризации не способствуют визуализации данных. Поэтому в настоящее время широкое развитие получил метод нейронных сетей.

Определение 2: Нейронные сети – самообучающиеся системы, имитирующие деятельность человеческого мозга.

Определение 3: Нейронные сети, обучаемые без учителя, служат средством для решения задачи классификации, организации и визуального представления больших объемов данных.

Одним из классов нейронных сетей являются самоорганизующиеся карты Кохонена (*SOM – Self-Organizing Map*). Особенностью *SOM* является то, что метод не требует никаких априорных предположений о распределении данных. Алгоритм *SOM* основывается на соревновательном обучении без учителя. Он обеспечивает сохраняющее топологию отображение из пространства большей размерности в элементы карты. Элементы карты, или нейроны, обычно образуют двумерную решетку. Таким образом, это отображение является отображением пространства большей размерности на плоскость. Свойство сохранения топологии означает, что *SOM* распределяет сходные векторы входных данных по нейронам, т.е. точки, расположенные в пространстве входов близко друг к другу, отображаются на близко расположенные элементы *SOM*. Таким образом, *SOM* может служить как средством кластеризации, так и средством визуального представления данных большой размерности.

Каждый нейрон в алгоритме *SOM* представляет собой n-мерный вектор-столбец весовых коэффициентов $m_i = (\mu_1, \mu_2, \dots, \mu_n)$, n определяется размерностью исходного пространства. Нейроны образуют сеть прямоугольного, гексагонального или

регулярного вида. Величина взаимодействия нейронов определяется расстоянием между нейронами на карте.

Алгоритм обучения:

1. Инициализация весовых коэффициентов нейронов малыми случайными величинами;
2. На все входы сети подается один случайный вектор данных;
3. Вычисляется расстояние от данного вектора до каждого нейрона. В качестве функции расстояния обычно используется евклидово расстояние $d_j = \sqrt{\sum_{i=1}^{n}\left(x_i(t) - m_{ij}(t)\right)^2}$;
4. Выбор нейрона j^* с наименьшим расстоянием d_j;
5. Для модификации весовых коэффициентов используется формула $m_i(t+1) = m_i(t) + \alpha(t)h_{j^*}(t)\big(x_i(t) - m_i(t)\big)$. Функция $h_{j^*}(t)$ называется функцией соседства нейронов, обычно применяется функция Гаусса $h_{j^*}(t) = e^{\frac{-\|r_{j^*} - r_i\|^2}{2\sigma^2(t)}}$ или, более простая, $h_{j^*}(t) = 1$, если $j^* \in \sigma$, $h_{j^*}(t) = 0$, иначе. $\sigma(t)$ – окрестность победившего нейрона, ее значение уменьшается со временем. Функция $\alpha(t)$ называется функцией скорости обучения, она также убывает со временем.

2. Описание статистики

Решение задачи основывается на на реальной статистике, полученной в результате опроса потребителей растворимого кофе из 8 регионов РФ. Заказчика исследования интересовали предпочтения как продавцов кофе (розничных и мелкооптовых), так и непосредственно потребителей. Опрос проводился в форме интервью с 1400 покупателями кофе. Анкета состояла из 10 основных вопросов и 3 дополнительных, посвященных социально-демографической информации о покупателе. Первые 6 основных вопросов позволяли уточнить частоту потребления кофе покупателем, частоту покупок кофе, места приобретения продукта, список марок растворимого кофе, с которым знаком потребитель, а также основные характеристики продукта, которые являются наиболее ценными для потребителя.

Следующие 4 показателя позволяли оценить характеристики каждой марки кофе: вкус кофе, аромат кофе, консистенция кофе, привлекательность марки кофе.

Именно эти показатели использовались для решения задачи кластеризации. Вкус, аромат и консистенция каждой марки кофе оценивались каждым потребителем по шкале от 1 до 10 (наихудшее и наилучшее значения, соответственно). Последний показатель является множественным. Привлекательность марки оценивалась с помощью следующих параметров: известность бренда/производителя, привлекательная цена, положительный опыт употребления данной марки продукта, удобный вес упаковки, привлекательный (оригинальный) дизайн упаковки. Каждому покупателю предлагалось указать параметры, влияющие на привлекательность данной марки (предлагалось выбрать любое множество вариантов из 5 предложенных).

3. Метод двудольных множеств случайных событий

Как видно из описания статистики каждая марка кофе характеризуется разнотипными данными – числовыми и множественными. Метод самоорганизующихся карт не умеет работать с нечисловыми данными, поэтому предлагается разработанный аналог алгоритма Кохонена, использующий аппарат эвентологии и метод двудольных множеств событий.

Основная идея метода двудольных множеств случайных событий заключается в представлении любой сложной системы с помощью двудольной эвентологической

модели, в которой каждый элемент системы характеризуется двудольным множеством событий: его первая доля определяется случайными величинами, а вторая - случайными множествами событий. Затем анализ поведения элементов системы сводится к анализу эвентологических распределений соответствующих им двудольных множеств событий.

Определение 4: Двудольное множество случайных событий представляет собой объединение двух множеств - множества событий, которое определяется случайными величинами, и множества событий, которое определяется случайными множествами событий:

$$\{Y, \mathfrak{X}\} = \{Y_a, \mathfrak{X}_a, a \in A, \beta \in B\}$$

Определение 5: Двудольной эвентологической моделью сложной системы будем называть такую систему, для которой поведение каждого элемента характеризуется двудольным множеством случайных событий $\{Y, \mathfrak{X}\}$, его первая доля Y, определяется случайными величинами $\boldsymbol{\xi}$, а вторая доля $\mathfrak{X}$ – случайными множествами событий $\mathbf{K}$.

Тогда каждый вектор данных будет характеризоваться двудольным множеством случайных событий $s^i = \left(p_a^i, p_\beta^i\right) = \left\{Y_a^i, \mathfrak{X}_\beta^i, a \in A, \beta \in B\right\}$, где $\mathbf{p_a^i}$ – вектор, составленный из значений вероятностей, который соответствует числовой доле Y_a^i, $\mathbf{p_\beta^i}$ – вектор вероятностей, который соответствует множественной доле $\mathfrak{X}_\beta^i$. Каждый нейрон также представим в виде двудольного множества случайных событий $m^i = \left(\mu_a^i, \mu_\beta^i\right) = \left\{Y_a^i, \mathfrak{X}_\beta^i, a \in A, \beta \in B\right\}$, где μ_a^i – весовые коэффициенты, соответствующие числовой доле Y_a^i, μ_β^i – весовые коэффициенты, соответствующие множественной доле $\mathfrak{X}_\beta^i$.

Расстояние между векторами данных и каждым нейроном будет вычисляться по формуле:

$$\mathbf{P}\left(s^i(\Delta)m^i\right) = \frac{1}{|A|}\sum_{a\in A}\frac{1}{|Y_a|}\sum_{r_a\in R_a}^{N_a}\mathbf{P}\left(Y_a^i(r_a)(\Delta)Y_a^i(r_a)\right) + \frac{1}{|B|}\sum_{\beta\in B}\frac{1}{|\mathfrak{X}_\beta|}\sum_{X_\beta\subseteq\mathfrak{X}_\beta}\mathbf{P}\left(X_\beta^i(\Delta)X_\beta^i\right).$$

4. Результаты

В результате кластеризации марок кофе по числовым показателям с помощью классического алгоритма самоорганизующихся карт, получили следующие результаты, представленные в таблице 1 и на рис. 1.

№ класса	Наименование марок
1 класс	Nescafe Gold, Nescafe Classic, Tchibo Exclusive, Jacobs Monarch, Carte Noire, Черная карта
2 класс	Московская кофейня на паях, Maxwell House, Tchibo Mild
3 класс	Ambassador, Tchibo Mocca, Millagro Aroma, Pele Royal, Elite Platinum
4 класс	La Café, Черный консул, Русский продукт, Ruscafe
5 класс	Moccona Continental Gold, Elgresso, Moccona Excellent, Café Pele

Таблица 1. Результат кластеризации 22 марок кофе по числовым показателям.

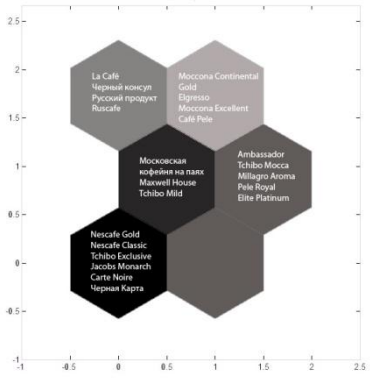

Рис. 1. Самоорганизующаяся карта для 22 марок кофе

На рис. 1 нейроны, окрашенные темным оттенком серого, соответствуют наибольшим значениям, светлым оттенком серого – наименьшим.

В результате кластеризации марок кофе по числовым и множественному показателям, получили результаты, представленные в таблице 2 и на рис. 2.

№ класса	Наименование марок
1 класс	Nescafe Gold, Nescafe Classic, Tchibo Exclusive, Jacobs Monarch, Carte Noire, Черная карта
2 класс	Московская кофейня на паях, Maxwell House, Ambassador, Tchibo Mocca, Tchibo Mild, Millagro Aroma
3 класс	Pele Royal, Moccona Continental Gold, Moccona Excellent, Elite Platinum, Черный консул
4 класс	La Café, Русский продукт, Ruscafe
5 класс	Elgresso, Café Pele

Таблица 2. Результат кластеризации 22 марок кофе по числовым и множественному показателю.

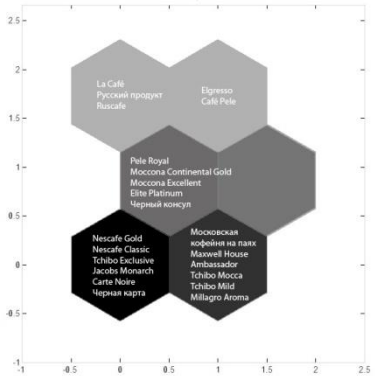

Рис. 2. Самоорганизующаяся карта для 22 марок кофе.

Сравнивая результаты, полученные классическим методом самоорганизующихся карт и разработанным аналогом, можно заметить, что первый класс, соответствующий наибольшим значениям показателей и в том и другом случае состоит из тех же марок кофе. Однако в других классах есть различия. Например, марки Tchibo Мосса и Tchibo Mild в первом случае принадлежат разным классам, а в другом случае принадлежат одному и тому же классу. Так как эти марки принадлежат одному бренду Tchibo, то на результат повлияли значения множественного показателя как, например, известность бренда. Марки Moccona Continental Gold и Moccona Excellent в случае использования классического алгоритма Кохонена принадлежат 5 классу с наименьшими значениями показателей, в случае использования аналога алгоритма принадлежат третьему классу.