

УДК 519.862.5

АНАЛИЗ МАРКЕТИНГОВЫХ КОММУНИКАЦИЙ С ПОМОЩЬЮ НЕЛИНЕЙНЫХ МЕТОДОВ ОТОБРАЖЕНИЯ ДАННЫХ

Михалькевич И.С.

Финансовый университет при Правительстве Российской Федерации (Финансовый университет),
Москва, e-mail: iliyesku@gmail.com

Для эффективного поиска закономерностей в данных, отбора целевых признаков, формального обоснования целесообразности их имплементации в регрессионную модель, интерпретации признаков для поддержки принятия решений применяются различные методы отображения данных. В данной статье приведено исследование отклика на маркетинговые коммуникации с использованием различных методологий: простые статистические методы, самоорганизующиеся карты Кохонена, а также упругие карты. Произведено качественное сравнение методологий путём оценки полученных кластеров. В работе показано, как нелинейные методы отображения данных позволяют оценить количество кластеров и выбрать подходящее разбиение. В ходе анализа выявлены основные закономерности и ключевые характеристики, непосредственно влияющие на отклик в маркетинговых коммуникациях. Результаты исследования могут использоваться при построении регрессионной модели отклика на маркетинговые коммуникации, а также при принятии управленческих решений.

Ключевые слова: отображение данных, визуализация данных, маркетинговые коммуникации

ANALYSIS OF MARKETING COMMUNICATIONS WITH NON-LINEAR DATA VISUALIZATION METHODS

Mikhailkevich I.S.

Financial University under the Government of the Russian Federation (Financial University),
Moscow, e-mail: iliyesku@gmail.com

There are different data mapping methods useful in different cases such as regularity detection in the data, target predictors selection, formal rationale of its implementation in regression model, and interpretation of characteristics for decision-making support. This article presents a study of marketing communications outcome using variety of methodologies: simple statistical methods, kohonen's self-organizing maps and the elastic maps. Author produce qualitative comparison of methodologies by cluster evaluation. The paper shows how non-linear data mapping techniques allow us to estimate the number of clusters and select the appropriate partitioning. The analysis revealed basic laws and key characteristics, which directly affect the marketing communications outcome. The results can be used to predict marketing communications outcome, as well as to support of decision-making.

Keywords: data mapping, data visualization, marketing communications

Простейшие методы анализа маркетинговых коммуникаций

С точки зрения маркетинговых коммуникаций целевым оптимизируемым показателем является конверсия. Конверсия понимается, как отношение числа участников маркетинговой коммуникации, выполнивших целевое действие, к их общему числу [7]:

$$Conv_i = \frac{TS_i}{TS_i + TF_i},$$

где $Conv_i$ – конверсия в группе, обладающей i -м признаком; TS_i – количество успешных попыток коммуникации по группе; TF_i – количество безуспешных попыток коммуникации по группе.

Отображение значимости признаков с точки зрения конверсии можно представить графически с использованием системы простых показателей.

Количество клиентов, обладающих данным признаком, или их доля:

$$P_i = \frac{V_i}{V},$$

где P_i – доля клиентов, обладающих признаком i ; V_i – количество клиентов, обладающих признаком i ; V – общее количество клиентов в группе.

Распространённость признака в целевой группе – доля целевых клиентов, обладающих данным признаком:

$$S_i = \frac{P_i}{TS},$$

где S_i – доля клиентов, обладающих признаком i , в целевой группе; p_i – количество клиентов в целевой группе, обладающих признаком i ; TS – количество клиентов в целевой группе (успешных попыток коммуникации).

Таким образом, можно вывести коэффициент характерности признака, показывающий, насколько этот признак более или менее характерен для целевого клиента, чем для группы в целом:

$$C_i = \frac{S_i - P_i}{P_i} = \frac{\frac{P_i}{TS} - \frac{V_i}{V}}{\frac{V_i}{V}} = \frac{V \times P_i}{V_i \times TS} - 1,$$

где C_i – коэффициент характерности признака i для целевой группы, показывает ценность обладания информацией о факте присутствия признака i .

По результатам проведения маркетинговой кампании был составлен портрет потенциального лизингополучателя, который был описан при помощи полученных выше метрик (рис. 1).

Помимо основных признаков были выявлены синтетические признаки, состоящие из совокупности условий (рис. 2).

Очевидно, что таких клиентов либо крайне мало, либо не существует вовсе. Однако такой портрет позволяет дать ориентиры в стратегическом планировании и при подготовке очередных маркетинговых кампаний.

Использование нелинейных методов отображения данных в анализе маркетинговых коммуникаций

Целесообразно рассмотреть методы, которые в настоящее время позволяют визуально представить сразу всю структуру многомерного набора данных.

Признак	Условие	S	P	C	
Количество проектов (интерес)	[0 - 17]	88%	97%		-0,09
Дней с последней покупки	[0 - 190]	86%	59%		0,46
Дней между покупками группа 1	[476 - inf)	59%	15%		2,90
Дней между покупками группа 2	[19 - 77]	12%	2%		4,83
Дней между покупками группы 1 и 2	[19 - 77] U [476 - inf)	71%	17%		3,14
Времени с последнего проекта (интереса)	[0 - 128]	82%	64%		0,28
Дней между проектами	[1075 - 2182]	49%	26%		0,85
Покупал более 1 раза	ИСТИНА	100%	51%		0,97
Пришёл от дилера	ИСТИНА	61%	52%		0,19
Пришёл с сайта	ИСТИНА	90%	92%		-0,02
Категория по выручке группа 1	[8,5М - 34М)	29%	14%		1,05
Категория по выручке группа 2	[0 - 0,85М)	22%	40%		-0,44
Категория по выручке группы 1 и 2	[0 - 0,85М) U [8,5М - 34М)	51%	54%		-0,06
Находится в Москве	ЛОЖЬ	100%	90%		0,11
Средний возраст транспортных средств	[0 - 11]	96%	89%		0,08
Работает в разных отраслях	ИСТИНА	96%	96%		0,00

Рис. 1. Анализ основных характеристик клиента

Синтетический признак	Условие	S	P	C	
Дней с последней покупки	[0-146]	45%	30%		0,4966
Пришёл от дилера	ИСТИНА				
Средний возраст транспортных средств	[0-11]	71%	36%		0,97921862
Совершено более двух покупок	ИСТИНА				

Рис. 2. Анализ синтетических характеристик клиента

Таким образом, самая ценная информация о клиенте с точки зрения отклика на маркетинговые коммуникации звучит следующим образом «клиент совершает покупки в среднем раз в 19–77 дней», однако таких клиентов всего 2%.

Информация о том, что выручка клиента составляет менее 850 000 рублей, говорит о том, что, скорее всего, клиент не склонен к покупке, причём данная информация имеется у 40% респондентов, что является относительно хорошим показателем.

Итоговый портрет целевого клиента будет звучать следующим образом «клиент часто совершает покупки, недавно что-либо покупал или интересовался покупкой, пришёл от дилера и имеет выручку от 8,5 до 34 млн руб.».

Под визуализацией данных в таком случае будет пониматься способ отображения многомерного массива данных на двумерной плоскости, где по меньшей мере качественно выражены основные закономерности исходного распределения. Это могут быть такие закономерности, как кластерная структура, топология, различные зависимости между признаками, расположение точек данных в пространстве и пр. [5].

Основные цели применения нелинейных методов визуализации:

- наглядное геометрическое представление данных;
- описание закономерностей в данных;
- сжатие информации;
- восстановление пропущенных значений;

– прогнозирование и построение регрессионных моделей.

Традиционными методами решения поставленных задач являются целенаправленное проецирование и многомерное шкалирование. Методы целенаправленного проецирования сводятся к поиску отображения данных из многомерного пространства на двумерную плоскость, оптимизирующему функционал от координат точек данных [1]. В задачах, решаемых методами многомерного шкалирования, отсутствуют исходные данные о координатах точек данных, известны только расстояния между ними. Задача сводится к поиску таких координат точек, которые будут сохранять матрицу расстояний [2, 4, 8].

Относительно новыми средствами визуализации данных являются самоорганизующиеся карты Кохонена [9], а также упругие карты, разработанные в Институте вычислительного моделирования г. Красноярск [6]. Методы, используемые этими средствами, основаны на поиске оптимальной ориентации вложенных поверхностей в многомерной структуре данных.

В методе самоорганизующихся карт Кохонена точки данных проецируются на сетку узлов, итеративно приближенных к сгущениям точек данных, при достижении минимума ошибки аппроксимации:

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - y_{BMU}(X_i))^2},$$

где $y_{BMU}(X_i)$ – ближайший узел к точке данных X_i .

Узлы сетки инициализируются и перемещаются по направлению к точкам данных по следующему правилу:

$$(y_j)' = y_j + h(r(y_j, y_{BMU}), t)(X_i - y_j), \quad j = 1 \dots p,$$

где $h(x, t)$ – функция соседства узлов; $r(y_1, y_2)$ – расстояние между узлами сетки y_1 и y_2 ; p – количество узлов сетки.

Ниже описаны часто используемые функции соседства:

Гауссова функция

$$h(x, t) = \alpha(t) e^{-\frac{x^2}{2\sigma^2(t)}};$$

Bubble-функция:

$$h(x, t) = \begin{cases} \alpha(t), & x \leq \sigma(t); \\ 0, & x > \sigma(t), \end{cases}$$

где $\alpha(t)$ – так называемый темп обучения; $\sigma(t)$ – радиус захвата соседей [5].

Для анализа данных выполним нормализацию по формуле

$$X_i \rightarrow X'_i \mid x'_{ij} = x_{F(x_{ij})},$$

где n – объем данных, $x_{i1}, \dots, x_{in}$ – набор уникальных значений признака X_i ; $x'_{i1}, \dots, x'_{in}$ – набор уникальных значений признака X'_i ; $F(x_{ij})$ – эмпирическая функция распределения признака X_i :

$$F(x_{ij}) = \overline{P(x < x_{ij})} = P(x < x_{ij}) + \frac{P(x = x_{ij})}{2},$$

где $x_{F(x_{ij})}$ – есть квантиль уровня $F(x_{ij})$ для нормального распределения

$$F(x) = N = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}.$$

Построим карту Кохонена, по очереди отбрасывая неинформативные признаки с точки зрения конверсии (рис. 3).

Самыми информативными оказались следующие признаки: «Имеет проект, заведённый по собственной инициативе» (характерно для группы с высокой конверсией), «Дней между передачами» (низкие значения характерны для группы с низкой конверсией), «Интересовался ли разными транспортными средствами» (не характерно для группы с низкой конверсией), «Пришёл от дилера» (не характерно для группы с высокой конверсией), «Имеет упрощённую систему налогообложения» (весьма характерно для группы с низкой конверсией), «Дней прошло с последней покупки» (высокие значения не характерны для группы с низкой конверсией).

Конкретные диапазоны, соответствующие нормированным значениям, можно вывести обратным преобразованием признаков.

Метод упругих карт может служить обобщением метода главных компонент. Задача построения вложенного многообразия является оптимизационной и состоит в поиске такой нелинейной (упругой) поверхности, дисперсия проекций точек данных на которую будет минимально искажена по сравнению с дисперсией исходных точек данных. Изначально упругая сетка может располагаться в плоскости первых двух главных компонент.

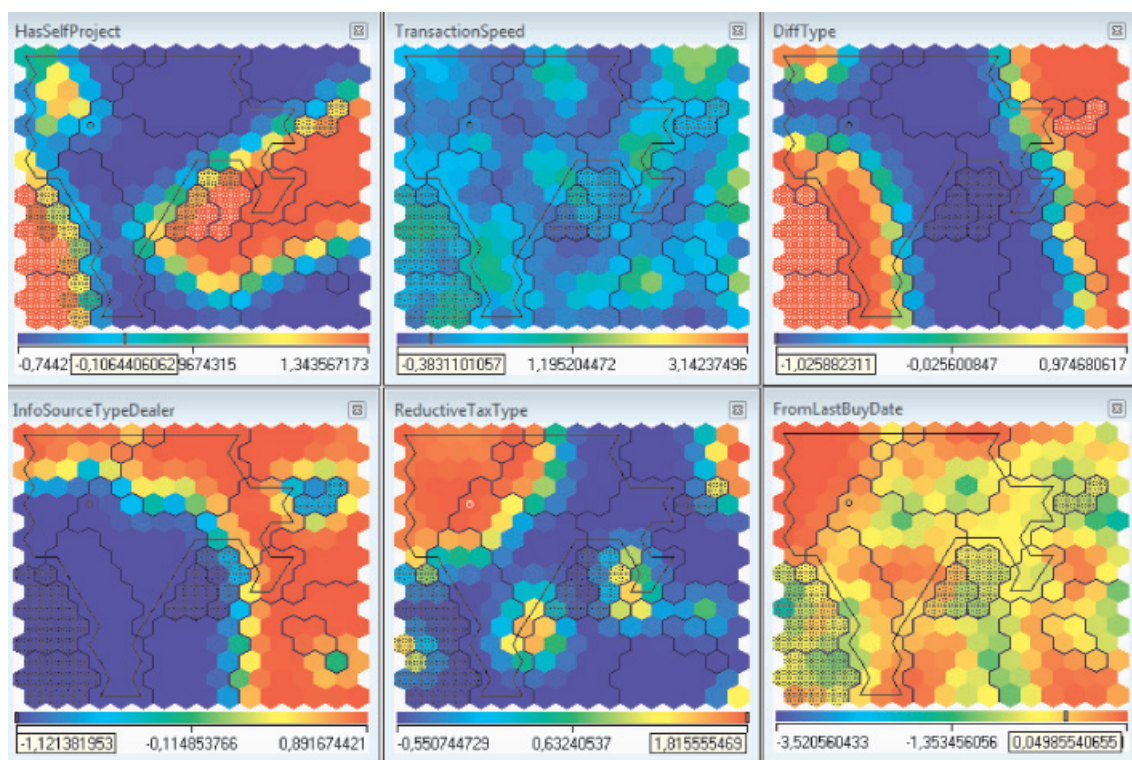


Рис. 3. Карта Кохонена: Закрашенная область – кластеры с высокой конверсией (2, 3, 4), обведённая контуром область – кластеры с низкой конверсией (0, 1, 6, 7, 9)

Номер кластера	Y=1	Количество	Конверсия
3	48	408	11,8%
2	87	812	10,7%
4	16	180	8,9%
5	34	803	4,2%
10	78	1912	4,1%
8	61	1513	4,0%
6	21	704	3,0%
9	30	1075	2,8%
0	21	870	2,4%
7	31	1363	2,3%
1	24	1183	2,0%

Рис. 4. Статистика по кластерам

Группа кластеров	%Количество	%(Y=1)	Средняя конверсия	Превышение средней конверсии
С высокой конверсией	13%	33,5%	10,5%	105%
С низкой конверсией	48%	28,2%	2,5%	-51%

Рис. 5. Статистика по группам кластеров

Для метода главных компонент задача формально ставится следующим образом
Старые признаки должны как можно точнее линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^m g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X$$

как можно точнее на обучающей выборке $x_1, \dots, x_l$:

$$\sum_{i=1}^l \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 \rightarrow \min_{\{g_s(x_i)\}_{s,j}},$$

где $f_1(x), \dots, f_n(x)$ – исходные числовые признаки; $g_1(x), \dots, g_m(x)$ – новые числовые признаки, $m \leq n$.

Исходная матрица «объекты-признаки»

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix}.$$

Новая матрица «объекты-признаки»

$$G = \begin{pmatrix} g_1(x_1) & \cdots & g_m(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_l) & \cdots & g_m(x_l) \end{pmatrix}.$$

Матрица линейного преобразования признаков:

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nm} \end{pmatrix}.$$

Матричная запись линейного преобразования

$$\hat{F} = GU^T \approx F.$$

Матричная запись оптимизируемого функционала

$$\sum_{i=1}^l \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\| \rightarrow \min_{GU}.$$

3. Упругость по отношению к изгибу:

$$D_3 = \sum_{i=1}^p \sum_{j=2}^{q-1} \|2y^{ij} - y^{i,j-1} - y^{i,j+1}\| + \sum_{i=2}^{p-1} \sum_{j=1}^q \|2y^{ij} - y^{i-1,j} - y^{i+1,j}\|.$$

Таким образом:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{pq} + \mu \frac{D_3}{pq} \rightarrow \min,$$

где $|X|$ – число точек данных; λ, μ – коэффициенты «упругости» сетки.

Построим двумерную упругую карту с параметрами $\lambda = 0,01$, $\mu = 2$ и развернём на плоскости (рис. 6).

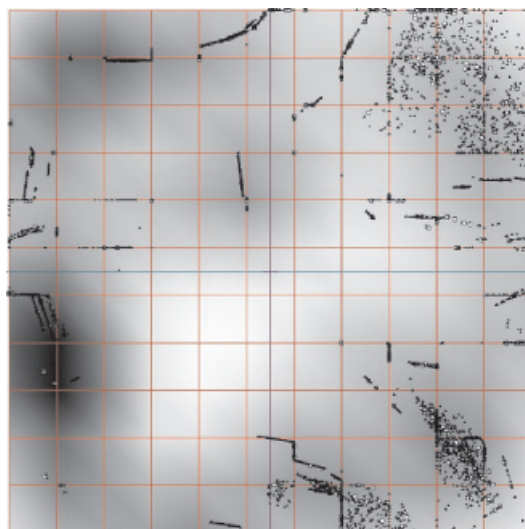


Рис. 6. Упругая карта. Распределение плотности данных. Белым отмечены точки данных, соответствующие успешным коммуникациям

Решением задачи метода главных компонент является нахождение матрицы U . Столбцы U – это собственные векторы матрицы $F^T F$, соответствующие максимальным собственным значениям $\lambda_1, \dots, \lambda_m$ [3].

$$G = FU.$$

Упругая сетка имеет узлы, нумеруемые индексами y^{ij} , $i = 1 \dots p$, $j = 1 \dots q$. Каждый узел y^{ij} ставится в соответствие подмножеству точек данных K_{ij} ($i = 1 \dots p$, $j = 1 \dots q$) так, что этот узел для каждой точки из этого подмножества является ближайшим:

$$K_{ij} = \{x \in X | \|y^{ij} - x\| \rightarrow \min_{i,j}\}.$$

Конечное положение упругой сетки достигается оптимизацией линейной комбинации функционалов, отражающих следующий набор свойств:

1. Близость к точкам данных:

$$D_1 = \sum_{ij} \sum_{X_k \in K_{ij}} \|X_k - y^{ij}\|.$$

2. Упругость по отношению к растяжению:

$$D_2 = \sum_{i=1}^p \sum_{j=1}^{q-1} \|y^{ij} - y^{i,j+1}\| + \sum_{i=1}^{p-1} \sum_{j=1}^q \|y^{ij} - y^{i+1,j}\|.$$

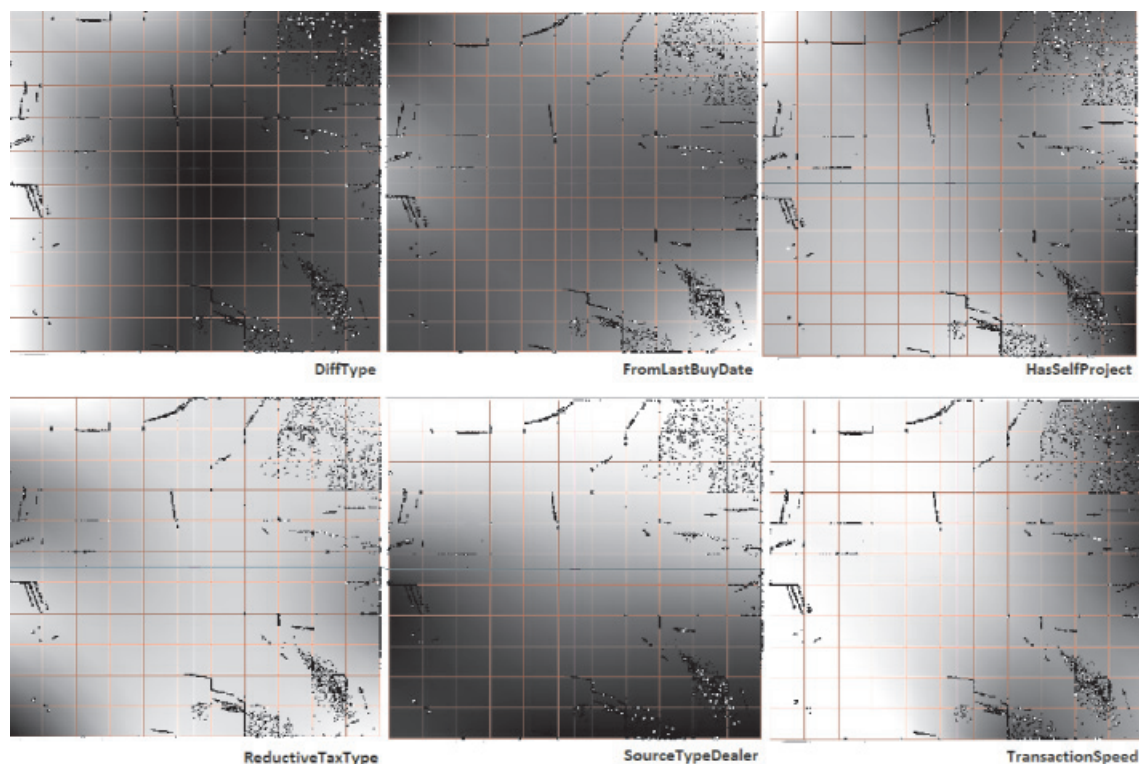


Рис. 7. Упругая карта. Дано 6 раскрасок в разрезе информативных признаков:

- 1 – «Интересовался разными ТС»; 2 – «Прошло дней с последней покупки»; 3 – «Имеет самостоятельно заведённый проект»; 4 – «Имеет упрощённую систему налогообложения»; 5 – «Пришёл от дилера»; 6 – «Частота покупок»

Отразим данные на карте в разрезе информативных признаков, аналогично карте Кохонена (рис. 7).

Выполним кластеризацию методом k-means с пятью кластерами (рис. 8) и составим статистику (рис. 9).

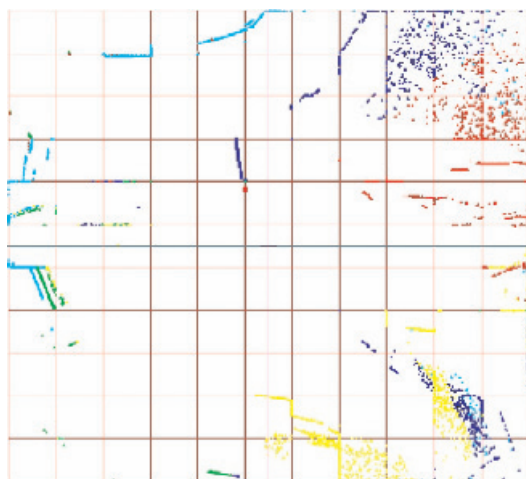


Рис. 8. Кластеризация после построения упругой карты

Номер кластера	Avg Y	Count %	DiffType	FromLastBuyDate	HasSelfProject	ReductiveTaxType	InfoSourceTypeDealer	TransactionSpeed
1	8,26%	16,33%	0,974680617	-0,830773756	0,918211849	-0,329782911	-0,280614775	-0,176794014
2	5,52%	15,39%	-1,025882311	-0,701516086	0,483890702	-0,111857117	-0,553472892	-0,475491992
3	0,89%	39,31%	-0,446970002	0,602897661	-0,581769772	0,23913451	0,271289629	-0,672283852
4	5,54%	17,35%	0,611426113	-0,081218502	0,055098779	-0,08580076	0,891674421	1,294086541
5	5,64%	11,62%	0,588244376	0,178461624	-0,045523959	-0,069208282	-1,121381953	1,219527554

Рис. 9. Статистика кластеров после применения упругой карты

Кластером с наиболее высокой конверсией (8,26 %) оказался кластер № 1, содержащий 16,33 % испытаний, для которого характерны признаки: «Скорее интересуется разными ТС», «Имеет недавнюю покупку», «Имеет проект, заведённый по собственной инициативе».

Кластером с наиболее низкой конверсией (0,89 %) оказался кластер № 3, содержащий 39,31 % испытаний, для которого характерны признаки: «Скорее не интересовался разными ТС», «Не имеет недавнюю покупку», «Не имеет проект, заведённый по собственной инициативе», «Имеет низкую частоту покупок»

Заключение

Применение упругих карт позволило сделать лучшую кластеризацию, по сравнению с картами Кохонена: гораздо лучше были классифицированы клиенты с очень низкой конверсией и сравнительно неплохо были классифицированы клиенты с высокой конверсией. Были получены наглядные результаты в виде плотности распределения данных, распределения значимых признаков, а также распределения кластеров. Полученные данные могут быть применены при построении регрессионных решений и планировании маркетинговых коммуникаций.

Список литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности: справочное издание / под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Статистическое оценивание зависимостей. – М.: Финансы и статистика, 1985.

3. Воронцов К.В., лекции курса «Машинное обучение», yandexdataschool.ru, 2016.

4. Дейвсон М. Многомерное шкалирование: Методы наглядного представления данных. – М.: Финансы и статистика, 1988.

5. Зиновьев А.Ю. Визуализация многомерных данных. – Красноярск: Изд-во КГТУ, 2000.

6. Зиновьев А.Ю. Метод упругих карт для визуализации данных: алгоритмы, программное обеспечение и приложения в биоинформатике. – Красноярск: ИВМ СО РАН, 2001.

7. Котлер Ф. Маркетинг менеджмент / пер. с англ. под ред. О.А. Третьяк, Л.А. Волковой, Ю.Н. Кантуревского. – СПб.: Питер, 1999.

8. Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука, 1986.

9. Kohonen T. Self-Organizing Maps, Berlin. – New York: Springer-Verlag, 2001.

References

1. Ajvazjan S.A., Buhstaber V.M., Enjukov I.S., Meshalkin L.D. Prikladnaja statistika. Klassifikacija i snizhenie razmernosti. Spravochnoe izdanie. Pod red. S.A. Ajvazjana. M.: Finansy i statistika, 1989.
2. Ajvazjan S.A., Enjukov I.S., Meshalkin L.D. Prikladnaja statistika. Statisticheskoe ocenivanie zavisimostej. M.: Finansy i statistika, 1985.
3. Voroncov K.V., lekcii kursa «Mashinnoe obuchenie», yandexdataschool.ru, 2016.
4. Dejvson M. Mnogomernoe shkalirovanie: Metody nagljadnogo predstavlenija dannyh. M.: Finansy i statistika, 1988.
5. Zinov'ev A.Ju. Vizualizacija mnogomernyh dannyh. Krasnojarsk: Izd-vo KGTU, 2000.
6. Zinov'ev A.Ju., Metod uprugih kart dlja vizualizacii dannyh: algoritmy, programmnoe obespechenie i prilozhenija v bioinformatike. Krasnojarsk: IVM SO RAN, 2001.
7. Kotler F. Marketing menedzhment / Per. s angl. pod red. O.A. Tretjak, L.A. Volkovoj, Ju.N. Kanturevskogo. SPb.: Piter, 1999.
8. Terehina A.Ju. Analiz dannyh metodami mnogomernogo shkalirovanija. M.: Nauka, 1986.
9. Kohonen T. Self-Organizing Maps, Berlin. New York: Springer-Verlag, 2001.