

Применение методов нечеткого кластерного анализа к задаче сегментации потребителей рынка изоляционных материалов

А.А. Акишев
alexander.akishev@urfu.me

М.Г. Близоруков
mikhail.blizorukov@urfu.ru

Уральский федеральный университет (Екатеринбург)

Аннотация

Работа посвящена применению методов нечеткого кластерного анализа к решению задачи сегментации потребителей изоляционных материалов города Екатеринбурга с построением профилей этих сегментов. Для проведения сегментации используется алгоритм нечеткой кластеризации методом c -средних. В работе была проведена сегментация групп «любители» и «профессионалы». В результате были выделены подгруппы покупателей со схожими характеристиками, оценен размер этих подгрупп и дано их описание.

Ключевые слова: сегментация рынка; кластерный анализ; нечеткая логика; метод нечетких c -средних.

1 Введение

Сегментация рынка — разделение рынка на отдельные подгруппы покупателей, каждая со своими особыми характеристиками спроса. Благодаря этому разделению фирмы имеют возможность проводить стратегию дифференциации продукта, чтобы удовлетворять потребности различных групп покупателей, увеличивая возможности сбыта своей продукции.

Кластерный анализ — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Кластерный анализ — это многомерная статистическая процедура обработки информации о выборке объектов, и упорядочивающая объекты в сравнительно однородные группы.

Кластерный анализ позволяет рассматривать большой (хотя и имеющий ограничения) объем информации и сокращать массивы информации, делать их компактными и наглядными. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество практически произвольно взятых исходных данных [1].

Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: G.A. Timofeeva, A.V. Martynenko (eds.): Proceedings of 3rd Russian Conference "Mathematical Modeling and Information Technologies" (MMIT 2016), Yekaterinburg, Russia, 16-Nov-2016, published at <http://ceur-ws.org>

Методы кластеризации можно разделить на четкие и нечеткие. В нечетких методах кластеризации используются идеи нечеткой логики, кластер рассматривается как нечеткое множество. Результат кластеризации M элементов в c кластерах удобно рассматривать, используя характеристическую функцию $U = [\phi_{ki}], \phi_{ki} \in \{0, 1\}, k \in \overline{1, M}, i \in \overline{1, c}$, где k -ая строчка матрицы U указывает на принадлежность k -го объекта i -му кластеру (0 - не принадлежит, 1 - принадлежит). Функция принадлежности элемента к нечеткому множеству может принимать любые значения из отрезка $[0, 1]$, а не только 0 или 1.

Впервые понятие нечеткого множества ввел Лотфи Заде в 1965 году, опубликовав работу по теории нечётких множеств, в которой изложил математический аппарат теории нечетких множеств. Нечеткие методы позволяют одному и тому же объекту одновременно принадлежать нескольким, или даже всем кластерам, но с разными степенями принадлежности. Во многих случаях нечеткая кластеризация более естественная и наглядная, чем четкая.

2 Постановка задачи

Цель данной работы - решение задачи сегментации потребителей изоляционных материалов на примере города Екатеринбурга с дальнейшим построением профилей этих сегментов.

Исходными данными являются реальные данные, полученные с рынка города Екатеринбурга некоторой компанией, в которую занесены ответы на анкету респондентов, занимающихся ремонтными работами с использованием различных изоляционных материалов. Все респонденты поделены на две группы в зависимости от того, связана ли их профессиональная деятельность со строительством и ремонтом: «профессионалы» и «любители». Для каждой группы будет проведена своя сегментация. Анкета содержала в себе вопросы касательно образования респондентов, их предпочтений в выборе фирм-производителей изоляционных материалов, семейного положения, доходов, рода занятий, и демографический блок вопросов.

3 Реализация

Были произведены центрирование и нормирование данных, а также удаление данных с некорректными значениями.

Для проведения сегментации используется один из наиболее известных методов нечеткой кластеризации – алгоритм нечетких c -средних. В ходе данного алгоритма решается задача минимизации критерия

$$\sum_{i=\overline{1, c}} \sum_{k=\overline{1, M}} (\mu_{ki})^m \|V_i - X_k\|^2, \quad (1)$$

где V_i – центры кластеров, X_k - элементы кластеризации, $m \in (1, \infty)$ – экспоненциальный вес, μ_{ki} - степень принадлежности k -го элемента i -му кластеру.

Параметрами алгоритма нечетких c -средних являются: c – число кластеров, m – экспоненциальный вес $m \in (1, \infty)$. Экспоненциальный вес m влияет на матрицу степеней принадлежности. Чем больше m , тем конечная матрица c -разбиения становится более “размазанной”, и при $m \rightarrow \infty$ она примет вид $M = 1/c$, что является очень плохим решением, т. к. все объекты принадлежат ко всем кластерам с одной и той же степенью [2].

Каждый кластер является нечетким множеством. Множеством альфа-уровня нечеткого множества называют множество, состоящее из элементов, для которых степень принадлежности больше значения альфа уровня $\alpha \in (0, 1)$ [3]. Будем считать что элемент принадлежит сегменту, если он принадлежит множеству альфа-уровня.

При $m \rightarrow 1$ результаты кластеризации приближаются к результатам четкой кластеризации, каждый элемент имеет степень принадлежности близкую к единице к какому-то одному кластеру, и почти нулевую к остальным.

Решение задачи кластеризации принципиально неоднозначно, и тому есть несколько причин: не существует однозначно наилучшего критерия качества кластеризации. Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом. При нечеткой кластеризации также необходимо определить экспоненциальный вес m .

Для предварительной обработки данных использовался Microsoft Excel 2016. Кластеризация была проведена в пакете Fuzzy Logic Toolbox вычислительной системы MATLAB R2016a. Анализ результатов кластеризации проведен в системе вычислений Wolfram Mathematica 9.

4 Результаты

4.1 Сегментация профессионалов

В качестве примера была проведена сегментация группы респондентов профессионально занимающихся строительством и ремонтом по следующим переменным: пол, возраст, род занятий, образование, семейное положение, среднемесячный доход на одного члена семьи, кем работает респондент, длительность профессиональной деятельности в области строительства и ремонта (в частности работ, связанных с изоляцией), связь образования со строительством, архитектурой и дизайном.

Число сегментов было выбрано равным 4, экспоненциальный вес $m = 1.6$. На рисунке 1 приведен график размера сегмента в зависимости от значения α -уровня.

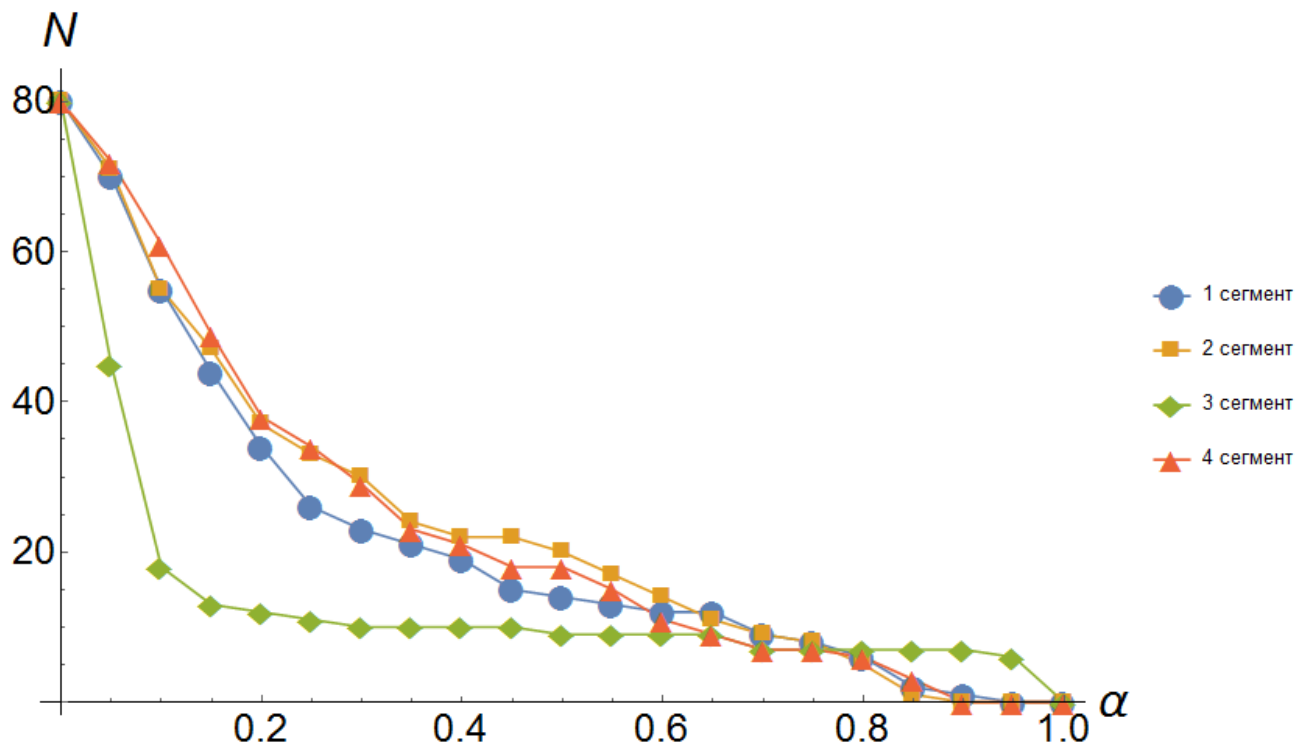


Рис. 1: Зависимость размера сегмента от значения α -уровня

При увеличении значения α –уровня размеры сегментов уменьшаются, при $\alpha = 1$, размеры всех сегментов равны 0, а при $\alpha = 0$ каждый сегмент содержит в себе все элементы.

Для дальнейшего анализа было выбрано значение альфа уровня равное 0.5. Причины:

- при данном уровне α любой элемент гарантированно принадлежит не более чем одному сегменту
- почти все элементы распределены по сегментам, количество элементов не попавших в сегменты 20 из 80 (25 %) . Элементы, не попавшие в сегменты, являются граничными и носят характеристики промежуточные между разными сегментами. Исключение таких элементов из рассмотрения является целесообразным и позволяет более качественно определить характеристики сегментов.
- по графику видно, что в окрестности $\alpha = 0.5$ размеры сегментов меняются незначительно, что говорит об устойчивости данного значения.

На основании кластеризации были получены следующие сегменты:

Таблица 1: Первый сегмент

Пол	Женский
Возраст	30-40 лет
Образование	Высшее
Род занятий	Работа по найму/полная занятость
Должность	Специалист
Доход на члена семьи	20000 - 24999 руб
Связь образования со строительством	Да
Лучшая марка по качеству	Кнауф
Лучшая марка по цене-качеству	Кнауф
Размер сегмента	18 представителей из 80 (около 20%)

Первый сегмент составляет группа людей, типичным представителем является женщина, имеющая образование в области строительства, архитектуры или дизайна. Фактически представители этого сегмента могут решать функции снабженца или консультанта в области отделки и дизайна, осуществляя приобретение изоляционных материалов в интересах заказчика в рамках работы над разработанным проектом.

Таблица 2: Второй сегмент

Пол	Мужской
Возраст	Около 36 лет
Образование	Средне-специальное
Род занятий	Краткосрочные контракты/неполная занятость
Должность	Рабочий
Доход на члена семьи	Около 30 000 руб
Связь образования со строительством	Нет
Лучшая марка по качеству	Порилекс
Лучшая марка по цене-качеству	Кнауф
Размер сегмента	9 представителей из 80 (около 11%)

Сегмент с условным названием «Мастера» описывается профилем человека среднего возраста, который сам осуществляет работы как отдельный специалист, и возможно, как сотрудник строительной компании. В ряде случаев это может быть сотрудник службы снабжения.

Таблица 3: Третий сегмент

Пол	Мужской
Возраст	около 33 лет
Образование	Высшее
Род занятий	Работа по найму
Должность	Рабочий
Доход на члена семьи	20000 - 24999 руб
Связь образования со строительством	Да
Лучшая марка по качеству	Изовер
Лучшая марка по цене-качеству	Изовер
Размер сегмента	20 представителей из 80 (около 25%)

Представители этого сегмента демонстрируют устойчивую в профессиональной среде модель, когда выбор поставщика материала происходит не от покупки к покупке, а один раз на достаточно продолжительный период. Это объясняется обычной привычкой, а также соображениями о стабильности условий производства работ.

Таблица 4: Четвертый сегмент

Пол	Мужской
Возраст	Около 36 лет
Образование	Высшее
Род занятий	Неполная занятость/частный бизнес
Должность	Руководитель
Доход на члена семьи	30 000 руб
Связь образования со строительством	Нет преобладания
Лучшая марка по качеству	Кнауф
Лучшая марка по цене-качеству	Кнауф
Размер сегмента	14 представителей из 80 (около 18%)

Потребители, сгруппированные в этом сегменте, отличаются большим опытом работы, чем представители других групп. Высокая квалификация представителей этой группы позволяет им очень точно подходить к оценке цены товара. Четко понимая особенности материалов, они могут заменить одну марку другой, просто выбрав материал дешевле, при условии, что знакомы с ним («Кнауф» на «Изолон»).

Отличительной чертой всех профессиональных сегментов является модель выбора, основанная на самостоятельном принятии решения, без участия персонала магазина. Для них важно, чтобы товар был в наличии. Профессионалы не склонны к коммуникации в магазине, они часто выбирают одну марку и предпочитают работать с ней постоянно.

4.2 Сегментация любителей

Сегментация группы «любители» была проведена по следующим переменным: пол, возраст, род занятий, образование, семейное положение, среднемесячный доход на одного члена семьи, а также кем работает респондент.

Число сегментов было выбрано равным 3, экспоненциальный вес $m = 1.3$. На рисунке 2 приведен график размера сегмента в зависимости от значения α -уровня.

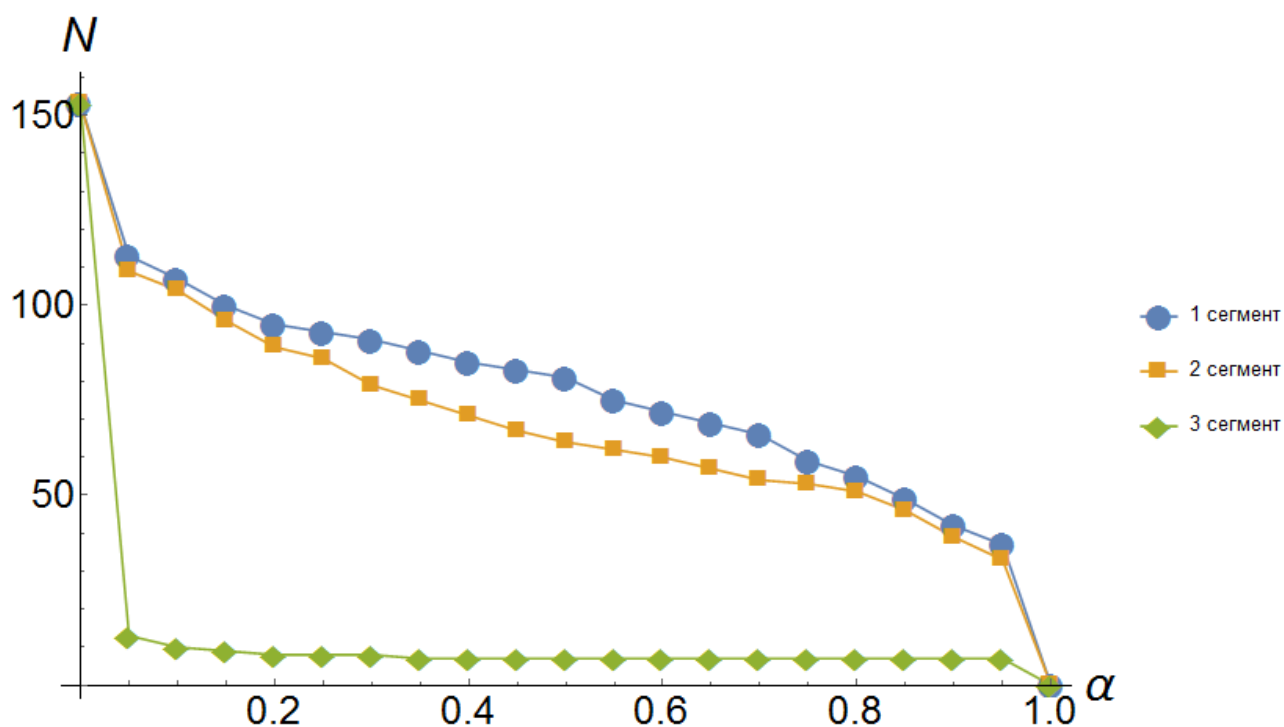


Рис. 2: Зависимость размера сегмента от значения α -уровня

Для дальнейшего анализа было выбрано значение α -уровня равное 0.6. Причины:

- при данном уровне α любой элемент гарантированно принадлежит не более чем одному сегменту.
- почти все элементы распределены по сегментам, количество элементов не попавших в сегменты 14 из 153 (около 10 %). Элементы, не попавшие в сегменты, являются граничными и носят характеристики промежуточные между разными сегментами. Исключение таких элементов из рассмотрения является целесообразным и позволяет более качественно определить характеристики сегментов.
- по графику (рис.2) видно, что в окрестности $\alpha = 0.6$ размеры сегментов меняются незначительно, что свидетельствует об устойчивости данного значения.

На основании кластеризации были получены следующие сегменты:

Таблица 5: Первый сегмент

Пол	Чуть больше представителей мужского пола
Возраст	Около 40 лет
Образование	Высшее
Род занятий	Работа по найму/полная занятость
Должность	Специалист
Доход на члена семьи	Около 30 000 руб
Лучшая марка по качеству	Порилекс
Лучшая марка по цене-качеству	Порилекс
Размер сегмента	72 представителей из 153 (Около 47%)

Таблица 6: Второй сегмент

Пол	Мужской
Возраст	Около 35 лет
Образование	Неоконченное высшее
Род занятий	Частный бизнес
Должность	Руководитель высшего звена
Доход на члена семьи	25 000 - 30 000 руб
Лучшая марка по качеству	Изовер
Лучшая марка по цене-качеству	Изовер
Размер сегмента	7 представителей из 153 (Около 4%)

Таблица 7: Третий сегмент

Пол	Мужской
Возраст	Около 36 лет
Образование	Средне-специальное
Род занятий	Работа по найму/полная занятость
Должность	Специалист
Доход на члена семьи	25 000 - 30 000 руб
Лучшая марка по качеству	Порилекс
Лучшая марка по цене-качеству	Изолон
Размер сегмента	60 представителей из 153 (Около 39%)

При кластеризации респондентов, непрофессионально занимающихся ремонтными работами, получилось три сегмента. Первый и третий сегменты схожи, но имеют различия в образовании и должности респондентов.

Второй сегмент сильно отличается от первого и третьего. Его составляют руководители имеющие свой бизнес. Представители этой группы попадают под влияние марок с агрессивной маркетинговой коммуникацией («Изовер»).

Модель выбора материала для любителей одинакова, в ней чаще всего задействован продавец, а сам выбор происходит в месте продаж. Очевидно, что на текущий момент это точка наиболее эффективной коммуникации с представителями целевой аудитории. В сегмент входят чувствительные к цене потребители.

Список литературы

- [1] В. Duran, P. Odell. *Cluster analysis*. Berlin, Springer-Verlag, 1974.
- [2] S. D. Stovba. *Designing fuzzy systems using MATLAB*. Moscow, Hotline - Telecom, 2007. (in Russian) = С. Д. Штовба. *Проектирование нечетких систем средствами MATLAB*. М., Горячая линия - Телеком, 2007.
- [3] L. K. Konysheva, D. M. Nazarov. *Basics of the theory of fuzzy sets: Tutorial*. Saint Petersburg, Piter, 2011. (in Russian) = Л. К. Коньшева, Д. М. Назаров. *Основы теории нечетких множеств: Учебное пособие*. СПб., Питер, 2011.

Application of fuzzy clustering methods to the market segmentation of insulation materials' consumers.

Alexander A. Akishev

Ural Federal University (Yekaterinburg, Russia)

Mikhail G. Blizorukov

Ural Federal University (Yekaterinburg, Russia)

Abstract. This paper presents the application of the fuzzy clustering methods to the market segmentation of consumers of the insulation materials as an the example of Yekaterinburg and the description of these segments. The segmentation was performed using the fuzzy c -means algorithm. The segmentation was made for the groups called "amateurs" and "professionals". As a result the segments of customers with similar characteristics were found, the size of the subgroups was estimated and the description was given.

Keywords: market segmentation; cluster analysis; fuzzy logic; fuzzy c -means algorithm.