УДК 004.94:336.221

АНАЛИЗ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДЛЯ ЭФФЕКТИВНОГО УПРАВЛЕНИЯ ПРОЦЕССАМИ В НАЛОГОВОЙ СЛУЖБЕ

Заболотникова В.С., Ромашкова О.Н.

ГАОУ ВО города Москвы «Московский городской педагогический университет», Москва, e-mail: zabolotnikovavs@ya.ru

В статье рассматривается вопрос необходимости обработки большого объема информации о налогоплательщиках в условиях наличия неполных или нечетких данных, имеющих не только количественный, но и качественный характер. Для сжатия информации и выделения налогоплательщиков в категории предложено использование методов кластерного анализа. Проведена классификация методов кластеризации Анализ достоинств и недостатков существующих методов кластерного анализа позволил выделить метод нечеткой кластеризации как наиболее эффективный при разбиении налогоплательщиков по категориям. Обоснована необходимость и актуальность применения кластерного анализа и подтверждена целесообразность использования метода нечеткой кластеризации для распределения налогоплательщиков по категориям для эффективного управления процессами в налоговой службе. Предложено использование двух алгоритмов нечеткой кластеризации с целью дальнейшей реализации и сравнения полученных результатов.

Ключевые слова: кластер, кластерный анализ, нечеткая кластеризация, налогоплательщик, налоговая служба

ANALYSIS OF CLUSTERING TECHNIQUES FOR EFFECTIVE PROCESS MANAGEMENT IN THE TAX SERVICE

Zabolotnikova V.S., Romashkova O.N.

The State Educational Government-Financed Institution of Higher Professional Education of the City of Moscow «Moscow City Teacher Training University», Moscow, e-mail: zabolotnikovavs@ya.ru

The article discusses the necessity of processing large amounts of information about taxpayers in the presence of incomplete or fuzzy data, which are not only quantitative but also qualitative. For the compression information and the allocation of taxpayers in the categories suggested the use of methods of cluster analysis. The classification methods, clustering Analysis of the advantages and disadvantages of existing methods of cluster analysis allowed to identify a method of fuzzy clustering as the most effective in the partitioning of taxpayers by category. The necessity and relevance of applying the cluster analysis and confirmed the expediency of using the method of fuzzy clustering for the distribution of taxpayers on categories for effective process management in the tax service. Proposed the use of two algorithms of fuzzy clustering with the aim of further implementation and comparison of the obtained results.

Keywords: cluster, cluster analysis, fuzzy clustering, the taxpayer, the tax office

Стратегическая цель экономической политики государства в целом и налоговой политики в частности заключается в создании в России стабильной налоговой системы, которая бы обеспечила достаточный объем налоговых поступлений в бюджеты всех уровней за счет формирования эффективных механизмов налогообложения всех категорий налогоплательщиков, а также информационного обеспечения процесса принятия управленческих решений в процессе функционирования налоговой системы. Возникает необходимость обработки огромного количества информации о налогоплательщиках, имеющей количественный и качественных характер, связанной с неформализуемыми и плохо формализуемыми факторами, которую невыгодно хранить и трудно перерабатывать традиционными методами. Поэтому целесообразно воспользоваться сжатием больших объемов информации путем нахождения некоторых общих закономерностей [1]. Распределить налогоплательщиков по категориям внимания, к которым будут применяться соответствующие комплексы мероприятий с целью эффективного управления процессами в налоговой службе. Наилучший способ применить методы кластерного анализа [2, 3]. Исследованиям в данной области посвящены работы известных зарубежных и российских ученых: J.C. Bezdek, L.A. Zadeh, А.Н. Аверкина, Д.А. Вятченина, В.Д. Штовбы, А.В. Леоненкова, А.Н. Борисова, М.П. Деменкова и др.

Целью работы является проведение анализа методов кластеризации для дальнейшего применения его результатов в эффективном управлении процессами в налоговой службе.

Центральное место среди методов кластерного анализа данных занимает кластерный анализ, являющийся одним из наиболее многообещающих и наиболее интересных подходов к исследованию многомерных процессов и явлений и относящийся к одному из основных методов интеллектуальной обработки данных [4].

Его использование оправдано везде, где требуется многомерный анализ разнокачественной информации.

Задача кластеризации заключается в разбиении объектов X на несколько подмножеств, объекты в которых имеют большие сходства между собой, чем с объектами из других кластеров. «Сходство» в метрическом пространстве определяют через расстояние [5, 6].

Алгоритмы кластеризации оперируют с объектами. С каждым объектом X отождествляется вектор характеристик $X_i = (x_1, ..., x_d)$.

Компоненты x_i, $i = 1, ..., d$ являются отдельными характеристиками объекта. Размерность пространства характеристик определяет количество характеристик d.

Состоящее из всех векторов характеристик объектов, множество обозначается $M = (X_1, ..., X_n)$, где $X_i = (X_{i1}, ..., X_{id})$.

Подмножество «близких друг к другу» объектов из M представляет собой кластер. Расстояние $D(X_i, X_j)$ между объектами X_i и X_j определяется в пространстве характеристик на основе выбранной метрики [7, 8].

Требования, предъявляемые к выявлению кластеров в исследуемой совокупности объектов:

– представление каждого кластера должно быть выражено однородной категорией, содержащей похожие объекты по близким значениям свойств или признаков;

– исследуемая совокупность всех объектов должна быть распределена по всем кластерам, иными словами, быть исчерпывающей;

– каждый объект исследуемой совокупности не должен принадлежать одновременно двум разным кластерам, т.е. кластеры должны быть взаимно исключающими.

Исходной причиной разработки большого многообразия алгоритмов и методов кластеризации послужила возможность использования различных подходов к формальному определению кластеров. Алгоритмы и методы кластерного анализа как инструмент предварительного анализа данных незаменимы при поиске закономерностей в больших наборах многомерных данных, таких как хранилища данных [9]. Именно поэтому для распределения налогоплательщиков по категориям внимания в данной работе предложен метод кластерного анализа. Несмотря на то, что на сегодняшний день известны сотни алгоритмов и методов кластеризации, не существует «универсального решения», поскольку специфика поставленной задачи характеризуется спецификой объекта кластеризации.

Исходя из того, что объектом кластеризации являются данные о налогоплательщиках, требования, которым должен удовлетворять используемый метод кластеризации, сформулированы следующим образом:

– Высокая размерность пространства данных – объекты описываются большим количеством атрибутов, следовательно, должна быть приспособленность алгоритма к работе в пространствах данных высокой размерности.

– Большой объем данных – в связи с тем, что информация о налогоплательщиках обновляется каждый отчетный период, увеличивая тем самым исходную выборку, необходимо, чтобы алгоритм был масштабируем для работы с большим объемом данных.

– Смешанный тип измерений – описание поведения налогоплательщиков включает в себя количественные и качественные характеристики, поэтому алгоритм должен быть приспособлен к использованию различных типов измерений.

По способу обработки данных методы кластерного анализа подразделяются на иерархические и неиерархические. В иерархических методах выполняется последовательное объединение меньших кластеров в большие или же наоборот, большие разделяются на меньшие. Из таких методов выделяют агломеративные, характеризующиеся последовательным объединением исходных элементов и уменьшением числа кластеров, выполняемым до тех пор, пока все элементы не будут принадлежать одному кластеру, и дивизимные, характеризующиеся последовательным разбиением исходного кластера, состоящего из всех объектов, с соответствующим увеличением числа кластеров. Наиболее широко известные из агломеративных методов CURE, ROCK, CHAMELEON и др., а среди дивизимных – BIRCH, MTS и др. Вышеперечисленные методы обладают рядом достоинств и недостатков. Так, например, CURE формирует кластеры несферической формы различных размеров и плотности, однако требует задания некоторых порогов плотности объектов и количества кластеров, а это не всегда приемлемо. ROCK применяет для кластеризации пространств со смешанными типами измерений понятие степени связи между объектами – количество их общих соседей, качество кластеризации определяется оценочной функцией, не масштабируем для большого количества объектов. CHAMELEON – производный метод кластеризации, основанный на устранении методов CURE и ROCK, использует динамическое моделирование для слияния двух кластеров, что облегчает исследование на-

туральных и гомогенных кластеров, однако сложен и зависим от времени.

BIRCH относится к дивизимным методам кластеризации и позволяет обрабатывать большие объемы данных, используя ограниченный объем памяти, в работе основывается на факте неоднородного распределения данных, а также обрабатывает области с большой плотностью как единый кластер, однако позволяет обрабатывать только числовые значения, выделяя кластеры выпуклой или сферической формы, а также существует необходимость задания пороговых значений. MTS позволяет выделять кластеры произвольной формы, в том числе кластеры выпуклой и вогнутой форм, выбирает из нескольких оптимальных решений самое оптимальное, но чувствителен к выбросам [10].

В неиерархических или итеративных методах кластеры формируются в зависимости от задаваемых условий разбиения с возможностью изменения пользователем с целью достижения желаемого результата. К таким методам относятся: k-средних (k-means), CLOPE, PAM, самоорганизующиеся карты Кохонена др. Метод k-средних – достаточно прост и быстр в использовании, алгоритм прозрачен и понятен, но слишком чувствителен к выбросам, на больших базах данных медленен в работе. Недостатком его является невозможность применения на пересекающихся кластерах и необходимость задания количества кластеров. У CLOPE можно отметить высокую масштабируемость и скорость работы, удобство подбирать автоматически количество кластеров, но но слишком чувствителен к выбросам. PAM – так же, как и k-means, прост и быстр в использовании, алгоритм менее чувствителен к выбросам, однако присутствует необходимость задания количества кластеров и медленная работа на больших базах данных. Самоорганизующиеся карты Кохонена используют универсальный аппроксиматор – нейронную сеть, метод прост в реализации, но позволяет работать только с числовыми данными, существует необходимость минимизации размеров сети.

По количеству применений алгоритмов кластеризации выделяют методы с одноэтапной и с многоэтапной кластеризацией.

По возможности расширения объема обрабатываемых данных различают методы масштабируемые и немасштабируемые.

По времени выполнения кластеризации методы можно разделить на потоковые (on-line) и непотоковые (off-line) [11].

По способу анализа данных методы кластеризации подразделяют на четкие и нечеткие. Четкие методы кластеризации разбивают первоначальное множество объектов X на несколько непересекающихся подмножеств. При этом любой объект с X принадлежит только одному кластеру. В результате использования нечетких методов кластеризации возможна одновременная принадлежность нескольким или даже всем кластерам, но с разной степенью принадлежности. Во многих ситуациях нечеткая кластеризация более «естественная», чем четкая, например, для объектов, расположенных на границе кластеров [6]. Нечеткая кластеризация получила широкое применение и развитие благодаря Бездеку и его метода нечетких с-средних (Fuzzy c-means – FCM), который позволяет выделять в кластер объекты, находящиеся на границе кластера, для него характерна меньшая чувствительность к выбросам, однако присутствует необходимость задания количества кластеров, вычислительная сложность кластеризации, а также возникает неопределенность с объектами, удаленными от всех центров кластеров [12].

Требование нахождения однозначной кластеризации элементов исследуемой проблемной области является достаточно грубым и твердым, особенно при решении плохо или слабо структурированных задач. Методы нечеткой кластеризации ослабляют это требование. При введении в рассмотрение нечетких кластеров и соответствующих им функций принадлежности, принимающих значения из интервала [0,1], происходит ослабление этого требования. А элементы матрицы степеней принадлежности четкого разбиения принимают значения из двухэлементного множества {0,1}, а не из интервала [0,1].

Алгоритм FCM для решения задачи нечеткой кластеризации имеет итеративный характер последовательного улучшения некоторой исходной нечеткой разметки, которая задается пользователем или формируется автоматически по некоторому эвристическому правилу. Значения функций принадлежности нечетких кластеров и их типичные представители пересчитываются на каждой итерации. Определяется в форме итеративного выполнения последовательности шагов. Алгоритм FCM закончит работу в случае, когда состоится выполнение заданного априори некоторого конечного числа итераций или когда минимальная абсолютная разница между значениями функций принадлежности на двух последовательных итерациях не станет меньше некоторого априори заданного значения [9].

В базовом алгоритме нечетких с-средних расстояние между объектом и центром кластера рассчитывается через

стандартную Эвклидову норму. В результате выполнения алгоритмов кластеризации с фиксированной нормой форма всех кластеров получается одинаковой. Алгоритмы кластеризации как бы навязывают данным несвойственную им структуру, что иногда приводит к неоптимальным результатам [6]. Для устранения этого недостатка существует несколько методов, среди которых выделим алгоритм Густафсона – Кесселя.

Относительно алгоритма с-средних главное изменение заключается во введении в формулу расчета расстояния между векторами масштабируемой матрицы *B*. Алгоритм Густафсона – Кесселя использует адаптивную норму для каждого кластера, то есть для каждого *i*-го кластера существует своя норм-порождающая матрица. В этом алгоритме при кластеризации оптимизируются не только координаты центров кластеров и матрица нечеткой разбивки, но также и норм-порождающие матрицы для всех кластеров. Это позволяет выделять кластеры различной геометрической формы [6].

Состоит алгоритм Густафсона – Кесселя из следующих шагов.

Шаг 1. Генерирование матрицы нечеткого разбиения.

Шаг 2. Расчет центров кластеров:

$$v_j^k = \frac{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m x_j^i}{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m}, j = \overline{1,q}, \qquad (1)$$

где $\mu_{A_k}(a_i)$ – степень принадлежности a_i-го элемента k-му кластеру;

n – общее количество объектов данных;

q – общее количество измеримых признаков объектов;

x_j^i – количественное значение признака $p_i \in P$;

$m \in (1,\infty)$ – экспоненциальный вес, определяющий нечеткость, размытость кластером и равный действительному числу ($m > 1$).

Шаг 3. Определение матрицы ковариации для k-го кластера:

$$B_j^k = \frac{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m \cdot (x_j^i - v_j^k)^T \cdot (x_j^i - v_j^k)}{\sum_{i=1}^{n} (\mu_{A_k}(a_i))^m}. \qquad (2)$$

Шаг 4. Расчет расстояния между объектами из X и центрами кластеров:

$$D_{B_j^k} = (x_j^i - v_j^k) \cdot [(\det(B_j^k)) \cdot (B_j^k)^{-1}] \cdot (x_j^i - v_j^k)^T. \qquad (3)$$

Шаг 5. Перерасчет элементов матрицы нечеткого разбиения.

Если $D_{B_j^k} > 0$, то проводим расчет по формуле

$$\mu'_{A_k}(a_i) = \left(\sum_{l=1}^{c} \left(\frac{\left(\sum_{j=1}^{q} \left(x_j^i - v_j^k \right)^2 \right)^{\frac{1}{2}}}{\left(\sum_{j=1}^{q} \left(x_j^i - v_j^l \right)^2 \right)^{\frac{1}{2}}} \right)^{\frac{2}{m-1}} \right)^{-1}. \qquad (4)$$

Если $D_{B_j^k} = 0$, то для соответствующего нечеткого кластера $\mu'_k(a_i) = 1$, а для других $\mu'_k(a_i) = 0$.

Шаг 6. Проверка условия

$$\left| f\left(A_k, v_j^k\right) - f'\left(A_k, v_j^k\right) \right| \le \varepsilon,$$

где

$$f(A_k, v_j^k) = \sum_{i=1}^{n} \sum_{k=1}^{c} (\mu_{A_k}(a_i))^m \sum_{j=1}^{q} (x_j^i - v_j^k)^2. \qquad (5)$$

Если условие выполняется, то «конец», а иначе – к шагу 2.

Для оценки качества кластеризации можно использовать величину силуэта *S*. При нечеткой кластеризации номер кластера определяется по максимальному значению степени принадлежности. Значение силуэта выражается для каждого объекта следующим образом:

$$S(x_i) = \frac{a(x_i) - b(x_i)}{\max(a(x_i), b(x_i))}, \qquad (6)$$

где $a(x_i)$ – среднее расстояние между объектом x_i ($x_i \in k, k = \overline{1,c}$) и объектами того же кластера k, к которому принадлежит x_i;

$b(x_i)$ – минимальное расстояние между объектом x_i и объектами в кластере, который ближе всего к кластеру k, то есть кластер, к которому x_i не принадлежит.

Значение силуэта лежит в интервале [–1; 1], если оно отрицательное, то налогоплательщик считается плохо кластеризованным.

Выводы

Таким образом, проведен анализ методов кластеризации и осуществлен выбор подходящего в соответствии с поставленными требованиями для распределения налогоплательщиков по категориям внимания. Следует отметить, что это лишь один из этапов обработки данных о налогоплательщиках. Для принятия эффективных управленческих решений в налоговой службе необходимо выполнение полного комплекса мероприятий обработки данных и создание информационной управленческой системы, что является перспективным направлением для дальнейшего исследования.

Список литературы

1. Чубукова И.А. Data Mining / И.А. Чубукова. – М.: Национальный Открытый университет «ИНТУИТ», 2016. – 471 с.

2. Заболотникова В.С. Учет субъектов предпринимательской деятельности в условиях неопределенности на основе метода нечеткой кластеризации / В.С. Заболотникова // Научные труды Донецкого национального технического университета. Серия «Информатика, кибернетика и вычислительная техника». – 2011. – № 14 (188). – С. 283–290.

3. Заболотникова В.С. Информационная управленческая система для налоговой службы / В.С. Заболотникова, О.Н. Ромашкова // Журнал «Современная наука: актуальные проблемы теории и практики». Серия: Естественные и технические науки. – 2017. – № 6. – С. 27–32.

4. Вятченин Д.А. Нечеткие методы автоматической классификации: монография / Д.А. Вятченин. – Мн.: УП «Технопринт», 2004. – 219 с.

5. Автоматическая классификация: Методы классификации [Электронный ресурс]. – Режим доступа: http://www.aiportal.ru/articles/autoclassification/methods-class.html (дата обращения: 10.05.2017).

6. Сокал Р.Р. Кластер-анализ и классификация: предпосылки и основные направления / Р.Р. Сокал; под ред. Дж. Вэн Райзина. – М.: Мир, 1980. – С. 7–19.

7. Круглов В.В. Интеллектуальные информационные системы: компьютерная поддержка систем нечеткой логики и нечеткого вывода / В.В. Круглов, М.И. Дли. – М.: Физматлит, 2002. – 554 с.

8. Левченко Т.А. Кластерные структуры: основные характеристики и генерируемый эффект / Т.А. Левченко, Е.В. Тунгусова // Фундаментальные исследования. – 2017. – № 3. – С. 144–148.

9. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH / А.В. Леоненков. – СПб.: БХВПетербург, 2005. – 736 с.

10. Ковалев С.С. Современные методы кластеризации в контексте задачи идентификации рассылок почтового спама / С.С. Ковалев, М.Г. Шишаев // Труды Кольского научного центра РАН. Информационные технологии. – 2012. – № 4(11). – С. 89–98.

11. Нейский И.М. Методика адаптивной кластеризации фактографических данных на основе интеграции методов минимального остовного дерева и нечетких k-средних: Автореф. дис. …канд. тех. наук: 05.13.17. – М., 2010. – 18 с.

12. Штовба С.Д. Введение в теорию нечетких множеств и нечеткую логику [Электронный ресурс]. – Режим доступа: http://matlab.exponenta.ru/fuzzylogic/book1/index.php (дата обращения: 15.05.2017).