МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение

высшего образования

«Национальный исследовательский университет

«Московский институт электронной техники»

Институт системной и программной инженерии и информационных технологий

(СПИНТех)

Мясников Максим Александрович

Пояснительная записка

Исследование и разработка алгоритмов и моделей оценки конкурентоспособности производственных предприятий

Студент       Мясников М. А.

Руководитель,   
{д.т.н., проф.}                 Гагарина Л. Г.

Москва, 2023 г.

Оглавление

[ВВЕДЕНИЕ 3](#_Toc153310508)

[ГЛАВА 1. АНАЛИТИЧЕСКИЙ ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И СРЕДСТВ НЕЧЕТКОЙ КЛАССТЕРИЗАЦИИ ДАННЫХ 6](#_Toc153310509)

[1.1. Кластеризация методом k-средних 6](#_Toc153310510)

[1.2. Нечеткая кластеризация C-средних 8](#_Toc153310511)

[1.3. Возможностный метод кластеризации 10](#_Toc153310512)

[Выводы по главе 1 11](#_Toc153310513)

[ГЛАВА 1. ФОРМАЛИЗОВАННОЕ ПРЕДСТАВЛЕНИЕ ВЫБОРА НАИБОЛЕЕ КОНКУРЕНТОСПОСОБНЫХ ПРЕДПРИЯТИЙ 12](#_Toc153310514)

[2.1. Декомпозиция задачи выбора 12](#_Toc153310515)

[2.2. Выделение необходимых признаков 12](#_Toc153310516)

[2.3. Кластеризация данных 13](#_Toc153310517)

[2.4. Выбор оптимального кластера 13](#_Toc153310518)

[2.5. Выбор оптимальных предприятий 14](#_Toc153310519)

# ВВЕДЕНИЕ

**Актуальность исследования и проблемная ситуация**

Процессы, протекающие на современном рынке динамичны и быстро изменчивы. Это приводит к тому, что под воздействием внешних факторов стабильные предприятия могут испытывать трудности и даже покидать рынок, а их место могут занимать новые предприятия. При этом, при заключении договоров и в процессе их исполнения, контрагентам важно понимать, что другая сторона стабильна и будет исполнять свои обязательства. Таким образом, исследование конкурентоспособности предприятий на рынке и оценка их экономических стабильности и безопасности является неотъемлемой частью экономических процессов.

Существуют различные рейтинговые агентства, оценивающие конкурентоспособность предприятий. Однако они характеризуются закрытостью методик проведения анализа, что делает вероятным предвзятость при составлении рейтингов. Это же относится и к агентствам, проводящим анализ на заказ. Примером такого агентства является сервис «РБК Pro». Он предоставляет такие данные, как рейтинг предприятий по выручке, чистой прибыли, капиталу, количеству сотрудников и т.д. Однако эти рейтинги не связаны между собой и не дают четкого представления о компании.

Провести анализ рынка можно и без обращения к сторонним агентствам по открытым данным. Так, собрав все критичные характеристики, можно обратиться к способу оценки конкурентоспособности предприятий известному как «Операционные методы». Для более точного вычисления данных способ можно модифицировать построением модели нечеткой системы.

В соответствии с операционным подходом, наиболее конкурентоспособны те предприятия, в которых лучше всего реализована работа всех служб. Для оценки конкурентоспособности необходимо определить перечень входных данных – операций и показателей, являющихся значимыми для обеспечения конкурентоспособности. Эти данные можно получить напрямую из государственных источников, таких как:

* Единый государственный реестр юридических лиц (ЕГРЮЛ),
* Единый государственный реестр индивидуальных предпринимателей (ЕГРИП),
* Статистический регистр хозяйствующих субъектов Главного Межрегионального Центра обработки и распространения статистической информации Федеральной службы государственной статистики (ГМЦ Росстата).

Либо получить уже собранные данные, с помощью агрегатора, предоставляемого информационным агентством «Интерфакс» - сервиса «СПАРК» («система профессионального анализа рынков и компаний»).

Существует множество способов анализа и обработки полученных данных. Необходимо разделить предприятия по имеющимся параметрам на классы различной степени конкурентоспособности. Эту задачу решает кластерный анализ, подкрепляемый для большей гибкости входных параметров и результатов инструментами нечеткой логики.

После распределения предприятий по группам, необходимо выбрать оптимальные группы и среди этих отфильтрованных предприятий выбрать наиболее оптимальные и конкурентоспособные, а это задача многокритериальной оптимизации.

Таким образом, диссертационное исследование направлено на изучение методов и алгоритмов нечёткого кластерного анализа и многокритериальной оптимизации.

**Объект исследования**

Объектом исследования диссертационной работы является задача поиска оптимального выбора.

**Предмет исследования**

Предметом исследования диссертационной работы являются алгоритмы кластеризации данных на основе методов машинного обучения с применением нечетких систем для решения задачи оценки конкурентоспособности производственных предприятий.

**Цель и задачи исследования**

Целью исследования магистерской диссертации является повышение точности выбора наиболее конкурентоспособного предприятия из представленного множества.

В соответствие с предметом и целью исследования диссертационной работы ставятся следующие задачи:

* анализ существующих методов и алгоритмов кластеризации данных, их преимущества и недостатки, взаимодействие с нечёткими системами;
* формализовать задачу оценки конкурентоспособности;
* разработка математической модели и алгоритмов, на основе нечеткой кластеризации;
* программная реализация разработанных модели и алгоритмов;
* сравнительный анализ результатов разработки с существующими решениями.

**Положения, выносимые на защиту**

На защиту выносятся:

* формализованное представление задачи оценки конкурентоспособности;
* модель и алгоритмы оценки конкурентоспособности на основе алгоритмов нечеткой кластеризации;
* программная реализация разработанной модели;
* оценка предлагаемой модели.

**Практическая значимость результатов**

Практическая значимость результатов, полученных в диссертационной работе, заключается в повышении точности выбора предприятия в производственной среде для заключения с ними договоров и контрактов при сравнении их на предмет конкурентоспособности.

# ГЛАВА 1. АНАЛИТИЧЕСКИЙ ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И СРЕДСТВ НЕЧЕТКОЙ КЛАССТЕРИЗАЦИИ ДАННЫХ

Нечеткая кластеризация – форма кластеризации, в которой каждая точка данных может принадлежать более чем одному кластеру. Кластерный анализ предполагает назначение точек данных таким образом, что элементы одного кластера должны быть как можно более похожи, а элементы, принадлежащие разным кластерам, – как можно более непохожи. Идентификация кластеров производится с помощью мер сходства. В эти меры входят такие параметры как расстояние, связанность и прочие. Меры сходства выбираются в зависимости от данных и целей модели.

При *не*-нечеткой кластеризации (жесткой кластеризации) данные разделяются на отдельные кластеры так, что каждая точка принадлежит только одному кластеру. При нечеткой (мягкой) кластеризации данные могут потенциально принадлежать нескольким кластерам.

Так, если мы имеем объект A, который может относиться к группе A1 или A2, то при жесткой кластеризации он может принадлежать либо исключительно группе A1, либо группе A2. Если же объект B обладает свойствами, позволяющими определить его как к группе B1, так и к группе B2, при мягкой кластеризации определяется степень принадлежности объекта к группам. Если обозначить степень принадлежности объекта как [B], то возможны ситуации:

* [B1] = 0, [B2] = 1 – объект B – принадлежит исключительно группе B2;
* [B1] = 1, [B2] = 2 – объект B – принадлежит исключительно группе B1;
* [B1] = α, [B2] = β – объект B принадлежит к группе B1 со степенью α и к группе B2 – со степенью β.

Здесь α, β – переменные, нормализованные от 0 до 1. При этом они выражают степень принадлежности, но не вероятность, поэтому α + β не обязательно равно 1.

* 1. **Кластеризация методом k-средних**

Одним из наиболее популярных методов кластеризации данных является метод k-средних. Алгоритм основывается на минимизации суммарного квадратичного отклонения (формула (1.1)) точек кластеров от центров этих кластеров.

*(1.1)*

где k – число кластеров, Si – полученные кластеры, μi – центрs масс всех векторов x кластера Si.

Алгоритм разбивает множество векторных элементов на заданное количество кластеров. На каждой итерации происходит перевычисление центра масс кластеров, полученных на предыдущем шаге и перераспределение векторов в соответствие с новыми центрами масс. Выходом из цикла является отсутствие перераспределений на очередном шаге. Так как количество разбиений конечного множества конечно, а на каждом шаге суммарное конечное отклонение уменьшается – гарантируется, что число итераций конечно.

На рисунке 1.1 приведен пример алгоритма для трех кластеров.

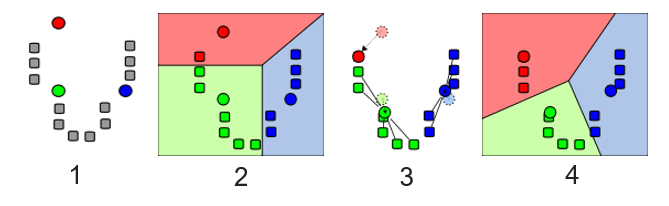


Рисунок 1.1 – этапы работы метода k-средних

На этапе 1 выбираются случайные точки-центры. На этапе 2 происходит распределение точек по кластерам. 3-1 этап – вычисление новых центров и перераспределение точек. Этапы 2 и 3 повторяются несколько раз, пока алгоритм не завершится – этап 4.

Однако метод k-средних не всегда дает точный результат – на рисунке 1.2. изображено сравнение кластеризации ирисов Фишера методом k-средних и их реальное распределение.

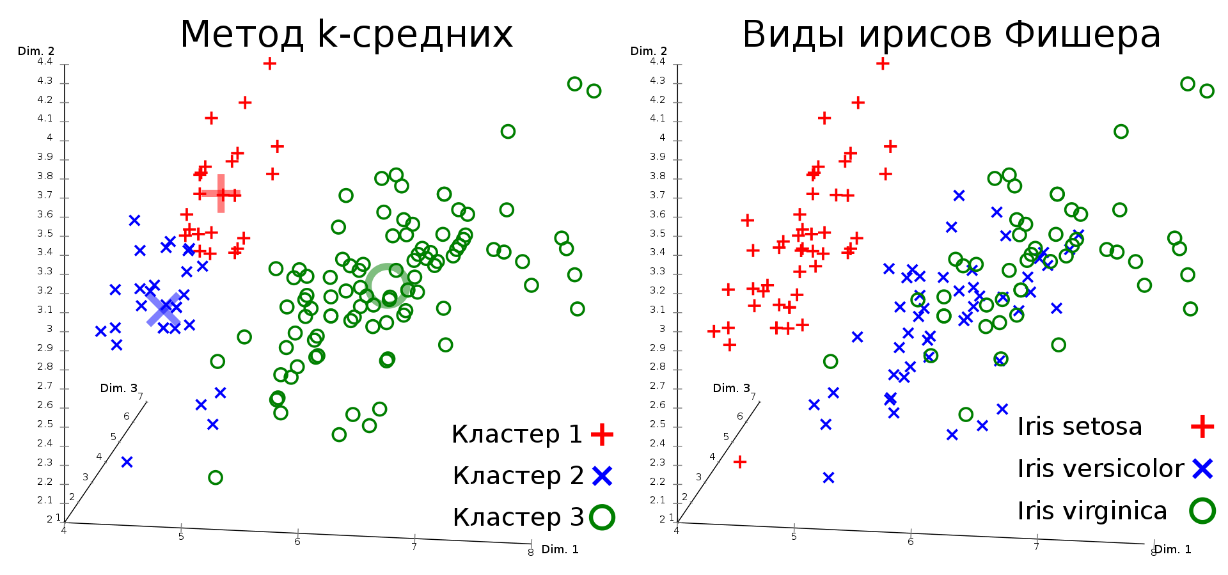


Рисунок 1.2 – сравнение результата кластеризации методом k-средних для ирисов Фишера с их реальными видами

Основные недостатки метода k-средних:

* достижение локального минимума суммарного квадратичного отклонения, но не глобального в общем случае;
* итоговый результат различен в зависимости от исходно выбранных центров кластеров;
* отсутствие универсального выбора оптимальных начальных центров;
* чувствительность к шуму – очень далеким точкам;
* необходимо заранее задавать число кластеров.
  1. **Нечеткая кластеризация C-средних**

Усовершенствованием вариантом метода k-средних является метод FCM (Fuzzy C-mean) известный также как метод C-средних. Если в методе k-средних определялась принадлежность точки к одному из кластеров, то в методе C-средних рассчитывается степень принадлежности точки каждому из кластеров.

В методе C-средних каждой точке задается степень принадлежности wk(x) в k-ом кластере. Центроидом кластера является среднее значение всех точек, взвешенных по степени их принадлежности к кластеру (формула (1.2)).

*(1.2)*

где m ∈ (1, ∞) – гиперпараметр, определяющий насколько нечетким будет кластер (чем m больше, тем более нечеткий кластер).

Минимизируемая функция J (W, C) для множества элементов X = {x1, …, xn} – формула (1.3).

*(1.3)*

где C = {c1, …, cc} – набор кластеров, W = wi,j ∈ [0, 1], - степень принадлежности элемента xi кластеру cj – формула (1.4).

*(1.4)*

такая, что

*(1.5)*

*(1.6)*

При m → 1 значения wi,j сходятся к 0 или 1, а целевая функция минимизации совпадает с минимизируемой функцией метода k-средних. В отсутствие каких-либо критериев и знаний о предметной области обычно устанавливается m = 2.

Т.к. алгоритм является расширение метода k-средних он обладает теми же недостатками:

* минимум – локальный минимум;
* чувствительность к шуму;
* результаты зависят от первоначального выбора весов.

Для лучшего понимая отличия метода C-средних от метода k-средних рассмотрим пример одномерных данных по оси X (рисунок 1.3).

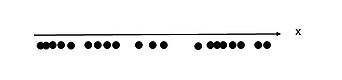


Рисунок 1.3 – точки на оси Х, представляющие данные

Разделим данные на два кластера «А» и «В» так, что каждая точка будет иметь коэффициент принадлежности 0 или 1, отображенный на ось Y (рисунок 1.5).

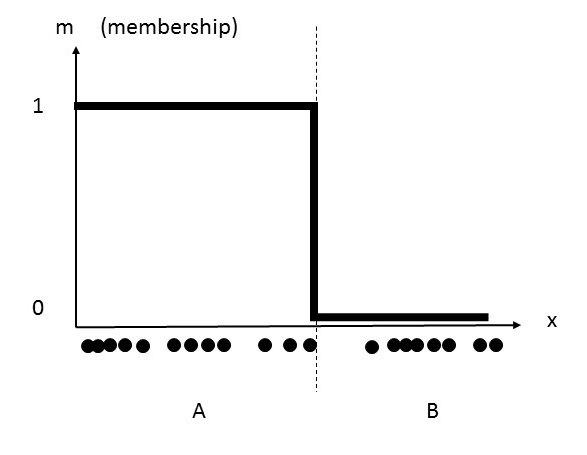


Рисунок 1.5 – принадлежность точек кластерам «А» и «В» при жесткой кластеризации

Для мягкой кластеризации каждая точка может принадлежать нескольким классам и коэффициенты принадлежности могут варьироваться от 0 до 1 (рисунок 1.6).

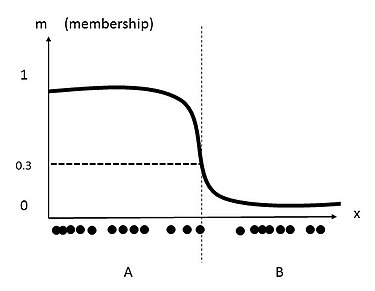


Рисунок 1.6 – принадлежность точек кластерам «А» и «В» при мягкой кластеризации

Как можно видеть, присутствует точка, принадлежащая обоим кластерам, такая что коэффициент принадлежности кластеру «А» равен 0.3.

* 1. **Возможностный метод кластеризации**

Возможностный метод кластеризации (PCM, Possibility C-mean) развивает идею FCM добавлением в формуле (1.3) второго слагаемого.

*(1.7)*

где ηi – положительные числа и

*(1.8)*

*(1.9)*

*(1.10)*

Таким образом, при малых значениях ηi малые и wij, и большие соответственно. Также ηi отвечает за степень сравнимости второго слагаемого в (1.7) с первым. Если оба слагаемых равновесны, то ηi ≈ ||xi – cj||2. Обычно сначала вычисляются wij с помощью FCM, а затем ηi – формула (1.11).

*(1.11)*

а затем применяется возможностный алгоритм, в котором ηi вычисляется по формуле (1.12).

*(1.12)*

где 0 < α < 1.

Наиболее подходящим значением параметра m для данного алгоритма является m = 1.5.

Преимуществом данного метода является победа чувствительности к шуму.

**Выводы по главе 1**

В ходе диссертационного исследования в первой главе были рассмотрены методы кластеризации данных с использованием классической и нечеткой логик, их преимущества и недостатки.

# ГЛАВА 1. ФОРМАЛИЗОВАННОЕ ПРЕДСТАВЛЕНИЕ ВЫБОРА НАИБОЛЕЕ КОНКУРЕНТОСПОСОБНЫХ ПРЕДПРИЯТИЙ

* 1. **Декомпозиция задачи выбора**

Для решения поставленных в диссертационной работе целей разложим задачу на более простые подзадачи. Для этого выделим следующие пункты:

1. Формирование базы о предприятиях выделив критичные для задачи выбора признаки.
2. Нормализация данных.
3. Группировка данных в кластеры по выделенным признакам.
4. Выбор наиболее подходящих кластеров.
5. Выбор внутри кластеров наиболее результативных элементов.
   1. **Выделение необходимых признаков**

Данные о предприятиях можно получить из открытых источников:

* ЕГРЮЛ;
* реестр исполнительных производств в отношении юридических лиц ФССП;
* реестр недобросовестных поставщиков ФАС;
* реестр госзакупок;

Характерными признаками для определения наиболее конкурентоспособных предприятий являются:

* данные, о недобросовестных поставщиках (количество недобросовестных поставок и цены контрактов этих поставок предприятий);
* долги предприятий;
* количество исполнительных производств (открытых, закрытых);
* госконтракты (выполненные, их цена).

Для дальнейших расчетов необходимо произвести нормализацию этих данных:

* абсолютное количество недобросовестных поставок;
* относительная цена недобросовестных поставок (формула 2.1, si – цена i-го контракта);
* абсолютный долг предприятия;
* абсолютное количество открытых делопроизводств;
* абсолютное количество закрытых делопроизводств;
* абсолютная цена выполненных контрактов;
* абсолютная цена текущих контрактов.

Признаки имеют область значений [0; + ∞), поэтому необходимо нормализовать признаки относительно наибольшего значения (формула 2.2, где P(i) – значение признака i-го предприятия)

.2)

* 1. **Кластеризация данных**

Для уменьшения итоговой выборки для задачи выбора оптимальных и наиболее конкурентоспособных предприятий сгруппируем данные в кластеры. Задачу кластеризации можно формально определить как функцию отображающую входное множество векторов признаков X = {x(1)(p1, p2, …, pN), …, x(M)(p1, p2, …, pN)} во множество пар вектор-кластер Y = { y(1), …, y(M)}, где y(i) = { x(1), kj} (kj – j-й кластер, j ∈ [1; K] – номер кластера) (формула 2.3).

*(2.3)*

* 1. **Выбор оптимального кластера**

Рассмотрим центры полученных кластеров как входные данные для задачи выбора оптимального решения.

Имеем C = { c(j) (p1, p2, …, pN), …, c(K) (p1, p2, …, pN)} – центры кластеров, W = { w1, w2, …, wN } – значимости признаков (область значений [0; 1]). Fc(C, W) – функция «прибыли» (формула 2.4).

Введем критерий оптимальности кластера c\*. Тогда оптимальными будут считаться кластеры такие что cj >= c\*.

* 1. **Выбор оптимальных предприятий**

Повторим алгоритм оптимальности для предприятий из оптимальных кластеров.

Имеем X\* = { x(k) (p1, p2, …, pN) | x(k) in j and cj > c\*} – центры кластеров, W = { w1, w2, …, wN } – значимости признаков (область значений [0; 1]). Fc(C, W) – функция «прибыли» (формула 2.4).

Введем критерий оптимальности кластера x\*. Тогда оптимальными будут считаться кластеры такие что xk >= x\*.

Множество X\*\* = {x(k)(p1, p2, …, pN) | k in L and xk >= x\*} – множество наиболее конкурентоспособных предприятий. Таким образом, задачу выбора наиболее конкурентоспособных предприятий можно выразить формулой (2.5).

(2.5)