Федеральное агентство по образованию РФ
Дальневосточный федеральный университет
Институт математики и компьютерных наук
Кафедра информатики

Кластеризация эмпирических данных ранговым методом

Курсовая работа
студента 238 группы
Гренкина Г. В.
Руководитель:
ст. преподаватель кафедры информатики
Черныш Е. В.

Владивосток, 2010

Содержание

Аннотация

В данной работе рассмотрена задача кластеризации эмпирических данных ранговым методом, основанным на модифицированном В.П. Масловым законе Ципфа. Предложен способ формализации исходной постановки задачи.

1. Введение

1.1. Описание предметной области

Эмпирические данные, т.е. данные, полученные опытным путём, являются основой для выделения информации о закономерностях изучаемых явлений, а также принятия решений в различных сферах человеческой деятельности. Подобные исследования актуальны в областях знаний, где имеются большие массивы данных. При анализе таких данных приходится решать задачу нахождения зависимости между значениями некоторого набора факторов и поведением исследуемого явления. Это часто приводит к необходимости структурировать, систематизировать, выделить полученные данные по тем или иным признакам, т.е. кластеризовать их [7].

Ещё задолго до создания ЭВМ ученые занимались построением классификаций как в естественных, так и в общественных науках. Например, это иерархическая классификация растений и видов М. Адансона (1757 г.), периодическая система элементов Д. И. Менделеева (1869 г.).

Однако до разработки аппарата многомерного статистического анализа и, главное, до появления и развития достаточно мощной электронно-вычислительной базы проблемы теории и практики классификации относились не к разработке методов и алгоритмов, а к полноте и тщательности отбора и теоретического анализа изучаемых объектов, характеризующих их признаков, смысла и числа градаций по каждому из этих признаков.

Все методы классификации сводились, по существу, к методу так называемой *комбинационной группировки*, когда все характеризующие объект признаки носят дискретный характер или сводятся к таковым, а два объекта относятся к одной группе только при точном совпадении зарегистрированных на них градаций одновременно по всем характеризующим их признакам.

По мере роста объемов перерабатываемой информации возможность эффективной реализации подобной логики исследования становилась всё менее реальной (так, если каждый объект характеризуется 5-ю признаками, а число градаций по каждому из признаков равно 3, то число групп или классов при комбинационной группировке оказывается равным $3^5 = 243$). Именно электронно-вычислительная техника стала тем главным инструментом, который позволил по-новому подойти к решению проблемы классификации и, в частности, конструктивно воспользоваться разработанным к этому времени мощным аппаратом многомерного статистического анализа: методами распознавания образов «с учителем» (дискриминантный анализ) и «без учителя» (автоматическая классификация, или кластер-анализ)[1].

В общей (нестрогой) постановке *проблема автоматической классификации объектов* заключается в том, чтобы всю анализируемую совокупность объектов разбить на сравнительно небольшое число (заранее известное или нет) однородных, в определённом смысле, групп или классов.

Для формализации этой проблемы удобно интерпретировать анализируемые объекты в качестве точек в соответствующем признаковом пространстве. Естественно предположить, что геометрическая близость двух или нескольких точек в этом пространстве означает близость «физических» состояний соответствующих объектов, их однородность. Тогда

[1] Если исследователь располагает не только классифицируемыми данными, но и так называемыми обучающими выборками, то говорят, что решается задача «классификации с обучением», в противном случае речь идет о задаче «классификации без обучения» [1, 2].

проблема классификации состоит в разбиении анализируемой совокупности точек-наблюдений на сравнительно небольшое число (заранее известное или нет) классов таким образом, чтобы объекты, принадлежащие одному классу, находились бы на сравнительно небольших расстояниях друг от друга. Полученные в результате разбиения классы часто называют *кластерами* (таксонами, образами), а методы их нахождения соответственно кластер-анализом, численной таксономией, распознаванием образов с самообучением [1, 2].

Однако, берясь за решение задачи классификации, исследователь с самого начала должен чётко представлять, какую именно из двух задач он решает. Рассматривает ли он обычную задачу разбиения статистически обследованного (обычно многомерного) диапазона изменения значений анализируемых признаков на интервалы (гипероблсти) группирования, в результате решения которой исследуемая совокупность объектов разбивается на некоторое число групп так, что объекты такой одной группы находятся друг от друга на сравнительно небольшом расстоянии (многомерный аналог задачи построения интервала группирования при обработке одномерных наблюдений). Либо он пытается определить *естественное расслоение* исходных наблюдений на чётко выраженные кластеры, сгустки, лежащие друг от друга на некотором расстоянии, но не разбивающиеся на столь же удалённые части.

Если первая задача — построение областей группирования — всегда имеет решение, то при второй постановке результат может быть и отрицательным: может оказаться, что множество исходных наблюдений не обнаруживает естественного расслоения на кластеры (например, образует один общий кластер).

Наиболее труден и наименее формализован в задаче автоматической классификации момент, связанный с определением *понятия однородности объектов*.

В общем случае [1, 2] понятие однородности объектов определяется заданием правила вычисления величины ρ_{ij}, характеризующей либо расстояние $d(O_i, O_j)$ между объектами O_i и O_j из исследуемой совокупности, либо степень близости (сходства) $r(O_i, O_j)$ тех же объектов. Если задана функция $d(O_i, O_j)$, то близкие в смысле этой метрики объекты считаются однородными, принадлежащими к одному классу. Естественно, при этом необходимо сопоставление $d(O_i, O_j)$ с некоторым пороговым значением, определяемым в каждом конкретном случае по-своему.

Конечно, *выбор метрики (или меры близости) является узловым моментом исследования*, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему. При этом решение данного вопроса зависит в основном от главных целей исследования, физической и статистической природы вектора наблюдений X, полноты априорных сведений о характере вероятностного распределения X.

В качестве примеров расстояний и мер близости, сравнительно широко используемых в задачах кластер-анализа, приведем здесь следующие [1, 2].

1. Обобщённое («взвешенное») расстояние махаланобисского типа:

$$d_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Lambda^T \Sigma^{-1} \Lambda (X_i - X_j)}.$$

Здесь Σ — ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения X_i, а Λ – некоторая симметричная неотрицательно определенная матрица «весовых» коэффициентов λ_{mq}, которая чаще всего выбирается диагональной.

2. Обычное евклидово расстояние:

$$d_E(X_i, X_j) = \sqrt{\sum_{k=1}^{p} \left(x_i^{(k)} - x_j^{(k)}\right)^2}.$$

К ситуациям, в которых использование этого расстояния можно признать оправданным, прежде всего, относят следующие:

- наблюдения X извлекаются из генеральных совокупностей, описываемыхмногомерным нормальным законом с ковариационной матрицей вида $\sigma^2 \cdot I$, т.е. компоненты X взаимно независимы и имеют одну и ту же дисперсию;
- компоненты $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ вектора наблюдений X однородны по своему физическому смыслу, причем установлено, например с помощью опроса экспертов,

что все они одинаково важны с точки зрения решения вопроса об отнесении объекта к тому или иному классу;

- признаковое пространство совпадает с геометрическим пространством нашего бытия, что может быть лишь в случаях $p = 1, 2, 3$, и понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

3. «Взвешенное» евклидово расстояние:

$$d_{\text{вЕ}}(X_i, X_j) = \sqrt{\omega_1\left(x_i^{(1)} - x_j^{(1)}\right)^2 + \omega_2\left(x_i^{(2)} - x_j^{(2)}\right)^2 + \ldots + \omega_p\left(x_i^{(p)} - x_j^{(p)}\right)^2}\ .$$

Обычно применяется в ситуациях, в которых так или иначе удаётся приписать каждой из компонент $x^{(k)}$ вектора наблюдений X некоторый неотрицательный «вес» ω_k, пропорциональный степени её важности с точки зрения решения вопроса об отнесении заданного объекта к тому или иному классу.

При конструировании различных процедур классификации в ряде ситуаций оказывается целесообразным введение понятия *расстояния между целыми группами объектов*. Приведём примеры наиболее распространённых расстояний, характеризующих взаимное расположение отдельных групп объектов [1, 2].

Пусть S_i — i-я группа (класс, кластер) объектов, n_i — число объектов, образующих группу S_i, вектор $\overline{X}(i)$ — среднее арифметическое векторных наблюдений, входящих в S_i (другими словами, $\overline{X}(i)$ — «центр тяжести» i-й группы), а $\rho(S_l, S_m)$ — расстояние между группами S_l и S_m.

1. Расстояние, измеряемое по принципу «ближнего соседа»:
$$\rho_{\min}(S_l, S_m) = \min_{X_i \in S_l, X_j \in S_m} d(X_i, X_j).$$
2. Расстояние, измеряемое по принципу «дальнего соседа»:
$$\rho_{\max}(S_l, S_m) = \max_{X_i \in S_l, X_j \in S_m} d(X_i, X_j).$$
3. Расстояние, измеряемое по «центрам тяжести» групп:
$$\rho(S_l, S_m) = d\left(\overline{X}(l), \overline{X}(m)\right).$$
4. Расстояние, измеряемое по принципу «средней связи», определяется как арифметическое среднее всевозможных попарных расстояний между представителями рассматриваемых групп, т.е.
$$\rho_{\text{ср}}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d(X_i, X_j).$$

Естественно попытаться определить сравнительное качество различных способов разбиения заданной совокупности элементов на классы, т.е. определить тот количественный критерий, следуя которому можно было бы предпочесть одно разбиение другому. С этой целью в постановку задачи кластер-анализа часто вводится понятие так называемого функционала качества разбиения $Q(S)$, определённого на множестве всех возможных разбиений. Тогда под наилучшим разбиением S^* понимается то разбиение, на котором достигается экстремум выбранного функционала качества. Выбор того или иного функционала качества, как правило, осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо строгую формализованную систему.

Приведём примеры наиболее распространённых функционалов качества разбиения [1, 2].

1. Функционалы качества разбиения при заданном числе классов.

Пусть исследователем уже выбрана метрика d и пусть $S = (S_1, S_2, \ldots, S_k)$ — некоторое фиксированное разбиение наблюдений $X_1, X_2, \ldots, X_n$ на заданное число k классов $S_1, S_2, \ldots, S_k$. За функционалы качества часто берутся следующие характеристики.

а) Сумма («взвешенная») внутриклассовых дисперсий:

$$Q_1(S)=\sum_{l=1}^{k}\sum_{X_i\in S_l} d^2\left(X_i,\overline{X}(l)\right).$$

б) Сумма попарных внутриклассовых расстояний между элементами:

$$Q_2(S)=\sum_{l=1}^{k}\sum_{X_i,X_j\in S_l} d^2\left(X_i,X_j\right) \text{ либо } Q_2'(S)=\sum_{l=1}^{k}\frac{1}{n_l}\sum_{X_i,X_j\in S_l} d^2\left(X_i,X_j\right).$$

2. Функционалы качества разбиения при неизвестном числе классов.

В ситуациях, когда исследователю заранее не известно, на какое число классов подразделяются исходные многомерные наблюдения $X_1, X_2, \ldots, X_n$, функционалы качества разбиения $Q(S)$ выбирают чаще всего в виде простой алгебраической комбинации (суммы, разности, произведения, отношения) двух функционалов $I_1(S)$ и $I_2(S)$, один из которых I_1 является убывающей (невозрастающей) функцией числа классов k и характеризует, как правило, внутриклассовый разброс наблюдений, а второй I_2 — возрастающей (неубывающей) функцией числа классов k. При этом интерпретация функционала I_2 может быть различной. Под I_2 понимается иногда и некоторая мера взаимной удалённости (близости) классов, и мера тех потерь, которые приходится нести исследователю при излишней детализации рассматриваемого массива исходных наблюдений, и величина, обратная так называемой «мере концентрации» всей структуры точек, полученной при разбиении исследуемого множества наблюдений на k классов [1, 2].

Вообще, методы кластеризации довольно разнообразны, в них по-разному выбирается способ определения близости между кластерами (и между объектами), а также используются различные алгоритмы вычислений. Заметим, что результаты кластеризации зависят от выбранного метода, и эта зависимость тем сильнее, чем менее явно изучаемая совокупность разделяется на группы объектов. Поэтому результаты вычислительной кластеризации могут быть дискуссионными и часто они служат лишь подспорьем для содержательного анализа [24].

Заметим также, что методы кластерного анализа не дают какого-либо способа для проверки статистической гипотезы об адекватности полученных классификаций. Иногда результаты кластеризации можно обосновать с помощью методов дискриминантного анализа [24].

В настоящее время существует несколько подходов к решению задачи кластерного анализа, которые основаны на различных представлениях о задаче, использовании специфичной для каждой предметной области дополнительной информации и т.д. Перечислим наиболее часто используемые подходы [4]:

- вероятностный подход;
- подход, использующий аналогию с центром тяжести;
- подход, основанный на теории графов;
- иерархический подход;
- подход, основанный на понятии ближайшего соседа;
- нечёткие алгоритмы кластерного анализа;
- подход, использующий искусственные нейронные сети;
- эволюционный (генетический) подход.

Трудности при решении задач кластеризации связаны с оптимальным выбором метрики, методов группировки объектов в пространстве признаков; существованием большого количества признаков, описывающих изучаемое явление, и другими факторами. Поэтому в кластерном анализе существует общая проблема формулировки алгоритма обработки эмпирических данных с целью извлечения информации [7].

В статье [7] предложен *ранговый метод кластеризации эмпирических данных*. Ниже указано, в чём заключается этот метод.

Поскольку при решении задач исследователям приходится изучать знаковые объекты произвольной природы, т.е. семиотические системы, то для анализа эмпирических данных естественным является использование подходов из лингвистики. В частности, для частотных словарей известен закон Ципфа [5, 12, 30], описывающий соотношение между частотой и

рангом слов в словаре. Согласно классическому подходу [30], частота совпадёт с вероятностью повторяемости знака, поэтому эта идея была использована при изучении объектов в других областях знаний [30]. В зарубежной научной литературе на соответствующее функциональное соотношение между вероятностью повторяемости знака и его рангом ссылаются как «power law» [7].

Однако ещё А.Н. Колмогоров в своих работах [11] наметил путь пересмотра теории вероятностей с точки зрения дискретного подхода. Согласно его концепции, случайность — это большая сложность. Тогда алгоритм возникновения случайного события будет очень сложным, а расшифровка этого алгоритма потребует очень длинного кода. Чем сложнее описана информация, тем длиннее необходим алгоритм расшифровки, а это, согласно Колмогорову, близко к случайному [7].

В настоящее время эта парадигма развита В.П. Масловым [13–17, 31]. В его работах получены формулы, которые точнее, чем закон Ципфа, описывают соотношение между частотой и рангом слов. Говоря другими словами, закон Ципфа справедлив для небольших текстов, а соотношение, полученное В.П. Масловым, работает для длинных текстов. Поэтому один из выводов, сделанных в [14, 17, 31], состоит в том, что функциональная зависимость между частотой и рангом слов является существенной характеристикой языка писателя. Следует отметить, что предлагаемый им подход был использован также для изучения экономических явлений [14]. В идейном плане, как указано в [17], его можно также использовать для анализа семиотических объектов [7].

В статье [7] подход В.П. Маслова использован для анализа эмпирических данных с целью выделения кластерных объектов.

Там этот подход реализован для решения задачи о выделении групп медицинских работников скорой помощи по их эмоциональному состоянию с учётом стажа работы и задачи выявления групп пациентов, страдающих хроническим заболеванием желудочно-кишечного тракта, по степени тяжести заболевания исходя из показателей различных видов анализа крови.

Как указано выше, важной характеристикой знака является его повторяемость в данном социуме, т.е. частота встречаемости, характеризующая активность использования знака. Известный для частотных словарей закон Ципфа [5, 12, 30], описывающий соотношение между частотой и рангом слов в словаре, обычно рассматривается в логарифмических координатах:

$$\ln r + \frac{1}{D}\ln w_r = const\,, \qquad (1)$$

где r — ранг слова, совпадающий с его номером в частотном словаре по убыванию частоты, w_r — частота встречаемости этого слова в тексте, D — константа. Поскольку Ципф рассматривал закономерность (1) на огромном числе словарей и частоты принимают достаточно большие значения порядка 10^{10}, то изменение этой величины, например, втрое для логарифма от неё меняется на величину логарифма 3, что является незначительной поправкой. Это означает, что в переменных без логарифмов формула (1) огрубляет соотношение между рангом и частотой, давая значительную ошибку [7].

Новый подход исследования статистических зависимостей в языке, предложенный В.П. Масловым [13–17, 31], позволяет получить более точное соотношение между рангом и частотой. Пусть рассматривается алфавитный словарь, в котором указаны частоты встречаемости w_i (i = 1, ..., s) каждого слова из некоторого корпуса текстов, n_i — число слов, имеющих одну частоту встречаемости, и $N = \sum_{i=1}^{s} n_i$ задаёт число слов словаря, а величина

$$\sum_{i=1}^{s} n_i w_i = M \qquad (2)$$

совпадает с объёмом текста. Справедливо следующее утверждение [17]: если варианты $\{n_i\}$ равноценны и удовлетворяют (2), то ранг r_l для l-го слова вычисляется по формуле

$$r_l = \sum_{i=1}^{l} \frac{1}{e^{\beta w_i + \sigma} - 1}. \qquad (3)$$

Эта формула аналогична распределению Бозе для тождественных частиц, для которого роль энергии частиц играет частота повторяемости w_i, а число частиц на заданном уровне совпадает с n_i [7].

Однако каждому реальному тексту соответствует более широкий (виртуальный) текст, поскольку язык позволяет заменять элементы текста словами-заместителями, а также пропускать в тексте легко подразумеваемые слова. Тогда в виртуальном тексте частота встречаемости $\tilde{w}_i > w_i$, а его виртуальный объём равен

$$\sum_{i=1}^{s} n_i \tilde{w}_i = \tilde{M}. \qquad (4)$$

В предположении равноценности вариантов $\{n_i\}$ и выполнения условия (4) справедлива формула (3) для ранга, в которой следует заменить $\tilde{w}_i \to w_i$ [7].

Практическое применение полученных соотношений связано с выбором виртуальной частоты и других феноменологических параметров. Простейшая параметризация для виртуальной частоты встречаемости имеет вид $\tilde{w}_i = w_i\left(1 + \alpha w_i^{\gamma}\right), \alpha > 0, \gamma > 0$. Если $w_i = i$, то, полагая $\beta << 1$ и $\sigma = 0$, можно перейти от суммы к интегралу, в результате имеем

$$r_l \cong \ln \frac{w_l^{\gamma}}{1 + \alpha w_l^{\gamma}} + c. \qquad (5)$$

Использование (5) для семиотических систем показало [16, 17, 31], что формула (5) точнее, чем закон Ципфа, описывает соотношение между рангом слова и частотой его встречаемости в словаре [7].

В.П. Масловым были предложены также другие параметризации для виртуальной частоты и рассмотрены задачи, связанные с анализом экономического риска при покупке товара.

Для товаров длительного пользования цена для покупателя выше, чем цена, выставленная продавцом. Это можно проследить на примере авторынка. Чем дороже автомобиль, тем больше риск, что его угонят, тем больше цена запасных частей в случае поломок и т.п. Следовательно, чтобы избежать этих рисков, нужно покупать страховку и учитывать хлопоты, которые связаны с выплатой страховой суммы, ликвидность, мгновенное падение цены покупки и т.п. Таким образом, цена автомобиля при подсчёте покупателем (виртуальная цена) возрастает прогрессивно, как прогрессивный налог. Чем дороже автомобиль, тем длиннее «хвост» дополнительных расходов и тем больше риск, который определяет его цену для покупателя, учитывающего сопутствующие товары и услуги [14].

Если цена на автомобиль невысокая, особенно если он подержанный, то здесь дополнительный шлейф расходов растёт по мере уменьшения цены: увеличивается риск попасть в аварию, дорожный налог на старую машину за рубежом также существенно увеличен, опасность выхода из строя деталей возрастает и т.д. Следовательно, цена автомобиля для покупателя с учётом этих обстоятельств складывается из номинальной цены p_i плюс цена шлейфа дополнительных расходов, для которых В.П. Маслов вводит четыре неизвестных параметра:

$$p_i + ap_i^{\gamma} + bp_i^{-\sigma}.$$

В.П. Маслов рассматривает два экспериментально построенных графика:

1) график зависимости N_{p_l} — числа проданных автомобилей по цене, меньшей или равной p_l, как функции от цены p_l (рис. 1);

2) график зависимости ранга (номера марки автомобиля по возрастанию цены) N_{p_l} от цены p_l (рис. 2) [14].

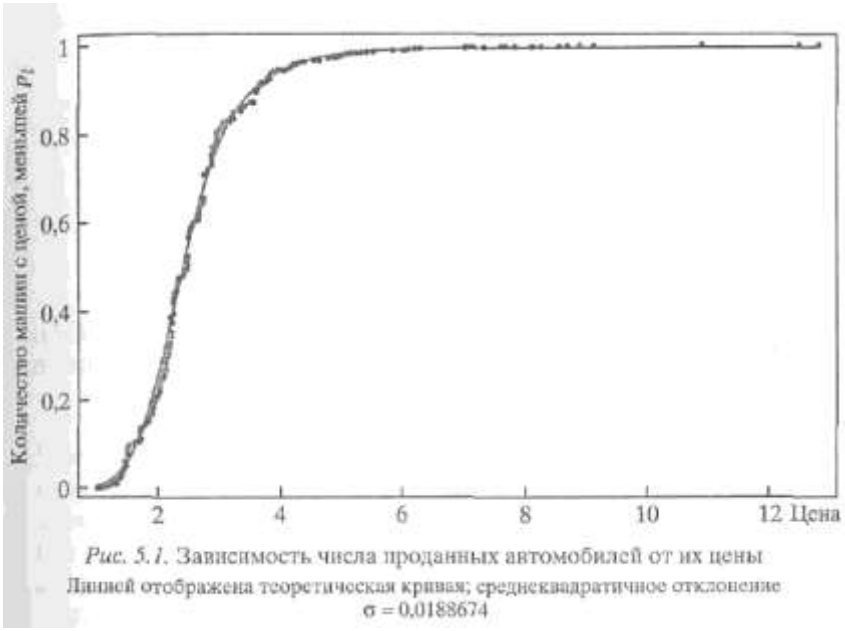

Рис. 5.1. Зависимость числа проданных автомобилей от их цены
Линией отображена теоретическая кривая; среднеквадратичное отклонение
σ = 0,0188674

Рис. 1. Рисунок 5.1 из книги [14]

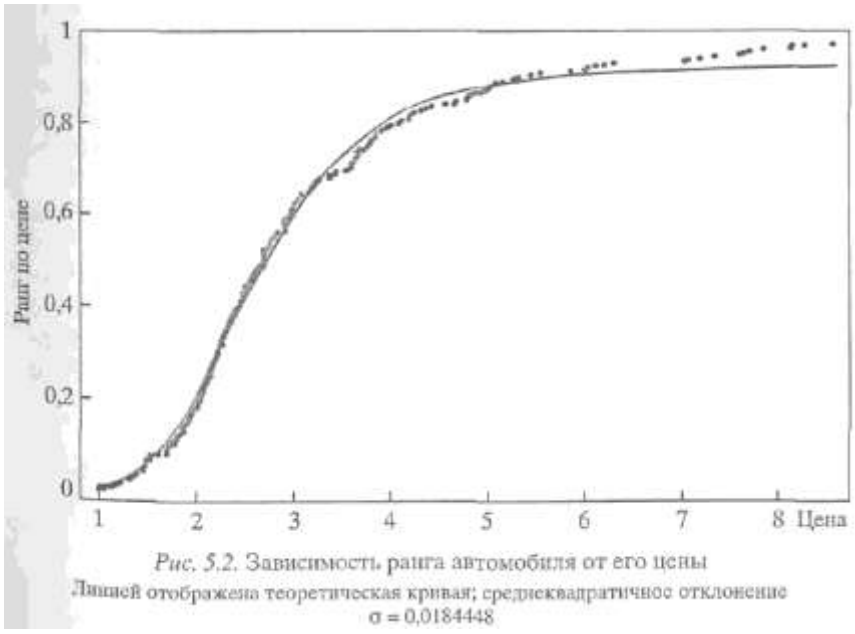

Рис. 5.2. Зависимость ранга автомобиля от его цены
Линией отображена теоретическая кривая; среднеквадратичное отклонение
σ = 0,0184448

Рис. 2. Рисунок 5.2 из книги [14]

Согласно В.П. Маслову, у этих графиков, вообще говоря, должен наблюдаться перегиб, причём (если цены установлены адекватно) точка перегиба обоих графиков должна быть около одной и той же цены x_0. При $b = 0$ и при $\sigma = \gamma$, $b = \frac{1}{a}$ берётся интеграл, что может помочь в счёте [14].

Пусть минимальная цена товара равна l денежных единиц. Тогда можно полагать, что остальные цены меняются на единицу, а значит, $p_i = i$ единиц. Отсюда в силу [13]:

$$N_p \sim \int_1^p \frac{c}{e^{\beta x\left(1+ax^{\gamma}+bx^{-\sigma}\right)}-1}\,dx\,.$$

Эти параметры определяются, прежде всего, из асимптотики приведённой формулы. Основной случай отвечает экспериментальным данным $\beta << 1$. При $\beta \to 0$ имеем

$$N_p \sim c_1 \int_1^p \frac{dx}{x\left(1 + ax^{\gamma} + bx^{-\sigma}\right)},\ c_1 = \text{const}. \qquad (6)$$

Одна из констант может быть определена из условия

$$\left(\frac{1}{x + ax^{\gamma} + bx^{-\sigma}}\right)' = 0 \qquad (7)$$

в точке перегиба экспериментального графика. Если «хвост» ниже x_0 недлинный, то можно для простоты положить $\sigma = \gamma,\ b = \frac{1}{a}$, отсюда

$$N_p = \frac{c_2}{1 + \kappa p^{\gamma}} + c_3,$$

κ, c_2, c_3, γ — константы. Тогда[2]

$$p \cong \alpha\left(\frac{N_p}{N_\infty - N_p}\right)^{\gamma}. \qquad (8)$$

Здесь N_p — номер (ранг) автомобиля, отсчитываемый от самых дешёвых, $(N_\infty - N_p)$ — номер автомобиля, отсчитываемый от самого дорогого в задаче 2. В задаче 1 N_p — это число проданных машин по цене, меньшей p; $(N_\infty - N_p)$ — число проданных машин по цене, равной или большей p [14].

Выполненные В.П. Масловым исследования показали, что для объектов, объединённых некоторым набором признаков, т.е. для определённой группы или кластера, существуют зависимости между соответствующими переменными модели, например, в виде (5) или (8). Тогда существенной характеристикой кластера являются параметры (γ, α, c), входящие в эти функциональные зависимости. Если данные следует выделить в несколько кластеров, то способ разбиения можно сформулировать следующим образом: на каждом из кластеров справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров, которые меняются при переходе от кластера к кластеру. Предварительный анализ показал, что функциональные зависимости (5) и (8) наиболее чувствительны к выбору γ. Поэтому естественное разбиение должно быть таковым, что для каждого кластера существует своё числовое значение степенного параметра γ, характеризующее соответствующий кластер [7].

В статье [7] модифицированный В.П. Масловым закон Ципфа использован для кластеризации медицинских данных.

Среди медицинских работников скорой помощи на примере нескольких городов Приморского края было проведено анкетирование, при котором выявлялась самооценка эмоционального состояния каждого работника (диагностика синдрома эмоционального выгорания, далее СЭВ, в структуре профессионально обусловленной патологии). На вопросы предполагались однозначные ответы («да» или «нет»). СЭВ включает три фазы: «напряжение», «резистенция» (сопротивление) и «истощение». По результатам опроса медицинских работников были подготовлены индивидуальные нейропсихические заключения, в которых на каждую фазу приходится некоторое количество баллов (абсолютное значение), рассчитанное по методике диагностики. Задача состоит в том, чтобы для отдельной специальности (врачи, фельдшеры, медицинские сёстры) выявить характерное разбиение на группы и сделать вывод о выраженности каждой фазы эмоционального состояния в группах риска [7].

В соответствии со сформулированной выше идеей рангового анализа медицинские работники каждой специальности были упорядочены в порядке возрастания абсолютного значения w показателя СЭВ отдельно по трём фазам и каждому значению w поставлен в соответствие порядковый номер — ранг r. Исходные точки анализировались с помощью соотношения (8), которое в логарифмических координатах записано в виде:

[2] В [7] в этой формуле стоит знак $\cong$, а в [14] — знак =.

$$\ln w \cong -\gamma \ln\left(\frac{N-r}{r}\right) + c \equiv -\gamma \ln R + c\,. \qquad (9)$$

Для обеспечения неотрицательности логарифма и удобства визуализации данных принято $N = 2n+1$, где n — количество диагностируемых работников данной специальности. Таким образом, естественными переменными для анализа эмпирических данных являются $\ln w$ и $\ln R$. В этих переменных данные о медицинских сёстрах разбиваются на кластеры в каждой фазе СЭВ. На рис. 3, 4 это представлено для фаз «резистенции» и «истощения», соответствующие кластеры обозначены как группа 1, группа 2, группа 3. Область особых точек составляют медработники с самым большим (48 лет) и самым маленьким (1 год) стажем работы. Для них профессиональные факторы не оказывают существенного влияния на их эмоциональную устойчивость [7].

Рецепт В.П. Маслова может быть применён для оценки влияющих факторов риска, таких как возраст и стаж медицинских работников, на выраженность СЭВ. В результате получены три характерных кластера работников, имеющих стаж до 10 лет, от 11 до 25 лет и свыше 25 лет (на рис. 3, 4 обозначены кружками). При этом оказалось, что фаза повышенной «резистенции» (группа 3) выявлена у контингента медицинских сестёр в возрасте от 38 до 55 лет и со стажем работы от 11 до 25 лет. Работники со стажем свыше 25 лет попали в фазу пониженной «резистенции», соответствующую группам 2 и 3. Фаза повышенного «истощения» выявлена у контингента медицинских сестёр (группа 3) в возрасте от 50 лет и со стажем работы от 25 лет и старше (рис. 4) [7].

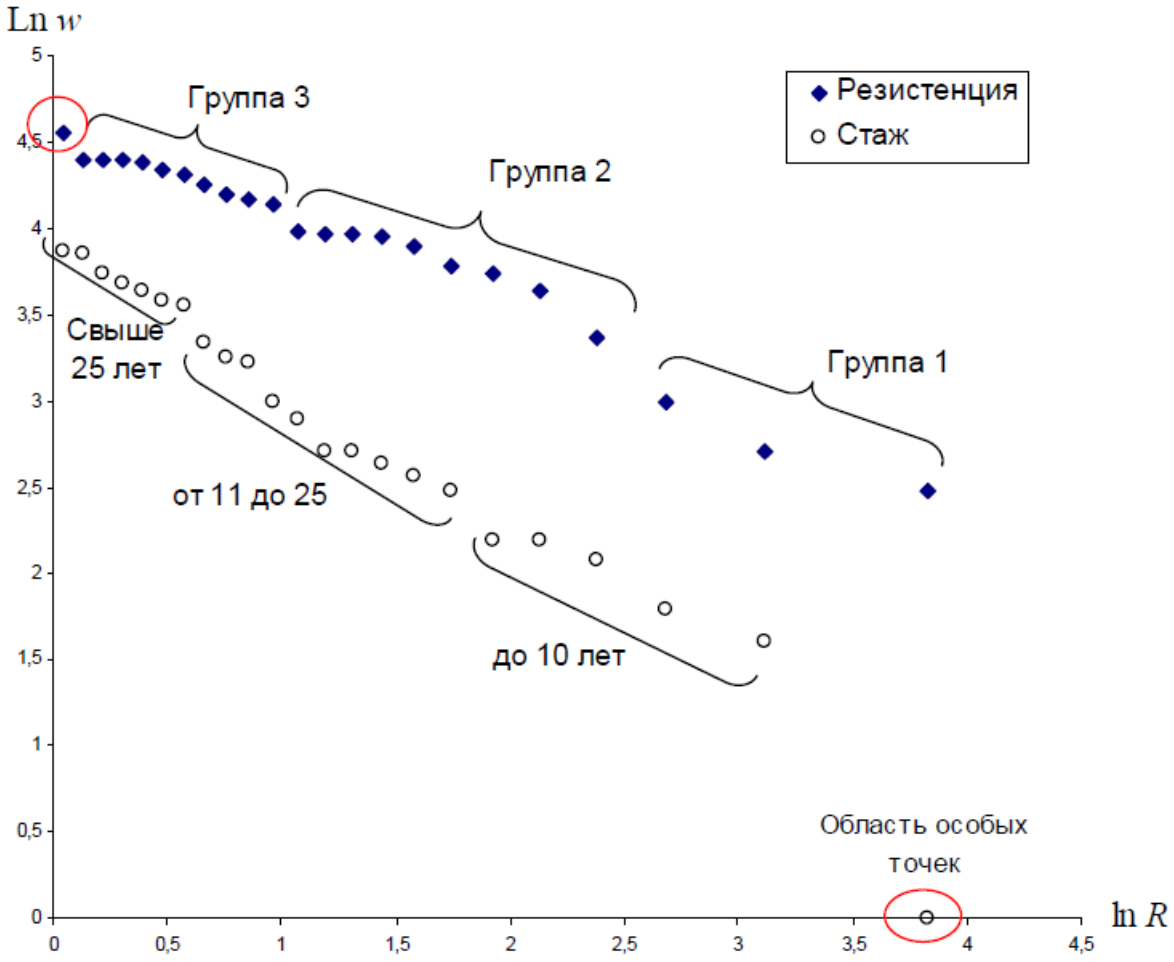

Рис. 3. Выраженность фазы «резистенции» синдрома эмоционального выгорания медицинских сестёр по стажу работы (рисунок из статьи [7])

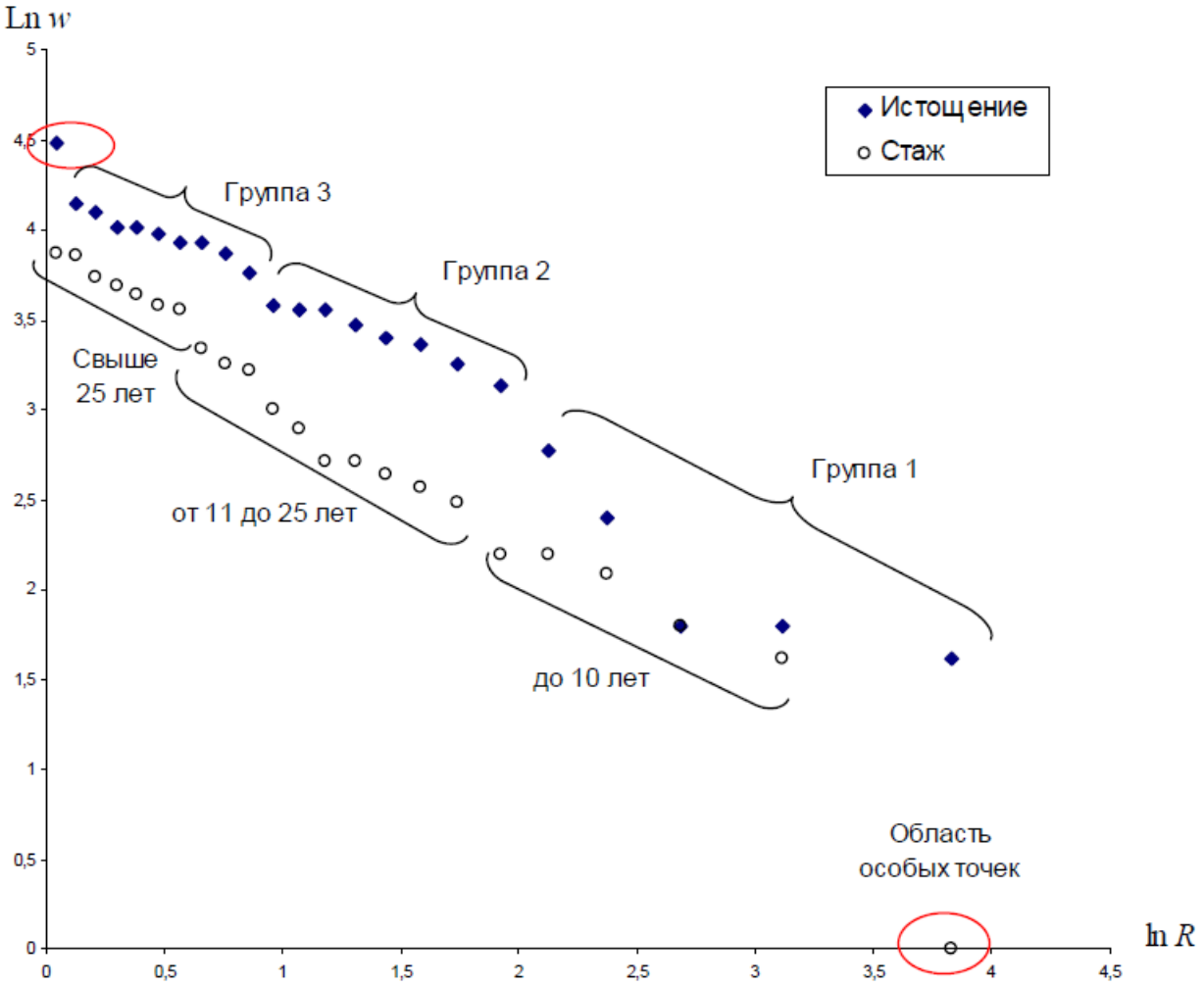

Рис. 4. Выраженность фазы «истощения» синдрома эмоционального выгорания медицинских сестёр по стажу работы (рисунок из статьи [7]).

Вторая задача связана с выявлением групп пациентов, страдающих хроническими заболеваниями желудочно-кишечного тракта, по степени тяжести заболевания исходя из показателей различных видов анализа крови. Сначала все пациенты были разделены на две части по степеням (средняя и тяжёлая) физического состояния больного на основе общего осмотра. Затем пациенты каждой степени состояния ранжированы по возрастанию абсолютных значений показателя выполненного химического анализа крови, и данные проанализированы с использованием ранговой зависимости (9). С помощью вышеизложенного подхода для кластеризации медицинских данных получено разбиение обследованных пациентов на кластеры. На рис. 5 для группы тяжёлой степени состояния отражены четыре кластера разными видами символов по данным анализа Prooxy. Группы, обозначенные светлыми и тёмными ромбиками, соответствуют пациентам, которым необходимо стационарное лечение. Группы светлых и тёмных кружков соответствуют пациентам, которым достаточно амбулаторное лечение [7].

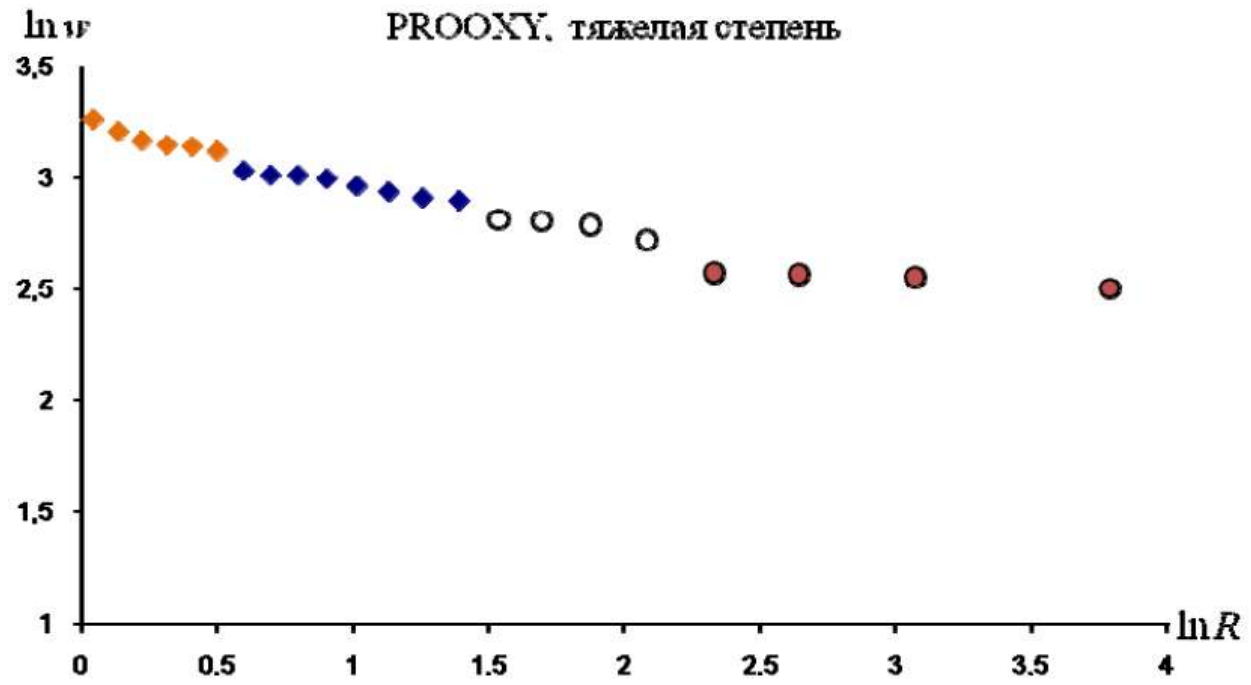

Рис. 5. Кластеризация «тяжёлых» пациентов по результатам анализа крови (рисунок из статьи [7])

1.2. Неформальная постановка задачи

Требуется автоматизировать процесс кластеризации эмпирических данных ранговым методом.

В статье [7] указана только общая идея кластеризации. Однако для вычислительной реализации рангового метода кластеризации необходимо уточнить общую идею, требуется построение математической модели.

Программа для кластеризации эмпирических данных ранговым методом должна принимать на вход исходные данные в формате, который похож на формат CSV, и выводить результаты в формате, удобном для чтения, например, в формате HTML.

Вначале предполагалось, что программа должна быть достаточно простой и доступной для любого пользователя, при этом число задаваемых пользователем параметров не должно быть слишком велико, чтобы пользователь не математик мог разобраться в программе.

Но потом от этого требования пришлось отказаться, так как сам ранговый метод кластеризации нуждается в исследовании и проверке (хотя программа всё равно должна быть простой и понятной).

Во-первых, для каких данных можно применять этот метод? В.П. Маслов в работе [14] рассматривал соотношение между рангом и ценой автомобиля, т.е. соотношение между рангом и целочисленными данными. А в статье [7] соотношение В.П. Маслова использовано для анализа не целочисленных данных. Какие ограничения нужно наложить на исходные данные, чтобы для них можно было бы применять ранговый метод, и в каких единицах измерения нужно задавать данные?

Во-вторых, какие входные параметры для программы являются оптимальными? Вероятно, нужно разработать рекомендации пользователям не математикам по заданию значений параметров.

В-третьих, каково минимальное число точек в кластере? Накладывается ограничение, что в кластере должно быть не менее трёх точек. Но в ранговом методе кластеризации каждый кластер рассматривается как система объектов, для которой справедлив модифицированный В.П. Масловым закон Ципфа, т.е. на каждом кластере должно быть ранговое распределение определённого вида. А можно ли вообще строить ранговое

распределение по трём точкам? Что, если минимальное число точек в кластере должно быть равно 10, 20, 50, 100?

В-четвёртых, как пользователь не математик должен интерпретировать результаты, которые выводит программа?

И наконец, в каких случаях можно говорить о действительно научных выводах, сделанных при помощи рангового метода кластеризации?

Итак, ранговый метод кластеризации следует считать ещё не до конца проверенным и опробованным. Поэтому разрабатываемая программная система предназначена в первую очередь для *исследования* рангового метода кластеризации и в настоящий момент разрабатывается для использования только на кафедре информатики.

Отметим, что предполагаемые пользователи владеют навыками программирования. Поэтому система разрабатывается, с одной стороны, как консольная программа с текстовым интерфейсом ввода, а с другой стороны, как совокупность модулей, которые можно подключать к другим программам, т.е. допускается возможность доработки программной системы пользователями.

Это не обязательно радикальная доработка. Можно, например, модифицировать функциональную зависимость, используемую для аппроксимации точек кластеров, путём передачи функции в качестве параметра шаблона. Такой способ ввода данных в программу, когда пользователь пишет небольшую подпрограмму, а затем вся система компилируется заново вместе с этой подпрограммой, используется в программе MINOS, которая предназначена для решения задач оптимизации [18, 32]. Это очень удобно как для пользователя, так и для разработчика системы. Также пользователь может изменить формат вывода данных, просто изменив модуль вывода.

1.3. Обзор существующих методов решения

По данной теме ранее был написан ряд работ. Это публикации [6, 25-29], а также курсовые и дипломные работы студентов: дипломная работа Зеленова А. [8], курсовая и дипломная работы Пиковой Т. [21, 22], дипломная работа Бидаевой Е. [3].

В дипломной работе Зеленова А. «Критериальная кластеризация квазиодномерных данных» [8] разработана программа, позволяющая находить разбиение данных на кластеры, удовлетворяющее определённому критерию. Входными данными для программы являются одномерные данные. Данные ранжируются, т.е. к координатам точек добавляется ещё одна координата — ранг,— в результате чего данные становятся двумерными.

В программе реализован ряд методов кластеризации. Каждому методу соответствует свой функционал качества разбиения (в [8] функционал качества разбиения называется общей оценкой). Программа вычисляет разбиение, на котором достигается экстремум функционала качества. Так как для всех методов общая оценка равна линейной комбинации частных оценок (в [8] частная оценка — это оценка отдельного кластера или пары соседних кластеров), то для нахождения искомого разбиения был использован метод динамического программирования.

В дипломной работе [8] реализовано много методов кластеризации. Для настройки методов требуется задание дополнительных параметров. Разные методы дают разные результаты. Возникает вопрос: какой метод выбрать?

Программа выводит только разбиение и его оценку. Не выводится никакой дополнительной информации, на основании которой пользователь мог бы сам принять решение, к какому кластеру отнести точку.

Кстати, в дипломной работе Зеленова А. [8] ни слова не сказано о ранговых распределениях. Работа выглядит так, как будто разработанная программа предназначена для кластеризации любых одномерных данных.

Однако из того, что программа Зеленова А. была проинтегрирована в инструментальное средство анализа эмпирических данных методами квантовой статистики SPED [3], мы можем заключить, что в его дипломной работе [8], наверное, речь идёт о

разбиении рангового распределения на участки, а не о кластеризации любых одномерных данных.

Следует отметить, что Зеленов А. не указал в своём отчёте, по какой формуле программа вычисляет ранг, хотя в программе значения рангов используются в вычислениях (например, в методах выпрямления). По картинкам, которые программа выводит на экран, видно, что значения рангов отличны от значений 1, 2, 3, ….

При выполнении обзора литературы нами не было найдено ни одного метода кластеризации одномерных данных, когда к данным добавлялся бы какой-нибудь ранг. В статье [23] проводится кластеризация данных с ранговыми признаками, но здесь других признаков, кроме ранговых, нет, кластеризация проводится по пяти ранговым признакам. В этой статье указано, что в случае ранговых признаков применение евклидовой метрики является некорректным ввиду того, что для ранговых признаков не определены алгебраические операции сложения и умножения, однако для них определена операция сравнения. В этой же статье использована манхэттенская метрика, представляющая собой сумму модулей разностей значений признаков двух объектов:

$$d\left(X_i, X_j\right) = \sum_{k=1}^{p} \left| x_i^{(k)} - x_j^{(k)} \right|.$$

Хотелось бы сделать ещё одно замечание по поводу кластеризации ранжированных данных. Е.В. Черныш была подготовлена презентация [27] о ранговом методе кластеризации. В этой презентации сравнивались результаты, полученные ранговым методом (кластеризация проводилась вручную), и результаты кластеризации, выполненной в статистических пакетах SPSS Statistics 17.0 и Statistica 6.1. Применены метод *k*-средних и евклидово расстояние. Данные двумерные: одна координата — это логарифм абсолютного значения показателя выполненного химического анализа крови $\ln w$, другая координата вычисляется по формуле $\ln R = \ln\left(\frac{N-r}{r}\right)$, где $N = 2n+1$ (n — количество исследуемых пациентов), r — порядковый номер пациента (если расположить пациентов по возрастанию значения показателя анализа крови). Тогда евклидово расстояние между двумя точками будет равно

$$\sqrt{\left(\ln w_1 - \ln w_2\right)^2 + \left(\ln\left(\frac{N-r_1}{r_1}\right) - \ln\left(\frac{N-r_2}{r_2}\right)\right)^2}.$$

Во-первых, чтобы избежать «доминирования» признаков с большим масштабом измерения, обычно проводят предварительную нормировку исходных признаков [4], т.е. вместо обычного евклидова расстояния было бы правильнее использовать «взвешенное» (см. раздел 1.1). Во-вторых, если даже произвести нормировку, складывать квадраты разнородных величин, по нашему мнению, не совсем верно.

В настоящее время сотрудниками кафедры информатики в научной работе используется инструментальное средство анализа данных методами квантовой статистики SPED, разработкой которого занимались студенты [3, 8, 21, 22]. Программа SPED позволяет проводить анализ эмпирических данных по методу В.П. Маслова: строить графики зависимости ранга от частоты, графики изменения различных параметров во времени, а также проводить аппроксимацию данных различными функциями.

Анализируются эмпирические данные, которые являются частотой встречаемости определённого явления. Набор частот ранжируется, т.е. частоты упорядочиваются по возрастанию и каждому элементу набора частот ставится в соответствие значение, равное порядковому номеру элемента в наборе — ранг [22].

Основная аппроксимирующая функция, используемая в программе, была получена путём уточнения параметров рангового распределения (см. формулу (6)):

$$r(x) \sim c_1 \int_1^x \frac{d\xi}{\xi\left(1 + a\xi^{\gamma} + b\xi^{-\sigma}\right)},\ c_1 = \text{const}.$$

В этой формуле полагается $\gamma = \sigma$, тогда результатом интегрирования является следующая функция:

$$r(x) = \begin{cases} \dfrac{c_1}{\gamma\sqrt{D}} \ln\left|\dfrac{(2ax^{\gamma}+1-\sqrt{D})(1+\sqrt{D})}{(2ax^{\gamma}+1+\sqrt{D})(1-\sqrt{D})}\right|, D>0, \\ \dfrac{2c_1}{\gamma\sqrt{-D}} arctg\,\dfrac{2ax^{\gamma}+1}{\sqrt{-D}}, D<0, \end{cases}$$

где $D = 1 - 4ab$.

Коэффициент c_1 находится из условия, что при $x \to \infty$ нормированный ранг $r(x) \to 1$:

$$c_1 = \begin{cases} \dfrac{\gamma\sqrt{D}}{\ln\left|\dfrac{\left(2ax_m{}^{\gamma}+1-\sqrt{D}\right)\left(1+\sqrt{D}\right)}{\left(2ax_m{}^{\gamma}+1+\sqrt{D}\right)\left(1-\sqrt{D}\right)}\right|}, D>0, \\ \dfrac{\sqrt{-D}}{2arctg\,\dfrac{2ax_m{}^{\gamma}+1}{\sqrt{-D}}}, D<0, \end{cases}$$

где x_m — максимальное значение частоты.

Коэффициенты a и b находятся из условия (7):

$$\left(\frac{1}{x+ax^{\gamma}+bx^{-\gamma}}\right)'\Bigg|_{x=x_0} = 0.$$

Отсюда

$$b = \frac{x_0}{\gamma-1}\left(ax_0^{\gamma}(\gamma+1)+1\right),\ a = \frac{1}{kx_0^{\gamma}},$$

где x_0 — точка перегиба экспериментальной кривой.

Таким образом, аппроксимирующая функция имеет три независимых параметра γ, k, x_0, определяющих её поведение. При этом на параметры действуют следующие ограничения: $\gamma > 1,\ 0 \le k \le \dfrac{2(1+\gamma)}{\gamma-1}$, x_0 определяется эмпирическими данными [22].

Производится поиск значений параметров, доставляющих минимум среднеквадратичному отклонению:

$$\text{СКО}^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - r(x_i))^2 \to \min,$$

где r_i — значение ранга для данных с частотой x_i, $r(x_i)$ — значение аппроксимирующей функции в точке с частотой x_i. Для нахождения значений параметров применены методы глобальной оптимизации.

2. Математические методы

2.1. Общая идея рангового метода кластеризации

В.П. Масловым для задачи о числе проданных машин [14] получена следующая теоретическая кривая для числа этих машин N_p по цене, меньшей p:

$$p \cong \alpha\left(\frac{N_p}{N_\infty - N_p}\right)^{\gamma}, \tag{8}$$

где $(N_\infty - N_p)$ — число автомобилей, проданных по цене, равной или большей p.

Выполненные В.П. Масловым исследования показали, что для объектов, объединённых некоторым набором признаков, т.е. для определённой группы или кластера, существуют зависимости между соответствующими переменными модели, например, в виде (5) или (8). Тогда существенной характеристикой кластера являются параметры (γ, α, c), входящие в эти функциональные зависимости.

Если данные следует выделить в несколько кластеров, то способ разбиения можно сформулировать следующим образом: на каждом из кластеров справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров, которые меняются при переходе от кластера к кластеру.

Предварительный анализ показал, что функциональные зависимости (5) и (8) наиболее чувствительны к выбору γ. Поэтому естественное разбиение должно быть таковым, что для каждого кластера существует своё числовое значение степенного параметра γ, характеризующее соответствующий кластер.

2.2. Исходные данные

Исходными данными являются одномерные эмпирические данные — набор положительных чисел $w_1, w_2, ..., w_n$.

Исходные данные упорядочиваются по возрастанию, в результате получается числовая последовательность $\{w_r\}_{r=1}^{n}$, в которой

$$0 < w_1 \le w_2 \le ... \le w_n.$$

Порядковый номер элемента последовательности носит название ранг и обозначается r.

Пусть задана функция $R(r, n)$, определённая на множестве

$$\{(r,n) \mid n \in \mathrm{N}, r \in \{1,2,...,n\}\},$$

принимающая только положительные значения, такая, что функция $R(\cdot, n)$ строго убывает при любом $n \in \mathrm{N}$:

1) $R(r, n) > 0 \ \forall\ n \in \mathrm{N}, \forall\ r \in \{1, 2, ..., n\}$;
2) $R(r, n) > R(r+1, n) \ \forall\ n \in \mathrm{N}, \forall\ r \in \{1, 2, ..., n-1\}$.

При фиксированном n функция $R(\cdot, n)$ представляет собой числовую последовательность $\{R_r\}_{r=1}^{n}$, где $R_r = R(r, n)$.

Частный случай функции $R(r, n)$:

$$R(r,n) = \frac{N(n) - r}{r} = \frac{N(n)}{r} - 1,$$

где $N(n)$ — функция, определённая на множестве N, принимающая значения из множества N, такая, что $N(n) > n \ \forall\ n \in \mathrm{N}$. Примеры таких функций: $N(n) = 2n+1$, $N(n) = n+1$.

2.3. Модифицированный В.П. Масловым закон Ципфа

В статье [7] исходные точки анализировались с помощью соотношения (8), которое в логарифмических координатах записано в виде:

$$\ln w \cong -\gamma \ln\left(\frac{N-r}{r}\right) + c \equiv -\gamma \ln R + c\,, \qquad (9)$$

где для обеспечения неотрицательности логарифма и удобства визуализации данных принято $N = 2n+1$.

Таким образом, естественными переменными для анализа эмпирических данных являются $\ln w$ и $\ln R$ [7].

Отметим, что из формулы (8), полученной В.П. Масловым, следует, что в формуле (9) нужно принять $N = n+1$. Вообще, в дальнейшем мы будем рассматривать обобщение модифицированного В.П. Масловым закона Ципфа с функцией $R(r, n)$, описанной в предыдущем разделе, потому что разрабатываемая программная система предназначена для исследования рангового метода кластеризации эмпирических данных, и может понадобиться попробовать применить для анализа эмпирических данных соотношения, отличные от (9). Например, в случае, когда $R(r, n) = n-r+1$, получим обычный закон Ципфа (см. формулу (1)).

Запишем модифицированный В.П. Масловым закон Ципфа в виде

$$\ln w_r \cong -\gamma \ln R_r + c\,. \qquad (10)$$

Приближённое соотношение (10) означает близость точек $\{(\ln R_r, \ln w_r)\}$ к функциональной зависимости

$$\ln w = l_{\gamma,c}(\ln R) = -\gamma \ln R + c\,. \qquad (11)$$

Здесь под близостью точки ($\ln R_r$, $\ln w_r$) к функциональной зависимости (11) понимается близость логарифма исходного значения $\ln w_r$ к значению линейной функции $l_{\gamma,c}$ от логарифма рангового значения $\ln R_r$.

Если модифицированный В.П. Масловым закон Ципфа рассматривается на промежутке [*a*...*b*] (т.е. для точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$), то этот закон удобно записать в векторной форме:

$$(\ln w_a, \ln w_{a+1}, ..., \ln w_b) \cong -\gamma(\ln R_a, \ln R_{a+1}, ..., \ln R_b) + (c, c, ..., c)\,.$$

Дадим модифицированному В.П. Масловым закону Ципфа геометрическую интерпретацию. Рассмотрим точки $\{(\ln R_r, \ln w_r)\}$ в логарифмических координатах ($\ln R$, $\ln w$). Графиком функции $\ln w = l_{\gamma,c}(\ln R)$ является прямая с угловым коэффициентом $(-\gamma)$ и начальной ординатой *c*. Как было сказано выше, приближённое соотношение (10), представляющее собой модифицированный В.П. Масловым закон Ципфа, означает близость точек $\{(\ln R_r, \ln w_r)\}$ к функциональной зависимости (11). Геометрически близость точки ($\ln R_r$, $\ln w_r$) к функциональной зависимости (11) означает, что ордината этой точки $\ln w_r$ близка к ординате точки графика функции $\ln w = l_{\gamma,c}(\ln R)$, имеющей абсциссу $\ln R_r$.

2.4. Формализация модифицированного В.П. Масловым закона Ципфа

Рассмотрим промежуток [*a*...*b*], содержащий не менее трёх точек[3]:

$$b–a+1 \geq 3 \ (b–a \geq 2).$$

Модифицированный В.П. Масловым закон Ципфа на промежутке [*a*...*b*] имеет вид

$$\ln w_r \cong -\gamma \ln R_r + c, r = a, ..., b\,, \qquad (12)$$

или, в векторной форме,

$$(\ln w_a, \ln w_{a+1}, ..., \ln w_b) \cong -\gamma(\ln R_a, \ln R_{a+1}, ..., \ln R_b) + (c, c, ..., c)\,.$$

Это означает, что точки $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ близки к функциональной зависимости (11), т.е. ординаты точек $\ln w_r$ близки к значениям функции $\ln w = l_{\gamma,c}(\ln R)$ от абсцисс этих точек.

Для измерения близости точек к функциональной зависимости (11) введём функцию $\delta(\ln R_0, \ln w_0, \gamma_0, c_0)$ — меру отклонения точки ($\ln R_0$, $\ln w_0$) от функциональной зависимости (11) со значениями параметров (γ_0, c_0). При этом мы требуем, чтобы функция $\delta(\ln R_0, \ln w_0, \gamma_0, c_0)$ была определена на множестве $R^2 \times R^2$ и принимала неотрицательные значения всюду на этом множестве.

Например,

$$\delta(\ln R_0, \ln w_0, \gamma_0, c_0) = d(\ln w_0, l_{\gamma_0,c_0}(\ln R_0))\,,$$

где $d(q_1, q_2)$ — функция расстояния на прямой,

- $d(q_1, q_2)$ определена на множестве R^2;
- $d(q_1, q_2) \geq 0 \ \forall \ q_1, q_2 \in R$,
- $d(q_1, q_2) = d(q_2, q_1) \ \forall \ q_1, q_2 \in R$.

Примеры таких функций:

1) $d(q_1, q_2) = |q_1 - q_2|$;
2) $d(q_1, q_2) = \left|e^{q_1} - e^{q_2}\right|$

[3] Это ограничение объясняется тем, что при кластеризации эмпирических данных ранговым методом наименьшее количество точек в кластере равно 3 (см. раздел 2.7).

(такое же расстояние в не логарифмических координатах).

Функция $\delta(\ln R_0, \ln w_0, \gamma_0, c_0)$ характеризует близость одной точки к функциональной зависимости. Нужно ввести характеристику близости набора точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ к функциональной зависимости.

Пусть $f\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right)$ — некоторая характеристика близости набора точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ к функциональной зависимости (11) со значениями параметров (γ_0, c_0).

Примеры таких характеристик:

1) $f_1\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right) = \dfrac{1}{b-a+1}\sum_{r=a}^{b}\delta^2(\ln R_r, \ln w_r, \gamma_0, c_0)$

(средний квадрат значения меры отклонения точки промежутка [a...b] от функциональной зависимости (11) со значениями параметров (γ_0, c_0);

2) $f_2\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right) = \sqrt{f_1\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right)}$;

3) $f_3\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right) = \dfrac{1}{b-a+1}\sum_{r=a}^{b}\delta(\ln R_r, \ln w_r, \gamma_0, c_0)$.

Зададим пороговое значение (порог) δ_0 ($\delta_0 \geq 0$) и будем сравнивать значение характеристики f с δ_0: будем считать, что на промежутке [a...b] справедлив модифицированный В.П. Масловым закон Ципфа со значениями параметров (γ_0, c_0) при пороге δ_0, если выполняется неравенство

$$f\left(\{(\ln R_r, \ln w_r)\}_{r=a}^{b}, \gamma_0, c_0\right) \leq \delta_0.$$

Обратимся теперь к ранговому смыслу соотношения (12): это соотношение представляет собой модифицированный закон Ципфа. По нашему мнению, закон Ципфа нужно рассматривать как системное свойство, которому удовлетворяют все объекты рассматриваемой совокупности, и усреднение здесь не приемлемо.

Поэтому потребуем, чтобы значение меры отклонения каждой точки от функциональной зависимости не превосходило пороговое значение:

$$\forall r = a,...,b : \delta(\ln R_r, \ln w_r, \gamma_0, c_0) \leq \delta_0,$$

или, что то же самое,

$$\max_{a \leq r \leq b} \delta(\ln R_r, \ln w_r, \gamma_0, c_0) \leq \delta_0, \qquad (13)$$

и будем считать, что на промежутке [a...b] справедлив модифицированный В.П. Масловым закон Ципфа со значениями параметров (γ_0, c_0) при пороге δ_0, если выполняется условие (13).

Величина

$$M_0(\gamma_0, c_0) = \max_{a \leq r \leq b} \delta(\ln R_r, \ln w_r, \gamma_0, c_0)$$

характеризует близость набора точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ к функциональной зависимости $\ln w = l_{\gamma_0, c_0}(\ln R)$.

Введём предикат

$$P_{\delta_0}(\gamma_0, c_0) = (M_0(\gamma_0, c_0) \leq \delta_0),$$

определяющий, справедлив ли на промежутке [a...b] модифицированный В.П. Масловым закон Ципфа при пороге δ_0 с заданными значениями параметров (γ_0, c_0).

Рассмотрим множество

$$\Phi_{\delta_0} = \{(\gamma, c) \mid P_{\delta_0}(\gamma, c) = 1\}.$$

Множество Φ_{δ_0} определяет множество функциональных зависимостей вида (11), которые годятся для промежутка [a...b] при пороге δ_0 (справедлив модифицированный В.П. Масловым закон Ципфа с такими значениями параметров).

Ещё введём в рассмотрение множество

$$\Gamma_{\delta_0} = \{\gamma \mid \exists c : (\gamma, c) \in \Phi_{\delta_0}\} —$$

множество значений параметра γ, которые годятся для промежутка [a...b] при пороге δ_0.

Теперь «забудем» про пороговое значение. Будем подбирать функциональную зависимость так, чтобы близость набора точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ к ней была наилучшей.

Введём величину

$$\delta_{\min}(0,\infty) = \inf_{(\gamma,c)\in R^2} M_0(\gamma, c) = \inf_{(\gamma,c)\in R^2} \max_{a\le r\le b} \delta(\ln R_r, \ln w_r, \gamma, c).$$

Величина $\delta_{\min}(0,\infty)$ характеризует качество промежутка [a...b].

Введём предикат

$$\Pi_{\delta_0} = (\delta_{\min}(0,\infty) \le \delta_0),$$

определяющий, справедлив ли на промежутке [a...b] модифицированный В.П. Масловым закон Ципфа при пороге δ_0. Заметим, что в случае, когда $\delta_{\min}(0,\infty) = \min_{(\gamma,c)\in R^2} M_0(\gamma, c)$,

$$\Pi_{\delta_0} = (\exists(\gamma, c): P_{\delta_0}(\gamma, c) = 1) = (\Phi_{\delta_0} \ne \varnothing).$$

Укажем геометрический смысл введённых величин.

Пусть

$$d(q_1, q_2) = |q_1 - q_2|,$$

$$\delta(\ln R_0, \ln w_0, \gamma_0, c_0) = d(\ln w_0, l_{\gamma_0, c_0}(\ln R_0)) = |\ln w_0 - (-\gamma_0 \ln R_0 + c_0)|. \qquad (14)$$

Геометрический смысл величины $\delta(\ln R_0, \ln w_0, \gamma_0, c_0)$ — отклонение точки ($\ln R_0$, $\ln w_0$) от прямой

$$\ln w = -\gamma_0 \ln R + c_0 \qquad (15)$$

по вертикали (см. рис. 6).

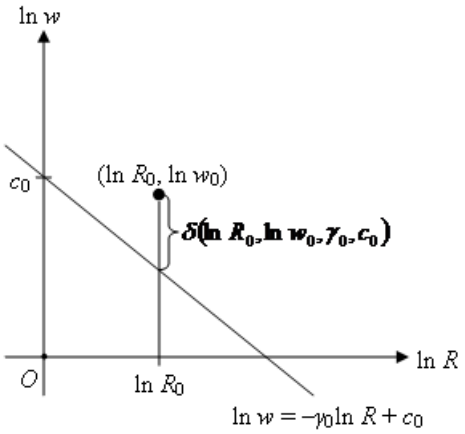

Рис. 6

Величина M_0:

$$M_0(\gamma_0, c_0) = \max_{a\le r\le b} \delta(\ln R_r, \ln w_r, \gamma_0, c_0) = \max_{a\le r\le b} |\ln w_r - (-\gamma_0 \ln R_r + c_0)|.$$

$M_0(\gamma_0, c_0)$ — наибольшее из отклонений точек промежутка [a...b] от прямой (15) по вертикали. Также $M_0(\gamma_0, c_0)$ — это половина минимальной высоты полосы со «средней линией» (15), содержащей все точки промежутка [a...b] (см. рис. 7).

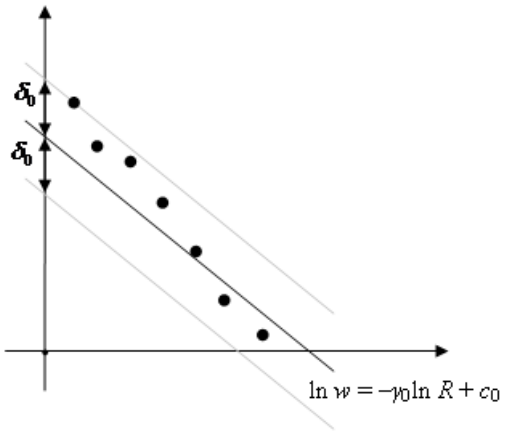

Рис. 7

$P_{\delta_0}(\gamma_0, c_0) = 1$ тогда и только тогда, когда полоса со «средней линией» (6*) высоты $2\delta_0$ содержит все точки промежутка [*a*...*b*].

$\delta_{\min}(0, \infty)$ — это половина наименьшей возможной высоты полосы, содержащей все точки промежутка [*a*...*b*].

$\Pi_{\delta_0} = 1$ тогда и только тогда, когда все точки промежутка [*a*...*b*] можно поместить в полосу высоты $2\delta_0$.

2.5. Множество максимальных промежутков

Пусть *T* — множество промежутков, количество точек в которых не меньше 3:

$$T = \{[a...b] \mid b - a + 1 \geq 3\}.$$

Рассмотрим множество

$$I_{\delta_0} = \{[a...b] \in T \mid \Pi_{\delta_0}^{a,b} = 1\} —$$

множество промежутков, на которых справедлив модифицированный В.П. Масловым закон Ципфа при пороге δ_0.

Введём на множестве *T* отношение ≤. Пусть

$$[a_1...b_1] \leq [a_2...b_2] \Leftrightarrow a_1 \geq a_2, b_1 \leq b_2.$$

Предложение 1. Отношение ≤ рефлексивно, транзитивно и антисимметрично.

Доказательство.

Рефлексивность. $[a...b] \leq [a...b] \Leftrightarrow a \geq a, b \leq b$.

Транзитивность. $[a_1...b_1] \leq [a_2...b_2], [a_2...b_2] \leq [a_3...b_3] \Leftrightarrow a_1 \geq a_2, b_1 \leq b_2, a_2 \geq a_3, b_2 \leq b_3 \Rightarrow$
$\Rightarrow a_1 \geq a_3, b_1 \leq b_3 \Leftrightarrow [a_1...b_1] \leq [a_3...b_3]$.

Антисимметричность. $[a_1...b_1] \leq [a_2...b_2], [a_2...b_2] \leq [a_1...b_1] \Leftrightarrow a_1 \geq a_2, b_1 \leq b_2, a_2 \geq a_1,$
$b_2 \leq b_1 \Rightarrow a_1 = a_2, b_1 = b_2 \Leftrightarrow [a_1...b_1] = [a_2...b_2]$.

Таким образом, отношение ≤ рефлексивно, транзитивно и антисимметрично, т.е. является частичным порядком на *T* [19].

Элемент *a* частично упорядоченного множества *A* называется максимальным, если $a \leq x \Leftrightarrow a = x$ [19]. Множество максимальных элементов множества I_{δ_0} обозначим $J_{\delta_0} = \max I_{\delta_0}$ и назовём множеством максимальных промежутков при пороге δ_0. Таким образом,

$$J_{\delta_0} = \{[a...b] \in I_{\delta_0} \mid \forall [a'...b'] \in I_{\delta_0} : [a...b] \leq [a'...b'] \Leftrightarrow [a...b] = [a'...b']\}.$$

Предложение 2. $\exists \inf_{x \in X} f(x), \inf_{x \in X} g(x), \forall x \in X : f(x) \leq g(x) \Rightarrow \inf_{x \in X} f(x) \leq \inf_{x \in X} g(x)$.

Доказательство.

$m_1 = \inf\limits_{x \in X} f(x), m_2 = \inf\limits_{x \in X} g(x)$.

По определению точной нижней грани [9]:

1) $\forall x \in X : f(x) \geq m_1$;

2) $\forall \varepsilon > 0 : \exists x' \in X : f(x') < m_1 + \varepsilon$.

Для функции $g(x)$:

1) $\forall x \in X : g(x) \geq m_2$;

2) $\forall \varepsilon > 0 : \exists x' \in X : g(x') < m_2 + \varepsilon$.

Предположим противное: $m_1 > m_2$.

Пусть $\varepsilon = \dfrac{m_1 - m_2}{2}$.

$$\exists x' \in X : g(x') < m_2 + \varepsilon = m_2 + \frac{m_1 - m_2}{2} = \frac{m_1 + m_2}{2},$$

$$f(x') \leq g(x') < \frac{m_1 + m_2}{2},$$

$$f(x') \geq m_1,$$

$$m_1 \leq f(x') < \frac{m_1 + m_2}{2} < m_1.$$

Противоречие.

Рассмотрим, как изменяется $\delta_{\min}^{a,b}(0,\infty)$ при изменении промежутка $[a...b]$.

Предложение 3. $[a_1...b_1] \leq [a_2...b_2] \Rightarrow \delta_{\min}^{a_1,b_1}(0,\infty) \leq \delta_{\min}^{a_2,b_2}(0,\infty)$.

Доказательство.

$$[a_1...b_1] \leq [a_2...b_2] \Rightarrow \forall(\gamma,c): \{\delta(\ln R_r, \ln w_r, \gamma, c) \mid a_1 \leq r \leq b_1\} \subset \{\delta(\ln R_r, \ln w_r, \gamma, c) \mid a_2 \leq r \leq b_2\}$$

$$\Rightarrow \forall(\gamma,c): \max_{a_1 \leq r \leq b_1} \delta(\ln R_r, \ln w_r, \gamma, c) \leq \max_{a_2 \leq r \leq b_2} \delta(\ln R_r, \ln w_r, \gamma, c) \Leftrightarrow \forall(\gamma,c): M_0^{a_1,b_1}(\gamma,c) \leq M_0^{a_2,b_2}(\gamma,c)$$

$$\Rightarrow \inf_{(\gamma,c) \in \mathrm{R}^2} M_0^{a_1,b_1}(\gamma,c) \leq \inf_{(\gamma,c) \in \mathrm{R}^2} M_0^{a_2,b_2}(\gamma,c) \text{ (по предложению 2)} \Leftrightarrow \delta_{\min}^{a_1,b_1}(0,\infty) \leq \delta_{\min}^{a_2,b_2}(0,\infty).$$

Предложение 4. $[a_1...b_1] \leq [a_2...b_2], [a_2...b_2] \in I_{\delta_0} \Rightarrow [a_1...b_1] \in I_{\delta_0}$.

Доказательство.

$$[a_1...b_1] \leq [a_2...b_2] \Rightarrow \delta_{\min}^{a_1,b_1}(0,\infty) \leq \delta_{\min}^{a_2,b_2}(0,\infty),$$

$$[a_2...b_2] \in I_{\delta_0} \Leftrightarrow \Pi_{\delta_0}^{a_2,b_2} = 1 \Leftrightarrow \delta_{\min}^{a_2,b_2}(0,\infty) \leq \delta_0 \Rightarrow \delta_{\min}^{a_1,b_1}(0,\infty) \leq \delta_0 \Leftrightarrow \Pi_{\delta_0}^{a_1,b_1} = 1 \Leftrightarrow$$

$$\Leftrightarrow [a_1...b_1] \in I_{\delta_0}.$$

2.6. Дальнейшая формализация с учётом аномальных точек

Пусть $\delta_0 \geq 0$ — пороговое значение. Точки ($\ln R_r$, $\ln w_r$), для которых

$$\delta(\ln R_r, \ln w_r, \gamma_0, c_0) > \delta_0,$$

назовём аномальными точками (аномалиями) относительно функциональной зависимости (11) со значениями параметров (γ_0, c_0).

Рассмотрим величину

$$V_{\delta_0}(\gamma_0, c_0) = \left|\{r \in \{a,...,b\} \mid \delta(\ln R_r, \ln w_r, \gamma_0, c_0) > \delta_0\}\right| \text{ —}$$

количество точек промежутка $[a...b]$, для которых значение меры отклонения от функциональной зависимости $\ln w = l_{\gamma_0,c_0}(\ln R)$ больше δ_0. $V_{\delta_0}(\gamma_0, c_0)$ — число аномалий промежутка $[a...b]$ относительно функциональной зависимости (11) со значениями параметров (γ_0, c_0).

Пусть $\{z_i\}_{i=1}^{s}$ — некоторая последовательность неотрицательных чисел и пусть j — натуральное число, удовлетворяющее неравенству $1 \leq j \leq s+1$. Обозначим

$$\max_j z_i$$

j-й по максимальности элемент последовательности $\{z_i\}_{i=1}^{s}$ (j-й по порядку элемент в отсортированной по убыванию последовательности $\{z_i\}_{i=1}^{s}$). В частности, $\max_1 z_i = \max z_i$, $\max_s z_i = \min z_i$. В случае $j = s+1$ примем $\max_{s+1} z_i = 0$.

Пусть ν — целое число, удовлетворяющее неравенству $0 \le \nu \le b-a+1$. Введём величину

$$M_\nu(\gamma_0, c_0) = \max_{a \le r \le b}{}_{\nu+1}\, \delta(\ln R_r, \ln w_r, \gamma_0, c_0),$$

характеризующую близость набора точек $\{(\ln R_r, \ln w_r)\}_{r=a}^{b}$ к функциональной зависимости $\ln w = l_{\gamma_0, c_0}(\ln R)$, допуская ν аномальных точек.

Фактически мы рассматриваем последовательность значений меры отклонения точек промежутка [a...b] от функциональной зависимости (11) со значениями параметров (γ_0, c_0):

$$\{\delta(\ln R_r, \ln w_r, \gamma_0, c_0)\}_{r=a}^{b}. \qquad (16)$$

$M_\nu(\gamma_0, c_0)$ равняется (ν+1)-му по максимальности элементу последовательности (16). $V_{\delta_0}(\gamma_0, c_0)$ равняется количеству элементов последовательности (16), превосходящих δ_0.

Предложение 5. Справедливы следующие свойства величин $V_{\delta_0}(\gamma_0, c_0)$ и $M_\nu(\gamma_0, c_0)$:

а) $\delta_0' < \delta_0'' \Rightarrow V_{\delta_0'}(\gamma_0, c_0) \ge V_{\delta_0''}(\gamma_0, c_0)$;

б) $\nu' < \nu'' \Rightarrow M_{\nu'}(\gamma_0, c_0) \ge M_{\nu''}(\gamma_0, c_0)$;

в) $M_\nu(\gamma_0, c_0) = \min\{\delta_0 \mid V_{\delta_0}(\gamma_0, c_0) \le \nu\}$;

г) $V_{\delta_0}(\gamma_0, c_0) = \min\{\nu \mid M_\nu(\gamma_0, c_0) \le \delta_0\}$;

д) $M_\nu(\gamma_0, c_0) \le \delta_0 \Leftrightarrow V_{\delta_0}(\gamma_0, c_0) \le \nu$.

Доказательство.

Пусть $\{\tilde{\delta}_i\}_{i=1}^{b-a+1}$ — последовательность (16), упорядоченная по убыванию, $\tilde{\delta}_{b-a+2} = 0$.

$\tilde{\delta}_i = \max_{a \le r \le b}{}_i\, \delta(\ln R_r, \ln w_r, \gamma_0, c_0), i = 1,2,...b-a+2$,

$M_\nu(\gamma_0, c_0) = \tilde{\delta}_{\nu+1}$,

$V_{\delta_0}(\gamma_0, c_0) = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right|$.

а) $\delta_0' < \delta_0'' \Rightarrow \left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0''\right\} \subset \left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0'\right\} \Rightarrow$

$$\Rightarrow V_{\delta_0''}(\gamma_0, c_0) = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0''\right\}\right| \le \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0'\right\}\right| = V_{\delta_0'}(\gamma_0, c_0).$$

б) Последовательность $\{\tilde{\delta}_i\}_{i=1}^{b-a+2}$ не возрастает (по определению).

$\nu' < \nu'' \Rightarrow M_{\nu'}(\gamma_0, c_0) = \tilde{\delta}_{\nu'+1} \ge \tilde{\delta}_{\nu''+1} = M_{\nu''}(\gamma_0, c_0)$.

в) $V_{\tilde{\delta}_{\nu+1}}(\gamma_0, c_0) = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \tilde{\delta}_{\nu+1}\right\}\right| \le \nu$.

Если $\delta_0 < \tilde{\delta}_{\nu+1}$, то $\{1,2,...,\nu+1\} \subset \left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\} \Rightarrow$

$\Rightarrow V_{\delta_0}(\gamma_0, c_0) = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right| \ge \nu+1 > \nu$.

Итак, $\tilde{\delta}_{\nu+1} \in \{\delta_0 \mid V_{\delta_0}(\gamma_0, c_0) \le \nu\}, \forall \delta_0 < \tilde{\delta}_{\nu+1} : \delta_0 \notin \{\delta_0 \mid V_{\delta_0}(\gamma_0, c_0) \le \nu\}$.

Следовательно, $M_\nu(\gamma_0, c_0) = \tilde{\delta}_{\nu+1} = \min\{\delta_0 \mid V_{\delta_0}(\gamma_0, c_0) \le \nu\}$.

г) Если $\nu < \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right|$, то $\nu+1 \le \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right| \Rightarrow$

$\Rightarrow \tilde{\delta}_{\nu+1} > \delta_0$. Если $\nu = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right|$, то $\Rightarrow \tilde{\delta}_{\nu+1} \le \delta_0$. Следовательно,

$$V_{\delta_0}(\gamma_0, c_0) = \left|\left\{i \in \{1,2,...,b-a+1\} \mid \tilde{\delta}_i > \delta_0\right\}\right| = \min\{\nu \mid \tilde{\delta}_{\nu+1} \le \delta_0\} = \min\{\nu \mid M_\nu(\gamma_0, c_0) \le \delta_0\}.$$

д) Пусть $M_\nu(\gamma_0,c_0)\le\delta_0$. По свойству в) $M_\nu(\gamma_0,c_0)=\min\{\delta_0\,|\,V_{\delta_0}(\gamma_0,c_0)\le\nu\}\Rightarrow$
$\Rightarrow V_{M_\nu(\gamma_0,c_0)}(\gamma_0,c_0)\le\nu$. $M_\nu(\gamma_0,c_0)\le\delta_0\Rightarrow$ по свойству а) $V_{\delta_0}(\gamma_0,c_0)\Rightarrow V_{M_\nu(\gamma_0,c_0)}(\gamma_0,c_0)\le\nu$.

Пусть $V_{\delta_0}(\gamma_0,c_0)\le\nu$. По свойству г) $V_{\delta_0}(\gamma_0,c_0)=\min\{\nu\,|\,M_\nu(\gamma_0,c_0)\le\delta_0\}\Rightarrow$
$M_{V_{\delta_0}(\gamma_0,c_0)}(\gamma_0,c_0)\le\delta_0$. $V_{\delta_0}(\gamma_0,c_0)\le\nu\Rightarrow$ по свойству б) $M_\nu(\gamma_0,c_0)\le M_{V_{\delta_0}(\gamma_0,c_0)}(\gamma_0,c_0)\le\delta_0$.

Изобразим связь между величинами $M_\nu(\gamma_0,c_0)$ и $V_{\delta_0}(\gamma_0,c_0)$ графически (см. рис. 8).

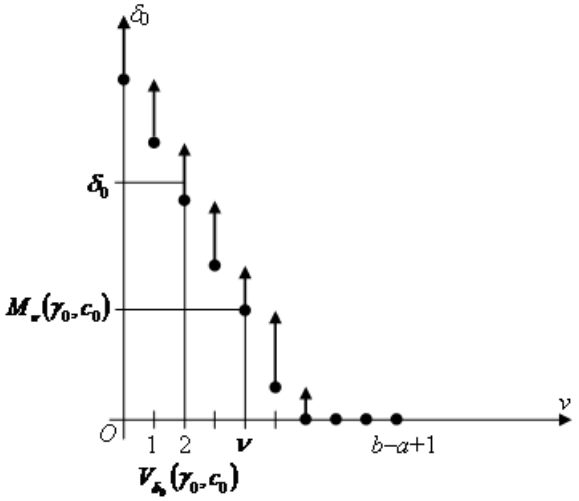

Рис. 8

Геометрический смысл величины $M_\nu(\gamma_0,c_0)$ — половина наименьшей возможной высоты полосы со «средней линией» (15), содержащей все точки промежутка [*a*...*b*], кроме *ν* штук точек, которые можно выкинуть (см. рис. 9). $M_\nu(\gamma_0,c_0)$ — половина наименьшей высоты полосы со «средней линией» (15), при которой число аномалий не превосходит *ν* (свойство в) предложения 5). $V_{\delta_0}(\gamma_0,c_0)$ — количество точек промежутка [*a*...*b*], находящихся вне полосы со «средней линией» (15) высоты $2\delta_0$.

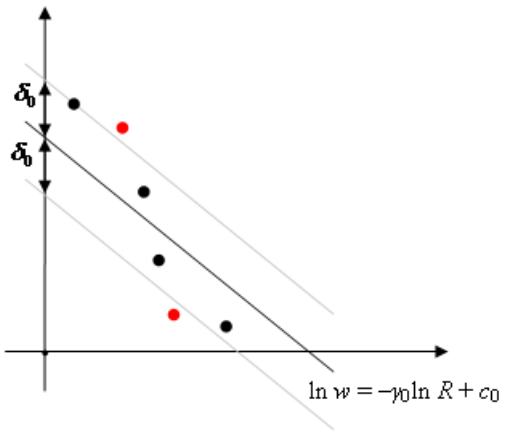

Рис. 9

Зададим дополнительное пороговое значение $\delta_0'>\delta_0$ (δ_0' может быть равно ∞) и потребуем, чтобы для всех аномальных точек промежутка [*a*...*b*] было выполнено неравенство

$$\delta(\ln R_r, \ln w_r, \gamma_0, c_0) \le \delta_0'. \qquad (17)$$

Тройку $L = \left(\delta_0, \nu_0, \delta_0'\right)$ назовём уровнем качества промежутка.

Заметим, что неравенство (17) выполнено для всех аномальных точек промежутка [*a*...*b*] тогда и только тогда, когда

$$M_0(\gamma_0, c_0) \le \delta_0'.$$

Поэтому в дальнейшем мы будем рассматривать пары значений параметров, принадлежащие множеству

$$\Phi_{\delta_0'} = \left\{(\gamma, c) \mid M_0(\gamma, c) \le \delta_0'\right\}.$$

Введём предикат

$$P_L(\gamma_0, c_0) = \left((\gamma_0, c_0) \in \Phi_{\delta_0'}\right) \wedge \left(M_{\nu_0}(\gamma_0, c_0) \le \delta_0\right) = \left(M_0(\gamma_0, c_0) \le \delta_0'\right) \wedge \left(M_{\nu_0}(\gamma_0, c_0) \le \delta_0\right),$$

определяющий, справедлив ли на промежутке [*a*...*b*] модифицированный В.П. Масловым закон Ципфа на уровне *L* с заданными значениями параметров (γ_0, c_0).

Рассмотрим множество

$$\Phi_L = \left\{(\gamma, c) \mid P_L(\gamma, c) = 1\right\},$$

которое определяет множество функциональных зависимостей вида (11), которые годятся для промежутка [*a*...*b*] на уровне *L*.

Рассмотрим множество

$$\Gamma_L = \left\{\gamma \mid \exists c : (\gamma, c) \in \Phi_L\right\} —$$

множество значений параметра γ, которые годятся для промежутка [*a*...*b*] на уровне *L*.

Введём величину

$$\delta_{\min}\left(\nu_0, \delta_0'\right) = \inf_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma, c) = \inf_{(\gamma,c)\in\Phi_{\delta_0'}} \max_{\nu_0+1 \le r \le b} \delta(\ln R_r, \ln w_r, \gamma, c),$$

причём если $\Phi_{\delta_0'} = \varnothing$, то величина $\delta_{\min}\left(\nu_0, \delta_0'\right)$ не определена. Величина $\delta_{\min}\left(\nu_0, \delta_0'\right)$ характеризует качество промежутка [*a*...*b*] с допущением ν_0 аномальных точек при дополнительном пороге δ_0'.

Введём величину

$$\nu_{\min}\left(\delta_0, \delta_0'\right) = \min_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}(\gamma, c) = \min_{(\gamma,c)\in\Phi_{\delta_0'}} \left|\left\{r \in \{a,...,b\} \mid \delta(\ln R_r, \ln w_r, \gamma, c) > \delta_0\right\}\right|,$$

причём если $\Phi_{\delta_0'} = \varnothing$, то величина $\nu_{\min}\left(\delta_0, \delta_0'\right)$ не определена. Величина $\nu_{\min}\left(\delta_0, \delta_0'\right)$ также характеризует качество промежутка [*a*...*b*] при пороге δ_0 и дополнительном пороге δ_0'.

Предложение 6. Справедливы следующие свойства величин $\delta_{\min}\left(\nu_0, \delta_0'\right)$ и $\nu_{\min}\left(\delta_0, \delta_0'\right)$:

а) $\delta_0^{(1)} < \delta_0^{(2)} \Rightarrow \nu_{\min}\left(\delta_0^{(1)}, \delta_0'\right) \ge \nu_{\min}\left(\delta_0^{(2)}, \delta_0'\right)$;

б) $\nu_0^{(1)} < \nu_0^{(2)} \Rightarrow \delta_{\min}\left(\nu_0^{(1)}, \delta_0'\right) \ge \delta_{\min}\left(\nu_0^{(2)}, \delta_0'\right)$;

в) $\delta_{\min}\left(\nu_0, \delta_0'\right) = \inf\left\{\delta_0 \mid \nu_{\min}\left(\delta_0, \delta_0'\right) \le \nu_0\right\}$;

г) если $\delta_{\min}\left(\nu_0, \delta_0'\right) = \min_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma, c)$, то $\nu_{\min}\left(\delta_0, \delta_0'\right) = \min\left\{\nu_0 \mid \delta_{\min}\left(\nu_0, \delta_0'\right) \le \delta_0\right\}$;

д) если $\delta_{\min}\left(\nu_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)$, то $\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0 \Leftrightarrow \nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0$.

Доказательство.

Предположим, что $\Phi_{\delta_0'}\neq\varnothing$. Тогда $\delta_{\min}\left(\nu_0,\delta_0'\right)$ и $\nu_{\min}\left(\delta_0,\delta_0'\right)$ определены.

а) По свойству а) предложения 5 $\forall(\gamma,c)\in\Phi_{\delta_0'}: V_{\delta_0^{(1)}}(\gamma,c)\ge V_{\delta_0^{(2)}}(\gamma,c)\Rightarrow$

$\nu_{\min}\left(\delta_0^{(1)},\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0^{(1)}}(\gamma,c)\ge\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0^{(2)}}(\gamma,c)=\nu_{\min}\left(\delta_0^{(2)},\delta_0'\right)$.

б) По свойству б) предложения 5 $\forall(\gamma,c)\in\Phi_{\delta_0'}: M_{\nu_0^{(1)}}(\gamma,c)\ge M_{\nu_0^{(2)}}(\gamma,c)\Rightarrow$ по предложению 2 $\delta_{\min}\left(\nu_0^{(1)},\delta_0'\right)=\inf\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0^{(1)}}(\gamma,c)\ge\inf\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0^{(2)}}(\gamma,c)=\delta_{\min}\left(\nu_0^{(2)},\delta_0'\right)$.

в) Пусть $\delta_0<\delta_{\min}\left(\nu_0,\delta_0'\right)$. $\delta_0<\inf\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)\Rightarrow\forall(\gamma,c)\in\Phi_{\delta_0'}: M_{\nu_0}(\gamma,c)>\delta_0\Rightarrow$

$\Rightarrow\forall(\gamma,c)\in\Phi_{\delta_0'}: V_{\delta_0}(\gamma,c)>\nu_0$ (по свойству д) предложения 5) $\Rightarrow$

$\Rightarrow\nu_{\min}\left(\delta_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}(\gamma,c)>\nu_0\Rightarrow\delta_0\notin\left\{\delta_0\mid\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\right\}$.

Пусть $\delta_0>\delta_{\min}\left(\nu_0,\delta_0'\right)$. $\delta_0>\inf\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)\Rightarrow\exists(\gamma_0,c_0)\in\Phi_{\delta_0'}: M_{\nu_0}(\gamma_0,c_0)<\delta_0\Rightarrow$

$\Rightarrow\exists(\gamma_0,c_0)\in\Phi_{\delta_0'}: V_{\delta_0}(\gamma,c)\le\nu_0$ (по свойству д) предложения 5) $\Rightarrow$

$\Rightarrow\nu_{\min}\left(\delta_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}(\gamma,c)\le\nu_0\Rightarrow\delta_0\in\left\{\delta_0\mid\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\right\}$.

Итак, $\forall\delta_0,\delta_0<\delta_{\min}\left(\nu_0,\delta_0'\right):\delta_0\notin\left\{\delta_0\mid\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\right\}$,

$\forall\delta_0,\delta_0>\delta_{\min}\left(\nu_0,\delta_0'\right):\delta_0\in\left\{\delta_0\mid\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\right\}$.

Следовательно, $\inf\left\{\delta_0\mid\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\right\}=\delta_{\min}\left(\nu_0,\delta_0'\right)$.

г) Пусть $\nu_0<\nu_{\min}\left(\delta_0,\delta_0'\right)$. $\nu_0<\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}(\gamma,c)\Rightarrow\forall(\gamma,c)\in\Phi_{\delta_0'}: V_{\delta_0}(\gamma,c)>\nu_0\Rightarrow$

$\Rightarrow\forall(\gamma,c)\in\Phi_{\delta_0'}: M_{\nu_0}(\gamma,c)>\delta_0$ (по свойству д) предложения 5) $\Rightarrow$

$\Rightarrow\delta_{\min}\left(\nu_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)>\delta_0\Rightarrow\nu_0\notin\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}$.

Пусть $\nu_0\ge\nu_{\min}\left(\delta_0,\delta_0'\right)$. $\nu_0\ge\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}(\gamma,c)\Rightarrow\exists(\gamma_0,c_0)\in\Phi_{\delta_0'}: V_{\delta_0}(\gamma_0,c_0)\le\nu_0\Rightarrow$

$\Rightarrow\exists(\gamma_0,c_0)\in\Phi_{\delta_0'}: M_{\nu_0}(\gamma_0,c_0)\le\delta_0$ (по свойству д) предложения 5) $\Rightarrow$

$\Rightarrow\delta_{\min}\left(\nu_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)\le\delta_0\Rightarrow\nu_0\in\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}$.

Итак, $\forall\nu_0,\nu_0<\nu_{\min}\left(\delta_0,\delta_0'\right):\nu_0\notin\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}$,

$\forall\nu_0,\nu_0\ge\nu_{\min}\left(\delta_0,\delta_0'\right):\nu_0\in\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}$.

Следовательно, $\min\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}=\nu_{\min}\left(\delta_0,\delta_0'\right)$.

д) Пусть $\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0$. $\delta_{\min}\left(\nu_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)\Rightarrow$

$\Rightarrow\exists(\gamma_0,c_0)\in\Phi_{\delta_0'}: M_{\nu_0}(\gamma_0,c_0)=\delta_{\min}\left(\nu_0,\delta_0'\right)\Rightarrow$ по свойству д) предложения 5

$V_{\delta_{\min}\left(\nu_0,\delta_0'\right)}(\gamma_0,c_0)\le\nu_0\Rightarrow\nu_{\min}\left(\delta_{\min}\left(\nu_0,\delta_0'\right),\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_{\min}\left(\nu_0,\delta_0'\right)}(\gamma,c)\le\nu_0$.

$\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0,\nu_{\min}\left(\delta_{\min}\left(\nu_0,\delta_0'\right),\delta_0'\right)\le\nu_0\Rightarrow$ по свойству а)

$\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_{\min}\left(\delta_{\min}\left(\nu_0,\delta_0'\right),\delta_0'\right)\le\nu_0$.

Пусть $\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0$. По свойству г) $\nu_{\min}\left(\delta_0,\delta_0'\right)=\min\left\{\nu_0\mid\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right\}\Rightarrow$

$\Rightarrow\delta_{\min}\left(\nu_{\min}\left(\delta_0,\delta_0'\right),\delta_0'\right)\le\delta_0$. $\nu_{\min}\left(\delta_0,\delta_0'\right)\le\nu_0\Rightarrow$ по свойству б)

$\Rightarrow\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_{\min}\left(\nu_{\min}\left(\delta_0,\delta_0'\right),\delta_0'\right)\le\delta_0$.

Введём предикат

$$\Pi_L=\left(\delta_{\min}\left(\nu_0,\delta_0'\right)\text{определена}\right)\wedge\left(\delta_{\min}\left(\nu_0,\delta_0'\right)\le\delta_0\right),$$

определяющий, справедлив ли на промежутке [*a*...*b*] модифицированный В.П. Масловым закон Ципфа на уровне *L*. Заметим, что в случае, когда $\delta_{\min}\left(\nu_0,\delta_0'\right)=\min\limits_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}(\gamma,c)$,

$$\Pi_L=\left(\exists(\gamma,c):P_L(\gamma,c)=1\right)=\left(\Phi_L\ne\varnothing\right).$$

Укажем геометрический смысл введённых величин.

$P_L(\gamma_0,c_0)=1$ тогда и только тогда, когда полоса со «средней линией» (15) высоты $2\delta_0$ содержит все точки промежутка [*a*...*b*], кроме не более ν_0 штук точек, которые содержатся в полосе со «средней линией» (15) высоты $2\delta_0'$ (см. рис. 10).

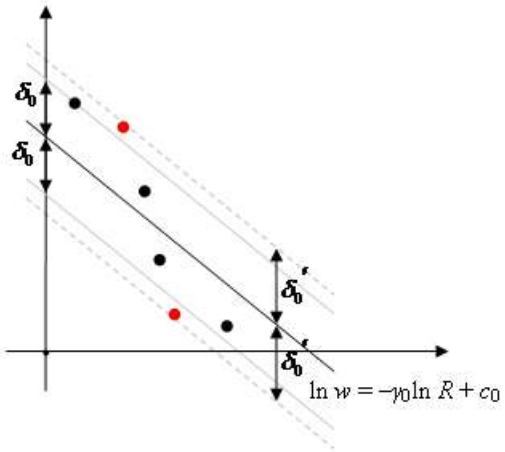

Рис. 10

$\delta_{\min}\left(\nu_0,\delta_0'\right)$ — это половина наименьшей возможной высоты полосы, при которой число точек, не попавших в полосу, не превосходит ν_0, причём эти точки содержатся в полосе с такой же «средней линией» высоты $2\delta_0'$.

$\nu_{\min}\left(\delta_0,\delta_0'\right)$ — это наименьшее возможное число аномалий при пороге δ_0 и дополнительном пороге δ_0', т.е. не более $\nu_{\min}\left(\delta_0,\delta_0'\right)$ штук точек находятся вне полосы высоты $2\delta_0$, но они содержатся в полосе с такой же «средней линией» высоты $2\delta_0'$.

Рассмотрим множество

$$I_L = \left\{[a...b] \in T \mid \Pi_L^{a,b} = 1\right\} —$$

множество промежутков, на которых справедлив модифицированный В.П. Масловым закон Ципфа на уровне *L*. Множество

$$J_L = \max I_L$$

назовём множеством максимальных промежутков на уровне *L*.

Рассмотрим, как изменяются $\delta_{\min}^{a,b}\left(\nu_0,\delta_0'\right)$ и $\nu_{\min}^{a,b}\left(\delta_0,\delta_0'\right)$ при изменении промежутка $[a...b]$.

Предложение 7. $[a_1...b_1] \le [a_2...b_2]$, $\delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right)$, $\delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right)$ определены

$$\Rightarrow \delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right) \le \delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right).$$

Доказательство.

$$\forall(\gamma,c): M_{\nu_0}^{a_1,b_1}(\gamma,c) = \max_{a_1 \le r \le b_1}{}_{\nu_0+1}\,\delta(\ln R_r, \ln w_r, \gamma, c) \le \max_{a_2 \le r \le b_2}{}_{\nu_0+1}\,\delta(\ln R_r, \ln w_r, \gamma, c) =$$

$$= M_{\nu_0}^{a_2,b_2}(\gamma,c) \Rightarrow \text{ по предложению 2 } \delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right) = \inf_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}^{a_1,b_1}(\gamma,c) \le \inf_{(\gamma,c)\in\Phi_{\delta_0'}} M_{\nu_0}^{a_2,b_2}(\gamma,c) =$$

$$= \delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right).$$

Предложение 8. $[a_1...b_1] \le [a_2...b_2]$, $\nu_{\min}^{a_1,b_1}\left(\delta_0,\delta_0'\right)$, $\nu_{\min}^{a_2,b_2}\left(\delta_0,\delta_0'\right)$ определены

$$\Rightarrow \nu_{\min}^{a_1,b_1}\left(\delta_0,\delta_0'\right) \le \nu_{\min}^{a_2,b_2}\left(\delta_0,\delta_0'\right).$$

Доказательство.

$$\forall(\gamma,c): V_{\delta_0}^{a_1,b_1}(\gamma,c) = \left|\left\{r \in \{a_1,...,b_1\} \mid \delta(\ln R_r, \ln w_r, \gamma, c) > \delta_0\right\}\right| \le$$

$$\le \left|\left\{r \in \{a_2,...,b_2\} \mid \delta(\ln R_r, \ln w_r, \gamma, c) > \delta_0\right\}\right| = V_{\delta_0}^{a_2,b_2}(\gamma,c) \Rightarrow \nu_{\min}^{a_1,b_1}\left(\delta_0,\delta_0'\right) = \min_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}^{a_1,b_1}(\gamma,c) \le$$

$$\le \min_{(\gamma,c)\in\Phi_{\delta_0'}} V_{\delta_0}^{a_2,b_2}(\gamma,c) = \nu_{\min}^{a_2,b_2}\left(\delta_0,\delta_0'\right).$$

Предложение 9. $[a_1...b_1] \le [a_2...b_2], [a_2...b_2] \in I_L \Rightarrow [a_1...b_1] \in I_L$.

Доказательство.

$$[a_2...b_2] \in I_L \Leftrightarrow \Pi_L^{a_2,b_2} = 1 \Leftrightarrow \left(\delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right) \text{определена}\right) \wedge \left(\delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right) \le \delta_0\right).$$

$\delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right)$ определена $\Rightarrow \Phi_{\delta_0'}^{a_2,b_2} \ne \varnothing$. $\Phi_{\delta_0'}^{a_2,b_2} \subset \Phi_{\delta_0'}^{a_1,b_1} \Rightarrow \Phi_{\delta_0'}^{a_1,b_1} \ne \varnothing \Rightarrow$

$\Rightarrow \delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right)$ определена .

По предложению 7 $\delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right) \le \delta_{\min}^{a_2,b_2}\left(\nu_0,\delta_0'\right) \Rightarrow \delta_{\min}^{a_1,b_1}\left(\nu_0,\delta_0'\right) \le \delta_0 \Leftrightarrow \Pi_L^{a_1,b_1} = 1 \Leftrightarrow$

$\Leftrightarrow [a_1...b_1] \in I_L$.

2.7. Формулировка рангового метода кластеризации эмпирических данных

В разделе 2.1 приведена формулировка рангового метода кластеризации эмпирических данных, которая дана в статье [7]. В такой формулировке задача кластеризации эмпирических данных ранговым методом является нечётко поставленной, т.е. нуждается в уточнении.

Формулировка рангового метода представляет собой два требования к искомому разбиению:

а) на каждом из кластеров справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров γ и c;

б) значения параметров меняются при переходе от кластера к кластеру.

Мы будем рассматривать случай, когда кластеры — промежутки без разрывов (в каждом кластере точки имеют последовательные ранги).

Накладывается ограничение на минимальное число точек в кластере: в каждом кластере должно быть не менее трёх точек.

2.8. Оценка разбиения данных на кластеры

Рассмотрим некоторое разбиение данных на кластеры

$$S = \{[1...m_1],[m_1+1...m_2],[m_2+1...m_3]...,[m_{K-1}+1...n]\}, \quad (18)$$

$$m_0 = 1, m_K = n, m_i - m_{i-1} \geq 3, i = 1,2,...,K, \quad (19)$$

K — число кластеров разбиения.

Нужно оценить это разбиение, т.е. определить, в какой мере оно удовлетворяет двум требованиям, которые указаны в предыдущем разделе.

Первое требование

Чтобы оценить меру справедливости на каждом из кластеров модифицированного В.П. Масловым закона Ципфа, используем величины, введённые в разделах 2.4 и 2.6.

Если задать уровень качества кластера $L = \left(\delta_0, \nu_0, \delta_0'\right)$, то с помощью предиката Π_L можно определить, справедлив ли на кластере модифицированный В.П. Масловым закон Ципфа на уровне L. Если нас интересуют значения параметра γ, которые годятся для кластера, воспользуемся множеством Γ_L.

Если задано пороговое значение δ_0 или допустимое число аномальных точек ν_0, а также задано дополнительное пороговое значение δ_0', то величины $\delta_{\min}\left(\nu_0, \delta_0'\right)$ и $\nu_{\min}\left(\delta_0, \delta_0'\right)$ позволяют оценить качество кластера.

Второе требование

Рассмотрим два соседних кластера [a...k] и [k+1...b] и предположим, что на каждом из них справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров.

Второе требование к разбиению можно трактовать так: если функциональная зависимость годится для некоторого кластера, то она не годится для соседнего кластера. Множество Φ_L задаёт множество функциональных зависимостей, которые годятся для кластера. Поэтому второе требование можно связать с предикатом

$$\Phi_L^{a,k} \cap \Phi_L^{k+1,b} = \varnothing.$$

Если данный предикат принимает значение 1, то не существует функциональной зависимости, которая годится для обоих кластеров [a...k] и [k+1...b] на уровне L. В противном случае такая функциональная зависимость существует.

Приведём общую схему (другую) трактовки второго требования к разбиению (см. рис. 11). На кластере справедлив модифицированный В.П. Масловым закон Ципфа со своими значениями параметров (γ_0, c_0), т.е. точки помещаются в полосу со «средней линией» (15) высоты $2\delta_0$, но возможны исключения из правила: допускается v_0 аномальных точек, при этом на аномальные точки накладывается ограничение: они должны содержаться в полосе со «средней линией» (15) высоты $2\delta_0'$. На другом кластере точки находятся вне полосы со «средней линией» (15) высоты $2\Delta_0$, но возможны исключения из правила: допускается, что не более U_0 точек могут содержаться в этой полосе.

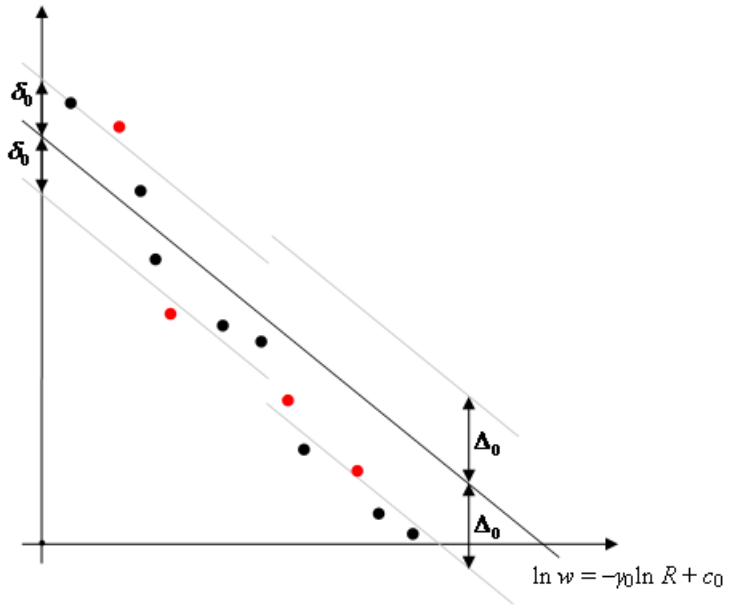

Рис. 11

Введём величину

$$W_{\Delta_0}(\gamma_0, c_0) = \left|\{r \in \{a,...,b\} \mid \delta(\ln R_r, \ln w_r, \gamma_0, c_0) \le \Delta_0\}\right|$$

и рассмотрим значения этой величины для кластера [k+1...b] ([a...k]) при $(\gamma_0, c_0) \in \Phi_L^{a,k}$ ($(\gamma_0, c_0) \in \Phi_L^{k+1,b}$). Нас будут интересовать следующие величины:

$$\min_{(\gamma,c)\in\Phi_L^{a,k}} W_{\Delta_0}^{k+1,b}(\gamma,c), \quad \max_{(\gamma,c)\in\Phi_L^{a,k}} W_{\Delta_0}^{k+1,b}(\gamma,c),$$

$$\min_{(\gamma,c)\in\Phi_L^{k+1,b}} W_{\Delta_0}^{a,k}(\gamma,c), \quad \max_{(\gamma,c)\in\Phi_L^{k+1,b}} W_{\Delta_0}^{a,k}(\gamma,c).$$

$\min\limits_{(\gamma,c)} W_{\Delta_0}(\gamma,c)$ — количество точек, попавших в полосу высоты Δ_0, в лучшем случае, $\max\limits_{(\gamma,c)} W_{\Delta_0}(\gamma,c)$ — количество точек, попавших в полосу высоты Δ_0, в худшем случае. В дальнейшем мы будем рассматривать худший случай.

Рассмотрим функциональные зависимости, которые годятся для кластера [a...k]. Будем рассматривать отдельные точки кластера [k+1...b] относительно этих функциональных зависимостей. Введём расстояние от точки ($\ln R_0$, $\ln w_0$) до кластера [a...k] на уровне L:

$$\rho_L^{a,k}(\ln R_0, \ln w_0) = \inf_{(\gamma,c)\in\Phi_L^{a,k}} \delta(\ln R_0, \ln w_0, \gamma, c).$$

Заметим, что если $\Phi_L^{a,k} = \varnothing$, то величина $\rho_L^{a,k}(\ln R_0, \ln w_0)$ не определена.

Будем сравнивать расстояния от точек кластера [k+1...b] до кластера [a...k] с Δ_0. Вместо величины

$$\max_{(\gamma,c)\in\Phi_L^{a,k}} W_{\Delta_0}^{k+1,b}(\gamma,c) = \max_{(\gamma,c)\in\Phi_L^{a,k}} \left|\{r \in \{k+1,...,b\} \mid \delta(\ln R_r, \ln w_r, \gamma, c) \le \Delta_0\}\right|$$

будем использовать величину

$$\left|\left\{r \in \{k+1,...,b\} \mid \rho_L^{a,k}(\ln R_r, \ln w_r) \le \Delta_0\right\}\right|$$

или будем рассматривать отдельные расстояния $\rho_L^{a,k}(\ln R_r, \ln w_r), r = k+1,...,b$.

Введём величину

$$\mu_{\delta_0,\Delta_0}^{a,k}(\ln R_0, \ln w_0) = \min_{(\gamma,c):\delta(\ln R_0, \ln w_0, \gamma, c) \le \Delta_0} V_{\delta_0}^{a,k}(\gamma, c)$$ —

расстояние от точки ($\ln R_0$, $\ln w_0$) до кластера [*a*...*k*], измеренное в количестве точек кластера [*a*...*k*].

2.9. Процесс кластеризации ранговым методом

Процесс кластеризации эмпирических данных ранговым методом мы будем рассматривать как процесс расширения кластеров.

Пусть задан уровень $L = \left(\delta_0, \nu_0, \delta_0'\right)$. Если кластер принадлежит множеству I_L, но не принадлежит множеству J_L, то его можно расширить. Если кластер принадлежит множеству J_L, то дальнейшее расширение кластера возможно при увеличении числа аномалий на этом кластере (точка, добавляемая в кластер при расширении, может стать аномальной), тогда требуется увеличить допустимое число аномалий ν_0. Также дальнейшее расширение кластера возможно при увеличении порогового значения δ_0 и дополнительного порогового значения δ_0'.

Множество J_L представляет собой результат расширения кластеров на уровне *L*. Если кластер [*a*...*b*] принадлежит множеству J_L, то как оценить возможность дальнейшего расширения этого кластера? Если задать только ν_0 и δ_0' (δ_0 и δ_0'), то для кластеров, которые получаются в результате расширения кластера [*a*...*b*], можно вычислить значение величины $\delta_{\min}\left(\nu_0, \delta_0'\right)$ ($\nu_{\min}\left(\delta_0, \delta_0'\right)$), характеризующей качество промежутка. Также можно вычислить расстояния от точек, не принадлежащих кластеру [*a*...*b*], до кластера [*a*...*b*].

2.10. Формальная постановка задачи

Требуется автоматизировать процесс кластеризации эмпирических данных ранговым методом.

Входная информация — исходные точки-данные $\{(\ln R_r, \ln w_r)\}$. Реализацию рангового метода в программной системе предлагается рассматривать как процесс измерения исходных точек-данных. Результатом измерения будет некоторая информация, которая должна позволить пользователю анализировать данные с точки зрения рангового метода кластеризации.

Возникает вопрос: какую информацию нужно получить, что нужно измерять?

Исходные данные — последовательность точек $\{(\ln R_r, \ln w_r)\}_{r=1}^n$, описанная в разделе 2.2. В качестве функции меры отклонения точки от функциональной зависимости будем использовать функцию (14).

1. Разбиение задано

Пусть задано разбиение (18) при выполнении условия (19). Требуется оценить это разбиение, т.е. определить, в какой мере оно удовлетворяет двум требованиям, которые указаны в разделе 2.7.

Пусть задан уровень $L = \left(\delta_0, \nu_0, \delta_0'\right)$, $\delta_0 \ge 0$, $\nu_0 \ge 0$ — целое число, $\delta_0' \ge \delta_0$ (δ_0' может быть равно ∞). Для оценки каждого кластера в отдельности вычислим значение предиката Π_L. Поскольку ещё нужно знать значение параметра γ, найдём множество Γ_L.

Задача 1. Найти $\Pi_L^{m_{i-1}+1,m_i}$, $\Gamma_L^{m_{i-1}+1,m_i}$, причём эти величины определены только для кластеров $[m_{i-1}+1...m_i]$, для которых $v_0 \le m_i - m_{i-1}$.

Пусть δ_0 не задано. Величина $\delta_{\min}\left(v_0, \delta_0'\right)$ характеризует качество кластера.

Задача 2. Найти $\delta_{\min}^{m_{i-1}+1,m_i}\left(v_0, \delta_0'\right)$ для кластеров, для которых эта величина определена[4] ($v_0 \le m_i - m_{i-1}$, $\Pi_{\delta_0'}^{m_{i-1}+1,m_i} = 1$).

Пусть v_0 не задано. Величина $v_{\min}\left(\delta_0, \delta_0'\right)$ характеризует качество кластера.

Задача 3. Найти $v_{\min}^{m_{i-1}+1,m_i}\left(\delta_0, \delta_0'\right)$ для кластеров, для которых эта величина определена ($\Pi_{\delta_0'}^{m_{i-1}+1,m_i} = 1$).

Теперь рассмотрим второе требование к разбиению (значения параметров меняются при переходе от кластера к кластеру). Это требование мы будем трактовать так: если функциональная зависимость годится для некоторого кластера, то точки соседних кластеров находятся далеко от этой функциональной зависимости.

Если рядом с кластером есть точки, близкие к функциональной зависимости, которая годится для кластера, то эти точки можно добавить к кластеру, т.е. кластер можно расширить. Спрашивается: можно ли расширить кластеры разбиения? А, может, какие-то кластеры, наоборот, надо сузить?

Пусть задан уровень L. Рассмотрим множество максимальных промежутков на уровне L.

Задача 4. Сопоставить множество J_L с разбиением S.

Пусть δ_0 не задано. Для количественной характеристики расширения или сужения кластеров разбиения найдём матрицу $\left\{\delta_{\min}^{a,b}\left(v_0, \delta_0'\right)\right\}_{[a...b]\in T}$.

Задача 5. Найти матрицу $\left\{\delta_{\min}^{a,b}\left(v_0, \delta_0'\right)\right\}_{[a...b]\in T}$ ($v_0 \le b-a+1$).

Пусть v_0 не задано. Для количественной характеристики расширения или сужения кластеров разбиения найдём матрицу $\left\{v_{\min}^{a,b}\left(\delta_0, \delta_0'\right)\right\}_{[a...b]\in T}$.

Задача 6. Найти матрицу $\left\{v_{\min}^{a,b}\left(\delta_0, \delta_0'\right)\right\}_{[a...b]\in T}$.

Для того чтобы оценить близость точек соседних кластеров к кластеру, найдём расстояния от точек соседних кластеров до кластера. Пусть задано натуральное число Q. Для каждого кластера разбиения найдём расстояния от Q точек «слева» и от Q точек «справа» от кластера до этого кластера.

Задача 7. Для каждого кластера $[m_{i-1}+1...m_i]$, для которого $v_0 \le m_i - m_{i-1}$, $\Pi_L^{m_{i-1}+1,m_i} = 1$, найти $\rho_L^{m_{i-1}+1,m_i}(\ln R_r, \ln w_r)$, $m_{i-1}+1-Q \le r < m_{i-1}+1$, $m_i < r \le m_i+Q$ (величина $\rho_L^{m_{i-1}+1,m_i}$ определена, если $\Pi_L^{m_{i-1}+1,m_i} = 1$).

Пусть заданы δ_0 и Δ_0 ($\delta_0 \ge 0, \Delta_0 \ge 0$).

Задача 8. Для каждого кластера $[m_{i-1}+1...m_i]$ найти $\mu_{\delta_0,\Delta_0}^{m_{i-1}+1,m_i}(\ln R_r, \ln w_r)$, $m_{i-1}+1-Q \le r < m_{i-1}+1$, $m_i < r \le m_i+Q$.

---

[4] В случае, когда мера отклонения точки от функциональной зависимости определяется формулой (14), $\delta_{\min}(0,\infty) = \min\limits_{(\gamma,c)\in R^2} M_0(\gamma,c)$, поэтому $\Pi_{\delta_0'} = \left(\Phi_{\delta_0'} \ne \varnothing\right)$.

2. Разбиение не задано

Найдём множество максимальных промежутков на уровне L.

Задача 9. Найти множество J_L.

Найдём матрицы $\left\{\delta_{\min}^{a,b}\left(\nu_0,\delta_0{}'\right)\right\}_{[a...b]\in T}$ и $\left\{\nu_{\min}^{a,b}\left(\delta_0,\delta_0{}'\right)\right\}_{[a...b]\in T}$. См. задачи 5, 6.

3. Прочее

Пусть задан кластер $[a...b]$, $b–a+1 \ge 3$.

Задача 10. Найти $\Pi_L^{a,b}$, $\Gamma_L^{a,b}$ ($v_0 \le b–a+1$).

Задача 11. Найти $\delta_{\min}^{a,b}\left(\nu_0,\delta_0{}'\right)$ (если эта величина определена: $v_0 \le b–a+1$, $\Pi_{\delta_0{}'}^{a,b}=1$).

Задача 12. Найти $\nu_{\min}^{a,b}\left(\delta_0,\delta_0{}'\right)$ (если эта величина определена: $\Pi_{\delta_0{}'}^{a,b}=1$).

Задача 13. Найти $\delta_{\min}^{a,b}\left(\nu,\delta_0{}'\right)$, $\nu = 0, 1, 2, \ldots v_0$ ($v_0 \le b–a+1$).

Задача 14. Q — натуральное число, $v_0 \le b–a+1$, $\Pi_L^{a,b}=1$. Найти $\rho_L^{a,b}\left(\ln R_r,\ln w_r\right)$, $a–Q \le r < a$, $b < r \le b+Q$.

Задача 15. Q — натуральное число. Найти $\mu_{\delta_0,\Delta_0}^{a,b}\left(\ln R_r,\ln w_r\right)$, $a–Q \le r < a$, $b < r \le b+Q$.

Пусть требуется оценить пару соседних кластеров: $[a...k]$, $[k+1...b]$. Найдём расстояния от точек одного кластера до другого кластера.

Задача 16. $v_0 \le b–a+1$, $\Pi_L^{a,k}=1$, $\Pi_L^{k+1,b}=1$. Найти $\rho_L^{a,k}\left(\ln R_r,\ln w_r\right)$, $r = k+1, ..., b$. Найти $\rho_L^{k+1,b}\left(\ln R_r,\ln w_r\right)$, $r = a, ..., k$.

Задача 17. Найти $\mu_{\delta_0,\Delta_0}^{a,k}\left(\ln R_r,\ln w_r\right)$, $r = k+1, ..., b$. Найти $\mu_{\delta_0,\Delta_0}^{k+1,b}\left(\ln R_r,\ln w_r\right)$, $r = a, ..., k$.

Таким образом, построена математическая модель. Требуется разработать программную систему, позволяющую решать указанные задачи.

Заключение

Таким образом, в ходе выполнения курсовой работы была построена математическая модель, произведена формальная постановка задачи.

Список литературы

[1] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. и др. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2 т. 2-е изд., испр. — Т. 1: Айвазян С. А., Мхитарян В. С. Теория вероятностей и прикладная статистика. — М.: ЮНИТИ-ДАНА, 2001. — 656 с.

[3] Бидаева Е. В. Инструментальное средство анализа эмпирических данных методами квантовой статистики. Дипломная работа. — Владивосток, 2008. URL: http://imcs.dvgu.ru/works/work?wid=4525

[4] Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе / Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению "Информационно-телекоммуникационные системы", 2008. - 26 с. URL: http://window.edu.ru/window/library?p_rid=56161

[5] Буховец А. Г. Системный подход и ранговые распределения в задачах классификации. // Вестник ВГУ, серия: экономика и управление, 2005, № 1. URL: http://www.ebiblioteka.lt/resursai/Uzsienio%20leidiniai/Voronezh/eko/2005-01/eko0501_21.pdf

[6] Гузев М. А., Никитина Е. Ю., Черныш Е. В. Решение некоторых задач анализа эмпирических данных в медицине и криминологии. // Математическое моделирование и биомеханика в современном университете. Тезисы докладов IV Всероссийской школы-семинара. Ростов-на-Дону: Изд-во «Терра Принт». 2008. С. 40–41.

[7] Гузев М. А., Черныш Е. В. Ранговый анализ в задачах кластеризации. // Информатика и системы управления. 2009. №3(21). — Благовещенск: Изд-во Амурского гос. ун-та. С. 13–19.
URL: http://www.khstu.ru/rus/ics/ics_pdf/N21_02.pdf

[8] Зеленов А. С. Критериальная кластеризация квазиодномерных данных. Дипломная работа. — Владивосток, 2008.
URL: http://imcs.dvgu.ru/works/work?wid=4523

[9] Ильин В. А., Позняк Э. Г. Основы математического анализа: В 2-х ч. Часть I: Учеб.: Для вузов. — 7-е изд., стер. — М.: ФИЗМАТЛИТ, 2008. — 648 с.

[10] Кленин А. С. Методические указания по подготовке и защите отчётов на специализации «Прикладная математика. Системное программирование» (версия 1.0 от 17.06.2009). — Владивосток, 2002–2009.
URL: http://imcs.dvgu.ru/lib/repplan/RepPlan.7z

[11] Колмогоров А. Н. Теория передачи информации. — М.: Изд-во АН СССР, 1956.

[12] Крашаков С. А., Теслюк А. Б., Щур Л. Н. Об универсальности рангового распределения популярности веб-серверов.
URL: http://www.rfbr.ru/pics/17664ref/file.pdf

[13] Маслов В. П. Закон «отсутствия предпочтения» и соответствующие распределения в частотной теории вероятностей // Мат. Заметки. — 2006. — Т. 80, вып. 2. — С. 220–230.

[14] Маслов В. П. Квантовая экономика. Рос. академия наук. — 2-е изд., доп. — М.: Наука, 2006. — 92 с.

[15] Маслов В. П. Об одной общей теореме теории множеств, приводящей к распределению Гиббса, Бозе-Эйнштейна, Парето и закону Ципфа-Мандельброта для фондового рынка // Мат. заметки. — 2005. — Т. 78, № 6. — С. 870–877.

[16] Маслов В. П. Фазовые переходы нулевого рода и квантование закона Ципфа // Теоретическая и математическая физика. — 2007. — Т. 150, № 1. — С. 118–142.

[17] Маслов В. П., Маслова Т. В. О законе Ципфа и ранговых распределениях в лингвистике и семиотике // Мат. Заметки. 2006. — Т. 80, вып. 5. — С. 718–732.

[18] Нурминский Е. А. Методы оптимизации. Курс лекций ДВГУ.
URL: http://elis.dvo.ru/~nurmi/edu/optimization.pdf

[19] Пак Г. К. Дискретная математика. Учеб. пособие. — Находка: Институт технологии и бизнеса, 2001. — 109 с.

[20] Пак Г. К. Лекции по аналитической геометрии. — Владивосток, 2005.

[21] Пиковая Т. В. Аппроксимация эмпирических данных методами квантовой статистики. Курсовая работа. — Владивосток, 2008.
URL: http://imcs.dvgu.ru/works/work?wid=4067

[22] Пиковая Т. В. Оптимизация поиска параметров аппроксимирующей функции в системе анализа эмпирических данных SPED. Дипломная работа. — Владивосток, 2010. URL: http://imcs.dvgu.ru/works/work?wid=10665

[23] Третьяков Н. П. Применение кластерного анализа к мировой статистике пожаров. URL: http://agps-2006.narod.ru/ttb/2009-2/08-02-09.ttb.pdf

[24] Тюрин Ю. Н., Макаров А. А. Анализ данных на компьютере / Под ред. В. Э. Фигурнова — 3-е изд., перераб. и доп. — М.: ИНФРА-М, 2003. — 544 с.

[25] Черныш Е. В. Некоторые особенности процедуры кластеризации медико-экологических данных. // Информатика и системы управления. Приложение к

журналу. 2007. № 1(13). — Благовещенск: Изд-во Амурского гос. ун-та. С. 80–82.

[26] Черныш Е. В. Некоторые подходы в решении задач кластеризации эмпирических данных. // XXXIV Дальневосточная математическая школа-семинар им. академика Е.В. Золотова: Тезисы докладов. — Владивосток: Изд-во «Дальнаука». 2009.

[27] Черныш Е. В. Особенности кластеризации медико-экологических данных. Презентация. VI Дальневосточный региональный Конгресс «Человек и лекарство»: образовательный семинар «Информационные технологии в медицине, биологии, экологии». Владивосток, 24–25 сентября 2009 г.

[28] Черныш Е. В. Предмодельный анализ медико-экологических данных. // Математическое моделирование и биомеханика в современном университете. Труды III Всероссийской школы-семинара. Ростов-на-Дону: Изд-во «Терра Принт». 2007. С. 84–85.

[29] Черныш Е. В. Принципы классификации заболеваемости по степени экологического напряжения. // XXXII Дальневосточная математическая школа-семинар им. академика Е.В. Золотова: Тезисы докладов. — Владивосток: Изд-во «Дальнаука». 2007. С. 108.

[30] Clauset A., Shalizi C. R., Newman M. E. J. Power-law distributions in empirical data. // SIAM Review. — 2007.
URL: http://tuvalu.santafe.edu/%7Eaaronc/courses/readings/Clauset_Shalizi_Newman_09_PowerlawDistributionsInEmpiricalData.pdf

[31] Maslov V. P. Quantum linguistic statistics // Russ. J. Math. Phys. — 2006. — Vol. 13, No. 3. — P. 315–325.

[32] Murtagh B. A., Saunders M. A. MINOS 5.5 USER'S GUIDE, Stanford University, 1983. URL: http://www.sbsi-sol-optimize.com/manuals/Minos%20Manual.pdf