УДК 519.7

© 2009 г. **М.А. Гузев**, д-р физ.-мат. наук, чл.-корр. РАН
(Институт прикладной математики ДВО РАН, Владивосток),
Е.В. Черныш
(Дальневосточный государственный университет, Владивосток)

РАНГОВЫЙ АНАЛИЗ В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ[1]

В данной работе рассматривается проблема кластеризации эмпирических данных. Для ее решения предлагается использовать модифицированный В.П. Масловым закон Ципфа. Данный подход реализован для задач кластеризации медицинских данных.
***Ключевые слова*:** кластеризация, частотный словарь, закон Ципфа, семиотические системы.

Введение

Эмпирические данные, полученные опытным путем, 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. Зачастую это приводит к необходимости структурировать, систематизировать, выделить полученные данные по тем или иным признакам, т.е. кластеризовать их.

Процедура кластеризации (см., например, [1]) включает выделение характеристик объектов, затем выбор метрики, на основе которой можно определить меру сходства объектов. Это позволяет исходное множество однородных объектов разбить в признаковом пространстве на кластеры (группы). Трудности при решении задач кластеризации связаны с оптимальным выбором метрики, методов группировки объектов в пространстве признаков, существованием большого количества признаков, описывающих изучаемое явление, и др.

Поэтому в кластерном анализе существует общая проблема формулировки алгоритма обработки эмпирических данных с целью извлечь информацию. Поскольку при решении задач исследователям приходится изучать знаковые объекты произвольной природы, т.е. семиотические системы, то для анализа эмпирических данных естественно использование подходов из лингвистики. В частности, для частотных словарей известен закон Ципфа [2], описывающий соотношение

[1] Работа выполнена при финансовой поддержке гранта НШ 2810.2008.1.

между частотой и рангом слов в словаре. Согласно классическому подходу [2] частота совпадет с вероятностью повторяемости знака, поэтому эта идея была использована при изучении объектов в других областях знаний [2]. В зарубежной научной литературе на соответствующее функциональное соотношение между вероятностью повторяемости знака и его рангом ссылаются как на power law.

Однако еще А.Н. Колмогоров в своих работах [3] наметил путь пересмотра теории вероятностей с точки зрения дискретного подхода. Согласно его концепции, случайность – это большая сложность. Тогда алгоритм возникновения случайного события будет весьма сложным, а расшифровка его потребует очень длинного кода. Чем сложнее описана информация, тем длиннее необходим алгоритм расшифровки, а это, согласно Колмогорову, близко к случайному.

В настоящее время данная парадигма развита В.П. Масловым [4 – 9]. 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 [4, 7, 8], 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 [7]. В идейном плане, как указано в [4], его можно также применять для анализа семиотических объектов. Поэтому мы воспользуемся подходом В.П. Маслова для анализа эмпирических данных с целью выделить кластерные объекты.

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

Подход В.П. Маслова и общая идея кластеризации

Как указано выше, важной характеристикой знака является его повторяемость в данном социуме, т.е. частота встречаемости, характеризующая активность использования. Известный для частотных словарей закон Ципфа [2], 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 нее – меняется на величину логарифма 3, что является незначительной поправкой. Это означает, что в переменных без логарифмов формула (1) огрубляет соотношение между рангом и час-

тотой, давая значительную ошибку.

Новый подход исследования статистических зависимостей в языке, предложенный В.П. Масловым [4 – 9], позволяет получить более точное соотношение между рангом и частотой. Пусть рассматривается алфавитный словарь, в котором указаны частоты встречаемости $w_i = 1, \ldots, s$ каждого слова из некоторого корпуса текстов; n_i – число слов, имеющих одну частоту встречаемости; $N = \sum n_i$ задает число слов словаря, а величина

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

совпадает с объемом текста. Справедливо следующее утверждение [4]: если варианты $\{n_i\}$ равноценны и удовлетворяют (2), то ранг r_l для l-го слова вычисляется по формуле

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

Эта формула аналогична распределению Бозе для тождественных частиц [10], для которого роль энергии частиц играет частота повторяемости w_i, а число частиц на заданном уровне совпадает с n_i.

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 объем равен

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

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

Практическое применение полученных соотношений связано с выбором виртуальной частоты и других феноменологических параметров. Простейшая параметризация для виртуальной частоты встречаемости имеет вид

$$\tilde{w}_i = w_i(1 + \alpha w_i^{\gamma}),\ \alpha > 0,\ \gamma > 0.$$

Если $w_i = i$, то, полагая $\beta << 1$, $\sigma = 0$, можно перейти от суммы к интегралу, в результате имеем

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

Использование (5) для семиотических систем показало [4, 6, 8], что формула (5) точнее, чем закон Ципфа, описывает соотношение между рангом слова и частотой его встречаемости в словаре.

В.П. Масловым были предложены также другие параметризации для виртуальной частоты и рассмотрены задачи, связанные с анализом экономического риска при покупке товара. В частности, для задачи о числе проданных машин [7] получена следующая теоретическая кривая для числа этих машин N_p по цене, меньшей p:

$$p \cong \alpha \left(\frac{N_p}{N_\infty - N_p} \right)^{\gamma}, \qquad (6)$$

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

Как было замечено выше, стандартные методы кластеризации предполагают использование метрики для ранжированных эмпирических данных. Выполненные В.П. Масловым исследования показали, что для объектов, объединенных некоторым набором признаков, т.е. для определенной группы или кластера, существуют зависимости между соответствующими переменными модели, – например, в виде (5) или (6). 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 полученные выше функциональные зависимости наиболее чувствительны к выбору γ. Поэтому естественное разбиение должно быть таковым, чтобы для каждого кластера существовало свое числовое значение степенного параметра γ, характеризующее соответствующий кластер.

Использование модифицированного закона В.П. Маслова для кластеризации медицинских данных

Первая задача связана с выделением групп медицинских работников «скорой помощи» по их эмоциональному состоянию.

Среди этой категории медицинских работников 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 сестры) выявить характерное разбиение на группы и сделать вывод о выраженности каждой фазы эмоционального состояния в группах риска.

В соответствие с сформулированной выше идеей рангового анализа медицинские работники каждой специальности были упорядочены в порядке возрастания абсолютного значения w показателя СЭВ отдельно по трем фазам, и каждому значению w поставлен в соответствие порядковый номер – ранг r. Исходные точки анализировались с помощью соотношения (6), которое в логарифмических координатах запишем в виде:

$$\ln w \cong \ln\left(\frac{N-r}{r}\right) + c \equiv -\gamma \ln R + c. \qquad (7)$$

Для обеспечения неотрицательности логарифма и удобства визуализации данных примем $N = 2n+1$, 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 сестрах разбиваются на кластеры в каждой фазе СЭВ. На рис. 1, 2 представлены для фазы «резистенции» и «истощения», соответствующие кластеры, обозначенные как группы 1, 2, 3. Область особых точек составляют медработники с самым большим (48 лет) и самым маленьким (1 год) стажем работы. На их эмоциональную устойчивость профессиональные факторы не оказывают существенного влияния.

Рецепт В.П. Маслова может быть применен для оценки влияющих факторов риска, – таких как возраст и стаж медицинских работников, на выраженность СЭВ. В результате получаем три характерных кластера работников, имеющих стаж до 10 лет, от 11 до 25 лет и свыше 25 лет (на рис. 1, 2 обозначены кружками).

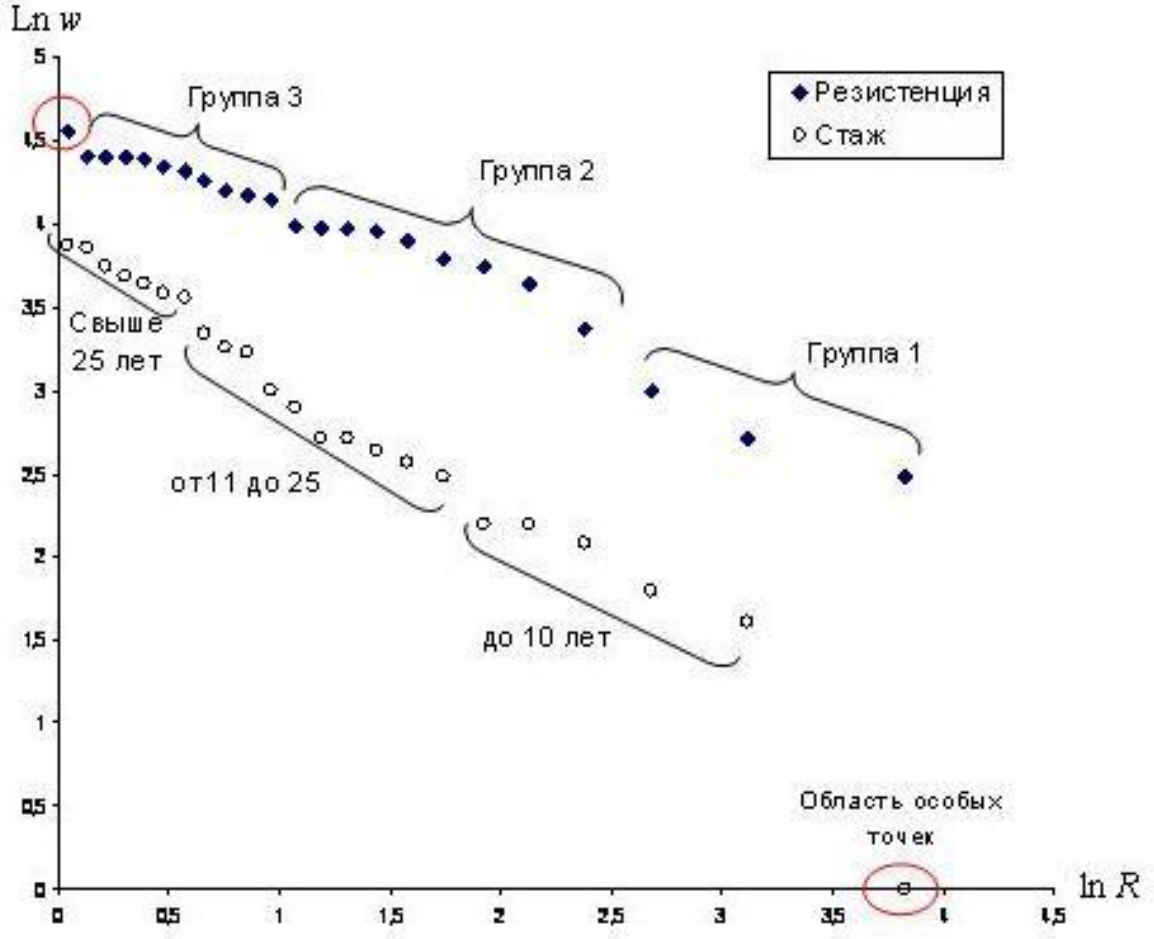

Рис. 1. Выраженность фазы «резистенции» *синдрома эмоционального выгорания.*

При этом оказалось, что фаза повышенной «резистенции» (группа 3) выяв-

лена у контингента медицинских сестер 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 сестер (группа 3) в возрасте от 50 лет и со стажем работы от 25 лет и старше (рис. 2).

Вторая задача связана с выявлением групп пациентов, страдающих хроническим заболеванием желудочно-кишечного тракта; степень тяжести заболевания выявлена исходя из показателей различных видов анализа крови. Сначала все пациенты были разделены на две части по степеням (средняя и тяжелая) физического состояния больного на основе общего осмотра. Затем пациенты каждой степени состояния ранжировались по возрастанию абсолютных значений показателя выполненного химического анализа крови и анализировались данные с использованием ранговой зависимости (7).

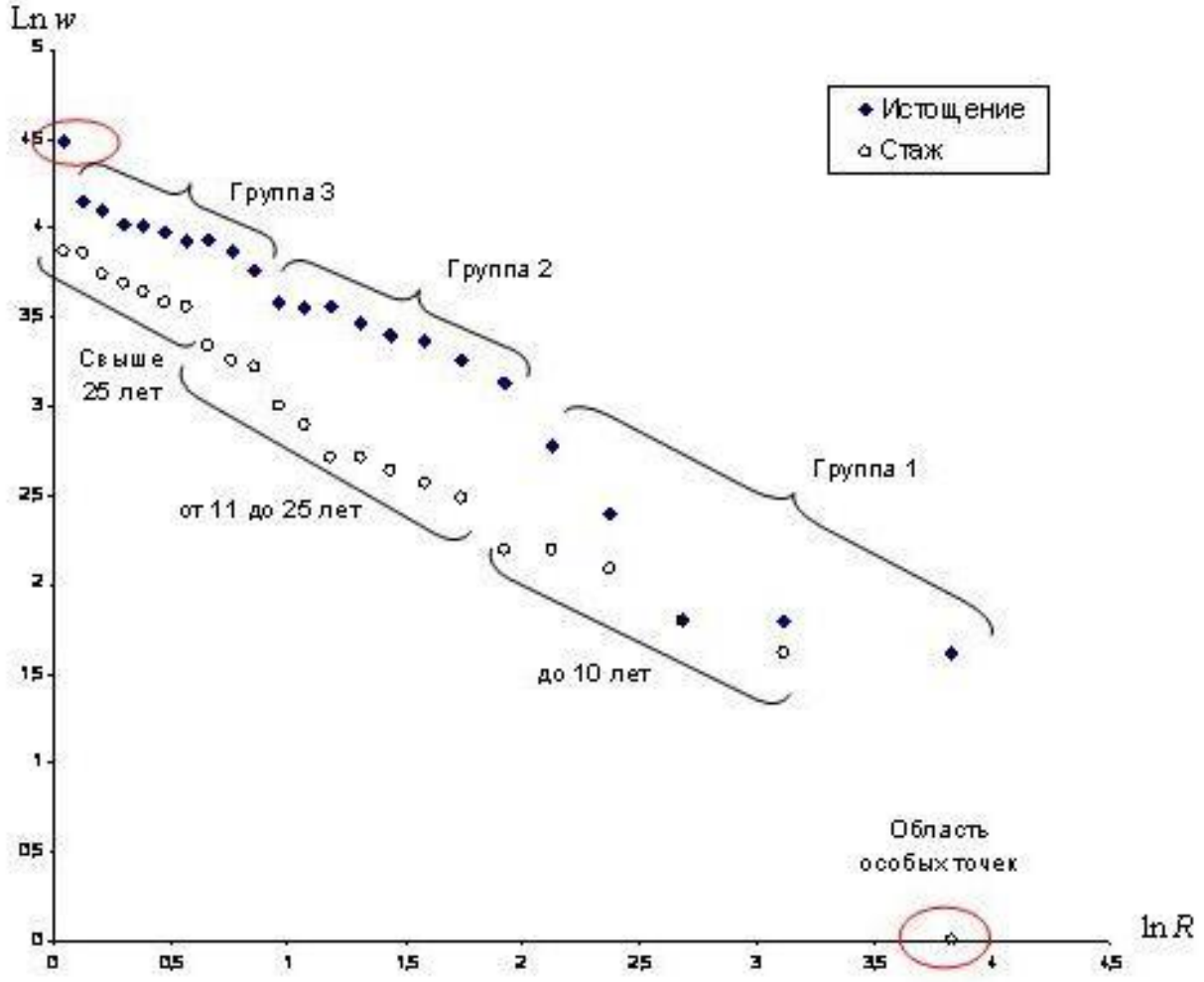

Рис. 2. Выраженность фазы «истощения» *синдрома эмоционального выгорания* медицинских сестер по стажу работы.

С помощью вышеизложенного подхода для кластеризации медицинских данных получим разбиение обследованных пациентов на кластеры. На рис. 3 для группы тяжелой степени состояния отражены четыре кластера разными видами символов по данным анализа Prooxy.

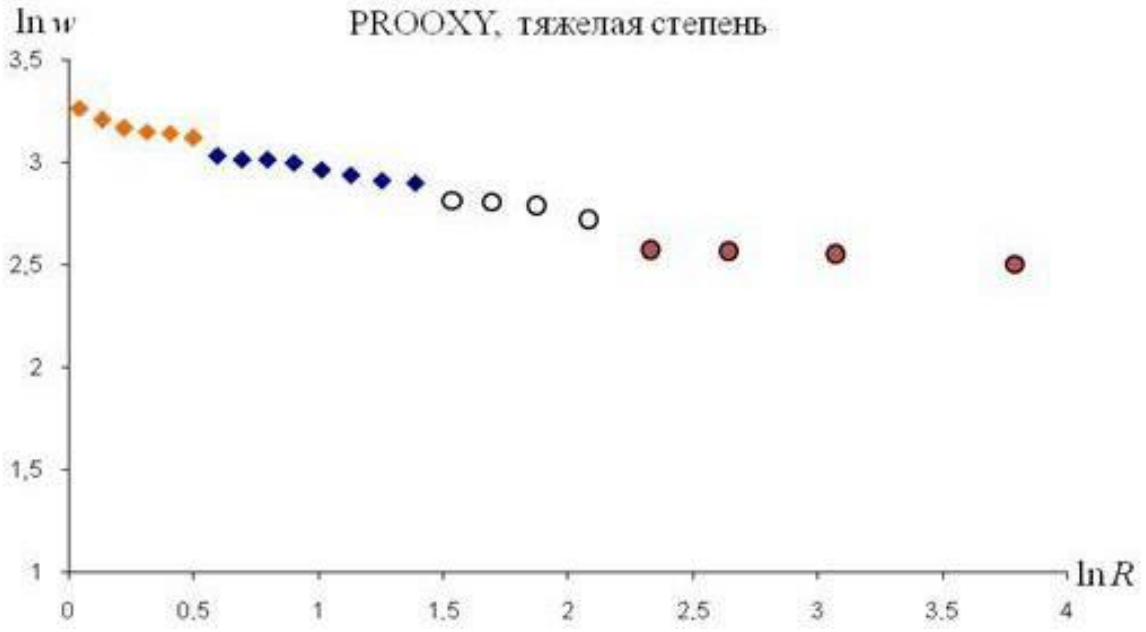

Рис. 3. Кластеризация «тяжелых» пациентов по результатам анализов крови.

Группы, обозначенные светлыми и темными ромбиками, соответствуют пациентам, которым необходимо стационарное лечение. Группы светлых и темных кружков соответствуют пациентам, которым достаточно амбулаторного лечения.

ЛИТЕРАТУРА

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. *Clauset A., Shalizi C.R., Newman M.E.J.* Power-law distributions in empirical data. // *SIAM Review,* 2007.
3. *Колмогоров А.Н.* Теория передачи информации. – М.: Изд-во АН СССР, 1956.
4. *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.
5. *Маслов В.П.* Закон «отсутствия предпочтения» и соответствующие распределения в частотной теории вероятностей // Мат. заметки, 2006. – Т. 80, вып. 2. – С.220-230.
6. *Маслов В.П.* Фазовые переходы нулевого рода и квантование закона Ципфа // Теоретическая и математическая физика. – 2007. – Т. 150, № 1. – С.118-142.
7. *Маслов В.П.* Квантовая экономика. – М.: Наука, 2006.
8. *Maslov V.P.* Quantum linguistic statistics // Russ. J. Math. Phys. – 2006. – Vol. 13, № 3. – P.315-325.
9. *Маслов В.П.* Об одной общей теореме теории множеств, приводящей к распределению Гиббса, Бозе-Энштейна, Парето и закону Ципфа-Мандельброта для фондового рынка // Мат. заметки. – 2005. – Т. 78, № 6. – С.870-877.
10. *Ландау Л.Д., Лифшиц Е.М.* Теоретическая физика. – Т. 5. Статистическая физика. Часть 1. – М.: Физматлит, 2004.

E-mail:
Гузев М.А. – *guzev@iam.dvo.ru;*
Черныш Е.А. – *cher@imcs.dvgu.ru.*