Экспериментальная психология
2018. Т. 11. № 3. С. 5—18
doi: 10.17759/exppsy.2018110301
ISSN: 2072-7593
ISSN: 2311-7036 (online)

Experimental Psychology (Russia)
2018, vol. 11, no. 3, pp. 5—18
doi: 10.17759/exppsy.2018110301
ISSN: 2072-7593
ISSN: 2311-7036 (online)

СТОХАСТИЧЕСКАЯ РОЕВАЯ КЛАСТЕРИЗАЦИЯ В ЗАДАЧАХ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ДАННЫХ, ПРЕДСТАВЛЕННЫХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

ЮРЬЕВ Г.А.*, *ФГБОУ ВО МГППУ, Москва, Россия,*
e-mail: g.a.yuryev@gmail.com

ВЕРХОВСКАЯ Е.К.**, *ФГБОУ ВО МГППУ, Москва, Россия,*
e-mail: katrin636bmw@yandex.ru

ЮРЬЕВА Н.Е.***, *ФГБОУ ВО МГППУ, Москва, Россия,*
e-mail: yurieva.ne@gmail.com

Рассматривается метод обработки данных, представленных на естественном языке, использующий стохастический алгоритм нелинейного снижения размерности многомерных данных, учитывающий дискриминирующую силу найденного решения для заданных значений категориальной переменной, связанной с каждым наблюдением. Для поиска характеристик, обеспечивающих наилучшее разделение наблюдений в смысле заданного функционала качества, предлагается использовать численную процедуру, основанную на методе оптимизации, известном как «Метод роя частиц». В основе оценки качества решения лежит чистота кластеров, полученных в найденном пространстве методом k-средних, либо с использованием самоорганизующихся карт Кохонена.

Ключевые слова: обработка данных, представленных на естественном языке, комбинаторная оптимизация, оптимизация методом роя частиц, нелинейное снижение размерности.

Введение

Автоматизация психолого-педагогических измерений приобрела широкое распространение с развитием вычислительной техники; очевидным преимуществом компьютерного тестирования перед традиционным является возможность проведения более масштабных выборочных исследований. Параллельно с внедрением компьютеризированных тестовых методик велась активная работа по формированию методологии проектирования те-

Для цитаты:
Юрьев Г.А., Верховская Е.К., Юрьева Н.Е. Стохастическая роевая кластеризация в задачах автоматизированной обработки данных, представленных на естественном языке // Экспериментальная психология. 2018. Т. 11. №. 3. С. 5—18. doi:10.17759/exppsy.2018110301

* *Юрьев Г.А.* Кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный психолого-педагогический университет». E-mail: g.a.yuryev@gmail.com
** *Верховская Е.К.* Сотрудник, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный психолого-педагогический университет». E-mail: katrin636bmw@yandex.ru
*** *Юрьева Н.Е.* Кандидат психологических наук, научный сотрудник, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный психолого-педагогический университет». E-mail: yurieva.ne@gmail.com

стов с вопросами закрытого типа, был разработан целый ряд концепций (Куравский, 2011; Тюменева, 2007), позволяющих повысить достоверность результатов тестирования, за счет применения оригинальных математических моделей процесса тестирования (Куравский, 2017; Куравский, 2012). До недавнего времени отсутствие подходящей программно-технической базы препятствовало развитию средств автоматизации оценивания ответов на задания открытого типа, с повышением интереса к семантическому анализу ситуация начала меняться. В данной статье изложена концепция параметризации данных, представленных на естественном языке, основанная на использовании современных технологий векторизации текстовой информации и методов нелинейного снижения размерности многомерных данных.

Приводится пример практического применения подхода к текстовым описаниям авиапроисшествий, этот подход легко может быть экстраполирован на задачи обработки ответов, описанных естественным языком на тестовые вопросы открытого типа.

Постановка задачи

Существует выборка наблюдений, характеризующихся набором категориальных переменных и соответствующих им текстов на естественном языке, задача заключается в поиске способа обработки текстовых данных позволяющего автоматически, на основе нового фрагмента текста определить значения соответствующих категориальных переменных, т. е. классифицировать наблюдение. Первичная обработка текстовых данных выполнялась с использованием программного инструмента Word2vecc (Mikolov, 2013), позволяющего получать контекстные вектора слов из векторного пространства, построенного по корпусу текстов при помощи рекуррентной нейронной сети. Параметризованное представление текста, соответствующего наблюдению, вычислялось как среднеарифметическое векторов слов, входящих в этот текст (такой способ представления документа является стандартным (Mikolov, 2013)). Предложенный метод стохастической роевой кластеризации может применяться в комбинации с любой аналогичной, построенной на основе концепции дистрибутивной семантики, технологией параметризации текстовых данных.

Выборку наблюдений обозначим как V_X^C, где $V = \{v_{x0}^c, ..., v_{xl\text{-}1}^c\}$ — множество из l наблюдений, харрактеризуемых классом c и многомерным параметрическим вектором X; $C \in \{c_0, ... , c_{n\text{-}1}\}$ — одно из n возможных значений категориальной переменной; $X = \{x_0, ... , x_{k\text{-}1}\}$ — параметрический вектор из k вещественных компонент (в контексте рассматриваемой задачи — результат параметризации текста). Необходимо определить множество компонент из X-$m = \{m_0, ... , m_{b\text{-}1}\}$ — где $m \in Z = \{0 ... k\text{-}1\}$, $b < k$, $m_{0 ... b}$ из m уникальны, такое что $\dot{X} = \{X_{m_0}, ... , X_{m_{b\text{-}1}}\}$ обеспечивает квазиоптимальные значения функционала качества $Q(V_{\dot{X}}^C)$.

Концепция функционала качества найденного признакового пространства

Если задана категориальная переменная, значение которой известно для всех наблюдений обучающей выборки, для любого результата кластеризации может быть задан функционал качества, связанный с «чистотой» найденных классов в контексте этой категориальной переменной. Тогда поиск подмножества компонент, оптимальных для дискриминации заданных классов, можно считать комбинаторной задачей на поиск одной из 2^N (где N — размерность исходного пространства признаков) комбинаций параметров, обеспечивающей наилучшие значения заданного функционала качества $Q(V_{\dot{X}}^C)$, бóльшие значения которого будут соответствовать результатам кластеризации с большей однородностью

кластеров. В качестве метода кластеризации использовался метод k-средних с инициализацией центройдов случайными значениями либо самоорганизующиеся карты Кохонена (Куравский, 2012).

Для формализации понятия «чистота кластера» определим понятие *однородность* как обобщенную меру отклонения количества наблюдений заданного класса $c \in \{c_0, \ldots, c_{n-1}\}$, «выигравших» в каждом кластере от общего числа отнесенных к данному классу наблюдений (далее *долевая однородность* $U_{partial}$), и обобщенное отклонение количества уникальных классов наблюдений в каждом кластере от 1 (далее *конструктивная однородность* $U_{formation}$). Если количества наблюдений каждого класса в V сопоставимы, целесообразно дополнить функционал качества характеристикой, отражающей обобщенное отклонение объема результирующих кластеров от ожидаемого объема (например, от среднеарифметического объема l/n), далее для ссылки на этот параметр будет использоваться термин *взвешенность* W.

Графическое представление структурных характеристик результатов кластеризации

В качестве иллюстраций к результатам применения предложенного алгоритма далее будет использоваться графическое представление чистоты кластерной структуры, предложенное исследователем Narayana Swamy (Swamy, 2016). Поскольку такая форма представления (рис. 1) не является стандартизованной, далее даются краткие пояснения по ее интерпретации.

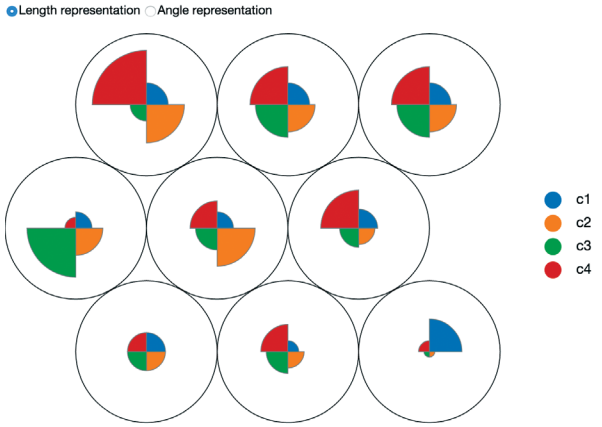


Рис. 1. Пример диаграммы кластерной чистоты

Предполагается, что заранее известны метки классов, кластеризованных наблюдений, их значения перечислены в легенде в правой части рис. 1. Каждому из классов сопоставлен цветовой код, также отраженный в легенде. В основной части диаграммы размещается набор концентрических окружностей, их число соответствует числу кластеров, полученных при выполнении процедуры кластеризации (9 для примера на рис. 1). Отношение площадей внутренних окружностей друг к другу соответствует отношению объемов соответствующих кластеров. Размер секторов во внутренних окружностях соответствует объему наблюдений класса с соответствующей цветовой кодировкой в соответствующем кластере; например, в кластере, представленном левой верхней окружностью, большая часть наблюдений отнесена к классу c4, имеющему красный цветовой код.

Практические оценки функционала качества

Могут быть даны формальные оценки каждой из компонент Q ($V_{\dot{X}}^{C}$), перечисленных ранее. Перед началом вычислений следует получить результат кластеризации $V_{\dot{X}}^{C}$ методом k-средних; множество из j результирующих кластеров далее будет обозначено как $G = \{g_0, \ldots, g_{j-1}\}$, где $g_{i=\{0\ldots j-1\}} = \{c_0, \ldots, c_{z-1}\}$ — множество из z меток классов, соответствующих наблюдениям, отнесенным к данному кластеру. Пусть $f_{maxFreq}(g_{i=\{0\ldots j-1\}})$ — функция, возвращающая абсолютное количество меток класса, имеющих максимальную долю в i-*ом* кластере. Тогда долевая однородность может быть оценена как

$$U_{partial} = \sum_{i=0}^{j-1} \frac{f_{maxFreq}(g_i)}{z_i} \Big/ j,$$

где z_i — число наблюдений, отнесенных к кластеру g_i по итогам кластеризации.

Результат максимизации долевой однородности на выборочных данных при фиксированном числе кластеров отражен на рис. 2

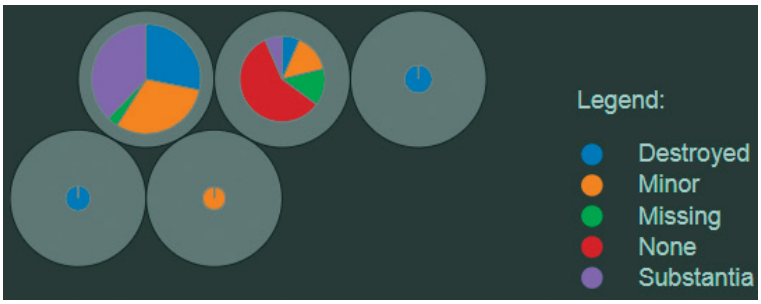


Рис. 2. Результат максимизации долевой однородности:
Destroyed — воздушное судно было полностью разрушено в результате происшествия; Minor — воздушное судно получило незначительные повреждения в результате происшествия; Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна; None — воздушное судно не было повреждено в результате происшествия; Substantia — воздушное судно получило значительное повреждения в результате происшествия

Пусть $f_{uniqueLabel}(g_{i=\{0\ldots j-1\}})$ — функция, возвращающая количество уникальных меток класса в i-*ом* кластере. Тогда конструктивная однородность

$$U_{formation} = \sum_{i=0}^{j-1} \frac{1}{f_{uniqueLabel}(g_i)} \Big/ j.$$

Результат максимизации конструктивной однород- ности на том же массиве данных отражен на рис. 3.

Несмотря на схожесть результатов максимизации обоих критериев, легко заметить, что в первом случае (рис. 2) предпочтение отдается увеличению доли выигравшего класса в каждом из кластеров, а во втором (рис. 3) меньшему числу классов, представленных в рамках одного кластера.

Оценка взвешенности, основанная на предположении о равных размерах результирующих кластеров,

$$W = 1 - \sum_{i=0}^{j-1} \sqrt[2]{(A - z_i)^2} \Big/ l,$$

где z_i — число наблюдений отнесенных к кластеру g_i по итогам кластеризации, l — общий объем выборки.

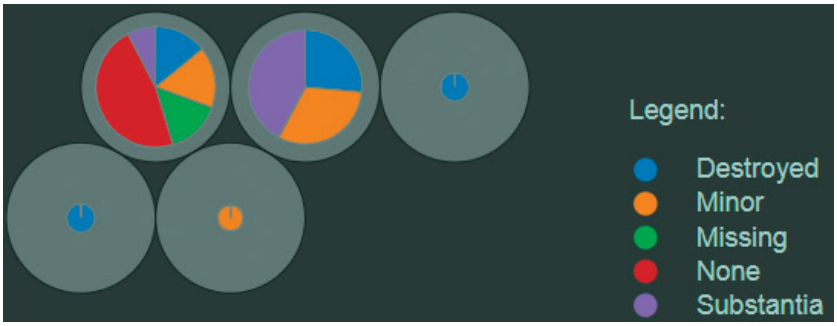


Рис. 3. Результат максимизации конструктивной однородности:
Destroyed — воздушное судно было полностью разрушено в результате происшествия; Minor — воздушное судно получило незначительные повреждения в результате происшествия; Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна; None — воздушное судно не было повреждено в результате происшествия; Substantia — воздушное судно получило значительное повреждения в результате происшествия

Результат максимизации взвешенности с теми же исходными данными показан на рис. 4.

Рис. 4. Результат максимизации взвешенности:
Destroyed — воздушное судно было полностью разрушено в результате происшествия; Minor — воздушное судно получило незначительные повреждения в результате происшествия; Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна; None — воздушное судно не было повреждено в результате происшествия; Substantia — воздушное судно получило значительное повреждения в результате происшествия

Совокупный функционал качества может быть записан следующим образом:

$$Q\left(V_{\dot{X}}^{C}\right) = \frac{U_{partial} \times a_{partial} + U_{formation} \times a_{formation} + W \times a_{balance}}{a_{partial} + a_{formation} + a_{balance}},$$

где $a_{partial}$, $a_{formation}$ и $a_{balance}$ — коэффициенты усиления соответствующих компонент заданного функционала качества, позволяющие обозначить желаемые характеристики результатов кластеризации, получаемых в процессе оптимизации. Результат максимизации совокупного функционала качества при попарно равных $a_{partial}$, $a_{formation}$ и $a_{balance}$ отражен на рис. 5.

Значения $Q\left(V_{\dot{X}}^{C}\right)$ — нормированы к единице, строгое равенство 1 достигается при описанных процедурах оценки в случае, когда все кластеры имеют равные размеры и внутри каждого из кластеров находятся наблюдения только одного класса.

Рис. 5. Результат максимизации трехкомпонентного функционала качества: Destroyed — воздушное судно было полностью разрушено в результате происшествия; Minor — воздушное судно получило незначительные повреждения в результате происшествия; Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна; None — воздушное судно не было повреждено в результате происшествия; Substantia — воздушное судно получило значительное повреждения в результате происшествия

Стохастическое решение задачи комбинаторной оптимизации

Как говорилось ранее, задачу поиска $\hat{X}$ можно рассматривать как задачу комбинаторной оптимизации. Для решения оптимизационных задач, не имеющих явной аналитической интерпретации, часто применяют методы поиска квазиоптимальных параметров, основанные на численных оценках градиента функционала качества, либо на стохастических методах направленного перебора (Гладков, 2009; Куравский, 2017; Формалев, Ревизников, 2004). В этом разделе приводится описание стохастического метода оптимизации, основанного на методе «роя частиц» в приложении к сформулированной проблеме поиска оптимального с точки зрения кластерной чистоты подмножества компонент исходного параметрического вектора, полученного из текстовых данных.

Идея метода «роя частиц» была предложена и развита в работах (3. Kennedy, 1995; 3. Kennedy, 2001; 6. Eberhart, 1995); в оригинальном приложении алгоритм использовался для моделирования социального поведения птиц, пчел и других животных, для которых характерно пространственное перемещение в рамках больших групп (стаи, роя). Позже была отмечена возможность эффективного использования такой модели для исследования признаковых пространств, в частности, поиска квазиоптимальных решений многомерных оптимизационных задач.

Роевые алгоритмы оптимизации можно отнести к эволюционным, общая схема перебора решений описывается следующей последовательностью шагов.

1. Генерируется популяция «особей», каждая из которых содержит некоторое случайное решение целевой задачи. Поиск решения представлен итерационным процессом (п. 2 и 3). В каждой итерации (эпохе) решения (позиции) всех особей незначительно модифицируются по правилам, обеспечивающим схождение итерационного процесса к квазиоптимальным решениям.

2. Вычисляется направление изменения позиции каждой особи, которое зависит от ее текущей позиции, наилучшего решения полученного данной особью за историю ее «существования» (локальным экстремумом) и наилучшим известным решением (полученным любой особью) для всей популяции (глобальным экстремумом).

3. Вычисляются новые позиции особей (их координаты) в соответствии с направлениями, полученными на шаге 2.

4. Проверяются критерии останова; если параметры решения им не соответствуют, переходят к шагу 2, в противном случае поиск прерывается.

Баланс между локальными и глобальными тенденциями в поведении особей определяется коэффициентами интерпретируемыми как ускорения движения особей в направлении локального и глобального экстремума. В первоначальной постановке задачи подразумевалось вещественное пространство решений, что не позволяло использовать метод в задачах линейного программирования, в частности, задачах комбинаторной оптимизации. Рядом авторов была предложена адаптация метода для линейных задач (Eberhart, 1995), при этом «ускорения» получили вероятностную интерпретацию. Рассмотрим более подробно модифицированную версию алгоритма, описанного в (Eberhart, 1995), которая может быть использована для поиска квазиоптимальных в смысле $Q(V^{C}_{\dot{X}})$ значений $\dot{X}= \{X_{m_0}, \dots, X_{m_{b-1}}\}$.

Поиск оптимальных комбинаций параметров из $X = \{x_0, \dots, x_{k-1}\}$ можно сформулировать как задачу о «сборке многомерного рюкзака» с однократным выбором. Ее решение $m = \{m_0, \dots, m_{b-1}\}$ представимо как вектор $P = \{p, \dots, p\}$, где $p \in \{0,1\}$, его компоненты с номерами из **m** равны 1, а все остальные 0, т. е. единица в *i-й* позиции P означает, что x_i входит в подмножество компонент $\dot{X}$, выбранных для кластеризации.

Популяция особей состоит из d решений, каждая из особей хранит текущее решение $P_{i=0 \dots d-1}$, лучшее из полученных ей решений P_{ibest} и два вектора R^0_{ie}, R^1_{ie}, определяющих частоту инверсии каждого из k бит в каждом из возможных направлений: R^0_{ie} — вероятность заменить 1 на 0 в позиции $e_{0 \dots k-1}$ для i-ой особи; R^1_{ie} — вероятность заменить 0 на 1 в позиции $e_{0 \dots k-1}$ для i-ой особи. Значения R^0_{ie}, R^1_{ie} будут изменяться на каждой итерации, как и позиции особей.

Для позиций каждой особи $p_{i=0 \dots d-1}$ на каждой итерации вероятность инверсии каждого бита $e_{0 \dots k-1}$ обозначим $R^{inversion}_{ie}$, она будет определяться на основе ее текущей позиции по следующему правилу:

$$R^{inversion}_{ie} = \begin{cases} R^0_{ie}, \text{если } p_{ie} = 1 \\ R^1_{ie}, \text{если } p_{ie} = 0 \end{cases}.$$

На производные ускорения $\ddot{R}^0_{le}$, $\ddot{R}^1_{le}$ будут влиять их текущие значения, лучшее локальное решение P_{ibest}, лучшее глобальное решение P_{gbest} и коэффициент инерции $a_{intertia}$:

$$\ddot{R}^0_{ie} = a_{inertia} \times R^0_{ie} + D\left(P_{ibest,e}\right)^0 + D(P_{gbest,e})^0;$$

$$\ddot{R}^1_{ie} = a_{inertia} \times R^1_{ie} + D\left(P_{ibest,e}\right)^1 + D(P_{gbest,e})^1.$$

Для вычисления $D(P_{ibest})^{0,1}$ и $D(P_{gbest})^{0,1}$ используются два масштабирующих коэффициента $a_{globlAttraction}$ и $a_{localAttraction}$, значения которых лежат в пределах {0...1}, они формируют баланс между локальными и глобальными тенденциями в полученной производной позиции. Кроме этого, на каждой итерации рандомизированно генерируется другая пара масштабирующих значений — $a^{stochastic}_{globlAttraction}$ и $a^{stochastic}_{localAttraction}$ — в пределах {0...1}. Итоговые значения формируются по следующему правилу:

$$\text{Если } P_{ibest,e} = 0 \text{ то } \begin{cases} D\left(P_{ibest,e}\right)^0 = a_{localAttraction} \times a^{stochastic}_{localAttraction} \\ D\left(P_{ibest,e}\right)^1 = -(a_{localAttraction} \times a^{stochastic}_{localAttraction}) \end{cases};$$

$$\text{Если } P_{ibest,e} = 1 \text{ то } \begin{cases} D\left(P_{ibest,e}\right)^0 = -(a_{localAttraction} \times a_{localAttraction}^{stochastic}) \\ D\left(P_{ibest,e}\right)^1 = a_{localAttraction} \times a_{localAttraction}^{stochastic} \end{cases};$$

$$\text{Если } P_{gbest,e} = 0 \text{ то } \begin{cases} D\left(P_{gbest,e}\right)^0 = a_{globlAttraction} \times a_{globlAttraction}^{stochastic} \\ D\left(P_{gbest,e}\right)^1 = a_{globlAttraction} \times a_{globlAttraction}^{stochastic} \end{cases};$$

$$\text{Если } P_{gbest,e} = 1 \text{ то } \begin{cases} D\left(P_{gbest,e}\right)^0 = -(a_{globlAttraction} \times a_{globlAttraction}^{stochastic}) \\ D\left(P_{gbest,e}\right)^1 = a_{globlAttraction} \times a_{globlAttraction}^{stochastic} \end{cases}.$$

Фактически, вероятность инверсии каждой компоненты вектора снижается при условии, что ее значение совпадает с соответствующим значением известного оптимального решения и возрастает — в противном случае.

Для вычисления новых позиций каждой особи i компоненты соответствующего ей решения e инвертируются с вероятностями $R_{ie}^{inversion}$, для этого значения $R_{ie}^{inversion}$ сопоставляются со случайными значениями RND_{ie}, генерируемыми при каждом сравнении. К вектору предварительно применяется следующее нормализующее условие:

$$R_{ie}^{inversion} = \frac{1}{1+e^{-R_{ie}^{inversion}}}.$$

Тогда правило определения производных позиций будет выглядеть так:

$$\ddot{p_{i,e}} = \begin{cases} \overline{p_{i,e}} \text{ если } \text{RND}ie < R_{ie}^{inversion} \\ p_{i,e} \text{если } \text{RND}ie \geq R_{ie}^{inversion} \end{cases}.$$

Содержательно приведенный алгоритм соответствует описанному в (Eberhart, 1995). Очевидно, что подобная процедура после определенного числа шагов значительно снизит вероятность появления в популяции новых решений, т. е. алгоритм «застрянет» в точке найденного локального экстремума. Для преодоления проблемы локальных экстремумов предлагается ввести процедуру рандомизации позиций, основанную на критерии «плотности роя». Под плотностью предлагается понимать обобщенную оценку отклонения позиции каждой особи от P_{gbest}; если $f_{hammDist}(a, b)$ — функция, возвращающая расстояние Хэмминга между бинарными векторами a и b, плотность роя будет оцениваться как

$$SwarmDencity = 1 - \frac{\sum_{i=0}^{d-1}(f_{hammDist}(P_{gbest}, p_i)/k)}{d},$$

где k —длина бинарного вектора решения, d — количество особей в популяции.

Область значений *SwarmDencity* соответствует интервалу {0...1}, при этом значение равное единице указывает на то, что текущие позиции всех особей совпадают наилучшим из известных решений, найденных алгоритмом. Значение *SwarmDencity* оценивается в конце каждой итерации; в случае превышения заданного порога плотности предлагается выполнять рандомизацию позиций определенного процента особей и соответствующих векторов ускорений R_{ie}^0, R_{ie}^1, с сохранением сведений об их лучших локальных решениях. Такой подход позволяет автоматически выводить численную процедуру из локальных экстремумов

без потери общего направления поиска определенного в процессе оптимизации. Мéньшие значения порога для *SwarmDencity* будут приводить к менее интенсивному поиску решения в области текущего экстремума, и наоборот.

Оценка эффективности при кластеризации данных на естественном языке

Предложенная концепция была протестирована на данных, представленных на естественном языке, об авиапроисшествиях, взятых из открытой базы «Aviation safety network» (Aviation safety network, 2017; https://aviation-safety.net/database/). В исходных данных содержались описания происшествий на английском языке и категориальные переменные, связанные с этим происшествием (уровень повреждений, тип судна, фаза полета и т. д.). Описания происшествий были параметризованны с использованием технологии word2vec и свободно распространяемого словаря, обученного на текстах агрегатора новостной информации Google News; каждому наблюдению был сопоставлен 300-мерный числовой вектор Эти векторы рассматриваются как интегральные оценки семантики описаний, не имеющие явной интерпретации в контексте указанных категориальных переменных. Для проверки предложенного алгоритма был проведен вычислительный эксперимент, целью которого ставилось снижение размерности многомерных векторных описаний с максимизацией их дискриминирующей силы в отношении уровня повреждений, полученных в результате инцидента.

Рассматривались следующие уровни значений повреждения:

• Serious — воздушное судно было существенно повреждено в результате происшествия;

• Minor — воздушное судно получило незначительные повреждения в результате происшествия;

• None — воздушное судно не было повреждено в результате происшествия;

• Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна.

Обучение выполнялось на выборке из 600 наблюдений — 300 использовались в качестве обучающей выборки, 300 — в качестве контрольной. Разделение выполнялось на 4, 8 и 12 кластеров с использованием при вычислении функционала качества метода k-средних и самоорганизующихся карт Кохонена в качестве алгоритмов кластеризации.

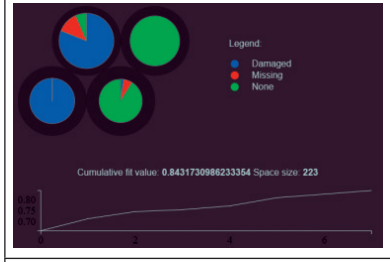
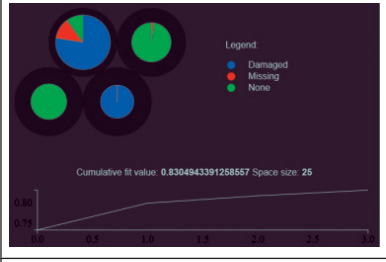
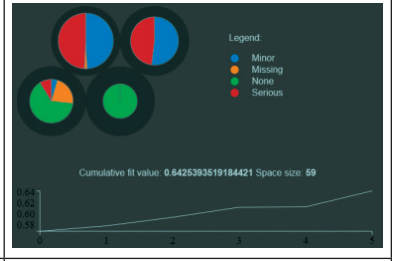
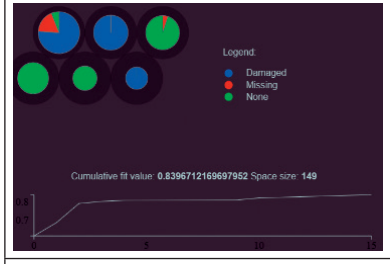
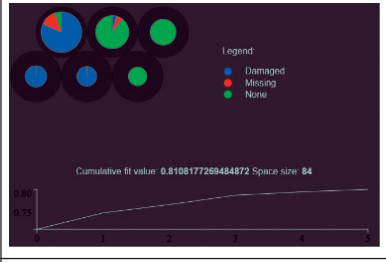
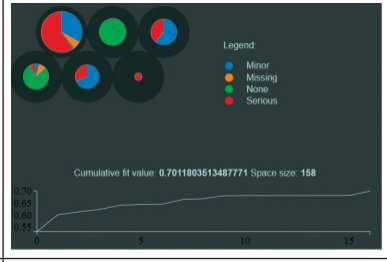
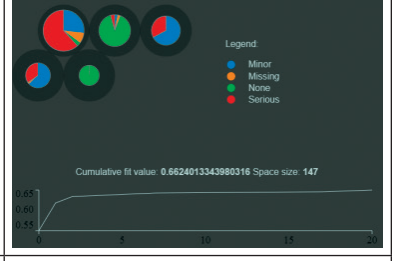
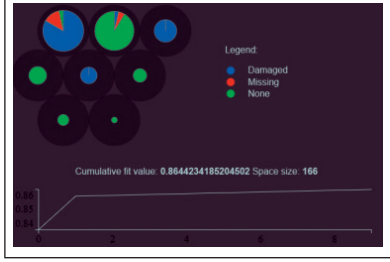
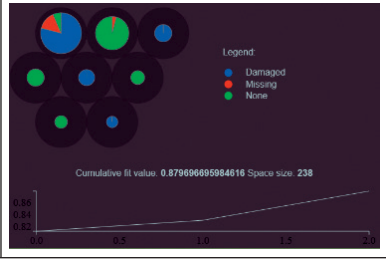
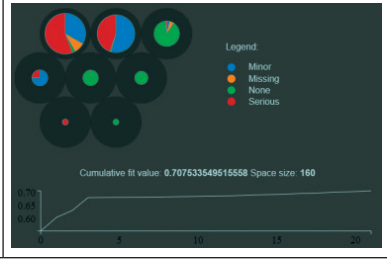
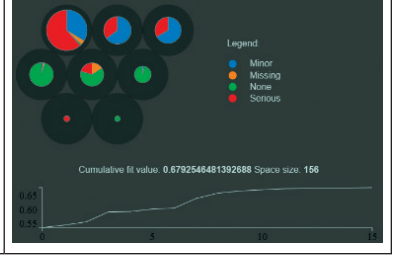
Графически результаты этого исследования представлены в табл. 1.

В результатах кластеризации полученных с использованием 4 уровней значений категориальной переменной (уровень повреждения) заметно, что случаи с незначительными повреждениями (minor, синий сектор) и с серьезными повреждениями (serious, красный сектор) регулярно объединяются в один кластер. В действительности, эти описания близки содержательно: использование способа нелинейного снижения размерности для визуализации расположения многомерных (300-компонентных) точек наблюдений в трехмерном пространстве — t-SNE показывает, что точки с соответствующими уровнями повреждений слабо разделимы (рис. 6).

При агрегации категорий Minor и Serious в один класс Damaged (табл. 1, колонки 1 и 2) результат разделения существенно улучшается. В терминах ошибки определения класса на контрольной выборке ошибка составляет ~12% ошибок для разбиения на 4 кластера и ~8% ошибок при разбиении на 6 и 8 кластеров. Результаты слабо зависят от выбранного метода кластеризации. Этот вывод может быть специфичен для данной конкретной задачи.

Таблица 1

Визуализация чистоты кластерных структур, полученных при снижении размерности текстовых описаний, при использовании 3 уровней повреждений (3 types) и 4 уровней повреждений (4 types) методами k-средних (k-mean) и с использованием самоорганизующихся карт Кохонена (SOM)

k-mean (3 types)	SOM (3 types)	k-mean (4 types)	SOM (4 types)
			
			
			

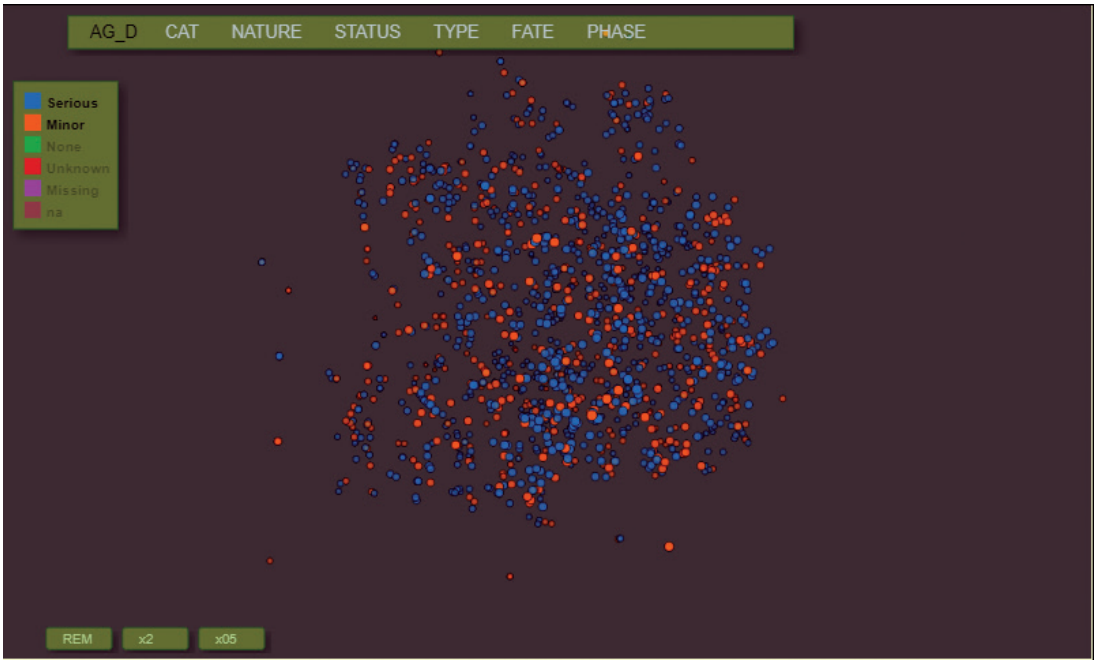


Рис. 6. Отображение распределения наблюдений
категорий Minor и Serious в 3-х мерное пространство:
CAT — категория событий; NATURE — характер полета; STATUS — статус расследования происшествия; TYPE — производитель и модель, участвующего в аварии воздушного судна; FATE — последствия для самолета, потерпевшего крушение; PHASE — фазы полета. Serious — воздушное судно получило серьезные повреждения в результате происшествия; Minor — воздушное судно получило незначительные повреждения в результате происшествия; None — воздушное судно не было повреждено в результате происшествия; Unknown — уровень повреждения не определен; Missing — в результате происшествия воздушное судно было полностью утрачено, его судьба не известна

Размерность результирующего пространства признаков в каждом из случаев, представленных в табл. 1, составила от трети до одной десятой от исходного числа компонент.

Выводы

Разработан и апробирован новый метод снижения размерности данных, обеспечивающий решения квазиоптимальные с точки зрения дискриминации заданных классов. Метод, в сочетании с технологией параметризации текстов, может использоваться для обработки записей на естественном языке в произвольных прикладных областях.

1. Результатом работы предложенного алгоритма является не только комбинация исходных признаков, но и координаты центров кластеров, объединяющих наблюдения в найденном пространстве (либо обученная сеть Кохонена в случае выбора ее в качестве метода кластеризации).

2. Предложенный метод не требуют предварительных предположений относительно вида исходного распределения наблюдаемых признаков.

3. Было выполнено прототипирование описанной процедуры снижения размерности, подтвердившее ее практическую применимость.

4. Эффективность предложенной технологии оценивалась на примере задач с вещественным исходным пространством. Метод может быть расширен на случай любых исходных пространств, в которых возможно выполнение кластеризации точек методом k-средних.

5. Сформулирован многокомпонентный функционал качества, позволяющий управлять процессом снижения размерности признакового пространства и формировать различные характеристики результирующего пространства.

6. Для комбинаторной оптимизации методом «роя частиц» предложен критерий «застревания» алгоритма в области локального экстремума. Описана процедура вывода алгоритма из этой области, выполняемая по результатам проверки критерия.

Финансирование

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации в рамках Соглашения о предоставлении субсидии от «26» сентября 2017 г. № 14.576.21.0092 (Уникальный идентификатор соглашения RFMEFI57617X0092) на выполнение прикладных научных исследований по теме: «Разработка нейросетевой системы прогнозирования авиапроисшествий и управления рисками безопасности полетов на основе ретроспективных данных, включающих множество параметров и текстовых описаний событий».

Литература

1. *Гладков Л.А.* Биоинспирированные методы в оптимизации: монография / Л. А.Гладков [и др.]. М.: Физматлит, 2009. 384 с.
2. *Куравский Л.С., Артеменков С.Л., Юрьев Г.А., Григоренко Е.Л.* Новый подход к компьютеризированному адаптивному тестированию // Экспериментальная психология. 2017. Т. 10. № 3. С. 33—45. doi:10.17759/exppsy.2017100303
3. *Куравский Л.С., Мармалюк П.А., Алхимов В.И., Юрьев Г.А.* Математические основы нового подхода к построению процедур тестирования // Экспериментальная психология. 2012. Т. 5. № 4. С. 75—98.
4. *Куравский Л.С., Мармалюк П.А., Алхимов В.И., Юрьев Г.А.* Новый подход к построению интеллектуальных и компетентностных тестов // Моделирование и анализ данных. 2013. № 1. С. 4—28.
5. *Куравский Л.С., Юрьев Г.А.* Probabilistic artifact filtration in adaptive testing // Моделирование и анализ данных. 2012. № 1. С. 70—81.
6. *Куравский Л.С., Юрьев Г.А.* Использование марковских моделей при обработке результатов тестирования // Вопросы психологии. 2011. № 2. С. 98—107.
7. *Куравский Л.С., Мармалюк П.А., Юрьев Г.А., Думин П.Н.* Численные методы идентификации марковских процессов с дискретными состояниями и непрерывным временем // Матем. моделирование. 2017. Том 29. № 5. С. 133—146.
8. *Куравский Л.С., Баранов С.Н.* Компьютерное моделирование и анализ данных: Конспекты лекций и упражнения: учеб. Пособие. М.: РУСАВИА, 2012. 18 с.
9. *Тюменева Ю.А.* Психологическое измерение. М.: Аспект-Пресс, 2007.
10. *Формалев В.Ф., Ревизников Д.Л.* Численные методы, Физматлит. М., 2004. 400 с.
11. Aviation safety network [Электронный ресурс]. — URL: https://aviation-safety.net/database/ (дата обращения: 06.12.2017).
12. *Kennedy J., Eberhart R.* Swarm Intelligence // Morgan Kaufmann Publishers, Inc. San Francisco, CA, 2001.
13. *Kennedy J., Eberhart R.* Particle Swarm Optimization // IEEE International Conference on Neural Networks (Perth, Australia). IEEE Service Center, Piscataway. NJ, 1995. P. 1942—1948.
14. *Khanesar M.A.* Novel Binary Particle Swarm Optimization, Particle Swarm Optimization [Электронный ресурс] / M.A. Khanesar, H. Tavakoli, M. Teshnehlab, M.A. Shoorehdeli, A. Lazinica (Ed.) // InTech, DOI: 10.5772/6738. 2009. URL: https://www.intechopen.com/books/particle_swarm_optimization/novel_binary_particle_swarm_optimization (дата обращения: 06.12.2017).
15. *Swamy N.* Cluster Purity Visualizer [Электронный ресурс] / N. Swamy. 2016. URL: https://bl.ocks.org/nswamy14/e28ec2c438e9e8bd302f (дата обращения: 06.12.2017).

16. *Eberhart R., Kennedy J.* A New Optimizer Using Particles Swarm Theory // Proc. Sixth International Symposium on MicroMachine and Human Science (Nagoya, Japan). NJ, 1995. IEEE Service Center, Piscataway. P. 39—43.
17. *Mikolov T., Yih W., Zweig G.* Linguistic Regularities in Continuous Space Word Representations // In Proceedings of NAACL HLT. 2013.

STOCHASTIC SWARM CLUSTERIZATION METHOD IN NATURAL LANGUAGE DATA PROCESSING

YURYEV G.A.*, *MCUPE, Moscow, Russia,*
e-mail: g.a.yuryev@gmail.com

VERKHOVSKAYA E.K.**, *MCUPE, Moscow, Russia,*
e-mail: katrin636bmw@yandex.ru

YURYEVA N.E.,*** *MCUPE, Moscow, Russia,*
e-mail: yurieva.ne@gmail.com

Consider natural language data processing technology based on non-linear dimensionality reduction method which takes into account the discriminating power of the solution found for given values of the categorical variable associated with each observation. Stochastic optimization method known as the "Particle swarm optimization" is proposed to found characteristics that ensure the best separation of observations in terms of a given quality functional. The basis for evaluating the quality of the solution lies in the purity of the clusters obtained with the k-means method, or with using self-organizing Kohonen feature maps.

Keywords: combinatorial optimization, particle swarm optimization, non-linear dimensionality reduction.

Funding

The study was supported by the Russian Ministry of Education and Science № 14.576.21.0092 (RFMEFI57617X0092).

References

1. Aviation safety network. URL: https://aviation-safety.net/database/ (06.12.2017).
2. Eberhart R. Kennedy J. A New Optimizer Using Particles Swarm Theory. *Sixth International Symposium on MicroMachine and Human Science (Nagoya, Japan)*. NJ, 1995. IEEE Service Center, Piscataway, pp. 39—43.
3. Formalev V.F., Reviznikov D.L. *Chislennye metody [Mathematical methods]*. Moscow, Fizmatlit. 2004. 400 p.
4. Gladkov L.A. *Bioinspirirovannye metody v optimizacii: monografiya [Bioinspiration methods in optimization]*. Moscow, Fizmatlit, 2009. 384 p.
5. Kennedy J., *Swarm Intelligence*. Morgan Kaufmann Publishers, Inc. San Francisco, CA, 2001.

For citation:
Yuryev G.A., Verkhovskaya E.K., Yuryeva N.E. Stochastic swarm clusterization method in natural language data processing. *Eksperimental'naya psikhologiya = Experimental psychology (Russia)*, 2018, vol. 11, no. 3, pp. 5—18. doi:10.17759/exppsy.2018110301

* *Yuryev G.A.* PhD, Docent (Associate Professor), MCUPE. E-mail: g.a.yuryev@gmail.com
** *Verkhovskaya E.K.* Researcher, MCUPE. E-mail: katrin636bmw@yandex.ru
*** *Yuryeva N.E.* PhD, Research Associate, MCUPE. E-mail: yurieva.ne@gmail.com

6. Kennedy J., Eberhart R. Particle Swarm Optimization. *IEEE International Conference on Neural Networks (Perth, Australia)*. IEEE Service Center, Piscataway. NJ, 1995, pp. 1942—1948.
7. Khanesar M.A. *Novel Binary Particle Swarm Optimization, Particle Swarm Optimization*. In M.A. Khanesar, H. Tavakoli, M. Teshnehlab, M.A. Shoorehdeli, A. Lazinica (Ed.). InTech, DOI: 10.5772/6738. 2009. URL: https://www.intechopen.com/books/particle_swarm_optimization/novel_binary_particle_swarm_optimization (06.12.2017).
8. Kuravsky L.S., Artemenkov S.L., Yuriev G.A., Grigorenko E.L. Novyj podhod k komp'yuterizirovannomu adaptivnomu testirovaniyu [New approach to computer adaptive testing]. *Eksperimental'naya psihologiya [Experimental Psychology],* 2017, vol. 10, no. 3, pp. 33—45. doi:10.17759/exppsy.2017100303
9. Kuravsky L.S., Marmalyuk P.A., Alhimov V.I., Yuriev G.A. Matematicheskie osnovy novogo podhoda k postroeniyu procedur testirovaniya [Mathematical basis of a novel approach to testing]. *Eksperimental'naya psihologiya [Experimental Psychology*], 2012, vol. 5, no. 4, pp. 75—98.
10. Kuravsky L.S., Marmalyuk P.A., Alhimov V.I., Yuriev G.A. Novyj podhod k postroeniyu intellektual'nyh i kompetentnostnyh testov [Novel approach to intellectual testing]. *Modelirovanie i analiz dannyh* [*Modeling and data analysis],* 2013, no. 1, pp. 4—28.
11. Kuravsky L.S., Yuriev G.A. Probabilistic artifact filtration in adaptive testing. *Modelirovanie i analiz dannyh* [*Modeling and data analysis*], 2012, no. 1, pp. 70—81.
12. Kuravskiy L.S., Yuriev G.A. Ispol'zovanie markovskih modelej pri obrabotke rezul'tatov testirovaniya [Markov models in testing data analysis]. *Voprosy psihologii [Issues in Psychology],* 2011, no 2, pp. 98—107.
13. Kuravsky L.S, Marmalyuk P.A., Yuriev G.A., Dumin P.N. Chislennye metody identifikacii markovskih processov s diskretnymi sostoyaniyami i nepreryvnym vremenem [Mathematical methods of markov processes in discrete state in time]. *Matem. Modelirovanie [Mathematical modeling*], 2017, vol. 29, no. 5, pp. 133—146.
14. Kuravsky L.S., Baranov S.N. *Komp'yuternoe modelirovanie i analiz dannyh: Konspekty lekcij i uprazhneniya: ucheb. Posobie [Computer modeling and data analysis].* Moscow, Rusavia, 2012. 18 p.
15. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations. *Proceedings of NAACL HLT*, 2013.
16. Swamy N. *Cluster Purity Visualizer*. 2016. URL: https://bl.ocks.org/nswamy14/e28ec2c438e9e8bd302f
17. Tyumeneva Y.A. *Psihologicheskoe izmerenie [Psychological measurement].* Moscow, Aspekt-Press, 2007.