КЛАСТЕРНЫЙ АНАЛИЗ ВИЗУАЛЬНОГО ВОСПРИЯТИЯ СТРУКТУРЫ ДАННЫХ

В.В. ЛАПТЕВ
кандидат искусствоведения, доцент кафедры инженерной графики и дизайна Института металлургии, машиностроения и транспорта, Санкт-Петербургский политехнический университет Петра Великого
Адрес: 195251, г. Санкт-Петербург, ул. Политехническая, д. 29
E-mail: laptevsee@yandex.ru

П.А. ОРЛОВ
старший преподаватель кафедры инженерной графики и дизайна Института металлургии, машиностроения и транспорта, Санкт-Петербургский политехнический университет Петра Великого; Исследователь Университета Восточной Финляндии
Адрес: 195251, г. Санкт-Петербург, ул. Политехническая, д. 29
E-mail: paul.a.orlov@gmail.com

Структуры данных являются распространенными показателями в среде управления бизнес-проектами. Инфографика как особое направление коммуникационного дизайна предусматривает ряд графических способов, позволяющих визуализировать информацию такого рода. Применение каждого из имеющихся типов диаграмм сопряжено с определенными ограничениями, связанными с особенностями визуального восприятия и семиотической спецификой. Из-за недостаточной степени изученности был выбран тип структурной диаграммы – потоковая диаграмма Сэнкей, которая часто используется в бизнес-процессах для представления структуры данных. Для выявления методов оценки формы графического образа визуализации структуры данных был проведен эксперимент, в котором в качестве стимула выступала 4-потоковая диаграмма. Результаты глазодвигательной активности человека фиксировались с помощью системы видеоокулографии или ай-трекера.

В качестве метода анализа были приняты иерархические дивизимные алгоритмы, работающие с универсальным кластером, состоящим из всех зрительных фиксаций, с последующим пошаговым разбиением его на меньшие части. Было обнаружено, как минимум, четыре кластера, основанных на координатах. В найденной модели присутствовал «входной» кластер и «выходная группа кластеров» и явно определился центральный кластер зрительных фиксаций. При дальнейшем увеличении числа кластеров картина менялась в сторону большей детализации. Очевидно, что прослеживается определенный нарратив при рассматривании диаграммы, выявляющий последовательность «движения» потока, от целого к его структурным частям. В итоге кластерная алгоритмизация их анализа позволяет перевести визуальную интерпретацию структур числовых данных в круг задач поддержки принятия решений, решаемых с помощью программных средств.

Ключевые слова: кластерный анализ, инфографика, визуализация данных, структура данных, диаграмма, видеоокулография, ай-трекер.

Цитирование: Лаптев В.В., Орлов П.А. Кластерный анализ визуального восприятия структуры данных // Бизнес-информатика. 2015. № 3 (33). С. 34–43.

Введение

Управление информацией, которая подразумевает построение структур различного рода и сценариев их восприятия, является важной прикладной задачей. Структуры данных являются распространенными показателями в среде управления бизнес-проектами. Вопросы, связанные с их визуальной интерпретацией, являются актуальными в связи с вариативностью формы представления. До сих пор существуют проблемы визуализации числовых данных, возникающие при выборе соответствующих типов диаграмм. Так, один и тот же пример структуры данных может быть представлен в различных паттернах: брусковых, секторных, плоскостных или потоковых диаграммах. Правильный выбор формы основывается не только на учете контекста и семантических связей между числовым массивом и его графическим образом, но и на удобочитаемости графика и простоте его восприятия. Это касается визуализации не только структуры числовых данных, но и классификационных или иерархических схем с количественным анализом.

1. Изучение вопросов формообразования при визуализации структуры данных

Графическое представление таких результатов, а значит, и определение формы сообщения, относится к вопросам коммуникативного дизайна. Инфографика, как его составная часть, рассматривается как возможный объект прикладной информатики [1, 2]. Необходимо рассмотреть особенности формирования структурных диаграмм с точки зрения удобства визуального восприятия и точности анализа представляемых данных. В работе предлагаются методы определения эффективности выбора формы визуального образа, основанные на кластерном анализе глазодвигательной активности человека с помощью системы видеоокулографии или ай-трекера (*eye-tracker*). Это позволяет не только включить в исходные параметры точность и скорость определения параметров, но и зарегистрировать направление взгляда наблюдателя, длительность фиксации и протяженность саккад.

Первоначально исследования, касающиеся определения эффективности вида структурных диаграмм, рассматривали в качестве оппозиции дилемму «что лучше: столбик или сектор?», оставляя за скобками другие формы представления структуры. Данные виды формы графического образа являлись основными для представления структуры числовых данных. В многочисленных работах исследовались вопросы применения брусковых (столбиковых и полосовых) и секторных диаграмм [3–6]. В основе проводимых экспериментов лежала оценка скорости и точности определения структуры (в процентах от целого). Собранные данные анализировались с помощью статистических методов оценки количественных отклонений показаний испытуемых с предъявляемыми стимулами. На основании этого было подтверждено предположение, что точность и скорость оценки структуры зависит, в первую очередь, от пропорционального состава долей, а уже во вторую – от вида диаграммы. Например, при визуализации соотношения 50% и 50% предпочтение отдавалось секторной диаграмме, для иных размеров долей – брусковой и т. п.

Продолжение дискуссии о соответствии формы диаграммы структуре числовых данных велась попрежнему на основе экспериментальных данных о точности и скорости визуального восприятия. Однако во главу угла был поставлен статистический метод качественной оценки распознавания, когда респондент отмечал соотношение долей и их совокупностей: «меньше, больше, равно» [7]. Эти методы использовались при расширении поля исследований визуальной структуры данных на другие типы формы графического образа: линейные и брусковые [8], брусковые и плоскостные [9], секторные, брусковые, кольцевые, плоскостные диаграммы [10]. Кроме того, в последнее время из-за популярности оперативных диаграмм управления бизнес-процессами (*dashboards*) значительное внимание стало уделяться исследованиям принципов визуализации сложных иерархических структур [11, 12]. При этом следует отметить, что изучению восприятия потоковых диаграмм (или диаграмм Сэнкей), которые зачастую входят в состав графических комплексов контроля, уделяется недостаточное внимание.

2. Условия применения потоковых диаграмм

Возникновение потоковых диаграмм относится к середине XIX в., когда возникла потребность в количественной информации о трафике для постройки новых дорог, мостов, каналов и предприятий. Требовались разнообразные данные о физических, технических, политических и геостратегических условиях в различных точках государства. Кроме того, возникла необходимость учитывать распре-

деление и мобильность людей и капитала. Такая корреляция экономических ресурсов и демографии была бы наиболее наглядна на картах, где с помощью графического языка выражалась бы количественная информация.

Примером такого подхода служит британский атлас (*Atlas to accompany the second report of the railway commissioners,* 1838). Его автором был железнодорожный инженер Генри Харнесс (*Henry Drury Harness*), член комиссии по изучению железных дорог. В 1837 г. он проиллюстрировал доклад комиссии серией плоскостных и потоковых картодиаграмм, представляющих распределение населения в городах Великобритании и соответствующее перемещение товаров и пассажиров железных дорог. Способ линейного изображения количественных показателей мог наглядно показать перемещение на карте того или иного экономического объекта: пассажиров, грузов, капитала, электроэнергии и т. п. Потоковые диаграммы могли соединять точки на карте прямыми, но чаще в качестве оси абсцисс использовались определенные топографические линии: реки, морские пути, железные или шоссейные дороги, трубопроводы, высоковольтные линии.

Наиболее известными статистиками прошлого, широко применявшими этот метод на практике, были французский инженер Шарль Минар (*Charles Joseph Minard*) и бельгийский железнодорожный инженер Альфред Бельпер (*Alfred Jules Belpaire*). В 1845 г. Минар показал возможность потоковых диаграмм на примере пассажирского трафика между городами Дижон и Мюлуз. Посредством толщины линии он выразил количественные показатели, которые были перенесены в координатную систему, где ось абсцисс выполняли железные дороги. Каждый миллиметр толщины означал тысячи перевезенных пассажиров. В русских экономических картах потоковые диаграммы с «масштабными полосками» начал применять И.Ф. Борковский в 1870-х гг. в отчетах экспедиции, снаряженной Вольным экономическим и Русским географическим обществами для исследования хлебной торговли и производства в России.

Потоковые диаграммы использовались для визуализации связей с количественными характеристиками не только на картах, но и на блок-схемах процесса, иерархических графах и т.п. В них ширина линий пропорциональна количеству потока, визуализирующего, например, баланс, переводы между процессами или структуру затрат. Такой специфический тип потоковой диаграммы получил название «диаграмма Сэнкей» (*sankey diagram*), которое происходит от имени Мэтью Сэнкея (*Matthew Henry Phineas Riall Sankey*), ирландского инженера XIX в. Он в 1898 г. использовал этот способ графического представления информации для сравнения эффективности использования энергии парового двигателя.

При визуализации структуры целого потоковые диаграммы расставляют визуальные акценты на динамике передачи данных, т.е. на потоках внутри системы. Они выявляют доминирующие части, полезны в поиске «слабого звена», показывают балансы показателей в системе. К подобному способу визуализации структуры можно отнести и графики параллельных координат, которые очень близки по графическому образу к потоковым диаграммам и служат для количественной характеристики связей. Здесь, как и в потоковых диаграммах, происходит не только разделение целого на доли, но и их визуальное обособление. Разделение долей столбика или полосы происходит и в так называемой структурной диаграмме водопада (*waterfall chart*). В общем, потоковую диаграмму можно интерпретировать как динамическую модификацию брусковой диаграммы. Следует отметить, что уже проводилось исследование вопроса восприятия структурных диаграмм с объединенными и разделенными долями [13]. В результате было получены данные о том, что элементы визуальной структуры могут в определенной степени предсказуемо влиять на семантическую интерпретацию данных, которая выходит за пределы простого считывания данных. Это необходимо учитывать при разработке и оценке методов визуализации. Однако методы, в основе которых лежит анализ точности и скорости восприятия данных или качественная (но приблизительная) оценка структуры, не рассматривают набор данных в целом на основе элементов визуального дизайна. Для выяснения того, как пользователь оценивает композицию визуализации, должна быть принята во внимание глазодвигательная активность человека: направление взгляда, длительность фиксаций и протяженность саккад.

3. Постановка эксперимента

Для определения условий выбора формы структурной диаграммы была выдвинута следующая гипотеза: «Паттерн зрительных фиксаций рассма-

тривания имеет связь с структурой потоковой диаграммы в условиях соответствующей задачи». Для уточнения гипотезы введем вопрос исследования: «Является ли семантическая основа потоковой диаграммы детерминирующим фактором для положения фиксаций взора?» Другими словами, проверяется, будут ли фиксации взора испытуемого группироваться в левой части диаграммы («входная часть» – целое) и в правой части («выходная часть» – доли); будет ли переходная фаза потоковой диаграммы («центральная часть») оставлена без внимания.

Человеческий глаз постоянно (за исключением некоторых фаз сна) находится в движении. Принято различать в глазодвигательной активности определенные фазы: дрейфы, фиксации, варианты саккад, нистагмы [14–16]. Интерес представляют зрительные фиксации – дрейф, медленное, плавное перемещение глаза в небольшой зоне и саккады – скачкообразные движения высокой скорости, при которых резко изменяется позиция глаза. Считается, что зрительная информация обрабатывается в момент фиксации [17]. Проверка гипотезы потребовала регистрации зрительных фиксаций, для чего в настоящей работе была использована технология видеоокулографии.

Для получения первоначальных данных и апробации математических алгоритмов было принято решение остановиться на типе «*case study*» [18–20] и пригласить одного испытуемого, не ознакомленного с целями исследования. Испытуемому в случайном порядке был предъявлен стимульный материал в виде структурных диаграмм, состоящих из четырех потоков. Данная инфографика демонстрировала процентное разделение инвестиций по отдельным проектам. На экране монитора с разрешением вывода 1280 на 1024 пикселей в центре испытуемый видел потоковую диаграмму размером 500 на 500 пикселей, с направлением потоков слева направо (*рис. 1*). В нижней части экрана расположены интерактивные элементы для выбора ответа. На втором и третьем варианте выбора показаны фиксации взора испытуемого. Примером правильного ответа будет служить 4-й вариант

Испытуемому предлагалось сделать выбор из пяти вариантов ответа. Каждый вариант представлял собой вертикальный набор процентного соотношения веса выходного потока к входному. Таким образом, испытуемому необходимо было соотнести цифровые значения с шириной (весом) выходных потоков или частей структуры.

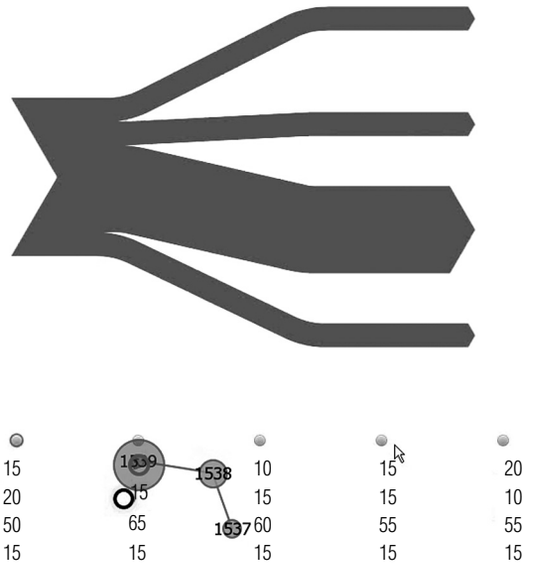

Рис. 1. Пример стимульного материала: четырехпотоковая диаграмма Сэнкей, демонстрирующая процентное разделение инвестиций по отдельным проектам

Испытуемый производил выбор с помощью компьютерного манипулятора типа мышь. Подразумевалось, что он является уверенным пользователем компьютера и задание с точки зрения человеко-компьютерного взаимодействия не является для него новым. Выбор производился кликом мышки по соответствующему варианту. Значения нерелевантных результатов подбирались таким образом, чтобы не было явного ошибочного результата, вызывающего отторжение варианта.

В процессе решения задачи глазодвигательная активность регистрировалась на оборудовании айтрекер *SMI RED250*. Кроме того, автоматически фиксировалась правильность ответа и длительность решения задачи. Для проведения эксперимента было разработано программное обеспечение на базе платформы *NetBeans*. После совершения выбора испытуемому предъявлялся стимул другой природы на 15 секунд, после чего снова предъявлялся стимул с четырехпотоковой диаграммой. Общее количество полезных стимулов – 30 штук. Однако, анализировались только те задания, с которыми испытуемый справился верно, общим числом 25 штук.

4. Анализ векторов кластеризации

Анализ глазодвигательной активности является нетривиальной задачей, в ходе решения которой требуется учесть несколько факторов одновре-

менно. Это координатное положение точки взора, протяженность саккады, длительность фиксации, размер зрачка и другие. Кластерный анализ – технология группирования объектов в ранее неизвестные группы. Он отличается от дискриминантного анализа тем, что не известны ни число кластеров, ни их характеристики. Проблема определения числа кластеров является одной из основных нерешенных до настоящего времени задач кластерного анализа. В рамках наших условий задача осложняется еще и тем, что необходимо определить число векторов кластеризации.

В качестве метода анализа были приняты иерархические дивизимные алгоритмы, работающие с универсальным кластером, состоящим из всех зрительных фиксаций, с последующим пошаговым разбиением его на меньшие части. «Удобство таких методов состоит в том, что процесс деления можно в любой момент остановить. Три наиболее популярных дивизимных метода – бисекция *k*-средних, бисекция главной компоненты и концептуальный кластер-анализ» [21, с. 10]. С помощью кластерного анализа была предпринята успешная попытка группировки результатов по назначенным признакам.

Для математического анализа были взяты данные только правильно решенных задач. Причем фиксации учитывались только в области диаграмм и не учитывались в области выбора ответа. В качестве аппарата кластеризации был выбран метод *k*-средних с многомерным вектором (под вектором кластеризации мы понимаем параметры глазодвигательной активности, которые выбираются для анализа).

Для определения качества векторов кластеризации были рассмотрены два алгоритма: *Davies Bouldin* [22, 23] и *Average Within Distance* [24]. Анализ оценки качества кластеризации с разным количеством кластеров приведен на *рис. 2*, где представлена визуализация расчетов для нескольких наборов векторов кластеризации:

– ALL 4 – для вектора, основанного на «координате X», «координате Y», «Длительности фиксации» и «Размере зрачка»;
– PUPIL – для вектора, основанного на «координате X», «координате Y» и «Размере зрачка»;
– DURATION – для вектора, основанного на «координате X», «координате Y» и «Длительности фиксации»;
– XY – для вектора, основанного на «координате X» и «координате Y»;

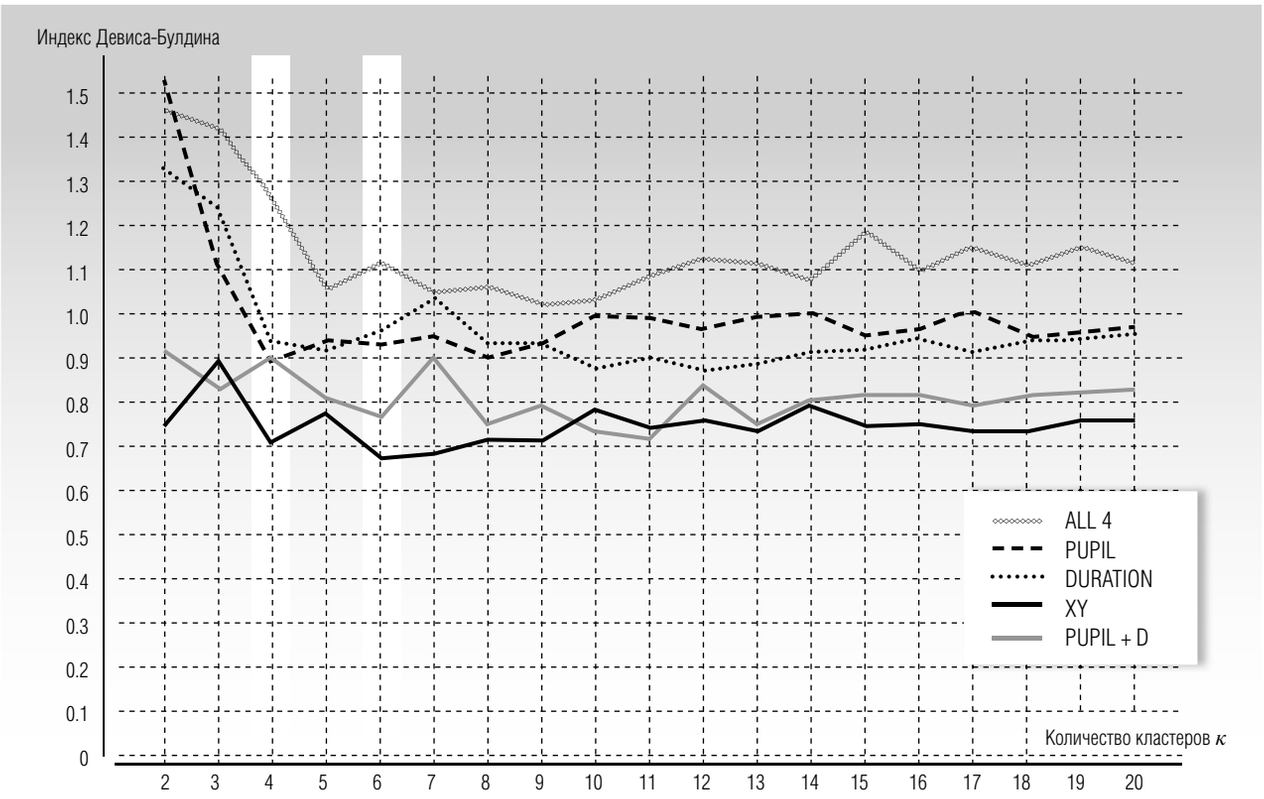

Рис. 2. Показатели качества кластеризации при количестве кластеров от 2 до 20 по алгоритму *Davies Bouldin*

– PUPIL+D – для вектора, основанного на «Длительности фиксации» и «Размере зрачка» (без «координаты X» и «координаты Y»).

График показывает, что наибольший интерес, в зависимости от состава вектора кластеризации, представляет группировка в четыре и шесть кластеров. По индексу Девиса-Булдина нас интересуют точки с наименьшим показателем. Однако для дальнейших исследований были также взяты другие размеры кластеризации: 3, 5, 7 и 8. Для большей наглядности и интерпретации предполагаемых результатов рассмотрен вектор кластеризации, содержащий два аргумента: «координату X» и «координату Y».

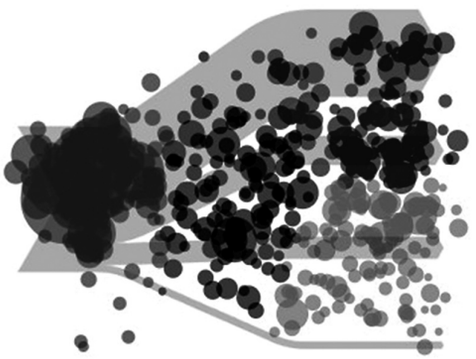

Рис. 3. Визуализация кластерной модели соотнесенной с фиксациями взора

Рабочим был принят вектор параметров для кластеризации, включающий в себя «координату X» и «координату Y». Визуализация результата при четырехкластерном разбиении представлена на *рис. 3* с демонстрацией стимульного материала одного из заданий. При этом точки фиксации показаны со всех 30 стимулов и наложены друг на друга.

5. Сравнение кластерных моделей

Для визуальной оценки кластеризации при использовании моделей с разным числом кластеров были построены соответствующие диаграммы визуализации (*рис. 4 и 5*). При разбиении на четыре кластера можно увидеть, что одна группа фиксаций сконцентрирована во входящей части стимула (потоковой диаграммы) с левой стороны, а две «выходящие» группы располагаются в его правой части. Однако обнаруживается устойчивый кластер, который не принимался во внимание при первоначальной постановке задачи: это четвертая группа, которая располагается в центральной части стимула.

При сравнении двух вариантов кластеризации можно отметить, что кластеры при группировке на шесть частей во входной части диаграммы ведут себя объяснимо. Положение нижнего левого кластера можно интерпретировать условиями эксперимента: в нижней части под диаграммой располагались варианты ответа, с которыми испытуемый должен был сверяться. Несмотря на то, что при обработки данным методом часть стимула с вариантами ответа была отсечена, результат этой сверки все же был получен в виде отдельного кластера. Особый интерес представляет шестой кластер, который располагается в правой части диаграммы, превратив два кластера на выходе в три.

На *рис. 5*, где представлены варианты моделей с 3, 5, 7 и 8 группировками, видно, что при кластеризации отсекаются вспомогательные фиксации, которые принадлежат задаче переключения внимания на сравнение с ответом.

Для проверки качества моделей было проведено тестирование моделей. В качестве алгоритма проверки

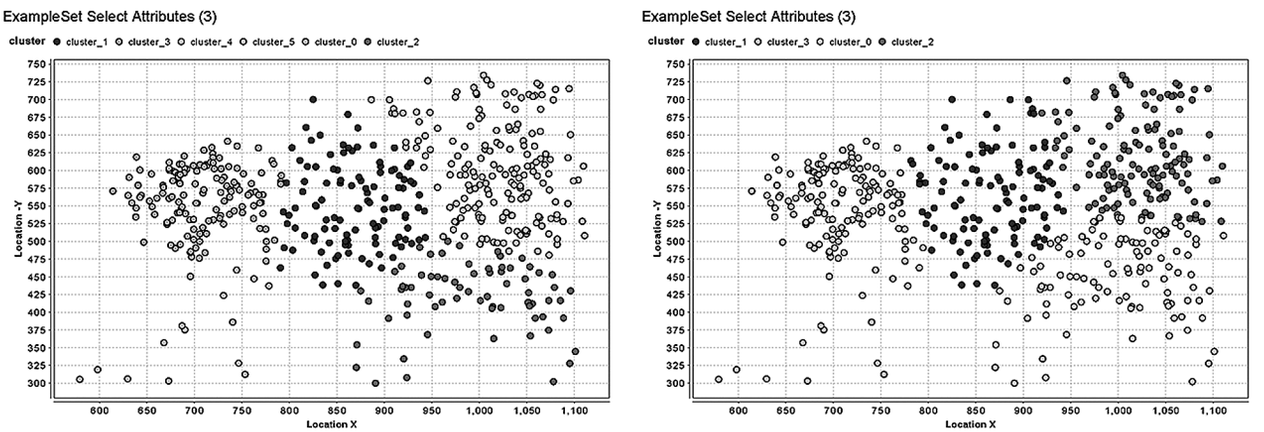

Рис. 4. Диаграммы расположения фиксаций при группировке в 4 и 6 кластеров

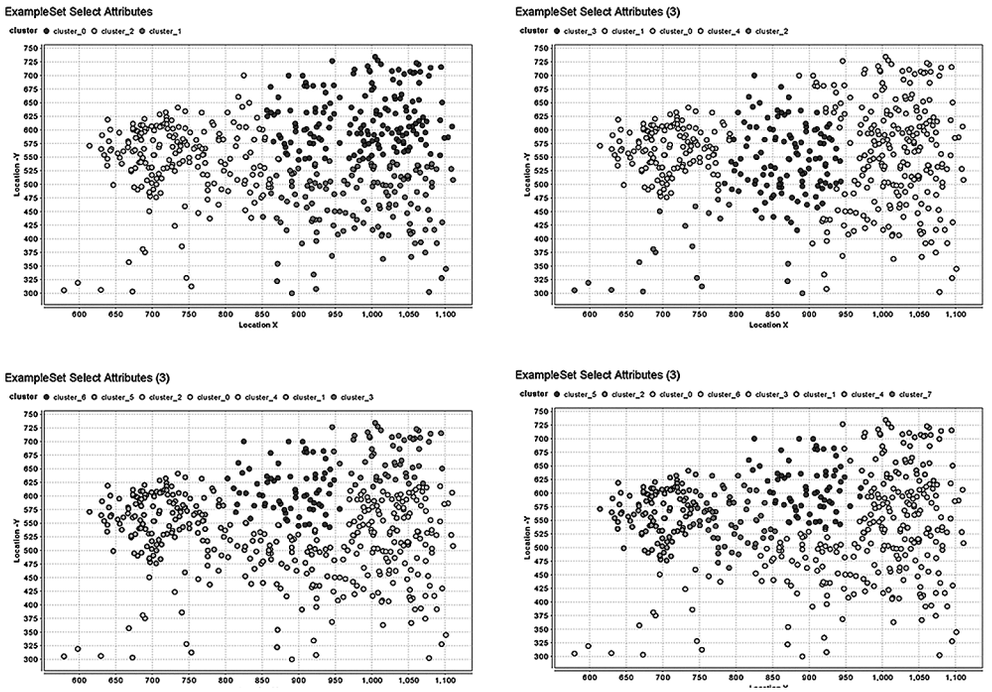

Рис. 5. Кластерные модели с 3, 5, 7 и 8 группировками

был выбран алгоритм *Map Clustering on Labels* из пакета *RapidMiner Core* [25]. Суть его работы заключается в следующем: используя как входной параметр кластерную модель с данными, он оценивает соответствие между данной моделью и моделью прогнозирования. Он настраивает входную модель и оценивает наиболее подходящие пары «элемент – кластер». Результат имеет вероятностный атрибут предсказания, который является производным от атрибута кластера.

Таблица 1.

Проверка кластерной модели с группировкой на четыре кластера

	true cluster_1	true cluster_3	true cluster_0	true cluster_2	class precision
pred. cluster_1	**117**	0	0	0	100%
pred. cluster_3	0	**151**	0	0	100%
pred. cluster_0	0	0	**109**	0	100%
pred. cluster_2	0	0	0	**135**	100%
class recall	100%	100%	100%	100%	

Результаты тестирования моделей с четырьмя и шестью кластерами приведены в *табл. 1* и *2*. Общая точность предсказания пар элементов для 4- и 6-кластерной модели составляют соответственно 100% и 98,44%. Значения точности для моделей кластеризации представлены в *табл. 3*.

Таблица 2.

Проверка кластерной модели с группировкой на шесть кластеров

	true cl._1	true cl._3	true cl.r_4	true cl._5	true cl.r_0	true cl._2	class precision
pred. cluster_1	**111**	0	0	0	0	1	99,11%
pred. cluster_3	0	**138**	0	0	0	0	100%
pred. cluster_4	0	0	**70**	1	0	0	98,59%
pred. cluster_5	0	0	0	**104**	0	4	96,30%
pred. cluster_0	1	1	0	0	**15**	0	88,24%
pred. cluster_2	0	0	0	0	0	**66**	100%
class recall	99,11%	99,28%	100%	99,05%	100%	92,96%	

Таблица 3.

Точность предсказания соотнесения элемента с кластером

Параметр\Число кластеров модели	3	4	5	6	7	8
accuracy	100%	100%	97,85%	98,44%	93,16%	90,43%

Из *табл. 3* видно, что кластеризация до шести групп включительно дает приемлемые результаты для анализа, но при увеличении числа кластеров вероятность верного предугадывания падает. Инте-

ресным представляется то, что при четырех кластерах модель демонстрирует 100% результат.

Заключение

В результате анализа визуальной оценки кластерной модели можно сделать заключение, что кластеры располагаются в соответствии со структурой стимула, т.е. четырехпотоковой диаграммы Сэнкей. Принято считать, что глазодвигательная активность детерминирована текущей задачей или общей целью [26]. Исходя из этого, полученный результат может быть интерпретирован только единственным способом: не стимульный материал сам по себе влияет на глазодвигательную активность, а задача, которую решает при этом испытуемый. Однако, по условиям эксперимента, стимульный материал является частью задачи и не может быть исключен. В таком случае решение прямой задачи по извлечению информации из потоковой диаграммы имеет связь с паттерном зрительных фиксаций. Алгоритмизация анализа кластерных моделей позволяет перевести визуальную интерпретацию структур числовых данных в круг задач поддержки принятия решений, решаемых с помощью программных средств.

Гипотеза относительно двух групп фиксаций во входной и в выходной части потоковой диаграммы не подтвердилась. Было обнаружено как минимум четыре кластера, основанных на координатах взгляда и длительности фиксации. В найденной модели присутствовал «входной» кластер и «выходная группа кластеров», также явно определился и центральный кластер зрительных фиксаций. При дальнейшем увеличении числа кластеров картина меняется в сторону большей детализации. Каждая часть структуры сравнивается с представленной таблицей данных последовательно от верхней правой части и далее вниз. Этим обусловлена правая «выходная группа кластеров». Также осуществляется сравнение каждой части с целым, представляемым «входным» кластером. Центральный кластер интерпретирует сравнение частей между собой в динамической части, когда разделение на части уже есть, но еще незначительное. В этом случае потоки композиционно объединяются в единый блок.

Дополнительное значение диаграмме дает обозначенное стрелками направление движения слева направо. Очевидно, что прослеживается определенный нарратив при рассматривании диаграммы, выявляющий последовательность «движения» потока от целого к его структурным частям. Вероятно, зритель включается в ее «потоковый» смысл. Для подтверждения данного вывода требуется дополнительные эксперименты, для анализа результатов которых можно использовать кластерный метод. ■

Литература

1. Лаптев В.В. Инфографика: основные понятия и определения // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Гуманитарные и общественные науки. 2013. № 4 (184). С. 180–187.
2. Орлов П.А. Инфографика и программирование. СПб.: Эйдос, 2013. 351 с.
3. Eells W.C. The relative merits of circles and bars for representing component parts // Journal of the American Statistical Association. 1926. Vol. 21. P. 119–132.
4. von Huhn R. A discussion of the Eells' experiment // Journal of the American Statistical Association. 1927. Vol. 22, No. 160. P. 31–36.
5. Croxton F.E., Stein H. Graphic comparisons by bar, squares, circles, and cubes // Journal of the American Statistical Association. 1932. Vol. 27, No. 177. P. 54–60.
6. Cleveland W.S., McGill R. Graphical perception: Theory, experimentation, and application to the development of graphical methods // Journal of the American Statistical Association. 1984. Vol. 79, No. 387. P. 531–554.
7. Spence I., Lewandowsky S. Displaying proportions and percentages // Applied Cognitive Psychology. 1991. No. 5. P. 61–77.
8. Zacks J., Tversky B. Bars and lines: A study of graphic communication // Memory & Cognition. 1999. Vol. 27, No. 6. P. 1073–1079.
9. Heer J., Bostock M. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design // Proceedings of ACM Human Factors in Computing Systems (CHI 2010). April 10–15, 2010, Atlanta, USA. P. 203–212.
10. Kosara R., Ziemkiewicz C. Do Mechanical Turks dream of square pie charts? // Proceedings of Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV). ACM Press, 2010. P. 373–382.
11. Ziemkiewicz C., Kosara R. Preconceptions and individual differences in understanding visual metaphors // Proceedings of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis 2009). June 10-12, 2009, Berlin, Germany. 2009. Vol. 28. No. 3. P. 911–918.

12. Kosara R., Bendix F., Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data // Proceedings of the IEEE Symposium on Information Visualization 2005 (InfoVis 2005). October 23-25, 2005, Minneapolis, MN, USA. 2005. P. 133–140.
13. Ziemkiewicz C., Kosara R. Implied dynamics in information visualization // Proceedings of the 2010 International Conference on Advanced Visual Interfaces (AVI 2010), May 25-29, 2010, Rome, Italy. 2010. P. 215–222.
14. Ярбус А.Л. Роль движений глаз в процессе зрения. М.: Наука, 1965. 165 с.
15. Гиппенрейтер Ю.Б. Движения человеческого глаза. М.: МГУ, 1978. 256 с.
16. Барабанщиков В.А., Милад М.М. Методы окулографии в исследовании познавательных процессов и деятельности. М.: Ин-т психологии РАН, 1994. 87 с.
17. Величковский В.М. Когнитивная наука: основы психологии познания. Т. 2. М.: Академия, 2006. 432 с.
18. Ma H.-H. An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median // Behavior Modification. 2006. No. 30 (5). P. 598–617.
19. Crosbie J. Interrupted time-series analysis with brief single-subject data // Journal of Consulting and Clinical Psychology. 1993. No. 61 (6). P. 966–974.
20. The use of single-subject research to identify evidence-based practice in special education / R.N. Horner [et al.] // Exceptional Children. 2005. No. 71. P. 165–179.
21. Алескеров Ф.Т., Белоусова В.Ю., Егорова Л.Г., Миркин Б.Г. Анализ паттернов в статике и динамике. Часть 1: Обзор литературы и уточнение понятия // Бизнес-информатика. 2013. № 3 (25). С. 3–18.
22. Davies D.L., Bouldin D.W. A cluster separation measure // Pattern Analysis and Machine Intelligence. IEEE Transactions. 1979. No. 2. P. 224–227.
23. Ray S., Turi R.H. Determination of number of clusters in k-means clustering and application in colour image segmentation // Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT 1999), December 27-29, 1999, Calcutta, India. 1999. P. 137–143.
24. Petrovic S. A comparison between the Silhouette index and the Davies-Bouldin index in labeling IDS clusters // Proceedings of the 11th Nordic Workshop of Secure IT Systems, October 19-20, 2006, Linkoping, Sweden. 2006. P. 53–64.
25. Graczyk M., Lasota T., Trawinski B. Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA // Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Berlin, Heidelberg: Springer, 2009. P. 800–812.
26. Бернштейн Н.А. Очерки о физиологии движений и физиологии активности. М.: Наука, 1990. 496 с.

CLUSTER ANALYSIS OF VISUAL PERCEPTION OF DATA STRUCTURE

Vladimir V. LAPTEV
Associate Professor, Department of Engineering Graphics and Design, Institute of Metallurgy, Mechanical Engineering and Transport, Peter the Great St.Petersburg Polytechnic University
Address: 29, Politekhnicheskaya Street, St. Petersburg, 195251, Russian Federation
E-mail: laptevsee@yandex.ru

Paul A. ORLOV
Senior Lecturer, Department of Engineering Graphics and Design, Institute of Metallurgy, Mechanical Engineering and Transport, Peter the Great St.Petersburg Polytechnic University; PhD candidate University of Eastern Finland
Address: 29, Politekhnicheskaya Street, St. Petersburg, 195251, Russian Federation
E-mail: paul.a.orlov@gmail.com

Data structures are common indicators in the fields of management and business. Infographics (serious graphics), a special area of Communication Design, provides a number of graphical ways to visualize this type of data. The application of each available chart type corresponds to certain limitations, which are associated with features of visual perception and semiotic aspects. In our study, we chose the Sankey flow diagram because of an insufficient degree of

scrutiny. This type of diagram is often used to represent data structure in business processes. We built an eye-tracking study to identify the methods of assessment forms of graphical image of data structure visualization. In our experiment, we used a 4-flow Sankey diagram as a stimulus.

Hierarchical divisive algorithms were taken as the method of analysis. This method works with a universal cluster consisting of all gaze fixation, followed by step partitioning it into smaller pieces. It has been found that there are at least four clusters based on the coordinates. In the present model, we found an «input» cluster and an «output cluster group» and clearly defined the central cluster of gaze fixations. Increasing the number of clusters changes the picture in the direction of greater detail. We show a certain narrative that is traced when viewing charts. This narrative identifies the sequence of flows «movement» from the whole to its structural parts. As a result, cluster analysis allows the visual interpretation of numerical data structures in a range of tasks to support decision making that can be solved by software.

Key words: cluster analysis, infographics, data visualization, data structure, diagram, eye tracking, eye tracker.

Citation: Laptev V.V., Orlov P.A. (2015) Klasternyi analiz vizual'nogo vosprijatija structury dannyh [Cluster analysis of visual perception of data structure]. *Business Informatics*, no. 3 (33), pp. 34–43 (in Russian).

References

1. Laptev V.V. (2013) Infografika: osnovnye ponjatija i opredelenija [Infographics: Basic concepts and definitions]. *St. Petersburg State Polytechnical University Journal. Humanities and Social Sciences,* no. 4 (184), pp. 180–187 (in Russian).
2. Orlov P.A. (2013) *Infografika i programmirovanie* [Infographics and programming]. St. Petersburg: Jejdos (in Russian).
3. Eells W.C. (1926) The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association*, vol. 21, pp. 119–132.
4. von Huhn R. (1927) A discussion of the Eells' experiment. *Journal of the American Statistical Association*, vol. 22, no. 160, pp. 31–36.
5. Croxton F., Stein H. (1932) Graphic comparisons by bar, squares, circles, and cubes. *Journal of the American Statistical Association,* vol. 27, no. 177, pp. 54–60.
6. Cleveland W., McGill R. (1984) Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association,* vol. 79, no. 387, pp. 531–554.
7. Spence I., Lewandowsky S. (1991) Displaying proportions and percentages. *Applied Cognitive Psychology,* vol. 5, pp. 61–77.
8. Zacks J., Tversky B. (1999) Bars and lines: A study of graphic communication. *Memory & Cognition,* vol. 27, no. 6, pp. 1073–1079.
9. Heer J., Bostock M. (2010) Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. Proceedings of *ACM Human Factors in Computing Systems (CHI 2010), April 10–15, 2010, Atlanta, GA, USA,* pp. 203–212.
10. Kosara R., Ziemkiewicz C. (2010) Do Mechanical Turks dream of square pie charts? Proceedings of *Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)*. ACM Press, pp. 373–382.
11. Ziemkiewicz C., Kosara R. (2009) Preconceptions and individual differences in understanding visual metaphors. Proceedings of the *Eurographics/ IEEE-VGTC Symposium on Visualization, (EuroVis 2009). June 10-12, 2009, Berlin, Germany*, vol. 28, no. 3, pp. 911–918.
12. Kosara, R., Bendix F., Hauser H. (2005) Parallel sets: Interactive exploration and visual analysis of categorical data. Proceedings of the *IEEE Symposium on Information Visualization 2005 (InfoVis), (InfoVis 2005). October 23-25, 2005, Minneapolis, MN, USA*, pp. 133–140.
13. Ziemkiewicz C., Kosara R. (2010) Implied dynamics in information visualization. Proceedings of the *2010 International Conference on Advanced Visual Interfaces (AVI 2010), May 25-29, 2010, Rome, Italy*, pp. 215–222.
14. Jarbus A.L. (1965) *Rol' dvizhenij glaz v processe zrenija* [The role of eye movements in vision]. Moscow: Nauka (in Russian).
15. Gippenrejter Ju.B. (1978) *Dvizhenija chelovecheskogo glaza* [Movement of the human eye]. Moscow: MGU (in Russian).
16. Barabanshhikov V.A., Milad M.M. (1994) *Metody okulografii v issledovanii poznavatel'nyh processov i dejatel'nosti* [Oculography methods in the study of cognitive processes and activities]. Moscow: Institut psihologii RAN (in Russian).
17. Velichkovskij B.M. (2006) *Kognitivnaja nauka: osnovy psihologii poznanija* [Cognitive science: foundations of cognitive psychology]. Moscow: Akademija (in Russian).
18. Ma H.-H. (2006). An alternative method for quantitative synthesis of single-subject researches: percentage of data points exceeding the median. *Behavior Modification*, vol. 30 (5), pp. 598–617.
19. Crosbie J. (1993) Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, no. 61 (6), pp. 966–974.
20. Horner R.H., Carr E.G., Halle J., McGee G., Odom S., Wolery M. (2005) The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, no. 71, pp. 165–179.
21. Aleskerov F.T., Belousova V.Yu., Egorova L.G., Mirkin B.G. (2013) Analiz patternov v statike i dinamike. Chast' 1: Obzor literatury i utochnenie ponjatija [Methods of pattern analysis in statics and dynamics. Part 1: Examples of application for social and economic processes analysis]. *Business Informatics*, no. 3 (25), pp. 3–18 (in Russian).
22. Davies D. L., Bouldin D. W. (1979) A cluster separation measure. *Pattern Analysis and Machine Intelligence. IEEE Transactions*, no. 2, pp. 224-227.
23. Ray S., Turi, R. (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation. Proceedings of the *4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT 1999), December 27-29, 1999, Calcutta, India,* pp. 137–143.
24. Petrovic S. (2006) A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. Proceedings of the *11th Nordic Workshop of Secure IT Systems, October 19-20, 2006, Linkoping, Sweden,* pp. 53–64.
25. Graczyk M., Lasota T., Trawinski B. (2009) Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA. *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems,* Berlin, Heidelberg: Springer, pp. 800–812.
26. Bernshtejn N.A. (1990) *Ocherki o fiziologii dvizhenij i fiziologii aktivnosti* [Essays about the physiology of movements and physiology of activity]. Moscow: Nauka (in Russian).