

Метафоры визуализации в задачах разведочного анализа гетерогенных данных

Р.А. Исаев^{1,A}, А.Г. Подвесовский^{2,A}, А.А. Захарова^{3,B}

^A Брянский государственный технический университет

^B Институт проблем управления им. В.А. Трапезникова РАН

¹ ORCID: 0000-0003-3263-4051, ruslan-isaev-32@yandex.ru

² ORCID: 0000-0002-1118-3266, apodv@tu-bryansk.ru

³ ORCID: 0000-0003-4221-7710, zaawmail@gmail.com

Аннотация

Предметом исследования являются пути построения и применения визуальных моделей, использующих понятие метафоры визуализации, в задачах разведочного анализа гетерогенных данных. Рассмотрены усовершенствованные варианты предложенных ранее метафор визуализации, которые могут быть положены в основу построения визуальных моделей. Предложена технология разведочного анализа гетерогенных данных, основанная на совместном использовании различных метафор визуализации. Показано, что процесс визуального исследования данных на этапе разведочного анализа с применением предложенной технологии носит итеративный и мультисценарный характер, зависящий от целей анализа. Описан разработанный программный инструмент, реализующий предложенную технологию и дополнительно поддерживающий возможность расчета и экспорта количественных характеристик визуальной модели. Рассмотрены примеры применения программного инструмента в задаче разведочного анализа синтетического набора данных. Определены направления дальнейшего развития предложенного подхода к построению визуальных моделей, технологии разведочного анализа данных и программного средства ее поддержки.

Ключевые слова: разведочный анализ, визуализация, метафора визуализации, визуальный анализ, гетерогенные данные.

1. Введение

Под разведочным анализом данных понимается предварительный анализ с целью выявления наиболее общих свойств данных, их внутренних взаимосвязей, закономерностей и аномалий. Основные идеи разведочного анализа были изложены в классической работе [1], но в некоторых более поздних работах (например, [2]) предпринимались попытки дискуссии по ее положениям и выдвигались альтернативные представления. Результаты разведочного анализа, как правило, далее ложатся в основу углубленного анализа данных.

Большое количество публикаций прикладного характера свидетельствует, что разведочный анализ актуален во всех областях, которые связаны с обработкой слабо формализованных данных [3-5].

Одним из видов таких данных, обработка которых чаще всего требует первоначального проведения разведочного анализа, являются гетерогенные данные. Такие данные могут быть получены в процессе функционирования сложных, распределенных в пространстве, гетерогенных динамических систем. В частности, одним из классов таких систем являются киберфизические системы.

Задачи, которые решаются при разведочном анализе гетерогенных данных, могут иметь высокую степень абстрактности, слабый уровень формализации и поисковый характер. Поэтому эффективный подход к организации разведочного анализа данных должен быть ориентирован на использование когнитивного потенциала аналитика. Наиболее естественным из подходов, которые обеспечивают задействование когнитивных функций человека-аналитика, является подход, который связан с использованием возможностей визуализации и визуальной аналитики [6]. Существует множество исследований, в которых продемонстрировано успешное применение визуализации при решении задач, связанных с пониманием объектов или процессов различной природы. Известны примеры применения визуальной аналитики в задачах вычислительной механики жидкости и газа [7-9], при решении задач оптимизации параметров распределенных экспериментов области в физике высоких энергий и ядерной физики [10], при проектировании программного обеспечения [11], а также для анализа текстовых данных [12].

Для описания и решения задачи визуализации гетерогенных данных можно использовать подход, который основан на понятии метафоры визуализации [13]. Метафора – это набор принципов, который описывает перенос характеристик исследуемого объекта (например, это может быть набор данных) в пространство визуальной модели (оно может быть как двумерным, так и трехмерным). Метафора визуализации включает в себя два компонента, которые применяются последовательно:

- пространственная метафора описывает общие принципы построения визуальной модели (вид и размерность пространства визуализации, взаимное расположение элементов модели в нем);
- метафора представления отвечает за уточнение характеристик визуального образа (это необходимо для визуализации определенных свойств исследуемого объекта, которые являются наиболее значимыми на текущем этапе анализа).

Как правило, одной пространственной метафоре соответствуют несколько метафор представления. Это связано со сложностью исследуемого объекта и необходимостью последовательной визуализации различных его свойств и характеристик в процессе исследования [14].

При взаимодействии аналитика с визуальной моделью данных важную роль играет простота и удобство визуального восприятия модели. Для описания этого аспекта часто используется понятие когнитивной ясности [15]. Это понятие означает простоту интуитивного понимания и интерпретации некоторого объема данных, который представлен в визуальной модели. Недостаточная когнитивная ясность модели чаще всего приводит к затруднению в понимании данных, неполной или ошибочной интерпретации некоторых элементов данных и т.д. В то же время, высокий уровень когнитивной ясности визуальной модели, во время разведочного анализа позволяет исследователю «охватить одним взглядом» больше важных свойств набора данных, упрощает обнаружение в нем неполноты и различных аномалий, ускоряет выявление и интерпретацию закономерностей и внутренних взаимосвязей.

2. Исследуемые данные: особенности и задачи их анализа

В настоящей работе предполагается, что исследуемые данные имеют следующую общую структуру:

- набор данных представляет собой описание некоторого количества *объектов*;
- каждый объект описывается некоторым количеством *свойств* (атрибутов, характеристик, параметров и т.п.);
- описание объекта представляет собой перечисление *значений* всех или некоторых его свойств.

Подразумевается, что объекты принадлежат к одному классу или, по крайней мере, родственным классам. В первом случае это означает, что все объекты имеют одинаковые наборы свойств, во втором случае – что множества свойств для разных объектов могут не совпадать (однако их пересечение является не пустым). Такая ситуация представляет собой одно из возможных проявлений гетерогенности данных.

Свойства объектов, в зависимости от природы этих свойств, могут быть измерены в различных измерительных шкалах: качественных (таких как номинальная или порядковая) и количественных (таких как интервальная или абсолютная).

Одним из наиболее серьезных следствий гетерогенности данных является их потенциальная неполнота. Неполнота данных, в общем случае, характеризуется отсутствием известных значений (то есть пропусками значений) по некоторым свойствам для всех или, чаще, для некоторых объектов. Можно выделить следующие основные типы такой неполноты:

- значение свойства не было измерено для конкретного объекта (либо вообще, либо с требуемым уровнем достоверности измерения для использования этого значения при анализе);
- в силу гетерогенности исследуемого набора данных, некоторые объекты могут не иметь одного (или более) из тех свойств, которые имеются у других объектов в этом же наборе данных.

Учитывать тип неполноты данных необходимо для их эффективного визуального анализа. Это влияет на обоснованность выводов о достаточности имеющихся данных для проведения анализа, на возможность заполнения пробелов в данных по результатам анализа и т.п.

Среди общих классов исследовательских задач, которые могут возникать на практике при работе с такими данными, можно выделить следующие.

- Задачи, связанные с генерацией гипотез о закономерностях и взаимосвязях в данных. Примером такой гипотезы может быть гипотеза о наличии и характере взаимосвязи двух или более свойств.
- Задачи, связанные с выбором из общего массива данных некоторого подмножества элементов данных (объектов или свойств) на основании определенного критерия (или набора критериев). Примерами могут являться нахождение объектов с аномальными значениями свойств, выбор объектов с «достаточно хорошими» значениями свойств, нахождение наиболее информативных свойств, выделение кластеров схожих между собой объектов и т.п.

Из-за потенциальной неполноты данных, дополнительной задачей, которая возникает при решении названных классов задач, является оценка достаточности имеющихся данных для достижения целей исследования. Еще одна связанная с этим задача заключается в заполнении пропусков в имеющемся наборе данных синтетическими значениями, которые будут сочтены аналитиком наиболее правдоподобными.

3. Метафоры визуализации данных для разведочного анализа

Технология визуального анализа, которая предлагается в данной работе, может иметь в своей основе различные метафоры визуализации и различные сочетания метафор. Основная идея совместного использования метафор визуализации заключается в объединении их преимуществ. Это достигается путем размещения визуальных образов данных в одном визуальном поле – так формируется общая визуальная модель данных, которая может восприниматься аналитиком как единое целое. Это позволяет повысить когнитивную ясность визуализируемых данных и их свойств за счет их одновременной визуализации «в разных аспектах».

Далее опишем две известные метафоры визуализации, которые могут быть использованы совместно и положены в основу предлагаемой технологии.

3.1. Метафора пространственных связей

В качестве основной метафоры визуализации в рамках технологии визуального анализа предлагается использовать трехмерную метафору, основные идеи которой описаны в работе [16].

Выбор указанной метафоры визуализации обусловлен тем, что она обладает следующими полезными в контексте задачи визуализации гетерогенных данных свойствами.

- Она предполагает размещение набора гетерогенных данных в едином «безразмерном» визуальном пространстве. Это позволяет аналитику одновременно воспринимать визуальные образы изначально несопоставимых величин.
- Она организует размещение визуальных образов таким способом, что аналитик может исследовать данные с двух взаимодополняющих точек зрения: во-первых, как цельные образы объектов со всеми их свойствами, во-вторых, как цельные образы свойств на всем множестве объектов.
- Данная метафора остается работоспособной в случае неполноты исследуемых данных и, более того, может оказаться весьма полезной именно в таких ситуациях.

При этом предлагается ряд доработок метафоры для расширения ее возможностей по визуализации различных характеристик данных. Далее опишем компоненты данной метафоры, то есть пространственную метафору и возможные метафоры представления.

В соответствии с пространственной метафорой, визуальное пространство модели описывается в цилиндрической системе координат. *Основа визуальной модели* строится из параллельных плоскостей, где каждая плоскость соответствует отдельному свойству данных (рис. 1). Свойства по умолчанию размещаются в направлении снизу вверх (таким образом, первое из свойств размещается внизу визуальной модели и т.д.). При изменении множества отображаемых свойств (например, при «отсеивании» некоторых свойств) визуальная модель перестраивается таким образом, чтобы она не содержала разрывов и всегда оставалась в центре визуального пространства.

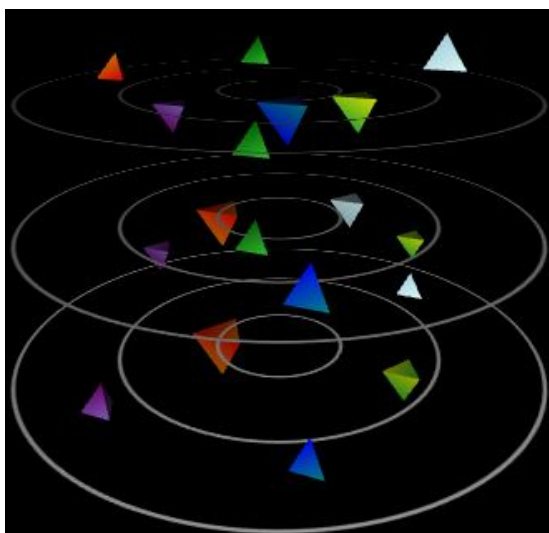


Рис. 1. Фрагмент основы визуальной модели и пример визуализации исходных данных

Концентрические окружности отражают разные уровни шкалы, которая измеряет характеристику, связанную с соответствующим свойством. На рис. 1 для всех свойств визуализированы три уровня: «начало шкалы» (наименьшая окружность), «середина шкалы» (средняя окружность) и «конец шкалы» (наибольшая окружность).

Предполагается, что все количественные свойства измерены в интервальной шкале или в шкале отношений, поэтому их исходные значения могут быть нормированы на безразмерный диапазон от 0 до 1.

За каждым объектом, который присутствует в исследуемом наборе данных, закрепляется конкретная угловая координата в цилиндрической системе координат. Эта координата рассчитывается для равномерного размещения объектов по пространству визуальной модели.

Данная метафора визуализации поддерживает ряд метафор представления, и некоторые из них могут применяться совместно.

Базовая метафора представления отвечает за визуализацию основных свойств исходных данных. Она использует идею визуальных маркеров. Визуальный маркер – это графический объект, который размещен на одной из плоскостей и связан с конкретным элементом данных. Этот объект имеет следующие параметры, которые могут соответствовать различным характеристикам данных.

- *Положение* маркера на шкале. В самом простом случае оно может отражать само значение конкретного свойства для конкретного объекта. Такой вариант метафоры представлен на рис. 1. Другой возможный вариант подразумевает, что положение маркера отражает отклонение значения свойства данного объекта от некоторой величины. В роли такой величины может выступать, например, среднее значение свойства по всем объектам или некоторое эталонное значение свойства.

- *Форма*. В предлагаемом варианте метафоры (рис. 1) маркер имеет форму тетраэдра. Он несет информацию об отклонении значения свойства данного объекта от заданной величины (в примере – от среднего значения свойства по всем объектам). Направление маркера «вверх» означает превосходство над опорным значением, «вниз» – наоборот. Другой вариант – использование формы для визуализации значения некоторого дискретного свойства объекта (его можно трактовать, например, как класс объекта). В таком случае объекты разных классов будут иметь маркеры разных форм (тетраэдр, куб, сфера и т.д.).

- *Размер*. В приведенном на рис. 1 варианте размер маркера пропорционален величине отклонения значения свойства объекта от среднего значения. Если объекты не различаются с точки зрения некоторого свойства (значения этого свойства у всех объектов равны), то маркеры объектов на соответствующей плоскости имеют форму кубов. Также размер маркеров может отвечать за визуализацию отклонений от других опорных величин или за визуализацию самих исходных значений свойств.

- *Цвет*. На рис. 1 цвет маркера выполняет функцию идентификатора объекта. Допустимы и другие способы использования цвета. Например, цвет может использоваться для визуализации значения дискретного свойства (то есть, для различения объектов различных классов), а также для визуализации знака и величины отклонения значения свойства от заданной величины. В последнем случае знак отклонения передается при помощи оттенка (красный/синий или зеленый/красный), а величина отклонения – при помощи интенсивности цвета.

Далее все метафоры описываются исходя из той интерпретации параметров маркеров, которой соответствует пример на рис. 1. Следует отметить, что другая интерпретация параметров маркеров может повлечь за собой другую интерпретацию этих метафор.

Метафора визуализации профилей объектов (рис. 2) дает аналитику возможность увидеть каждый объект как единое целое. Визуальные образы профилей представляют собой ломаные линии, формы которых позволяют визуально оценить сходство или различие объектов с точки зрения баланса значений их свойств. За счет этого возможно обнаружение пар и групп схожих между собой объектов. Так, на рис. 2 видно, что зеленый и фиолетовый объекты обладают схожими профилями свойств (взгляд аналитика быстро подмечает близость этих профилей к зеркальной симметрии), а оранжевый объект значительно отличается от них.

Метафора визуализации профилей свойств (рис. 3) изображает каждое свойство как единую характеристику совокупности данных. Замкнутый визуальный образ профиля помогает оценить величину разброса значений свойства на множестве объектов (по степени отклонения фигуры от правильного многоугольника). Также аналитик может обнаруживать объекты с аномальными значениями свойств. На рис. 3 видно, что с точки зрения среднего свойства объекты равнозначны, а с точки зрения других свойств объекты различаются, причем по верхнему свойству синий объект имеет аномальное значение.

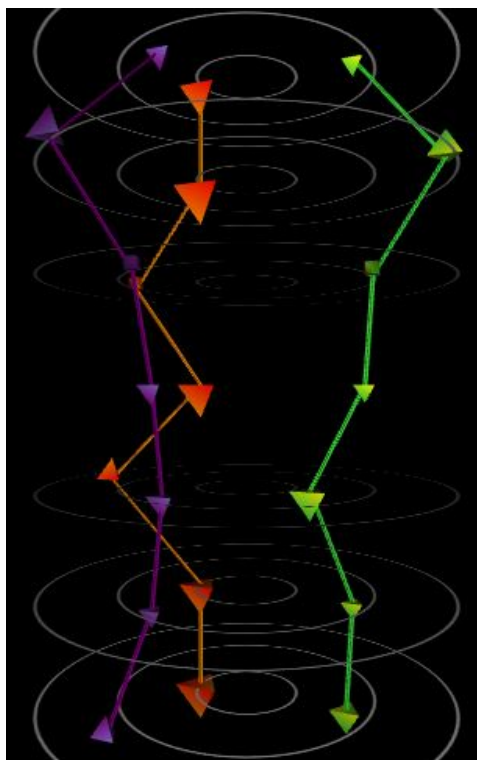


Рис. 2. Метафора визуализации профилей объектов

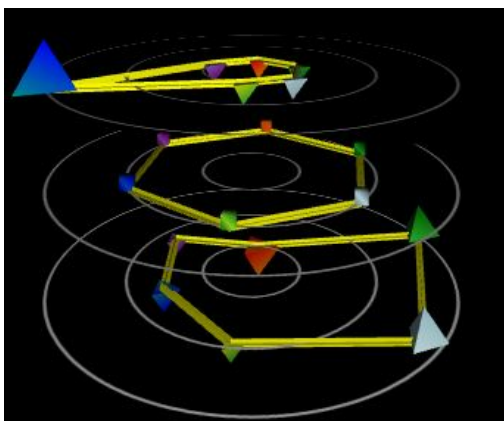


Рис. 3. Метафора визуализации профилей свойств

Метафора визуализации отклонения от эталона (рис. 4) показывает степень отклонения свойств объекта от свойств эталонного объекта. Эталонный объект, как правило, является абстракцией, которой не существует среди множества реальных объектов. В примере на рис. 4 эталон ассоциирован с максимальными значениями всех измерительных шкал (то есть, эталонный объект имеет максимально возможные значения по всем свойствам). Видно, что оранжевый объект имеет высокую степень отклонения от эталона, в то время как серый объект достаточно близок к эталону. Возможны и другие

интерпретации эталонного объекта – например, явное указание конкретных эталонных значений по каждому свойству.

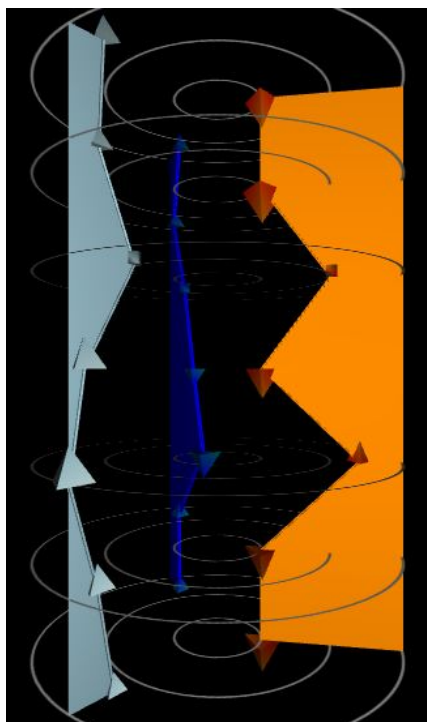


Рис. 4. Метафора визуализации отклонения от эталона

Метафора визуализации отклонения от среднего (рис. 5) демонстрирует степень отклонения свойств объекта от свойств «среднестатистического» объекта (то есть абстрактного объекта со средними значениями всех свойств). Это можно интерпретировать как степень нетипичности (аномальности) данного объекта. В качестве среднего значения по конкретному свойству может использоваться среднее арифметическое, среднее геометрическое или медианное значение (это зависит от типа измерительной шкалы). В примере на рис. 5 учитываются не только величины, но и знаки отклонений. Видно, что объект справа проигрывает «среднестатистическому» объекту почти по всем свойствам, а объект слева напротив, превосходит его. Также возможен вариант метафоры без учета знака.

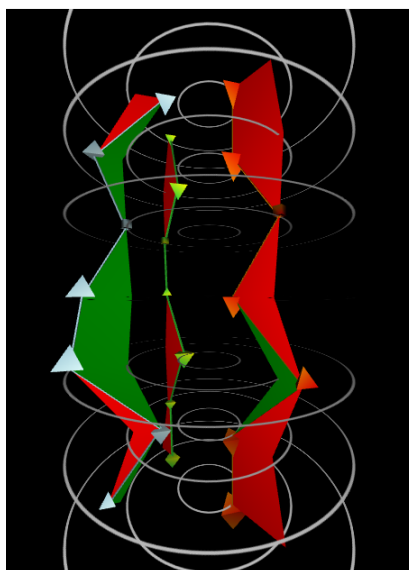


Рис. 5. Метафора визуализации отклонения от среднего

Метафора сравнения объектов (рис. 6) значительно упрощает визуальное сравнение двух объектов друг с другом. Такие визуальные образы показывают, насколько сильно каждый из объектов отличается от другого по каждому из свойств. На рис. 6 видно, что объект слева сильно превосходит объект справа по некоторым свойствам, при этом равен или немного уступает по остальным свойствам. В этом примере учитываются знаки отличий, но возможен вариант, показывающий лишь абсолютную величину отличий.

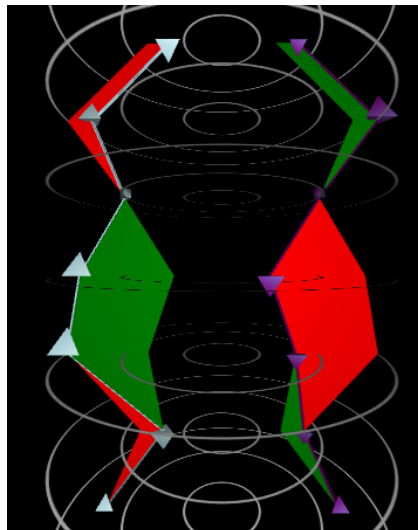


Рис. 6. Метафора сравнения двух объектов

Метафора визуализации пропусков в данных (рис. 7) используется, чтобы обратить внимание аналитика на наличие пропусков значений в анализируемых данных. Эта метафора изображает «неполные» профили объектов – то есть, участки профилей, которые соответствуют свойствам с пропусками значений, визуализируются красными линиями. Такие участки привлекают внимание аналитика к факту наличия пропуска в описании объекта и указывают на местоположение пропуска. Такой визуальный образ данных может быть полезен, если требуется вначале оценить достаточность имеющихся данных для дальнейшего визуального анализа с конкретной целью. На рис. 7 фиолетовый и зеленый объекты имеют очень мало известных данных в своем описании, а в описании остальных объектов имеются лишь отдельные пробелы.

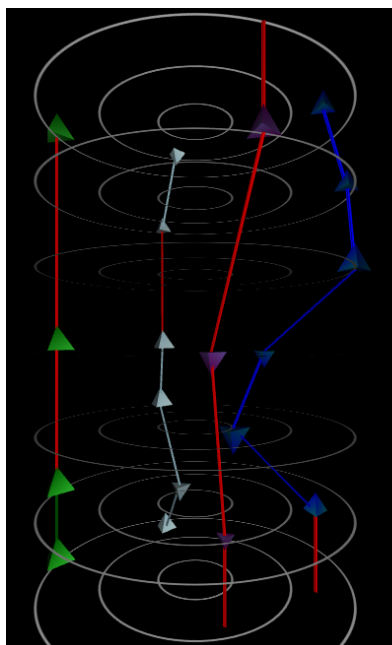


Рис. 7. Метафора визуализации пропусков в данных

Завершая описание метафоры пространственных связей, отметим ее ограничения с точки зрения данных, подлежащих визуализации.

- Во-первых, исследуемые данные должны иметь структуру «объекты-свойства», которая описана в предыдущем разделе статьи.
- Во-вторых, препятствием для использования метафоры является чрезмерно высокая степень гетерогенности данных, под которой здесь понимается малое количество общих свойств (при большом количестве уникальных свойств) у объектов. То есть, это ситуация, при которой объекты становятся сложно сопоставимыми, поскольку имеют слишком различную природу.
- В-третьих, еще одним возможным препятствием является чрезмерная неполнота данных. Несмотря на то, что метафора способна визуализировать пропуски в наборе данных, преобладание в нем пропусков над известными значениями сделает визуальный анализ неэффективным.

3.2. Лепестковая метафора

В дополнение к метафоре пространственных связей, в рамках технологии можно использовать двумерную лепестковую метафору визуализации, которая была описана в работе [17].

Эта метафора формирует отдельные визуальные образы исследуемых объектов. Визуальный образ объекта выглядит как круговая диаграмма с лепестками, при этом количество лепестков равно количеству тех свойств объекта, которые подлежат визуализации (рис. 8). Как правило, длина лепестка рассчитывается таким образом, чтобы площадь лепестка была пропорциональна значению соответствующего свойства (с учетом нормировки всех значений на единичный диапазон).

Полученный визуальный образ позволяет «охватить объект одним взглядом», визуально оценить его «сильные» и «слабые» стороны (если такая интерпретация свойств допускается в контексте решаемой задачи). Также эта метафора дает возможность сравнивать объекты друг с другом, в том числе искать сходства и различия между ними.

В приведенном примере цвет служит для идентификации объекта (разным объектам соответствуют образы разных цветов). Возможны и другие интерпретации цвета: например, для идентификации свойства (разным свойствам соответствуют лепестки разных цветов), а также для визуализации отклонения значения свойства от заданного значения.

При использовании этой метафоры для визуализации неполных данных можно указывать на пробелы в описании объекта при помощи специального визуального признака, как показано на рис. 9 (отсутствует информация по одному из 7 свойств объекта). Возможны также другие способы указания на неполноту данных, при этом должен соблюдаться общий принцип: отображение пробела в данных не должно быть идентично с отображением свойства с нулевым значением.

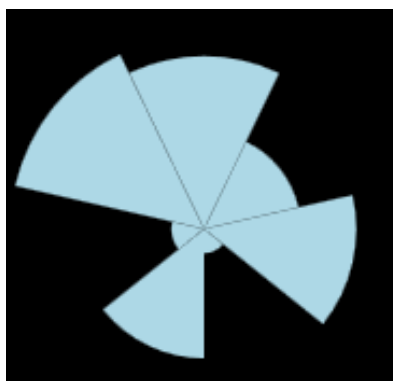


Рис. 8. Пример визуального образа объекта на основе лепестковой метафоры



Рис. 9. Представление пропусков в описании объекта на основе лепестковой метафоры

4. Технология разведочного анализа данных на основе метафор визуализации

Предлагаемая технология разведочного анализа гетерогенных данных с применением рассмотренных метафор визуализации представлена на рис. 10. Как следует из схемы, применение технологии подразумевает систематическое выполнение ряда этапов. Каждый этап имеет свой уровень баланса между ролью аналитика и ролью программного обеспечения в его выполнении.

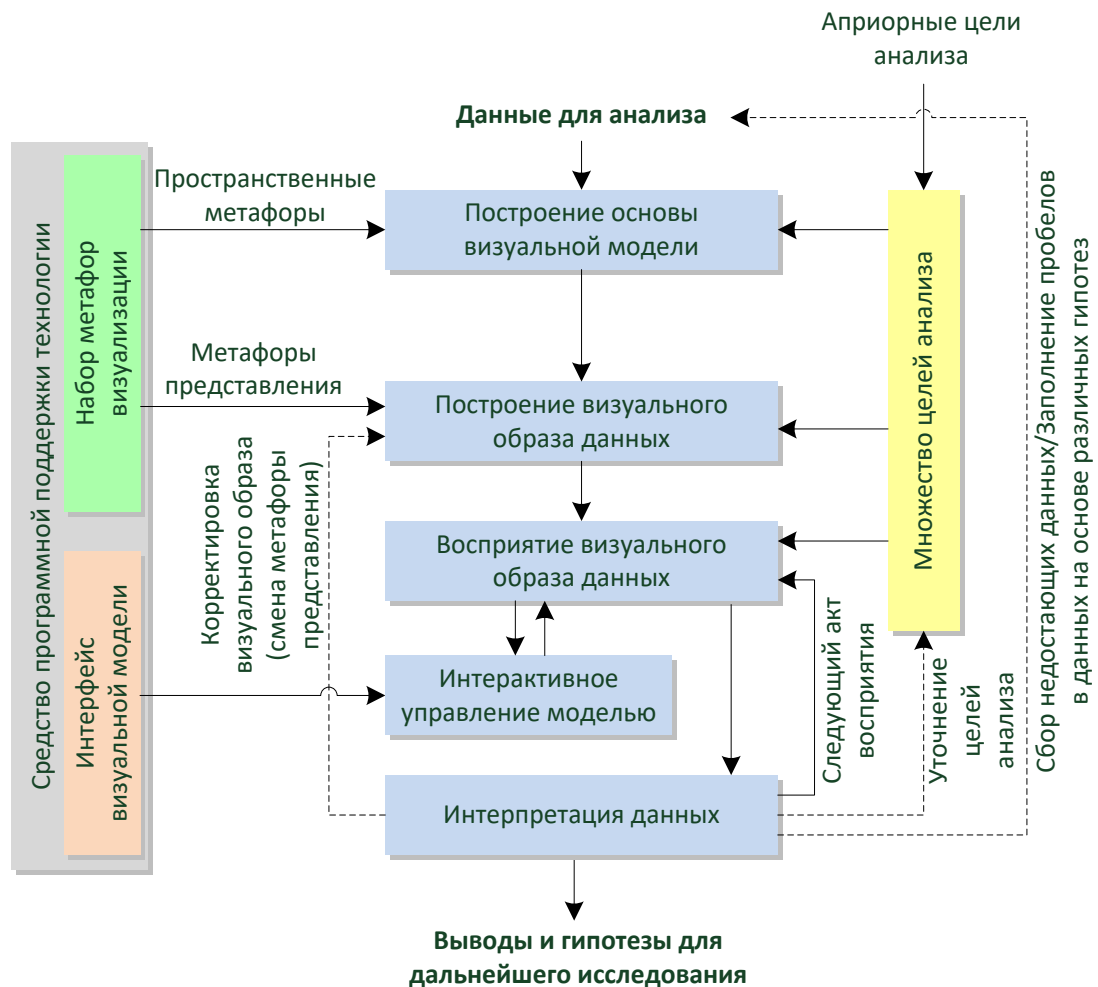


Рис. 10. Схема технологии разведочного анализа данных с применением метафор визуализации

Процесс применения представленной технологии характеризуется значительной степенью вариативности и мультисценарности. Смысл этапов анализа и количество их повторений, а также множество реально задействованных переходов между этапами, в конкретной ситуации зависят от следующих факторов.

- Множество изначально поставленных (априорных) целей анализа. Примерами таких целей являются оценка достаточности имеющихся данных для проведения анализа (это актуально в случае наличия пропусков в данных), поиск объектов с аномальными свойствами, выбор наилучшего в некотором отношении объекта или подмножества объектов и т.п.

- Результаты, уже полученные в процессе визуального анализа. Они могут влиять, во-первых, на необходимость и целесообразность выполнения следующих итераций анализа, а во-вторых, на характер этих итераций (то есть на то, какой вариант обратной связи на схеме технологии будет задействоваться).

Пунктирные линии на схеме обозначают не обязательные переходы между этапами. Такие переходы могут быть ни разу не задействованы в ходе применения технологии – например, при решении наиболее простых аналитических задач.

Рассмотрим смысл имеющихся на схеме циклов. Они соответствуют разным вариантам итеративных действий, которые совершаются в процессе использования технологии.

1. Цикл *«Восприятие визуального образа данных» – «Интерактивное управление моделью»*. Этот цикл соответствует рутинным действиям аналитика, которые связаны с решением конкретной задачи или подзадачи визуального анализа. Он отражает попытки аналитика приблизиться к пониманию исследуемых данных путем интерактивного управления их визуальным образом. Цикл завершается, когда было достигнуто понимание (интерпретация) фрагмента данных, относящегося к решаемой задаче. Длительность выполнения такого цикла (число его итераций) может зависеть от таких факторов, как качество программной поддержки интерактивного управления визуальной моделью, подготовленность аналитика (в частности, уровень освоения средств интерактивного управления), качество применяемых метафор визуализации.

2. Цикл *«Восприятие визуального образа данных» – «Интерпретация данных»*. Этот цикл отражает решение аналитиком ряда однотипных задач, которые связаны с исследованием конкретного аспекта (свойства) данных с помощью выбранной метафоры представления. При этом процесс решения одной такой задачи описывается циклом, рассмотренным выше. Цикл завершается, когда все задачи в рамках конкретной метафоры представления решены, то есть достигнуты интерпретации связанных с ними фрагментов данных. Длительность выполнения этого цикла естественным образом зависит от объема данных, которые подвергаются визуальному исследованию. Из-за ограниченности когнитивных возможностей аналитика, он будет вынужден делить процесс исследования на отдельные акты восприятия.

3. Цикл *«Построение визуального образа данных» – «Интерпретация данных» – «Корректировка визуального образа»*. Цикл описывает процесс смены метафор представления для решения новых типов задач, которые возникают перед аналитиком в ходе исследования данных. Цикл завершается, когда задачи всех требуемых типов решены (то есть цели исследования достигнуты) или сделан вывод, что в текущих условиях решение некоторых задач невозможно (например, по причине недостаточного объема данных).

4. Цикл *«Интерпретация данных» – «Уточнение целей» – «Построение визуального образа данных»*. Цикл отражает возможность корректировки априорных целей и задач анализа данных на основании промежуточных результатов, полученных в ходе анализа. Как правило, это сопровождается сменой метафоры представления в связи с переходом к исследовательской задаче нового типа.

5. Цикл *«Интерпретация данных» – «Сбор недостающих данных» – «Построение визуального образа данных»*. Этот цикл описывает ситуацию прерывания визу-

ального исследования в связи с невозможностью достижения всех или некоторых его целей. Это может случиться по причине нехватки данных, при этом после получения недостающих данных анализ может быть возобновлен.

В результате применения технологии визуального анализа формируются выводы об исследуемом наборе данных. Форма выводов определяется целями анализа – например, выводы могут включать в себя описания выявленных аномалий в данных, подмножество наиболее предпочтительных объектов и т.п. Частной формой таких выводов могут являться также гипотезы о данных (например, о статистически значимой взаимосвязи различных показателей), которые подлежат дальнейшей проверке другими методами (как правило, уже не визуальными, а более формальными – например, методами математической статистики).

Еще одной формой выводов могут быть рекомендации, направленные на повышение обоснованности результатов анализа. Такие рекомендации могут касаться, например, необходимости сбора большего объема данных по определенным объектам или свойствам.

Кроме того, результатом применения технологии к неполному набору данных может стать автоматизированное заполнение пропусков значений в исходном наборе данных. При этом предлагаются значения на основании некоторой гипотезы о распределении значений в данных, которая формулируется аналитиком в процессе визуального исследования.

5. Программная поддержка предлагаемой технологии

Программная поддержка технологии разведочного анализа гетерогенных данных реализована в форме Windows-приложения. Данное приложение разработано на платформе .NET Framework с применением технологии Windows Presentation Foundation (WPF), которая позволяет создавать приложения с насыщенным графическим пользовательским интерфейсом.

На рис. 11 представлен интерфейс приложения с визуальной моделью некоторого набора данных на основе двух рассмотренных метафор визуализации.

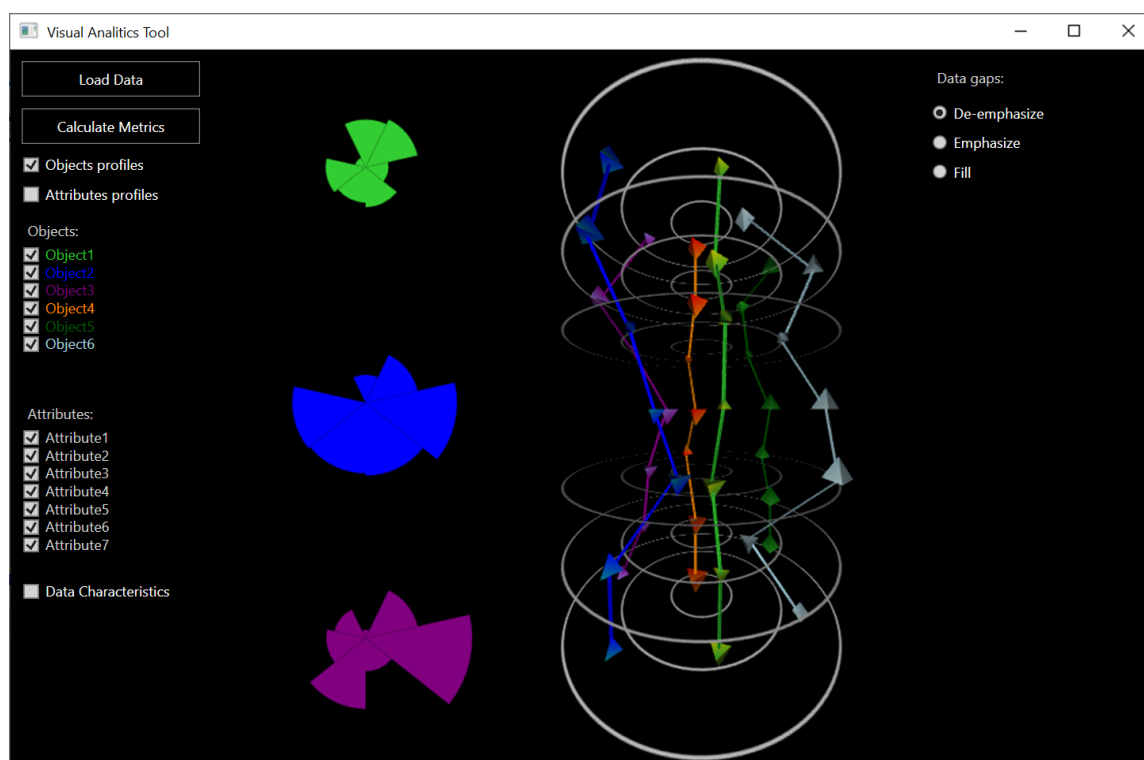


Рис. 11. Интерфейс приложения для программной поддержки технологии разведочного анализа гетерогенных данных

Приложение предоставляет пользователю следующие возможности управления визуальной моделью:

- интерактивную фильтрацию элементов данных (объектов и свойств), которые подлежат визуализации;
- задействие, в том числе в любых комбинациях, метафор визуализации профилей объектов и профилей свойств;
- активизация метафор представления для визуализации различных свойств данных (отклонение объектов от эталона, отклонение свойств объектов от средних значений, отличие пары сравниваемых объектов друг от друга);
- выбор метафоры визуализации пропусков в данных (актуально для наборов данных, в которых имеются пропуски значений свойств).

Важной особенностью программного средства является расчет количественных характеристик различных элементов визуальной модели. Это обеспечивает возможность интеграции реализуемой технологии в общий конвейер анализа данных. В качестве примеров таких характеристик можно назвать следующие:

- длины профилей объектов;
- длины (периметры) профилей свойств;
- площади фигур, ограниченных профилями свойств;
- площади фигур, которые визуализируют отклонения объектов от эталона и от среднего.

Программное средство поддерживает сохранение количественных характеристик визуальной модели для проведения их анализа «строгими» методами (например, с применением методов математической статистики). Это может быть сделано с целью подтверждения полученных выводов и проверки выдвинутых гипотез об исследуемой совокупности данных.

Разработанное приложение в настоящее время проходит государственную регистрацию. В дальнейшем планируется его интеграция в состав создаваемой при участии авторов программной платформы обработки и анализа гетерогенных данных функционирования объектов киберфизических систем, в качестве подсистемы разведочного анализа данных с применением визуальных моделей. Предполагается взаимодействие указанной подсистемы с другими аналитическими подсистемами, в том числе с целью построения единого конвейера анализа данных, а также с вспомогательными подсистемами, отвечающими за реализацию различных методов сбора данных, их хранение и предобработку.

6. Демонстрация применения технологии к анализу тестового набора данных

Продemonстрируем ряд возможностей технологии разведочного анализа. Для проведения демонстрации будем использовать тестовый синтетический набор данных, особенности которого позволят наглядно показать конкретные возможности используемых метафор. Тестовый набор представлен 9 объектами, каждый из которых описывается 7 свойствами. При этом у нескольких объектов пропущены значения некоторых свойств, то есть их описания неполные.

В частности, приведем примеры достижения таких целей, как:

- определение объектов с недостаточным объемом данных в их описании (то есть объектов, для которых рекомендуется дополнительный сбор недостающих данных);
- выявление свойств, которые не несут полезной для анализа информации (неинформативных свойств);
- определение объектов, которые проигрывают другим объектам по представленным характеристикам;
- выделение групп схожих между собой объектов;

- нахождение объектов-аномалий, которые не входят в выявленные группы.

При визуализации исходного набора данных с применением трехмерной метафоры был получен визуальный образ, представленный на рис. 12 (слева). При этом визуальный образ на рис. 12 (справа) позволил отдельно рассмотреть объекты с пропусками в данных и оценить эти объекты на предмет целесообразности дальнейшего исследования. Так, видно, что один из объектов описан слишком малым объемом данных, поэтому его исследование не приведет к получению достоверных выводов. При этом другие два объекта, даже при наличии некоторой неполноты знаний о них, могут рассматриваться далее.

Применение лепестковой метафоры дает другой взгляд на исследуемый набор данных (рис. 13). Полученные визуальные образы также отражают наличие и местонахождение пробелов в описании объектов.

Визуальный образ на рис. 14 был использован для оценки уровня информативности свойств объектов. Под неинформативным свойством здесь понимается такое свойство, с точки зрения которого исследуемые объекты не различаются или различаются незначительно. Так, профиль одного из свойств имеет форму правильного многоугольника, что сигнализирует о его неинформативности (это также подтверждается типом маркеров на соответствующей плоскости).

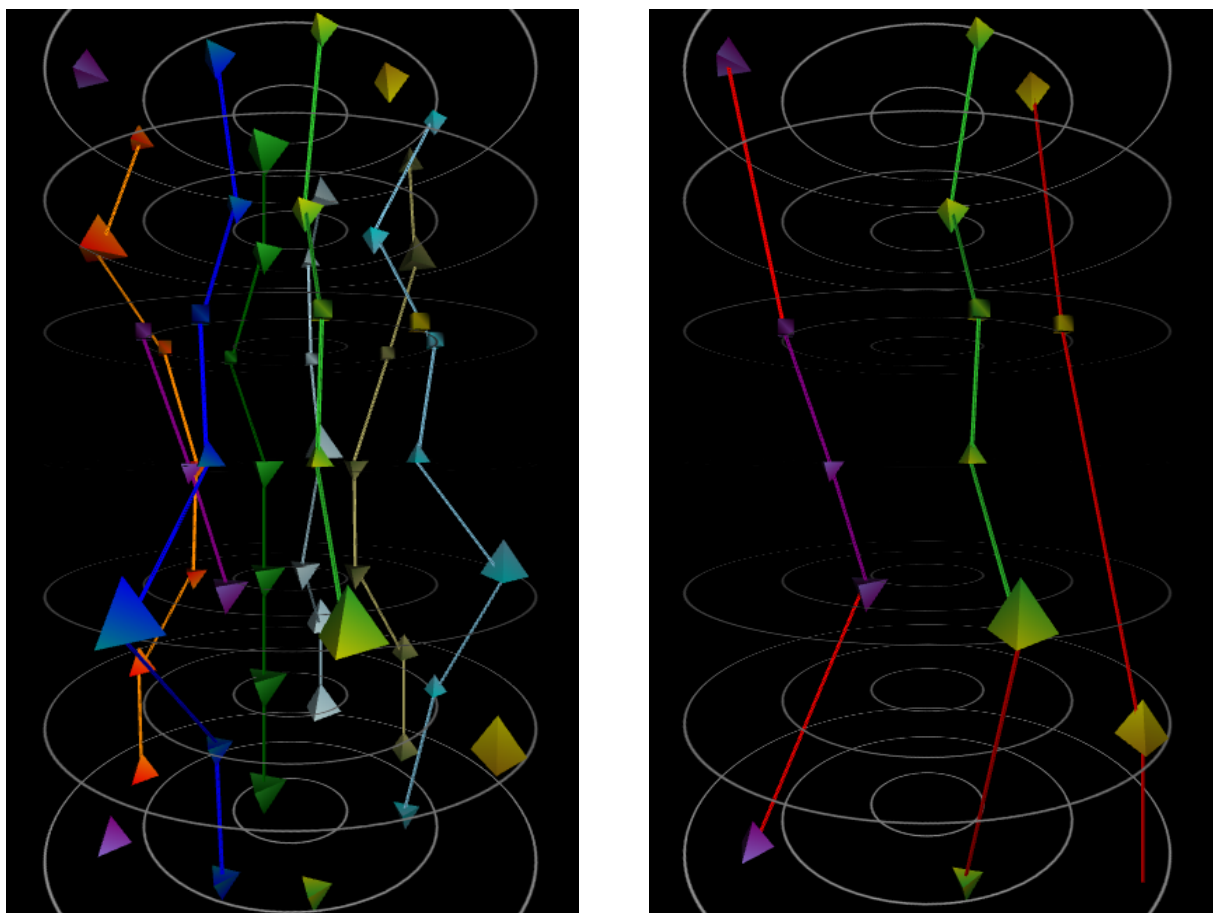


Рис. 12. Общий визуальный образ данных и визуализация объектов с пропусками в данных (метафора пространственных связей)

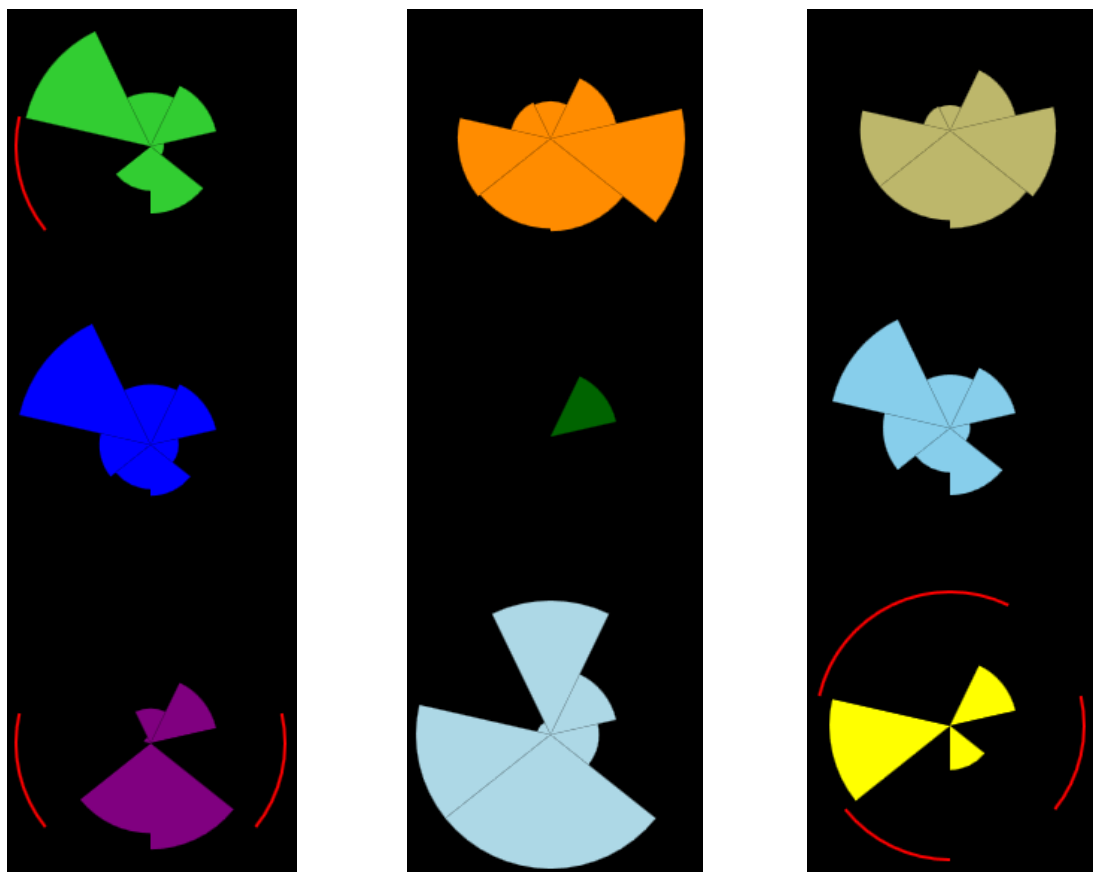


Рис. 13. Визуальный образ данных с указанием пропусков в данных (лепестковая метафора)

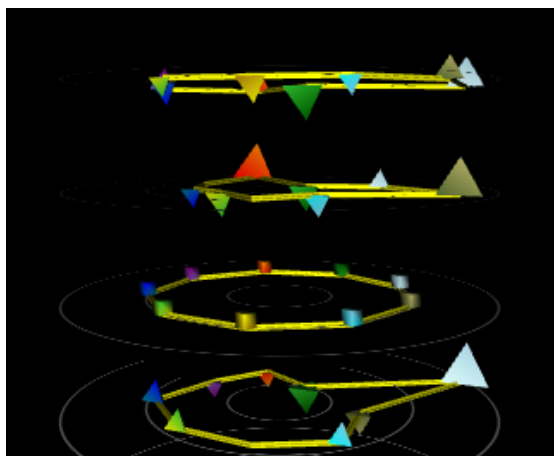


Рис. 14. Визуализация профилей свойств: обнаружение неинформативного свойства

После устранения из визуальной модели неинформативного свойства и плохо описанного объекта, были достигнуты другие поставленные цели. Визуальный образ на рис. 15 (слева) помог обнаружить объект с наихудшими (минимальными) значениями свойств. При этом была использована метафора отклонения от эталона, ассоциированного с максимальными значениями свойств. Таким образом, искомому объекту соответствует фигура наибольшей площади, которая в данном случае выявляется быстро и однозначно. На рис. 15 (справа) приведен образ, который способствует обнаружению объекта с наиболее нетипичными (аномальными) характеристиками. В данном случае площадь фигуры отражает степень отличия значений свойств объекта от средних значений этих свойств по всему набору данных.

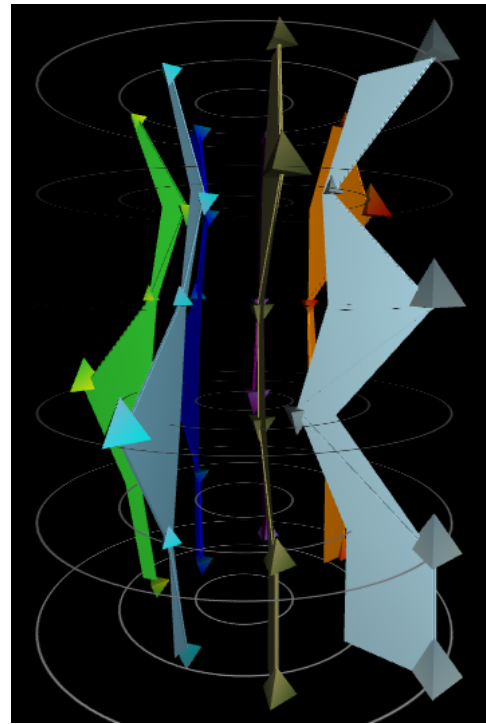
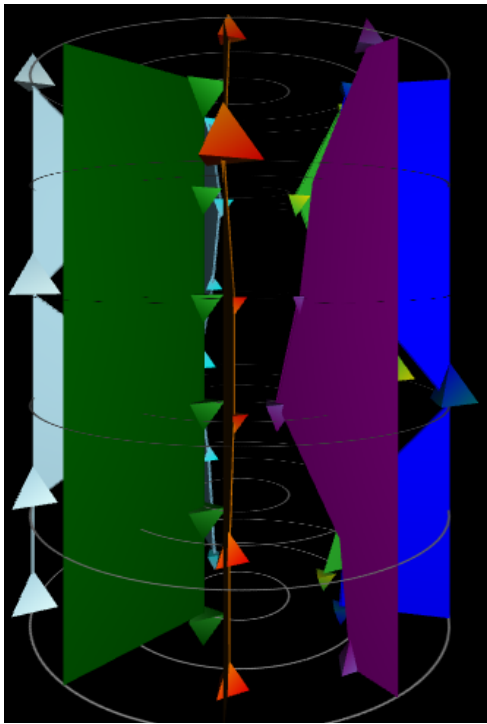


Рис. 15. Обнаружение объектов с неудовлетворительными и с аномальными характеристиками

Визуальное сравнение формы профилей объектов помогло выделить группы схожих между собой объектов. На рис. 16 приведены схожие объекты, которые были обнаружены за счет близости их профилей к зеркальной симметрии. Также в данную группу был отнесен объект с пробелами в описании – информации о нем оказалось достаточно для обоснования этого вывода (рис. 16, справа). Отметим, что в случае последующего искусственного заполнения пробелов в описании этого объекта целесообразно будет использовать значения схожих объектов.

Аналогичным способом была обнаружена другая группа схожих объектов (рис. 17).

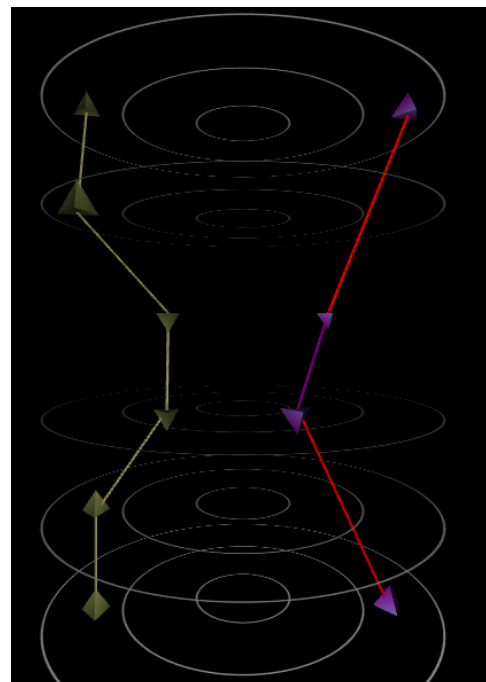
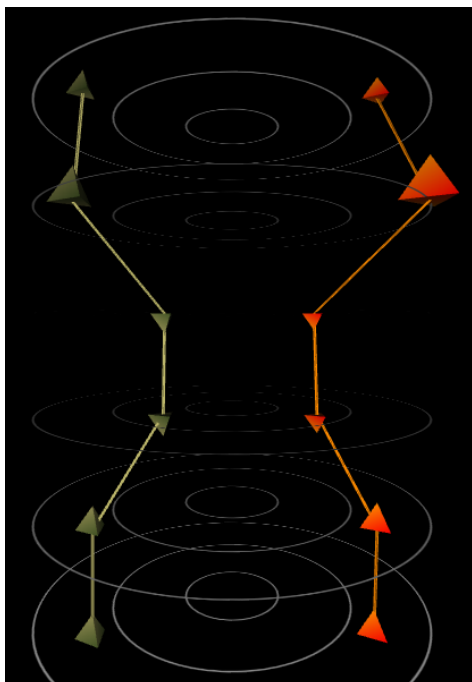


Рис. 16. Обнаружение схожих объектов, в том числе в условиях неполной информации

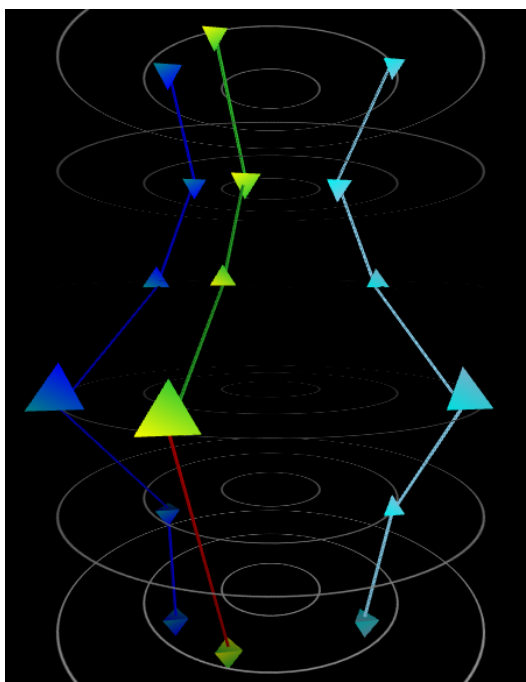


Рис. 17. Обнаружение другой группы схожих объектов

Среди возможностей интерактивного управления отметим весьма полезную возможность вращения визуальной модели вокруг своей оси. Таким образом формируется анимация, которая способствует ускоренному обнаружению схожих профилей объектов, более быстрому выбору фигур с наибольшими площадями и более точной оценке формы профилей свойств (в сравнении с исследованием неподвижной модели). Применение такой возможности соответствует внутреннему циклу («Восприятие визуального образа данных» – «Интерактивное управление моделью») на схеме технологии (рис. 10).

Кроме того, были рассчитаны метрики (количественные характеристики) визуальной модели. Значения некоторых метрик для всех элементов модели приведены в табл. 1-2. Отметим, что эти значения подтверждают выводы об объектах и их свойствах, сделанные в ходе визуального анализа. В дальнейшем эти и другие метрики могут быть проанализированы с применением методов математической статистики, для проверки выдвинутых гипотез о данных.

Табл. 1. Примеры метрик визуальной модели (для объектов)

	Объект								
Название метрики	1	2	3	4	5	6	7	8	9
Длина профиля	1,142	1,132	1,101	1,094	1,083	1,336	1,07	1,12	1,071
Площадь отклонения от эталона	0,507	0,517	0,583	0,425	0,917	0,34	0,492	0,523	0,5

Табл. 2. Примеры метрик визуальной модели (для свойств)

	Свойство						
Название метрики	1	2	3	4	5	6	7
Периметр свойства	1,34	1,441	1,384	1,355	1,026	1,546	1,236
Площадь свойства	0,418	0,514	0,393	0,29	0,435	0,293	0,469

Помимо тестирования разработанной технологии на синтетическом наборе данных, была выполнена ее апробация на данных реальной киберфизической системы. Были использованы данные экспериментов, которые связаны с проведением полетов беспилотных летательных аппаратов (БПЛА), а именно, с тестированием работы GPS в процессе полета. Всего было проанализировано 9 экспериментов, при этом для анализа были собраны такие данные, как планируемая и фактическая длительность эксперимента, емкость аккумуляторной батареи БПЛА, температура и влажность воздуха, атмосферное давление, скорость ветра и др.

Набор данных был визуализирован с применением разработанного программного средства. Результаты анализа визуальной модели позволили:

- обнаружить параметры, не информативные для анализа моделей полетных миссий в условиях проводившихся экспериментов;
- выявить эксперименты с аномально высоким значением фактической длительности, что, в свою очередь, позволило обнаружить техническую ошибку при фиксации времени окончания одного из экспериментов;
- обнаружить эксперименты, описываемые идентичными наборами значений, что также обусловлено техническими ошибками при фиксации их результатов.

Таким образом, подтверждено, что программное средство может эффективно применяться при разведочном анализе данных функционирования киберфизических систем.

7. Заключение

В работе предложена технология разведочного визуального анализа гетерогенных данных, которая основана на совместном применении двух метафор визуализации. Рассмотрены возможности этих метафор визуализации, приведены примеры визуальных образов данных, которые можно получить с помощью метафор.

Описана общая схема технологии и показано, что процесс визуального исследования данных носит итеративный характер. Рассмотрено смысловое содержание различных видов итеративных действий, которые аналитик совершает в ходе исследования.

Разработан и испытан программный инструмент для работы с визуальной моделью гетерогенных данных, который реализует представленную технологию. Важной функциональной возможностью инструмента является расчет и экспорт метрик (количественных характеристик) визуальной модели. Эти метрики могут использоваться для их последующего анализа другими методами с целью более строгой проверки выдвинутых гипотез о данных.

Приведен пример использования программного инструмента в задачах разведочного анализа синтетических наборов данных, для демонстрации различных аспектов технологии и возможностей метафор визуализации. Полученные результаты подтверждают возможность применения технологии и программного средства в задачах визуального анализа данных. Дополнительно проведенная апробация на данных эксперимента с реальной киберфизической системой подтверждает возможность использования разработанной технологии и программного средства при разведочном анализе данных функционирования киберфизических систем.

Перспективные исследования могут проводиться в следующих взаимосвязанных направлениях.

- Апробация представленной технологии на реальных гетерогенных данных из различных предметных областей. Это позволит дополнительно выявить особенности и ограничения технологии и метафор для их последующей модернизации.
- Расширение возможностей описанных метафор за счет разработки новых метафор представления для визуализации новых показателей и характеристик данных, в том числе системных. Также предполагается расширить состав метрик (количественных характеристик) визуальной модели, которые могут вычисляться.

- Развитие предложенной технологии за счет разработки для нее новых метафор визуализации (как двухмерных, так и трехмерных), в том числе для их совместного использования в различных сочетаниях.
- Разработка путей интеграции технологии в общий конвейер анализа данных – в том числе, на основе экспорта количественных характеристик визуальной модели для дальнейшего анализа.

8. Благодарности

Исследование выполнено при поддержке Российского научного фонда, проект 23-19-00342, <https://rscf.ru/project/23-19-00342/>.

Список литературы

1. Tukey J.W. Exploratory Data Analysis. Pearson, London, 1977.
2. Chatfield C. Exploratory data analysis. European Journal of Operational Research, 1986, Vol. 23(1), pp. 5-13. doi: 10.1016/0377-2217(86)90209-2
3. Komorowski M., Marshall D.C., Saliccioli J.D., Crutain Y. Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham, 2016. doi: 10.1007/978-3-319-43742-2_15
4. Verbeeck N., Caprioli R.M., Van de Plas R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. Mass Spectrometry Reviews 39(3), 245–291 (2020). doi: 10.1002/mas.21602
5. Wang G., Zhao B., Wu B., Zhang C., Liu W. Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases. International Journal of Mining Science and Technology, 33(1), 47–59 (2023). doi: 10.1016/j.ijmst.2022.07.002
6. Авербух В.Л. Семиотический подход к формированию теории компьютерной визуализации // Научная визуализация. 2013. Т. 5. № 1. С. 1-25.
7. Tricoche X., Garth C. Topological Methods for Visualizing Vortical Flows. In: Möller T., Hamann B., Russell R.D. (ed.), Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration. Mathematics and Visualization. Springer, Berlin, Heidelberg, pp. 89-108 (2009). doi: 10.1007/b106657_5
8. Bondarev A.E., Galaktionov V.A. Generalized Computational Experiment and Visual Analysis of Multidimensional Data. Scientific Visualization, 11(4), 102–114 (2019). doi: 10.26583/sv.11.4.09
9. Галкин В.А., Дубовик А.О. Визуализация течений вязкой проводящей жидкости с учетом наличия примесей в поле течения, соответствующих точным решениям уравнений МГД // Научная визуализация. 2021. Т. 13. № 1. С. 104-123. doi: 10.26583/sv.13.1.08
10. Галкин Т.П., Григорьева М.А. и др. Применение методов визуальной аналитики для кластеризации и категоризации задач анализа и обработки данных экспериментов в области физики высоких энергий и ядерной физики // Научная визуализация. 2018. Т. 10. № 5. С. 32-44. doi: 10.26583/sv.10.5.03
11. Намиот Д.Е., Романов В.Ю. 3D визуализация архитектуры и метрик программного обеспечения // Научная визуализация. 2018. Т. 10. № 5. С. 123-139. doi: 10.26583/sv.10.5.08
12. Bondarev A.E., Bondarenko V.A., Galaktionov V.A. Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives. Scientific Visualization, 12(4), 9–22 (2020). doi: 10.26583/sv.12.4.02
13. Захарова А.А., Шкляр А.В. Метафоры визуализации // Научная визуализация. 2013. Т. 5. № 2. С. 16-24.
14. Подвесовский А.Г., Исаев Р.А. Метафоры визуализации нечетких когнитивных карт // Научная визуализация. 2018. Т. 10. № 4. С. 13-29. doi: 10.26583/sv.10.4.02
15. Исаев Р.А., Подвесовский А.Г. Когнитивная ясность графовых моделей: подход к пониманию идеи и способ выявления влияющих факторов с использованием визуаль-

ного анализа // Научная визуализация. 2022. Т. 14. № 4. С. 38-51. doi: 10.26583/sv.14.4.04

16. Захарова А.А., Шкляр А.В. Визуальное представление разнотипных данных при помощи динамических знаковых структур // Научная визуализация. 2016. Т. 8. № 4. С. 28-37.

17. Захарова А.А., Коростелёв Д.А., Федонин О.Н. Алгоритмы визуализации для фильтрации многокритериальных альтернатив // Научная визуализация. 2019. Т. 11. № 4. С. 66-80. doi: 10.26583/sv.11.4.06

Development of the human face tracking algorithm based on the optical flow application

R.A. Isaev^{1,A}, A.G. Podvesovsky^{2,A}, A.A. Zakharova^{3,B}

^A Bryansk State Technical University, Bryansk, Russia

^B V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia

¹ ORCID: 0000-0003-3263-4051, ruslan-isaev-32@yandex.ru

² ORCID: 0000-0002-1118-3266, apodv@tu-bryansk.ru

³ ORCID: 0000-0003-4221-7710, zaawmail@gmail.com

Abstract

The subject of the study is the construction and application of visual models using the concept of visualization metaphors in the context of exploratory analysis of heterogeneous data. This study considers improved variants of the previously proposed visualization metaphors that can be used as a basis for building visual models. A technology for exploratory analysis of heterogeneous data based on the joint use of different visualization metaphors is proposed. The process of visual data exploration at the stage of exploratory analysis using the proposed technology is demonstrated to be iterative and multiscenary, contingent upon the analysis goals. The software tool developed to implement the proposed technology is described, along with its additional functionality to calculate and export quantitative characteristics of the visual model. The software tool is then considered in the context of exploratory analysis of a synthetic data set. The future direction of the proposed approach to the construction of visual models, the technology of exploratory data analysis and the software tool for its support are determined.

Keywords: exploratory analysis, visualization, visualization metaphor, visual analysis, heterogeneous data.

References

1. Tukey J.W. Exploratory Data Analysis. Pearson, London, 1977.
2. Chatfield C. Exploratory data analysis. European Journal of Operational Research, 1986, Vol. 23(1), pp. 5-13. doi: 10.1016/0377-2217(86)90209-2
3. Komorowski M., Marshall D.C., Saliccioli J.D., Crutain Y. Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham, 2016. doi: 10.1007/978-3-319-43742-2_15
4. Verbeeck N., Caprioli R.M., Van de Plas R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. Mass Spectrometry Reviews 39(3), 245–291 (2020). doi: 10.1002/mas.21602
5. Wang G., Zhao B., Wu B., Zhang C., Liu W. Intelligent prediction of slope stability based on visual exploratory data analysis of 77 in situ cases. International Journal of Mining Science and Technology, 33(1), 47–59 (2023). doi: 10.1016/j.ijmst.2022.07.002
6. Averbukh V.L. Semiotic Approach to Forming the Theory of Computer Visualization. Scientific Visualization, 5(1), 1–25 (2013).
7. Tricoche X., Garth C. Topological Methods for Visualizing Vortical Flows. In: Möller T., Hamann B., Russell R.D. (ed.), Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration. Mathematics and Visualization. Springer, Berlin, Heidelberg, pp. 89-108 (2009). doi: 10.1007/b106657_5

8. Bondarev A.E., Galaktionov V.A. Generalized Computational Experiment and Visual Analysis of Multidimensional Data. *Scientific Visualization*, 11(4), 102–114 (2019). doi: 10.26583/sv.11.4.09
9. Galkin V.A., Dubovik A.O. Visualization of flows of a viscous conductive liquid with the presence of impurities in the flow field corresponding to exact solutions of the MHD equations. *Scientific Visualization*, 13(1), 104–123 (2021). doi: 10.26583/sv.13.1.08
10. Galkin T.P. Grigoryeva M.A., et al. An Application of Visual Analytics Methods to Cluster and Categorize Data Processing Jobs in High Energy and Nuclear Physics Experiments. *Scientific Visualization*, 10(5), 32–44 (2018). doi: 10.26583/sv.10.5.03
11. Namiot D.E., Romanov V.Yu. 3D Visualization of Architecture and Metrics of the Software. *Scientific Visualization*, 10(5), 123–139 (2018). doi: 10.26583/sv.10.5.08
12. Bondarev A.E., Bondarenko V.A., Galaktionov V.A. Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives. *Scientific Visualization*, 12(4), 9–22 (2020). doi: 10.26583/sv.12.4.02
13. Zakharova A.A., Shklyar A.V. Visualization Metaphors. *Scientific Visualization*, 5(2), 16–24 (2013).
14. Podvesovskii A.G., Isaev R.A. Visualization Metaphors for Fuzzy Cognitive Maps. *Scientific Visualization*, 10(4), 13–29 (2018). doi: 10.26583/sv.10.4.02
15. Isaev R.A., Podvesovskii A.G. Cognitive Clarity of Graph Models: an Approach to Understanding the Idea and a Way to Identify Influencing Factors Based on Visual Analysis. *Scientific Visualization*, 14(4), 38–51 (2022). doi: 10.26583/sv.14.4.04
16. Zakharova A.A., Shklyar A.V. Visual Presentation of Different Types of Data by Dynamic Sign Structures. *Scientific Visualization*, 8(4). 28–37 (2016).
17. Zakharova A.A., Korostelyov D.A., Fedonin O.N. Visualization Algorithms for Multi-criteria Alternatives Filtering. *Scientific Visualization*, 11(4), 66–80 (2019). doi: 10.26583/sv.11.4.06