ЛЕКЦИЯ
ВИЗУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

ВОПРОСЫ:

1. Традиционные методы визуального анализа данных
2. Геометрические преобразования визуальных образов
3. Визуализация инструментов Data Mining
4. Основные тенденции в области визуализации

1. Традиционные методы визуального анализа данных

Любое средство визуализации может быть классифицировано по всем трём параметрам, т.е. по виду данных, с которым оно работает, по визуальным образам, которые оно может предоставлять, и по возможностям взаимодействия с этими визуальными образами. Очевидно, что одно средство визуализации может поддерживать разные виды данных, разные визуальные образы и разные способы взаимодействия с образами.

С возрастанием количества накапливаемых данных, даже при использовании сколь угодно мощных и разносторонних алгоритмов *Data Mining*, становится все сложнее "переваривать" и интерпретировать полученные результаты. А, как известно, одно из положений *Data Mining* - *поиск*практически полезных закономерностей. *Закономерность* может стать практически полезной, только если ее можно осмыслить и понять.

В 1987 году по инициативе *ACM SIGGRAPH IEEE Computer Society Technical Committee* of *Computer Graphics*, в связи с необходимостью использования новых методов, средств и технологий данных, были сформулированы соответствующие задачи направления визуализации.

К способам визуального или графического представления данных относят графики, диаграммы, таблицы, отчеты, списки, структурные схемы, карты, электронная переписка людей, гиперссылки документов и т.п.;

Визуализация традиционно рассматривалась как вспомогательное средство при анализе данных, однако сейчас все больше исследований говорит о ее самостоятельной роли.

Традиционные методы визуализации могут находить следующее применение:

- представлять пользователю информацию в наглядном виде;
- компактно описывать закономерности, присущие исходному набору данных;
- снижать размерность или сжимать информацию;
- восстанавливать пробелы в наборе данных;
- находить шумы и выбросы в наборе данных.

Существует достаточно большое количество средств визуализации данных, предоставляющих различные возможности.

Для выбора учитывают три основные характеристики средств визуализации данных:

- характер данных, которые нужно визуализировать с помощью данного средства;
- методы визуализации и образцы, в виде которых могут быть представлены данные;
- возможности взаимодействия с визуальными образами и методами для лучшего анализа данных.

Для визуализации перечисленных типов данных используются различные визуальные образы и методы их создания. Очевидно, что количество визуальных образов, которыми могут представляться данные, ограничиваются только человеческой фантазией. Основное требование к ним - это наглядность и удобство анализа данных, которые они представляют. Методы визуализации могут быть как самые простые (линейные графики, диаграммы, гистограммы и т.п.), так и более сложные, основанные на сложном математическом аппарате. Кроме того, при визуализации могут использоваться комбинации различных методов.

Выделяют следующие типы методов визуализации:

- стандартные 2D/3D-образы - гистограммы, линейные графики и т.п.;
- геометрические преобразования - диаграмма разброса данных, параллельные координаты и т.п.;
- отображение иконок - линейчатые фигуры (needle icons) и звёзды (star icons);
- методы, ориентированные на пикселы - рекурсивные шаблоны, циклические сегменты и т.п.;
- иерархические образы - древовидные карты и наложение измерений.

К простейшим методам визуализации относятся графики, диаграммы, гистограммы и т.п. Основным их недостатком является невозможность приемлемой визуализации сложных данных и большого количества данных.

2. Геометрические преобразования визуальных образов

Методы геометрических преобразований визуальных образов направлены на трансформацию многомерных наборов данных с целью отображения их в декартовом и в недекартовом геометрических пространствах. Данный класс методов включает в себя математический аппарат статистики.

Другим классом методов визуализации данных являются методы отображения иконок. Их основной идеей является отображение значений элементов многомерных данных в свойства образов. такие образы могут представлять собой: человеческие лица, стрелки, звёзды и т.п. Визуализация генерируется отображением атрибутов элементов данных в свойства образов.

Такие образы можно группировать для целостного анализа данных. Результирующая визуализация представляет собой шаблоны текстур, которые имеют различия, соответствующие характеристикам данных. Основной идеей методов, ориентированных на пикселы, является отображение каждого измерения значения в цветной пиксел и из группировка по по принадлежности к измерению. Так как один пиксел используется для отображения одного значения, то, следовательно, данный метод позволяет визуализировать большое количество данных (свыше одного миллиона значений).
Методы иерархических образов предназначены для представления данных, имеющих иерархическую структуру. В случае многомерных данных должны быть правильно выбраны измерения, которые используются для построения иерархии.

К методам геометрических преобразований относятся:

- Матрица диаграмм разброса;
- параллельные координаты;
- Методы, ориентированные на пикселы
- Рекурсивные шаблоны;
- циклические сегменты;
- Иерархические образы
- Наложение измерений.

В результате применения методов визуализации будут построены визуальные образы, отражающие данные. Однако этого не всегда бывает достаточно для полного анализа. Пользователь должен иметь возможность работать с образами: видеть их с разных сторон, в разном масштабе и т.п.

Для этого у него должны быть соответствующие возможности взаимодействия с образами:

- динамическое проецирование;
- интерактивная фильтрация;
- масштабирование образов;
- интерактивное искажение;
- интерактивное комбинирование.

Основная идея динамического проецирования заключается в динамическом изменении проекций при проведении исследования многомерных наборов данных. Примером может служить проецирование в двумерную плоскость всех интересующих проекций многомерных данных в виде диаграмм разброса (scatter plots).

Примером масштабирования образов является "магическая линза" (Magic Lenses), основная идея состоит в использовании инструмента, похожего на увеличительное стекло, чтобы выполнять фильтрацию непосредственно при визуализации. Данные, попадающие под увеличительное стекло, обрабатываются фильтром, и результат отображается отдельно от основных данных. Линза показывает модифицированное

изображение выбранного региона, тогда как остальные визуализированные данные не детализируются.

Масштабирование - известный метод взаимодействия, используемый во многих приложениях. При работе с большим объёмом данных этот метод хорош тем для представления данных в общем сжатом виде, и, в то же время, он предоставляет возможность отображения любой их части в более детальном виде. Масштабирование может заключаться не только в простом увеличении объектов, но в изменении их представления на разных уровнях.

Метод интерактивного искажения поддерживает процесс исследования данных с помощью искажения масштаба данных при частичной детализации. Основная идея этого метода заключается в том, что часть данных отображается с высокой степенью детализации, а одновременно с этим остальные данные показываются с низким уровнем детализации. Наиболее популярные методы - это гиперболическиое и сферическое искажения.

3. Визуализация инструментов Data Mining

Каждый из алгоритмов *Data Mining* использует определенный подход к визуализации. В предыдущих лекциях мы рассмотрели ряд методов *Data Mining*. В ходе использования каждого из методов, а точнее, его программной реализации, мы получали некие ***визуализаторы***, при помощи которых нам удавалось интерпретировать результаты, полученные в результате работы соответствующих методов и алгоритмов.

- Для деревьев решений это *визуализатор* дерева решений, список правил, таблица сопряженности.
- Для нейронных сетей в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.
- Для карт Кохонена: *карты входов*, выходов, другие специфические карты.
- Для линейной регрессии в качестве *визуализатора* выступает линия регрессии.
- Для кластеризации: *дендрограммы*, диаграммы рассеивания.

Диаграммы и графики рассеивания часто используются для оценки качества работы того или иного метода.

Все эти способы визуального представления или отображения данных могут выполнять одну из функций:

- являются иллюстрацией построения модели (например, представление структуры (графа) нейронной сети);
- помогают интерпретировать полученный результат;
- являются средством оценки качества построенной модели;
- сочетают перечисленные выше функции (дерево решений, дендрограмма).

Визуализация Data Mining моделей

Первая *функция* (иллюстрация построения модели), по сути, является ***визуализацией Data Mining модели***. Существует много различных способов представления моделей, но графическое ее *представление* дает пользователю максимальную "ценность". *Пользователь*, в большинстве случаев, не является специалистом в моделировании, чаще всего он эксперт в своей *предметной области*. Поэтому модель *Data Mining* должна быть представлена на наиболее естественном для него языке или, хотя бы, содержать минимальное количество различных математических и технических элементов.

Таким образом, доступность является одной из основных характеристик модели *Data Mining*. Несмотря на это, существует и такой распространенный и наиболее простой способ представления модели, как "*черный ящик*". В этом случае *пользователь* не понимает поведения той модели, которой пользуется. Однако, несмотря на непонимание, он получает результат - выявленные закономерности. Классическим примером такой модели является модель нейронной сети.

Другой способ представления модели - *представление* ее в интуитивном, понятном виде. В этом случае *пользователь* действительно может понимать то, что происходит "внутри" модели. Таким образом, можно обеспечить его непосредственное участие в процессе. Такие модели обеспечивают пользователю возможность обсуждать ее логику с коллегами, клиентами и другими пользователями, или объяснять ее.

Понимание модели ведет к пониманию ее содержания. В результате понимания возрастает *доверие* к модели. Классическим примером является *дерево* решений. Построенное *дерево* решений действительно улучшает понимание модели, т.е. используемого инструмента *Data Mining*.

Кроме понимания, такие модели обеспечивают пользователя возможностью взаимодействовать с моделью, задавать ей вопросы и получать ответы. Примером такого взаимодействия является средство "что, если". При помощи диалога "система-*пользователь*" *пользователь* может получить понимание модели.

Теперь перейдем к функциям, которые помогают интерпретировать и оценить результаты построения *Data Mining* моделей. Это всевозможные графики, диаграммы, таблицы, списки и т.д.

Примерами средств визуализации, при помощи которых можно оценить качество модели, являются *диаграмма* рассеивания, *таблица*сопряженности, *график* изменения величины ошибки.

Диаграмма рассеивания представляет собой *график* отклонения значений, прогнозируемых при помощи модели, от реальных. Эти диаграммы используют для непрерывных величин. Визуальная оценка качества построенной модели возможна только по окончанию процесса построения модели.

Таблица сопряженности используется для оценки результатов классификации. Такие таблицы применяются для различных методов классификации. Они уже использовались нами в предыдущих лекциях. Оценка качества построенной модели возможно только по окончанию процесса построения модели.

График изменения величины ошибки. *График* демонстрирует изменение величины ошибки в процессе работы модели. Например, в процессе работы нейронных сетей *пользователь* может наблюдать за изменением ошибки на обучающем и тестовом множествах и остановить обучение для недопущения "*переобучения*" сети. Здесь оценка качества модели и его изменения может оцениваться непосредственно в процессе построения модели.

Примерами средств визуализации, которые помогают интерпретировать результат, являются: линия тренда в линейной регрессии, карты Кохонена, *диаграмма* рассеивания в кластерном анализе.

Методы визуализации

Методы визуализации, в зависимости от количества используемых измерений, принято классифицировать на две группы:

- представление данных в одном, двух и трех измерениях;
- представление данных в четырех и более измерениях.

Представление данных в одном, двух и трех измерениях

К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением. Практически любой современный инструмент *Data Mining* включает способы визуального представления из этой группы.

В соответствии с количеством измерений представления это могут быть следующие способы:

- одномерное (univariate) измерение, или *1-D* ;
- двумерное (bivariate) измерение, или *2-D* ;
- трехмерное или проекционное (projection) измерение, или *3-D*.

Наиболее естественно человеческий глаз воспринимает двухмерные представления информации.

При использовании двух- и трехмерного представления информации *пользователь* имеет возможность увидеть закономерности набора данных:

- его кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);
- топологические особенности;
- наличие трендов;
- информацию о взаимном расположении данных;
- существование других зависимостей, присущих исследуемому набору данных.

Если набор данных имеет более трех измерений, то возможны такие варианты:

- использование многомерных методов представления информации (они рассмотрены ниже);
- снижение размерности до одно-, двух- или трехмерного представления. Существуют различные способы снижения размерности, один из них - факторный анализ - был рассмотрен в одной из предыдущих лекций. Для снижения размерности и одновременного визуального представления информации на двумерной карте используются *самоорганизующиеся карты* Кохонена.

Представление данных в 4 + измерениях

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

Наиболее известные способы многомерного представления информации:

- *параллельные координаты* ;
- " *лица Чернова* ";
- лепестковые диаграммы.

Параллельные координаты

В *параллельных координатах* переменные кодируются по горизонтали, вертикальная линия определяет *значение* переменной. Пример набора данных, представленного в декартовых координатах и *параллельных координатах*. Этот метод представления многомерных данных был изобретен Альфредом Инселбергом (Alfred Inselberg) в 1985 году.

"Лица Чернова"

Основная идея представления информации в " *лицах Чернова* " состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица [66]. Пример такого "лица" приведен на рис.16.2.

Для каждого наблюдения рисуется отдельное "лицо". На каждом "лице" относительные значения переменных представлены как формы и размеры отдельных черт лица (например, *длина* и ширина носа, размер глаз, размер зрачка, угол между бровями).

Анализ информации при помощи такого способа отображения основан на способности человека интуитивно находить сходства и различия в чертах лица.

Перед использованием методов визуализации необходимо:

- Проанализировать, следует ли изображать все данные или же какую-то их часть.
- Выбрать размеры, пропорции и масштаб изображения.
- Выбрать метод, который может наиболее ярко отобразить закономерности, присущие набору данных.

Многие современные средства анализа данных позволяют строить сотни типов различных графиков и диаграмм. Поэтому выбор метода визуализации, если он самостоятельно осуществляется пользователем, не так

прост и легок, как может показаться на первый взгляд. Наличие большого количества средств визуализации, представленных в инструменте, который применяет *пользователь*, может даже вызвать растерянность.

Одну и ту же информацию можно представить при помощи различных средств. Для того чтобы средство визуализации могло выполнять свое основное назначение - представлять информацию в простом и доступном для человеческого восприятия виде - необходимо придерживаться законов соответствия выбранного решения содержанию отображаемой информации и ее функциональному назначению. Иными словами, нужно сделать так, чтобы при взгляде на визуальное *представление* информации можно было сразу выявить закономерности в исходных данных и принимать на их основе решения.

Среди двухмерных и трехмерных средств наиболее широко известны линейные графики, линейные, столбиковые, круговые секторные и *векторные* диаграммы.

Приведем рекомендации по использованию этих наиболее простых и популярных средств визуализации.

При помощи **линейного графика** можно отобразить тенденцию, передать изменения какого-либо признака во времени. Для сравнения нескольких рядов чисел такие графики наносятся на одни и те же оси координат.

Гистограмму применяют для сравнения значений в течение некоторого периода или же соотношения величин.

Круговые диаграммы используют, если необходимо отобразить соотношение частей и целого, т.е. для анализа состава или структуры явлений. Составные части целого изображаются секторами окружности. Секторы рекомендуют размещать по их величине: вверху - самый крупный, остальные - по движению часовой стрелки в порядке уменьшения их величины. Круговые диаграммы также применяют для отображения результатов факторного анализа, если действия всех факторов являются однонаправленными. При этом каждый фактор отображается в виде одного из секторов круга.

Выбор того или иного средства визуализации зависит от поставленной задачи (например, нужно определить структуру данных или же динамику процесса) и от характера набора данных.

Качество визуализации

Современные аналитические средства, в том числе и *Data Mining*, немыслимы без качественной визуализации. В результате использования средств визуализации должны быть получены наглядные и выразительные, ясные и простые изображения, за счет использования разнообразных средств: цвета, контраста, границ, пропорций, масштаба и т.д.

В связи с ростом требований к средствам визуализации, а также необходимости сравнивания их между собой, в последние годы был

сформирован ряд принципов качественного визуального представления информации.

Принципы Тафта (Tufte's Principles) графического представления данных высокого качества гласят:

- предоставляйте пользователю самое большое количество идей, в самое короткое время, с наименьшим количеством чернил на наименьшем пространстве;
- говорите правду о данных.

Основные ***принципы компоновки визуальных средств*** представления информации:

1. Принцип лаконичности.
2. *Принцип обобщения* и унификации.
3. Принцип акцента на основных смысловых элементах.
4. Принцип автономности.
5. Принцип структурности.
6. Принцип стадийности.
7. Принцип использования привычных ассоциаций и стереотипов.

Принцип лаконичности говорит о том, что средство визуализации должно содержать лишь те элементы, которые необходимы для сообщения пользователю существенной информации, точного понимания ее значения или принятия (с вероятностью не ниже допустимой величины) соответствующего оптимального решения.

Кроме обозначенных выше принципов, средство визуализации должно обладать высокой надежностью и скоростью, которая устроит пользователя, принимающего на основе этой информации решения.

Представление пространственных характеристик

Отдельным направлением визуализации является наглядное *представление* пространственных характеристик объектов. В большинстве случаев такие средства выделяют на карте отдельные регионы и обозначают их различными цветами в зависимости от значения анализируемого показателя.

4. Основные тенденции в области визуализации

Как уже отмечалось, при помощи средств визуализации поддерживаются важные задачи бизнеса, среди которых - процесс *принятия решений*. В связи с этим возникает необходимость перехода средств визуализации на более качественный уровень, который характеризуется появлением абсолютно новых средств визуализации и взглядов на ее функции, а также развитием ряда тенденций в этой области.

Среди основных тенденций в области визуализации Филип Рассом (Philip Russom) выделяет:

- Разработка сложных видов диаграмм.
- Повышение уровня взаимодействия с визуализацией пользователя.

- Увеличение размеров и сложности структур данных, представляемых визуализацией.

1. Разработка сложных видов диаграмм.

Большинство визуализаций данных построено на основе диаграмм стандартного типа (секторные диаграммы, графики рассеяния и.т.д.). Эти способы являются одновременно старейшими, наиболее элементарными и распространенными. В последние годы перечень видов диаграмм, поддерживаемых инструментальными средствами визуализации, существенно расширился.

2. Повышение уровня взаимодействия с визуализацией пользователя.

Еще совсем недавно большая часть средств визуализации представляла собой статичные диаграммы, предназначенные исключительно для просмотра. Сейчас широко используются динамические диаграммы, уже сами по себе являющиеся пользовательским интерфейсом, в котором пользователь может напрямую и интерактивно манипулировать визуализацией, подбирая новое представление информации.

Например, базовое взаимодействие позволяет пользователю вращать диаграмму или изменять ее тип в поисках наиболее полного представления данных. Кроме того, пользователь может менять визуальные свойства - к примеру, шрифты, цвета и рамки. В визуализациях сложного типа (графиках рассеяния или диаграммах констелляции) пользователь может выбирать информационные точки с помощью мыши и перемещать их, облегчая тем самым понимание представления данных.

Более совершенные методы визуализации данных часто включают в себя диаграмму или любую другую визуализацию как составной уровень. Пользователь может углубляться (drill down) в визуализацию, исследуя подробности обобщенных ею данных, или углубляться в OLAP, *Data Mining* или другие сложные технологии.

Сложное взаимодействие позволяет пользователю изменять визуализацию для нахождения альтернативных интерпретаций данных. Взаимодействие с визуализацией подразумевает минимальный по своей сложности пользовательский интерфейс, в котором пользователь может управлять представлением данных, просто "кликая" на элементы визуализации, перетаскивая и помещая представления объектов данных или выбирая пункты меню. Инструменты OLAP или *Data Mining* превращают непосредственное взаимодействие с визуализацией в один из этапов итерационного анализа данных. Средства Text Mining или управления документами придают такому непосредственному взаимодействию характер навигационного механизма, помогающего пользователю исследовать библиотеки документов.

Визуальный запрос является наиболее современной формой сложного взаимодействия пользователя с данными. Пользователь использует информационные точки графика рассеяния, выбирать их мышкой и получать

новые визуализации, представляющие именно эти точки. Приложение визуализации данных генерирует соответствующий язык запроса, управляет принятием запроса базой данных и визуально представляет результирующее множество. Пользователь может сфокусироваться на анализе, не отвлекаясь на составление запроса. Увеличение размеров и сложности структур данных, представляемых визуализацией.

Визуализация поддерживает обработку структурированных данных, она также является ключевым средством представления схем так называемых неструктурированных данных, например текстовых документов, т.е. Text Mining. В частности, средства Text Mining могут осуществлять парсинг больших пакетов документов и формировать предметные указатели понятий и тем, освещенных в этих документах. Когда предметные указатели созданы с помощью нейросетевой технологии, пользователю непросто продемонстрировать их без некоторой формы визуализации данных. Визуализация в таком случае преследует две цели:

- визуальное представление контента библиотеки документов;
- навигационный механизм, который пользователь может применять при исследовании документов и их тем.

Визуальный анализ данных обычно выполняется в три этапа:

- беглый анализ - позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;
- увеличение и фильтрация - идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;
- детализация по необходимости - если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные.