2. ГИС как средство визуализации и анализа данных различной природы

2.1 Введение в ГИС

Геоинформационная система (ГИС) – программно-аппаратный комплекс, предназначенный для сбора, управления, анализа и отображения пространственно распределенной информации.

ГИС – не только и не столько информационные системы для географии, сколько информационные системы с географически организованной информацией. В простейшем варианте геоинформационные системы – сочетание обычных баз данных (атрибутивной информации) с электронными картами, то есть мощными графическими средствами.

Основная идея ГИС – связь данных на карте и в базе данных. ГИС – это и аналитические средства для работы с любой координатно-привязанной информацией. В принципе, ГИС можно рассматривать как некое расширение концепции баз данных. В этом смысле ГИС фактически представляет собой новый уровень и способ интеграции и структурирования информации [32].

ГИС предлагает совершенно новый путь развития картографии. Преодолеваются основные недостатки обычных карт – их статичность и ограниченная емкость как носителя информации. В последние десятилетия бумажные карты из-за перегруженности информацией становятся нечитабельными. ГИС же обеспечивает управление визуализацией информации. Появляется возможность выводить (на экран, на твердую копию) только те объекты или их множества, которые интересуют нас в данный момент. Фактически осуществляется переход от сложных комплексных карт к серии взаимоувязанных частных карт. При этом улучшается структурированность информации, а следовательно, повышается эффективность ее обработки и анализа. В ГИС карта оживает и становится действительно динамическим объектом в смысле:

- изменяемости масштаба;
- преобразования картографических проекций;
- варьирования объектным составом карты;
- возможности опрашивать через карту в режиме реального времени многочисленные базы данных;
- изменения способа отображения объектов (цвет, тип линии и т.п.), в том числе и определения символогии через значения атрибутов, то есть синхронизации визуализации с изменениями в базах данных;
- легкости внесения любых изменений.

Рассмотрим основные понятия ГИС, в том или ином виде присутствующие во всех современных геоинформационных системах.

Данные

В ГИС данные делятся на две категории:

- пространственные (местоположение);
- непространственные (атрибуты).

Объекты

Пространственные данные включают географические объекты, представляемые:

- точками;
- линиями;
- полигонами.

Дугами описываются те реальные объекты, которые можно рассматривать как линии. Дуга состоит из отрезков линий и дуг окружностей.

Полигоны – замкнутые области, которые представляют однородные по некоторым критериям участки.

Атрибутивные данные могут включать идентификатор объекта, любую описательную информацию из баз данных, изображение и многое другое.

Слой

Слои в карте подразделяются на два основных вида – растровые и векторные.

Векторные слои – это совокупность простых геометрических объектов (точка, дуга, полигон), которые представляют те или иные объекты на местности. Векторные слои могут также хранить топологию, т.е. информацию о взаимном расположении объектов.

Растровые слои представляют из себя сплошные изображения. Они не могут содержать объекты. Однако они могут служить фоном для векторных слоев

Объект слоя

Каждому объекту векторного слоя может соответствовать запись в базе данных, чем обеспечивается привязка информации к местности. Это соответствие может обеспечиваться в частности назначением каждому объекту соответствующего идентификатора.

Легенда карты

Легенда карты – свод *условных обозначений*, использованных на карте, с текстовыми пояснениями к ним. Обычно, легенды создаются на основе классификаций изображаемых объектов и явлений, они становятся их графической моделью и часто служат для построения классификаторов.

Карта

Представляет собой набор географических слоев, каждый из которых привносит в карту информацию по какой-либо определенной теме. Например, на слой границ некоторой территории может быть нанесен слой рек, затем слой, отображающий количество атмосферных осадков в процентном отношении и т.д.

Электронную карту в ГИС можно рассматривать как многокомпонентную модель реальности. Основными целями ее создания являются:

- графическая коммуникация пространственных отношений и распределений;
- улучшение возможности анализа, обработки и отображения геоинформационных данных;
- визуальное отображение цифровых моделей явлений, невидимых для человеческого глаза;
- автоматизация отображения и картографического анализа в системах управления; исследование объектов, явлений и процессов с учетом динамики их развития и возможного использования;
- получение аналитических решений в графическом виде в режимах реального и разделенного времени и т.д.

2.2 Модели ГИС

Основой визуального представления данных при помощи ГИС-технологий служит так называемая графическая среда. Основу графической среды и соответственно визуализации базы данных ГИС составляют векторные и растровые модели.

В общем случае модели пространственных (координатных) данных могут иметь векторное или растровое (ячеистое) представление, содержать или не содержать топологические характеристики. Этот подход позволяет классифицировать модели по трем типам:

- растровая модель;
- векторная нетопологическая модель;
- векторная топологическая модель.

Все эти модели взаимно преобразуемы. Тем не менее, при получении каждой из них необходимо учитывать их особенности. В ГИС форме представления координатных данных соответствуют два основных подкласса моделей – векторные и растровые (ячеистые или мозаичные). Возможен класс моделей, которые содержат характеристики как векторов, так и мозаик. Они называются гибридными моделями.

В дальнейшем под терминами решетка, мозаика, элемент растра будем понимать одно и то же. Основу такой классификации составляет атомарная единица (пространства), содержащая представления площадей, линий и точек.

Между векторными и растровыми изображениями имеется различие, характерное именно для ГИС. Растровые изображения отображают поля данных, т.е. носят полевой характер. Векторные изображения в ГИС, как правило, отображают геоинформационные объекты, т.е. носят объектный характер.

Растровые модели

Рассмотрим подробнее растровые модели данных, которые ближе касаются нашей основной задачи, нежели чем векторные. Напомним, что

модель данных представляет собой отображение непрерывных последовательностей реального мира в набор дискретных объектов.

В растровых моделях дискретизация осуществляется наиболее простым способом – весь объект (исследуемая территория) отображается в пространственные ячейки, образующие регулярную сеть. При этом каждой ячейке растровой модели соответствует одинаковый по размерам, но разный по характеристикам (цвет, плотность) участок поверхности объекта. В ячейке модели содержится одно значение, усредняющее характеристику участка поверхности объекта. В теории обработки изображений эта процедура известна под названием пикселизация или растеризация.

Если векторная модель дает информацию о том, где расположен тот или иной объект, то растровая – информацию о том, что расположено в той или иной точке территории. Это определяет основное назначение растровых моделей – непрерывное отображение поверхности.

В растровых моделях в качестве атомарной модели используют двумерный элемент – пиксель (ячейка). Упорядоченная совокупность атомарных моделей образует растр, который, в свою очередь, является моделью карты или геообъекта.

Растровые модели имеют следующие достоинства:

- растр не требует предварительного знакомства с явлениями, данные собираются с равномерно расположенной сети точек, что позволяет в дальнейшем на основе статистических методов обработки получать объективные характеристики исследуемых объектов. Благодаря этому растровые модели могут использоваться для изучения новых явлений, о которых не накоплен материал. В силу простоты этот способ получил наибольшее распространение;
- растровые данные проще для обработки по параллельным алгоритмам и этим обеспечивают более высокое быстродействие по сравнению с векторными;
- некоторые задачи, например создание буферной зоны, много проще решать в растровом виде;
- многие растровые модели позволяют вводить векторные данные, в то время как обратная процедура весьма затруднительна для векторных моделей;
- процессы растеризации много проще алгоритмически, чем процессы векторизации, которые зачастую требуют экспертных решений.

Данные для анализа могут быть получены из векторных слоев, отражающих поля тематических или/и временных характеристик, растеризацией и записаны в таблицу или напрямую занесены туда из отчетов. Таблица, содержащая атрибуты объектов, называется таблицей атрибутов. В таблице каждому объекту соответствует строка таблицы, каждому тематическому признаку – столбец таблицы. Каждая клетка таблицы отражает значение определенного признака для определенного

объекта.

В общем случае ввод информации для задач ГИС осуществляется комплексно: по данным дистанционного зондирования, со снимков спутников, аэроснимков, по материалам дешифрирования снимков, полевым измерениям, по информации с карт.

2.3 Основные идеи метода анализа данных в ГИС с помощью искусственных нейронных сетей

Далеко не все ГИС снабжены возможностями специализированного анализа, например геологического. Связано это с тем, что четкой схемы проведения таких работ, не существует и организации, занимающиеся ими, предпочитают производить анализ по собственным методикам и правилам. Работа со специфическими данными специфическим образом является характерной чертой этого типа анализа. Кроме того, взгляды на приемы его проведения могут меняться с течением времени. Поэтому такие возможности в ГИС представляются средствами создания приложений самими пользователями. Сложность состоит в том, для каждой специализированной области возникает необходимость создавать отдельное приложение к ГИС и часто даже свою методику обработки. Это не всегда возможно и часто дорого.

Нейронные сети претендуют на то, чтобы стать универсальным аппаратом решающим разные специфические задачи из разных проблемных областей в ГИС [33, 34]. Такая универсальность обуславливается тем, что нейросети дают стандартный способ решения многих нестандартных задач [35]. И неважно, что специализированная программа решит лучше один класс задач. Важнее что один нейроимитатор решит и эту задачу и другую и третью и не надо каждый раз создавать специализированные приложения для каждой специфической задачи [36].

Обобщение задач

Как правило, модули, реализующие специализированный анализ для разных проблемных областей, решают одинаковые качественные задачи. Перейдя от специфических частных аналитических задач к общему видению проблемы в целом можно увидеть одно важное обстоятельство. А именно, что большинство аналитических задач сводится к одной проблеме, которая легко формулируется, но сложно решается: к проблеме заполнения пропусков в таблице [35, 37-44].

Учитывая то, что часто методика обработки неизвестна, с этой задачей справляются лучше всего нейронные сети, которые позволяют строить эмпирические зависимости [45, 46] без привлечения дополнительной информации. Проблема заполнения пропусков в таблице тесно связана с задачами, такими как построение отношений на множестве объектов и построение функции по конечному набору значений [35, 47-55]. В такой постановке преследуемая цель – это восстановление пропущенных данных. В нашем случае наиболее общим способом проблема восстановления пропущенных данных формулируется как построение

(дополнение) одного из слоев по информации, имеющейся в других слоях карты. В такой постановке она является решением большинства классификационных задач в ГИС [23, 24].

Методы классификации используются в решении следующих основных задач:

- классификация процессов и явлений;
- районирование, типология;
- выявление определяющих факторов;
- временной анализ;
- интерполяция и создание моделей поверхности
- анализ и прогнозное картирование пространственно распределенных данных и т.д.

Формальная постановка

Пусть, существует набор пространственных данных (сеть мониторинга). Обычно, данные представляются в виде: X, Y – пространственные координаты, Z – зависящая от них переменная. Задачей картирования пространственных данных, как правило, является интерполяция неравномерных данных Z на равномерной координатной сетке.

Как уже отмечалось в первой главе, существует три вида постановки задач относительно расположения географических явлений в пространстве (рис.2.1).

Для исследователя географических комплексов интерес представляют все три модели. В данной работе акцент сделан на первой и второй модели, поскольку конечной целью обработки данных является

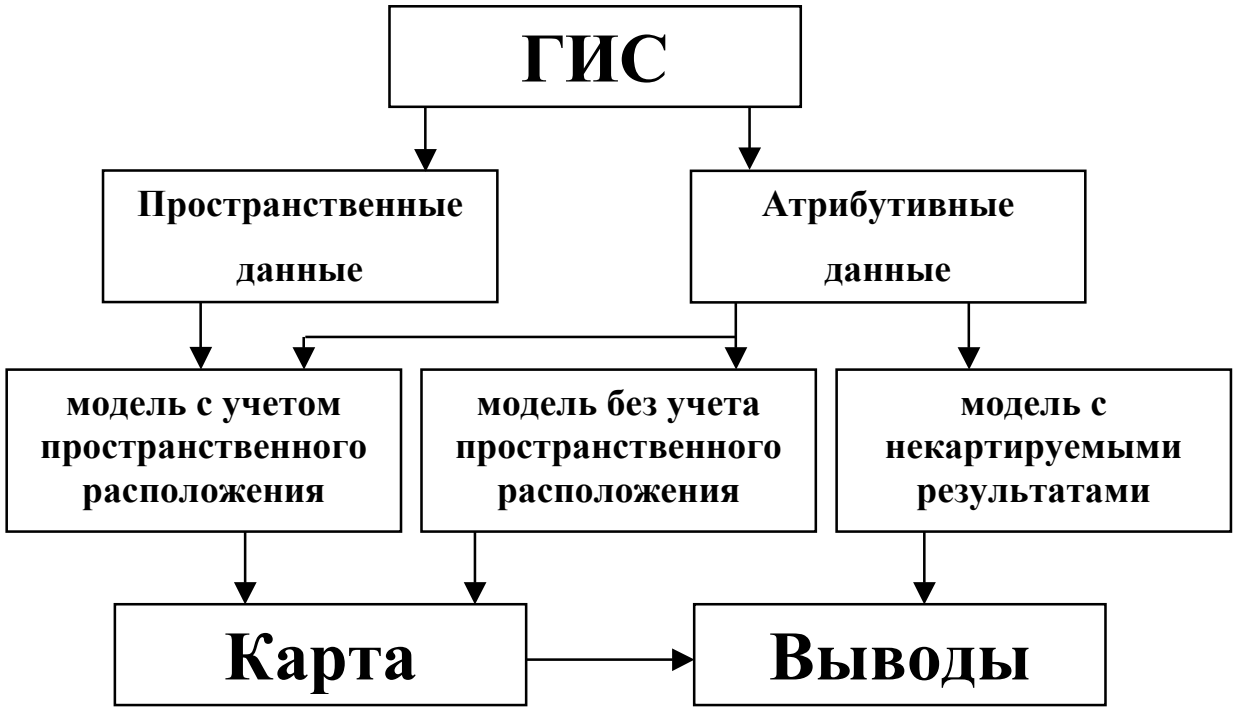

Рис. 2.1. Способы анализа данных ГИС

получение карты.

Рассмотрим более общий случай и введем фундаментальное предположение о фиксированной «вертикальной» связи между слоями. Пусть, как и прежде, существует набор пространственных данных. Предположим что каждую точку сетки с координатами (x,y) характеризует некоторый вектор состояний ($z_1,\ldots,z_n$). Для всей сетки получаем набор векторов $Z_1,\ldots,Z_n$ – параметров в точках сетки мониторинга. Часть параметров – координаты. В общем случае пространственное положение может быть выражено через относительные единицы, например, как обратно пропорциональное квадрату расстояний между объектами.

Данные легко представимы в виде двумерной таблицы, в которой столбцы – это структура параметров $Z_1,\ldots,Z_n$, а строки точки сетки. Показатели состояния $Z_1,\ldots,Z_n$ разделяются на входные переменные C_i ($i = 1,\ldots,p$), полученные тем или иным способом, и выходные D_j ($j = 1,\ldots,q$), ($p + q = n$) – те которые нужно выразить через входные. Т.е. найти функционал: $D_j = F(C_1,\ldots,C_p)$, ($j = 1,\ldots,q$). Выходные параметры могут быть какими-либо известными классификациями, непрерывными измерениями или другими значениями. Каждый параметр Z_k может быть отдельным атрибутивным слоем в ГИС. Предположение о фиксированной "вертикальной" зависимости между слоями состоит в следующем.

Упростим задачу, сведя ее к классификации на два класса с одним выходным параметром. Если такое разделение возможно, то F является решающим правилом отнесения к одному из классов в зависимости от входных данных. Следовательно, отнесение точки сетки к классу зависит только от параметров самой точки и не зависит напрямую от соседних точек. Все связи между соседями, в том числе пространственное положение, можно закодировать в параметрах $Z_1,\ldots,Z_n$ для каждой точки. Для большинства классификационных задач, например, поиска полезных ископаемых по косвенным признакам, прямую пространственную привязку можно исключить. Это позволяет знания о уже разведанной территории переносить на исследуемую. Естественно с учетом некоторой похожести. Пример – заполнение пробелов в данных на исследуемой территории.

Основная задача, которую нужно решить – это задача заполнения пробелов (восстановления, предсказания) в выходных параметрах. Иными словами построение (дополнение) одного или нескольких слоев по информации, имеющейся в других слоях карты. Попутно, возникает ряд проблем, связанных с заполнением пропусков данных во входных параметрах, исключением незначимых для решения основной задачи признаков и других.

2.4 Аналитические задачи в ГИС, решаемые с помощью искусственных нейронных сетей

Опишем круг задач, требующих решения в ГИС, для которых могут быть использованы нейросетевые технологии.

Построение (дополнение) слоя

Основная задача, к которой, так или иначе, относятся остальные,

описанные ниже, это построение слоя. Она означает заполнение его недостающих частей (или построение слоя полностью) по информации, имеющейся в других слоях, на основе нахождения некоторой функциональной зависимости между параметрами, полученными эмпирическим путем, и скрытыми теоретическими параметрами, определяющими сущностные характеристики каждой конкретной точки.

Даны слои качественных характеристик одной и той же территории. Слой, который необходимо восстановить, известен частично. Для восстановления слоя при обучении нейросети используется только та информация из слоев, которая покрывает известные участки слоя с пробелами. После обучения можно распространить знания о зависимости между слоями на отсутствующие области карты. Получившиеся знания обладают переносимостью за рамки данной территории. Все описанные ниже задачи можно рассматривать как частный случай данной.

Классификационные задачи. Поскольку при сборе информации для БД приходится иметь дело с результатами измерений, определим по этому показателю три типа задач классификации.

К задачам классификации первого типа относятся те, в которых исходные измерения требуется разделить на устойчивые группы. Их называют задачами классификации без учителя, кластеризации, таксономии, типизации [56-62]. Этот тип классификации применяется для обработки опытных данных.

Задачи классификации второго типа характеризуются тем, что исходные данные уже сгруппированы и требуется оценить их информативность (значимость) относительно совокупности известных эталонов. Такого рода задачи встречаются при распознавании образов [63-65], дешифрировании снимков и т.д.

Задачи классификации третьего типа – задачи разбиения. В них исходные измерения или их функции требуется разбить на устойчивые группы в зависимости от их величины (типичный пример – зонирование) [23, 30, 66].

В ГИС задачи классификации первого типа возникают и решаются при разработке классификаторов, т.е. при организации информационной основы, задачи второго типа – при сборе первичных данных и при использовании ГИС для экспертных решений или оценок. Задачи классификации третьего типа возникают в приложениях ГИС для решения проблем в области экологии, землепользования, статистики и т.п.

Восстановление легенды слоя

Вторая решаемая задача – восстановление легенды. Классификация с учителем – генерация объектов слоя по заданным классификационным правилам. Правила задаются во время обучения нейросети и остаются скрытыми от пользователя. Пользователь имеет возможность задавать, по его мнению, полезные для классификации признаки, выбрав слои, участвующие в обучении. Типичная задача поиск полезных ископаемых по косвенным признакам. Эта задача решается на основе информации об уже

разведанных месторождениях и полевых съемках косвенных признаков. Знания, полученные при обучении, переносимы на другую территорию с известными косвенными признаками.

Районирование и типология

Зонирование. Основное назначение функций этой группы состоит в построении новых объектов – зон до того на карте не существовавших, т.е. участков территорий, однородных в смысле некоторого критерия или группы критериев. Границы зон могут либо совпадать с границами ранее существовавших объектов (задача определения "нарезки" избирательных округов по сетке квартального деления), либо строиться в результате различных видов моделирования (зоны экологического риска). Типичные задачи этого типа: выделение зон градостроительной ценности территорий, зон экологического риска, зонирование урбанизированных территорий по транспортной доступности, построение зон обслуживания поликлиник и т.д. Работа может производиться как с растровыми, так и с векторными изображениями.

В сущности, зонирование это – классификация без учителя. Задан набор объектов, каждому объекту сопоставлен вектор значений признаков (строка таблицы). Требуется разбить эти объекты на классы эквивалентности.

Отнесение объекта к классу проводится путем его сравнения с типичными элементами разных классов и выбора ближайшего. Простейшая мера близости объектов – квадрат евклидового расстояния между векторами значений их признаков (чем меньше расстояние, тем ближе объекты). Соответствующее определение признаков типичного объекта – среднее арифметическое значение признаков по выборке, представляющей класс. Другая мера близости, естественно возникающая при обработке сигналов, изображений и т.п. – квадрат коэффициента корреляции (чем он больше, тем ближе объекты). Возможны и иные варианты – все зависит от задачи. Для каждого нового объекта нужно выполнить два действия:

1) найти класс, к которому он принадлежит;

2) использовать новую информацию, полученную об этом объекте, для исправления (коррекции) правил классификации.

В результате классификации как бы появляются новые имена и правила их присвоения.

Создание моделей поверхностей

Создание моделей поверхностей – это и построение моделей изолинейных изображений по регулярным и нерегулярным сеткам и создание модели трехмерной визуализации, например, построение панорамы города в аксонометрической или иной проекции. Расчет производится по содержащимся в базах данных численным характеристикам. Моделироваться могут, как изображения действительного рельефа или непрерывного поля, современного или с учетом динамических изменений, так и воображаемые поверхности,

построенные по одному или нескольким показателям, например, поверхность цен на землю, плотность дорожной сети или населения и т.п.

Интерполяция и прогнозное картирование

Задача – интерполяция пространственно распределенных данных. Сводиться к задаче построения функции по конечному набору значений и как следствие к задаче заполнения пробелов. Цель – извлечение максимума информации из набора данных, учитывая возможные ошибки измерений, неравномерную плотность сетки мониторинга, и прочие помехи, встречающиеся при реальных измерениях. Данные по окружающей среде обладают неоднородностью как на крупных, так и на мелких масштабах, что затрудняет анализ. Нейросетевая обработка обладает рядом преимуществ перед детерминистическими моделями.

Временной анализ

Временной анализ растровых изображений. В качестве таких изображений в ГИС обычно выступают снимки или растеризованные векторные изображения. Преимущество снимков – в их современности и достоверности, поэтому часто встречающийся вид анализа в этой группе – временной. Сравниваются и ищутся различия между снимками различной давности, таким образом, оценивается динамика произошедших изменений. Не менее часто анализируются пространственные взаимосвязи двух или нескольких явлений.

Анализ временных рядов содержит комплекс задач, которые сводятся к построению функций по конечным наборам значений и заполнению пробелов в таблицах. Временные ряды представляют собой специальный вид таблиц и заслуживают отдельного рассмотрения. Для каждого типа объектов выделяется набор постоянных признаков (констант) и множество свойств, меняющихся со временем (переменных признаков). Предполагается, что в любой момент времени для каждого объекта существуют свои значения переменных признаков. Вот, например, три задачи, специфичные для обработки временных рядов:

а) определение констант (всех или части) по известным значениям переменных в разные моменты времени;

б) предсказание значений части переменных в некоторый момент времени по известным значениям констант, переменных в нескольких предшествовавших моментах времени и части переменных в текущий момент;

в) определение объема данных о прошлом, достаточных для предсказания будущего на конкретное время и с заданной точностью.

Обычная задача при временном анализе – получение прогноза. Легко заметить, что решение такой задачи немногим отличается от решения задачи по восполнению пробелов в слое на основе информации, заключенной в других слоях. Единственное концептуальное отличие состоит в том, что слои вместо разных пространственных признаков содержат изменение во времени одного и того же слоя. Для примера, упростим описание анализа временных рядов. Возьмем за основу

известные слои одной и той же территории в количестве N-1 (без последнего), а в качестве восстанавливаемого слой с номером N. Произведем обучение нейросети. Для прогноза в качестве входных параметров возьмем слои в количестве N-1 (без первого) и подадим на вход нейросети. На выходе получим прогнозируемый слой N+1.

Выбор значимых признаков

Анализ значимости. Как уже было сказано, основной задачей является восполнение пробелов в данных и решается она применительно к данной области построением слоя по слою (или нескольким слоям). При этом исследуется вопрос, какие из входных сигналов являются доминирующими, (значимыми) при принятии нейросетью решения, а какие нет. Другими словами, насколько каждый слой участвующий в построении влияет на восполнение пробелов. Такая информация дает знание, например, о том какие признаки можно убрать из рассмотрения, а какие оставить. То есть решается задача нахождения оптимального набора исходных показателей, которые полностью описывают изучаемые явления. Это может помочь в понимании сущности географического комплекса.

Значимость по слою складывается из значимости точек сетки. Благодаря такой информации можно видеть, какие области из слоев участвующих в качестве входов были значимы при построении. Таким образом, получаем представление о территориальном распределении значимости.

Нейросетевые технологии анализа данных решают такие задачи и позволяют помочь в оптимальном выборе системы исходных показателей, исследовать признаки на дублирование, выяснить значимость исходных признаков для решения основной задачи.

Подробнее все поставленные задачи и методы их решения в нейросетевом базисе описаны в третьей главе.

2.5 Основные идеи визуализации и анализа данных произвольной природы

Пожалуй, главным преимуществом ГИС является наиболее естественное (для человека) представление как собственно пространственной информации, так и любой другой информации, имеющей отношение к объектам, расположенным в пространстве (т.н. атрибутивной информации). Пространством можно называть не только трехмерное пространство, в котором мы существуем, но и любое абстрактное пространство произвольной размерности.

Это свойство ГИС является определяющим для использования предлагаемого подхода визуализации данных, поскольку основное качество ГИС – это наглядность. Кроме того, современные ГИС имеют множество мощных инструментов для анализа. Отображение в готовых ГИС произвольных данных позволяет подключить для визуализации и анализа весь накопленный арсенал средств обработки пространственной информации.

Предлагаемый новый подход позволяет отображать многомерные

данные в общем случае различной природы и не обязательно числовые [19, 67, 68]. Например, можно представлять патенты, статьи, курсы акций, временные ряды, ... Из данных создается карта, куда они будут нанесены. Карта это трансформируемый объект – атлас и на этом атласе можно отразить существенные детали данных, дающие представление об их структуре.

Описание задачи

Обычно подразумевается, что данные при нанесении на карту отображаются на какую-либо подложку. Например, какие-либо характеристики накладываются на территорию. Посмотрим на это с новой стороны и предположим, что подложка порождается с использованием самих данных. Отобразив данные специальным образом, мы получим их визуальное представление в виде некоторого многообразия, например, в виде пленки натянутой на многомерные данные. Пленку затем спроецируем на поверхность. Данные же отображаются ближайшей точкой пленки. В результате и получим подложку для данных, функций над данными, производных, показателей значимости, классификаций, отношений данных, различных тематик и др. [69, 70]

Такая визуализация многомерных данных осуществляется понижением размерности с сохранением некоторых специфических особенностей исходного пространства данных.

Картографирование является прямым следствием визуализации данных [71]. Объекты, с которыми оперирует метод, отображаются в слои, если угодно одной территории данных, с которыми можно проводить аналитические операции, принятые в ГИС. Если существует географическая привязка, то ничто не мешает отображать результат анализа дополнительно ко всему прочему на реальные территории.

Перейдем к формальному описанию задачи визуализации и картографирования данных. Она заключается в отображении многомерных данных в представимую человеком размерность, например, на плоскость так, чтобы точки данных, близкие на плоскости (на карте), были близки и в исходном пространстве (обратное в общем случае неверно).

Понятно, что, визуализируя данные, мы можем получать большое количество информации о них без какой-либо обработки. Становятся видимыми области группировки данных и разреженные области. Например, упрощается решение задач классификации. Видно количество кластеров, их форма, взаимное расположение и т.д. Обратим внимание, что это естественная классификация данных.

Заметим, однако, что все это видно когда данные отображаются на многообразия малой размерности. Размерности один, два максимум три. Типичные же данные при решении серьезных задач это, например, 100 мерное пространство и 100 – 100000 точек в нем. Даже если размерность или объем выборки меньше, то все равно осмысленно ее представить человек не в состоянии за исключением двух- или трехмерных.

Для дальнейшего изложения необходимо описать сущности, с

которыми оперирует метод.

Объекты метода

Данные

Данные, которые можно картографировать могут быть любые, т.е. все. Слово "все" выступает здесь в трех значениях.

Все данные как любая информация о мире. Предлагаемый подход позволяет строить отображение многомерных данных, заключенных в таблицах, в "человеческом" виде. Отсюда следует, что если данные можно представить в таблице, то они могут быть картографированы. А такой информации в мире большинство.

Все для одной задачи, без изъятия, целиком, в полном составе. Нейросетевые технологии позволяют решать такую задачу как определение значимости входной информации для решения задачи. Поэтому можно давать все собранные данные по задаче и получать сокращенный набор признаков, факторов необходимых для ее решения.

Все в смысле, какие есть, с пробелами и неполные. Нейросетевые методы позволяют заполнять пробелы в данных [35, 72]. Для заполнения пропусков, как правило, решается или задача построения функции по конечному набору значений или задача построения отношений на множестве объектов. Для этого могут использоваться разные методы, например, линейная регрессия, транспонированная регрессия, нейросетевая нелинейная регрессия, линейный и квазилинейный факторный анализ, мозаичная регрессия [47, 48, 73-76].

В некотором смысле любая обработка данных заполняет области незнания. Считается, что нейросети делают это "хорошо". И большое преимущество имеют в области плохо формализуемых и нестандартных задач, а также в тех случаях, когда плотность пробелов высока, расположены они нерегулярно, а данных немного, например, число объектов (строк) примерно таково же, как и число признаков (столбцов).

Предложенный алгоритм картографирования данных большой размерности не требует предварительного априорного заполнения пробелов. В общем случае может быть несколько вариантов работы с неполными данными. Отображать с пропусками – в этом случае пропущенные данные не влияют на построение поверхности (карты). Заполнить пропуски перед отображением. Заполнить во время и путем отображения многомерного пространства данных.

Чаще всего данные должны быть предварительно нормированы (обезразмерены) – переходом в каждом столбце таблицы к "естественной" единице измерения. Обычно нормировка производится на единичное среднеквадратичное уклонение в столбцах или на единичный разброс данных в каждом столбце (если нет каких-либо специфических ограничений, связанных со смыслом задачи).

Графически облака данных представляются точками на одной из координатных плоскостей базового пространства. Об информативном

отображении данных будет сказано дальше.

Многообразия

Существует многомерное облако данных. Многообразия это построенные в этом облаке поверхности малой размерности, приближающие его.

Некоторое представление может дать описание самоорганизующейся карты (Self-Organizing Map – SOM). В 1982 году финский ученый Тойво Кохонен [77] предложил ввести в базовое правило обучения нейросети информацию о пространстве. Построение топографических карт (карт Кохонена) является методом, дающим оптимальное представление информации в виде координат двумерной сетки.

В многомерное пространство данных погружается двумерная сетка. Эта сетка изменяет свою форму таким образом, чтобы по возможности точнее аппроксимировать облако данных. Каждой точке данных ставится в соответствие ближайший к ней узел сетки. Таким образом, каждая точка данных получает некоторую координату на сетке.

Такое отображение локально непрерывно: близким точкам на карте соответствуют близкие точки в исходном пространстве (обратное, вообще говоря, не верно: близким точкам в исходном пространстве могут соответствовать далекие точки на карте). Таким образом, распределение данных на двумерной карте позволяет судить о локальной структуре многомерных данных.

Такая топографическая самоорганизующаяся карта дает наглядное представление о структуре данных в многомерном входном пространстве, геометрию которого мы не в состоянии представить себе иным способом. Визуализация многомерной информации является главным применением SOM.

Достоинства SOM начинают проявляться после нанесения на нее какой-либо графической информации. Различные раскраски топографической карты являются удобным средством для выявления взаимосвязей различных факторов. В принципе, любая характеристика порождает свою раскраску карты. Вместе подобные раскраски дают исчерпывающую и наглядную картину. Здесь имеется полная аналогия с географическими картами различных типов на одной и той же географической сетке, которые в совокупности дают полное представление о данной местности.

При построении многообразий можно пользоваться классическим методом главных компонент. Для определенности возьмем двумерный случай. При этом плоскости над данными строятся по двум главным компонентам. Также построения могут вестись по комбинациям пар компонент получающимся в результате дальнейшей обработки.

Предложенная технология моделирует данные (в общем случае – с пробелами) многообразиями (линейными и нелинейными) малой размерности. Для построения многообразий используется линейный метод

главных компонент, квазилинейный метод, надстраиваемый над линейным и использующий его результаты, существенно нелинейный метод, построенный с помощью формализма самоорганизующихся кривых [75]. Разработан метод построения ***упругой карты***, моделирующей данные [69-71].

Экстраполяция и интерполяция получаемых зависимостей производится линейно и с помощью формул Карлемана. Метод решает следующие задачи:

1) заполнение пробелов в данных;
2) ремонт данных – корректировка значений исходных данных так, чтобы наилучшим образом работали построенные модели;
3) построение вычислителя, заполняющего пробелы в поступающей на вход строке данных (в предположении, что данные в ней связаны теми же соотношениями, что и в исходной таблице).

Существует еще одна техника получения многообразий малой размерности моделирующих данные называемая "метод узкого горла". От нейросети требуется выдать те же вектора данных, которые были получены на входе, т.е. быть для них прозрачной. Эффект заключается в сокращении числа нейронов среднего слоя нейросети после ее обучения. Сеть таким образом можно разделить пополам по ее узкой части. Это будет напоминать кодирование-декодирование данных. Сократив до возможного минимума средний слой нейросети, получим на ее выходе внутренние координаты данных. По ним можно строить многообразия.

Применение многообразий малой размерности требует постановка задачи, а именно, визуально представить данные в естественном для человека виде. Иначе картография данных не имеет смысла. Поскольку не только отобразить, но и представить многомерное пространство данных в реальных задачах не представляется возможным.

Подведем итог. Многообразиями малой размерности могут быть в простейшем случае прямые, ломаные и надстройки над ними типа кривых и более сложные плоскости, пленки и упругие карты. Эти объекты располагаются в облаке данных, аппроксимируя их. Необходимо понимать, что вряд ли будет одно универсальное многообразие, поскольку решаемая задача в каждом конкретном случае накладывает условия на приближение. Поэтому их может быть несколько для одного и того же набора данных, и они могут составлять даже сообщества многообразий.

Проекции

С каждым *многообразием* связан проектор на него, с помощью которого данные отображаются на многообразии. Построение этого проектора может вестись различными способами [70].

"Образцовый" путь построения проектора – метод максимума правдоподобия (максимума вероятности, максимума энтропии…). Он предполагает, что плотность вероятности в точке на многообразии больше, чем в тех точках, которые в нее проецируются.

Также и с каждым *отображением данных* в меньшей размерности связано многообразие, погруженное в пространство данных и строящееся из тех же соображений. Эта двойственность "многообразие-проектор" является основой многих преобразований карт данных.

Построенный проектор дает знание об отображении точки данных из исходного пространства данных на многообразие малой размерности. Например, из стомерного пространства на плоскость или кривую. Каждая проекция – это новая подложка для данных и топологические свойства данных меняются от проекции к проекции. Зная правила проектирования, можно отображать дополнительные точки данных уже после построения карты.

Вновь поступившие данные занимают свое место в многомерном пространстве, проектор определяет их место на плоскости. Для прикладных задач это мгновенная классификация. Более того, при изменении некоторых характеристик новой точки данных проектор помогает отследить траекторию движения точки по плоскости. Также может решаться обратная задача, какие свойства и как нужно изменить, чтобы попасть в определенный класс.

Многообразие и проекция – это две взаимосвязанных вещи. У многообразия есть проектор, у каждого проектора многообразие. Иными словами, есть случаи, когда в облаке данных строиться многообразие малой размерности, например SOM или *упругая карта*, а затем определяется проекция, отображающая туда данные, и также каждый проектор определяет свое многообразие в данных.

Инъекции

Операция, обратная проекции. Объект, сопутствующий проекции. Оператор, отображающий точки с плоскости в многообразие R^n. Позволяет, выбрав точки на плоскости, узнать, где они находятся в пространстве. Например, определить пространственное расположение точек класса, точек разделяющей поверхности или выбрав на плоскости область, определить характеристики точки которые удовлетворяют условию выбора.

Развертки

Проекции многообразий малой размерности на стандартные многообразия. Многообразия в данных могут быть различной формы от плоских до сферических. Развертки – это отображения многообразий, которые уже имеют "хорошую" размерность, на некоторый набор стандартных, например, на прямую, плоскость, сферу (глобус), тор (глобус в форме бублика) ...

Слои

Теперь, опишем, пожалуй, главную сущность, в которой и заключается смысл картографирования данных. Слои имеют тот же смысл

что и слои в ГИС более того это они и есть. Существенная разница в том, что они отображают. Отображаться в слоях могут как все вышеописанные сущности, так и дополнительные характеристики данных которые и составляют основное информационное содержание карты [19].

Данные представляют собой при отображении точки. Они образуют в карте точечный слой. Многообразия отображаются сетками и образуют слои сеток. Проекции и инъекции выглядят как прямые, соединяющие точки данных с соответствующими узлами на многообразиях. Развертки образуют топографическую основу карты.

Удобным инструментом визуализации данных является раскраска описанных объектов аналогично тому, как это делают на обычных географических картах. Порождать свою раскраску ячеек сетки, проекций, данных и др. объектов могут различные характеристики данных. Это могут быть известные классификационные признаки, значимости, зависимости, производные. Любые функции над данными могут служить основой для раскраски.

Собрав воедино карты всех интересующих нас признаков, получаем топографический атлас, дающий интегральное представление о структуре многомерных данных.