Тема №13. Интеллектуальный анализ данных – технология Data Mining

— Задачи Data Mining.

— Специфические методы Data Mining-а

— Области применения методов Data Mining

— Средства реализации и задачи проектирования комплекса Data Mining

Задачи Data Mining

Следует различать два различных процесса получения знаний. Первый – это «извлечение» их из живого источника – эксперта, специалиста с целью их идентификации и возможной формализации, помещения в базу знаний и построения на этой основе экспертных систем, а также в других целях. Такой процесс относят к инженерии знаний. Другой – это «добыча» скрытых от пользователя знаний из данных, помещенных в различного рода компьютерные информационные системы, в том числе базы данных различного назначения, информационные хранилища. Процесс второго рода называют Data Mining – используют русский перевод «интеллектуальный анализ».

Предметом нашего изучения является Data Mining.

Для обработки накопленных в различных источниках и местах сбора и хранения данных и выполнения интеллектуального анализа используются все достижения математической науки и информационных технологий. В первую очередь используются методы линейной алгебры, классического математического анализа, дискретной математики, многомерного статистического анализа.

В экономической предметной области применение методов поиска решений, условий неотрицательности и других свойств математических моделей путем дедуктивного получения следствий, исходя из предварительно

сформулированных предпосылок, относится к разделу экономической науки, называемому математическая экономика.

Анализ количественных закономерностей и взаимозависимостей в экономике, который выполняется статистическими методами, относится к эконометрике.

Традиционная математическая статистика долгое время являлась основной методологией анализа данных в экономической и других предметных областях. Однако базовая концепция усреднения по выборке часто приводит к операциям над фиктивными величинами. В экономике средние значения ряда показателей по различным предприятиям иногда создают искаженное представление об отсталости или, наоборот, о незаурядных успехах ряда предприятий, отраслей или регионов – сглаживают их.

По этой причине появился ряд методик, которые относят к специфическим для Data Mining-а. Эти методики позволяют избежать таких ситуаций. В таблице приведены примеры постановок задач для методик, основанных на математической статистике и специфических методах, см. Табл. 8.:

Таблица №8. Специфика технологий =DataMiningиOLAP

OLAP	Data Mining
Каковы средние показатели рентабельности предприятий в регионе?	Какова характерная совокупность значений показателей финансово-хозяйственной деятельности предприятий в регионе?
Каковы средние размеры счетов клиентов банка – физических лиц?	Каков типичный портрет клиента – физического лица, отказывающегося от услуг банка?
Какова средняя величина ежедневных покупок по украденной или фальшивой кредитной карточке?	Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?

Выше показано, что работа по интеллектуальной обработке данных происходит в сфере закономерностей.

Основными задачами интеллектуального анализа являются:

– выявление взаимозависимостей, причинно-следственных связей, ассоциаций и аналогий, определение значений факторов времени, локализация событий или явлений по месту;

– классификация событий и ситуаций, определение профилей различных факторов;

– прогнозирование хода процессов, событий.

Главной задачей здесь является выявление закономерностей в исследуемых процессах, взаимосвязей и взаимовлияния различных факторов, поиск крупных «непривычных» отклонений, прогноз хода различных процессов в области мягких и глубинных знаний.

Одновременно с этим многомерный статистический анализ твердо удерживает свои позиции в жесткой области знаний. Он делится на: факторный, дисперсионный, регрессионный, корреляционный, кластерный анализ (является также сферой интересов Data Mining-а).

Эти методы позволяют решать многочисленные задачи в области экономики, менеджмента, юриспруденции, которые являются составной частью аналитической подготовки принятия решений.

Специфические методы и области применения Data Mining-а

Помимо перечисленных выше методов многомерного статистического анализа, ставших традиционными, все более широкое применение находят специфические методы интеллектуального анализа, происходящие из смежных областей информационных технологий (IT-систем) и достижений различных областей науки.

К специфическим методам интеллектуального анализа относятся:

– методы нечеткой логики;

– системы рассуждений на основе аналогичных случаев;

– классификационные и регрессионные деревья решений;

– нейронные сети;

– генетические алгоритмы;

– байесовское обучение (ассоциации);

– кластеризация и классификация;

– эволюционное программирование;

– алгоритмы ограниченного перебора.

Методы нечеткой логики используются для описания плохо формализуемых объектов из состава «мягких» знаний. Над ними также совершаются мягкие вычисления. Используется понятие «лингвистическая переменная», значения которой определяются через нечеткие множества, а они представляются базовым набором значений или базовой числовой шкалой.

Системы рассуждений на основе аналогичных случаев CaseBasedReasoning (CBR) основаны на том, что принятие решения осуществляется по прецеденту, наиболее подходящему к данной ситуации с учетом определенных корректив. Иногда решение принимается на основе учета всех примеров, находящихся в хранилище данных.

Деревья решений основаны на иерархической древовидной структуре классифицирующих правил. Решения об отнесении того или иного объекта или ситуации к соответствующему классу принимается по ответам на вопросы, стоящие в узлах дерева. Положительный ответ означает переход к правому узлу следующего уровня, отрицательный – к левому узлу. Процесс разделения продолжается до полного ответа на все поставленные вопросы.

Нейронные сети – это упрощенная аналогия нервной системы живого организма. Разработаны модели нейронных сетей. Распространенной моделью является многослойный персептрон с обратным распространением ошибки. Нейроны работают в составе иерархической сети, в которой нейроны нижележащего слоя своими выходами соединены с входами нейронов вышележащего слоя. На нейроны нижнего слоя подаются значения входных параметров, которые являются сигналами, которые передаются в следующий слой. При этом они ослабляются или усиливаются в зависимости от числовых значений, которые придаются межнейронным связям, называемых весами. На выходе нейрона верхнего слоя вырабатывается сигнал, являющийся

ответом сети на введенные значения входных параметров. Для получения необходимых значений весов сеть необходимо «тренировать» на примерах с известными значениями входных параметров и правильных ответов на них. Подбираются такие веса, которые обеспечивают наибольшую близость ответов нейронной сети к правильным.

Генетические алгоритмы представляют собой поисковый метод, используемый для нахождения наилучшего решения или совокупности решений. Он основан на идее естественного отбора. Начинается построение генетических алгоритмов с кодировки исходных логических закономерностей, называемых как и в биологии хромосомами. Набор таких кодов называют популяцией хромосом. Далее применяется функция пригодности, которая выделяет наиболее подходящие элементы для дальнейших операций. Это может быть отбор в какие-либо группы, но возможен и вариант применения скрещивания и мутации с целью получения «нового» поколения. Алгоритм работает над изменением старой популяции до тех пор, пока новая не будет отвечать заданным требованиям.

Байесовское обучение или ассоциации применяются в тех случаях, когда сложилась ситуация увязки между собой некоторых событий. Например, заселение новостроек сопровождается приобретением мебели и других предметов домашнего обихода. Необходимо выявить количественные характеристики этой связи.

Кластеризация и классификация. Слово кластеризация происходит от английского cluster – пучок, сгусток. Кластеризация предусматривает разделение совокупности схожих объектов на группы – кластеры по наибольшей близости их признаков. Проблема состоит в том, что оценка производится не по одному какому либо признаку, а одновременно по их совокупности. Разработаны алгоритмы кластеризации, которые пересчитывают значения признаков в некоторую величину, характеризующую «расстояние» между объектами рассматриваемой совокупности и объединяют близкие объекты в кластеры. Классификация отличается тем, что выявляются признаки, объеди-

няющие объекты, которые уже состоят в группах. Этими методами занимается также и эконометрика.

Эволюционное программирование. В этой методике предположения о виде аппроксимирующей функции строятся в виде программ на внутреннем языке программирования. Процесс построения программ выглядит как эволюция в среде программ. После нахождения в этой среде подходящей программы система начинает вносить в нее необходимые корректировки. Эта методика реализована российской системой Polyanalyst. Специальный модуль этой системы переводит найденные зависимости на доступный язык формул, таблиц.

Алгоритмы ограниченного перебора. Они вычисляют частоты комбинаций простых логических событий в группах данных. На основании оценки полученных частот делается заключение о полезности комбинаций для обнаружения ассоциаций в данных, прогнозирования и других целей.

Эти методы стали весьма широко и эффективно применяться в связи с бурным развитием в последнее десятилетие XX века самих методик и соответствующих инструментальных средств. Они находят применение в тех ситуациях, когда обычные методы анализа трудно или невозможно применить из-за отсутствия сведений о характере или закономерностях исследуемых процессов, взаимозависимостях явлений, фактов, поведении объектов и систем из различных предметных областей, в том числе в социальной и экономической.

Области применения методов Data Mining

С помощью этих методов при отсутствии априорной информации об объектах и их поведении и значительной ее неполноте решаются следующие задачи:

– выделение в данных групп, сходных по некоторым признакам записей;

– нахождение и аппроксимация зависимостей, связывающих анализируемые параметры или события;

– поиск наиболее значимых параметров в данной проблеме (задаче);

– выявление данных, характеризующих значительные или существенные отклонения от сложившихся ранее закономерностей (анализ отклонений);

– прогнозирование развития объектов, систем, процессов на основе хранящейся ретроспективной информации или с использованием принципов обучения на известных примерах и другие задачи.

Решение перечисленных задач может осуществляться каким-либо из перечисленных выше методов или комплексно для получения наиболее адекватного решения.

Средствами ИАС обеспечивается также оценка полученных результатов анализа и моделирования, в том числе оценка точности и устойчивости результатов, верификация моделей на тестовых наборах данных.