УДК 378.147:004.427

ПРОГНОЗИРОВАНИЕ СИСТЕМ И АНАЛИЗ ДАННЫХ В СРЕДЕ DATA MINING

ЦИЦИНА А.С.
ст.преп., магистр каф. «ИВС»,
ФАБЕР Е.Н.
ст.преп., магистр каф. «ИВС», Карагандинский экономический университет Казпотребсоюза

Сегодня наша жизнь практически немыслима без компьютера, интернета и других информационных технологий, которые с каждым днем становятся все более дружественными и удобными благодаря внедрению в них новейших технологических инноваций, в частности, элементов искусственного интеллекта. Вступление человечества в информационный век связано, прежде всего, с колоссальными изменениями в сфере информационной деятельности. В современном мире появляются новые инструменты анализа данных - компьютерные программы, реализующие не только статистические алгоритмы анализа данных, но и методы, основанные на недавних достижениях в области математики: нейронных сетях, генетических алгоритмах, теории нечетких множеств, синергетике, фрактальной математике, теории игр и др. Интеллектуализация, став императивом развития современных средств коммуникации, поиска информации, вычислений, обработки и анализа данных, значительно повышает доступность информационных технологий для пользователей, имеющих разные уровни компьютерной подготовки.

Цель интеллектуального анализатора решений - это определение верного предложенного решения, или нет; нахождение того, что конкретно неправильно или неполно в ответе; и, возможно, определение какие недостающие или неправильные знания могут быть ответственны за ошибку. Интеллектуальные анализаторы могут предоставлять далеко идущую обратную связь и обновлять модель. Интеллектуальный анализ имеет дело с конечными ответами на задачи.

В наше стремительно развивающееся время информационные технологии занимают значимое место. Любая технология является ключевым звеном в любой предметной области. Отличительной особенностью технологии методов интеллектуального анализа данных (ИАД) является то, что она является инструментом для специалистов, работающих в любой предметной области.

Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

В основу современных методов технологии Data Mining (discovery-driven data mining) положена концепция шаблонов, отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам (классам) данных, которые могут быть компактно выражены в понятной человеку форме.

Важным достоинством технологии Data Mining является нетривиальность разыскиваемых шаблонов, т.е. они должны отражать неочевидные, неожиданные регулярности в данных, составляющие, так называемые скрытые знания (hidden knowledge).

Существующие системы Data Mining [6] дорогостоящие и не ориентированы на решение задач принятия решений. Самыми известными являются See5^5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада). Известные продукты Data Mining приведены на рисунке 1.

Рисунок 1 - Известные продукты Data Mining

К методам и алгоритмам Data Mining относятся следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k- медианы; методы поиска ассоциативных правил, в том числе алгоритм Apriori; метод ограниченного перебора, эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов.

Метод (method) представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера.

Для автоматизированного извлечения знаний использовался метод CART (classification and regression trees) из класса методов деревьев решений. Данный подход является самым распространенным в настоящее время способом выявления, структурирования и графического представления логических закономерностей в данных. Его преимущества заключаются в следующем [16]:

- быстрый процесс обнаружения знаний;
- генерация правил в предметных областях, в которых трудно формализуются знания;

– извлечение правил на естественном языке;

– создание интуитивно понятной классификационной модели предметной области;

– прогноз с высокой точностью, сопоставимой с другими методами (статистическими и нейросетевыми);

– построение непараметрических моделей.

Хорошая эволюция и достигнутый уровень формализации методов послужили основанием использовать процедуру CART, как лучший из этого класса, в блоке извлечения знаний. В данном алгоритме можно выделить три операции, от реализации которых зависит его трудоёмкость и качество обнаружения знаний: сортировка источника данных при формировании множества условий U для атрибутов числового типа, вычисление критерия Gini [16] при разбиении узлов бинарного дерева, перемещение в таблице значительных объёмов информации при делении узла.

Покажем вычислительные затраты при классификации одного узла дерева. Пусть узлу, для которого осуществляется классификация, соответствует M объектов (строк) сводной таблицы. Каждая строка таблицы рассматривается как один пример обучающей выборки. Параметром N обозначим количество атрибутов таблицы без учёта целевого атрибута. Предположим, что в базе данных содержатся только атрибуты категорийного типа, имеющие в среднем Ncp значений.

Для определения необходимости последующего деления узла потребуется M проверок. Рассмотрим случай, когда из узла порождаются узлы-потомки. В этом случае для каждого атрибута формируются 2Ncp-1-1 возможных условий ui принадлежит U (|U|=2Ncp-1-1) (4), которые определяют варианты разбиения узла. Эта операция реализуется M проверками. Отбор наилучшего варианта разбиения узла дерева проводится по наибольшей классифицирующей силе, вычисляемой по критерию Gini, показан по формуле (1).

$$GINI = \frac{1}{|L|} \times \sum_{i=1}^{Ncc} l_i^2 + \frac{1}{|R|} \times \sum_{i=1}^{Ncp} r_i^2 \qquad (1)$$

Из формулы (1) видно, что её вычислительная сложность состоит из суммы следующих операций: подсчёт элементов li, ri класса i (i=1..Ncp) в множествах L и R и вычисление индекса Gini. Подсчёт объектов каждого класса занимает M операций, а вычисление индекса Gini выполняется за 2 × Ncp + 2 операций. Следовательно, классификация узла по условию ui и отбор наилучшего разбиения занимает в целом 2M + 2Ncp операций. Тогда для каждого категорийного атрибута потребуется (2M + 2Ncp) × (2Ncp-1-1) операций. А так как таблица имеет N атрибутов, то классификация одного узла без учёта разделения будет занимать (2M + 2Ncp) × (2Ncp – 1 - 1) × N + M условных операций. На примере таблицы, содержащей 1000 строк, 10 категорийных атрибутов с 5 возможными значениями, разбиение корневого узла дерева потребует приблизительно 300 000 условных операций, что значительно меньше полного перебора.

В большинстве случаев эти требования противоречат друг другу: чем точнее и быстрее работает алгоритм, тем он сложнее в вычислительном плане и более требователен к ресурсам компьютера. Поэтому каждый раз при выборе алгоритма нужно оценить все его преимущества и недостатки, «примерить» их к особенностям решаемой задачи. Так, если известно, что исходные данные низкого качества (то есть содержат аномалии и шумы, являются неполными и противоречивыми и т. д.), то предпочтение следует отдать более устойчивым методам, возможно, в ущерб скорости и точности; если объем исходных данных очень велик, то на первое место выходит производительность и т. д.

При выборе наилучшего алгоритма для решения конкретной задачи в первую очередь учитывается вычислительная сложность. Это вполне оправданно, поскольку большинство задач в бизнес-аналитике имеет дело с большими объемами данных и важно знать, как поведет себя алгоритм в условиях возрастания объемов обрабатываемых данных, какие ресурсы (объем памяти, дисковое пространство и т. д.) при этом потребуются.

Как основной критерий эффективности алгоритма используется трудоемкость — количество элементарных операций, которые необходимо выполнить для решения задачи с помощью данного алгоритма. При анализе трудоемкости рассматривается функция трудоемкости — отношение, связывающее объем входных данных алгоритма с количеством элементарных операций, которое требуется для их обработки.

Трудоемкость алгоритма может по-разному зависеть от входных данных. Трудоемкость одних алгоритмов зависит от количества входных данных, других — от значений. В некоторых случаях на трудоемкость может повлиять и порядок поступления данных.

Одним из наиболее простых видов анализа, используемых при сравнении трудоемкости алгоритмов, является асимптотический. Используемая в асимптотическом анализе оценка функции трудоемкости, называемая сложностью алгоритма, позволяет оценить, насколько быстро растет трудоемкость алгоритма с увеличением объема входных данных. Обычно эта оценка представляется в виде 0(f{N)), где f(N) — функция сложности, а N — число обрабатываемых наблюдений или примеров.

Можно выделить следующие функции сложности, которые позволяют оценить ожидаемые вычислительные затраты и требуемые ресурсы при реализации того или иного алгоритма. Наименее затратными являются алгоритмы, для которых функция сложности имеет вид f(N) = C и f(N) = CN, где C — константа. В первом случае вычислительные затраты не зависят от количества обрабатываемых данных, а во

втором — линейно возрастают с их увеличением. Встречаются функции сложности с логарифмической или экспоненциальной зависимостью типа f(N) = log(N) или f(N) = Л : 0, где С — константа. Самыми затратными являются алгоритмы, сложность которых имеет степенную зависимость от числа обрабатываемых наблюдений f(N) = СЛ и факториальную зависимость f(N) = N.

Условия функционирования рыночной экономики делают невозможным эффективное управление бизнесом без прогнозирования. От того, насколько прогноз будет точным и своевременным, а также от его соответствия поставленным задачам будет зависеть успех деятельности предприятия.

Прогнозирование — очень широкое понятие. Как отмечалось ранее, в большинстве случаев оно связывается с предвидением во времени, с предсказанием дальнейшего развития событий. В таком контексте прогнозирование понимается и в системах бизнес-аналитики. Поэтому далее при изложении аспектов прогнозирования речь будет идти именно о прогнозировании временных рядов; с этих позиций будут рассматриваться основные модели и методы прогнозирования [7,с.23].

Список литературы:

1. Послание народу Казахстана Президента РК Н.А. Назарбаева "Рост благосостояния граждан Казахстана - главная цель государственной политики" Февраль 2014.// Официальный сайт Президента РК www.akorda. kz
2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: ОЬАР и DataMining. СПб.: БХВ-Петрбург, 2006-336с.
3. Вячеслав Дюк, Санкт-Петербургский институт информатики и автоматизации РАН «DataMining - состояние проблемы, новые решения» http://inftech.webservis.ru/it/database/datamining/ar1.html.
4. Knowledge DiscoveryThroughData Mining: WhatIsKnowledge Discovery? - Tandem ComputersInc., 2006.
5. Дюк В.А. Data Mining - интеллектуальный анализ данных, 2010 http://www.iteam.ru/publications/it/section_92/article_1448/.
6. Дюк В.А., Самойленко А.П. DataMining: учебный курс. - СПб.: Питер, 2011.
7. Han J., Kamber M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2010.