ТЕХНИЧЕСКИЕ НАУКИ

Лутченко Игорь Аркадьевич

студент

Казиахмедов Туфик Багаутдинович

канд. пед. наук, доцент

ФГБОУ ВПО «Нижневартовский государственный университет»

г. Нижневартовск, ХМАО – Югра

АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА В ИНФОРМАЦИОННЫХ БАЗАХ ДАННЫХ

Аннотация: *в статье описываются основные существующие алгоритмы при построении моделей интеллектуального анализа данных в современных СУБД, их назначение и направление развития.*

Ключевые слова: *алгоритм, СУБД, интеллектуальный анализ.*

С увеличением количества знаний, накопленных человеком за время своего существования и развития, а также со вступлением человечества в период информационной эры, выросла и необходимость хранения знаний в систематизированном и структурированном виде. Появление баз знаний, или, или, иначе, баз данных, было неотъемлемой частью развития.

Но немаловажным оставалась не только возможность хранения больших объемов данных, но и проблема оперативного использования информации, то есть получения выборки части информации, необходимой в настоящий момент. Традиционная математическая статистика, служившая долгое время эталоном и основным инструментом для анализа данных, перестала справляться с возросшим количеством разнородной информации.

Именно поэтому разработчики систем управления базами данных особое внимание стали уделять интеллектуальному анализу.

В основу современных модулей интеллектуального анализа, встроенных в основные системы управления базами данных положены несколько алгоритмов:

– деревья решений (принятия решений). Представляет собой алгоритм регрессии и алгоритм классификации для использования как дискретных, так и непрерывных атрибутов. Для дискретных атрибутов алгоритм осуществляет прогнозирование на основе связи между входными данными. Осуществляет прогнозирование на основе корреляции входных и выходных данных в направлении конкретного результата. Для непрерывных атрибутов алгоритм использует линейную регрессию для определения места разбиения дерева решений. Модель строится путем создания ряда разбиений в дереве, когда вводные данные имеют значительную корреляцию с прогнозируемым результатом, используя при этом набор наиболее полезных атрибутов с целью предотвращения использования времени на менее значимые атрибуты;

– упрощенный алгоритм Байеса. Алгоритм классификации, основанный на теоремах Байеса, для использования в прогнозирующем моделировании. Не учитывает возможные зависимости данных. Алгоритм требует меньшего количества вычислений и применяется для быстрого формирования моделей обнаружения отношений между входными и прогнозируемыми столбцами. В основном, алгоритм применяется для первоначального исследования данных с целью применения результатов в других, более точных и требующих большего количества вычислений;

– нейронные сети. Алгоритм связывает все возможные состояния входного атрибута со всеми возможными состояниями прогнозируемого атрибута и использует эти данные для вычисления вероятностей для прогнозирования конечного значения прогнозируемого атрибута. Количество сетей в модели может быть разным и определяется количеством возможных состояний входных и прогнозируемых атрибутов. Алгоритм нейронной сети создает сеть, состоящую из двух или трех слоев нейронов. Такими слоями являются входной слой (значения входных атрибутов и их вероятности), необязательный скрытый слой (весовые коэффициенты, описывающие важность входного атрибута) и выходной слой

(значения прогнозируемых атрибутов). После обработки модели сеть и весовые коэффициенты можно использовать для составления прогнозов. Модель нейронной сети поддерживает регрессионный анализ, анализ взаимосвязей и классификационный анализ. Поэтому каждый прогноз может иметь различное значение;

– алгоритмы кластеризации. Алгоритм, использующий итерационные методы для группировки данных вариантов в наборы (кластеры) с одинаковыми характеристиками. Такая группировка используется для просмотра данных, выявления в них аномалий и создания прогнозов. Модели данного алгоритма определяют связи в наборе данных, которые невозможно получить случайным наблюдением. Алгоритм кластеризации создает модель строго на основе связей, существующих в данных и на основе полученных алгоритмом кластеров. Отличается от прочих алгоритмов необходимостью назначать прогнозируемый столбец, необходимый для создания модели кластеризации;

– временных рядов. Алгоритм обеспечивает алгоритмы регрессии для прогноза непрерывных значений во времени. С помощью модели временных рядов можно прогнозировать тенденции на основе только исходного набора данных. При этом итоговая модель одного ряда может строиться на основе изменения другого ряда, частично связанного с первым, формируя прогноз на основании временных срезов второго. Однако модель временных рядов должна всегда использовать дату, время или другое уникальное числовое значение как набор вариантов. Таким образом, для использования данного алгоритма необходимо иметь как минимум непрерывные значения времени или дат, уникальных для этого ряда;

– алгоритмы взаимосвязей. Алгоритм полезен для механизмов выработки рекомендаций. Модели взаимосвязей построены на наборах данных, содержащих идентификаторы для отдельных вариантов и элементов этих вариантов. Модель взаимосвязей состоит из рядов наборов элементов и правил, описывающих, как эти элементы группируются в вариантах. Для описания набора элементов и формируемых ими правил алгоритм использует два параметра: поддержка и вероятность. Алгоритм прослеживает набор данных для поиска элементов, которые находятся в варианте совместно, и группирует в наборы элементов любые

связанные элементы, после чего формируются правила из наборов элементов. Правила используются для прогнозирования наличия элемента в базе данных на основе наличия других определенных элементов, которые алгоритм определяет как значимые;

– алгоритмы регрессий. Алгоритм является разновидностью алгоритма дерева принятия решений для расчета связи между зависимой и независимой переменной и последующего использования связи при прогнозировании. Существуют разные типы регрессии, в которых используется несколько переменных, а также нелинейные методы регрессии. Линейная регрессия является наиболее известным методом моделирования ответа на изменение в каком-либо базовом факторе. Однако линейная регрессия может чрезмерно упростить связи в сценариях, в которых на результат влияют несколько факторов. При этом в модели линейной регрессии для вычисления связей при начальном проходе используется весь набор данных, тогда как в стандартной модели дерева решения данные многократно разбиваются на более малые подмножества или деревья.

Не существует универсального алгоритма интеллектуального анализа. Можно говорить лишь о наиболее подходящих группах алгоритмов, применяемых в зависимости от поставленной задачи. Причем часто одного алгоритма недостаточно для получения исчерпывающей информации. Для получения такой выборки необходимо использовать несколько алгоритмов в некоторой последовательности. Поэтому одним из направлений работы над системами управления базами данных является проектирование системы, которая определит последовательность выполняемых алгоритмов в рамках поставленной задачи.

Список литературы

1. Чубукова И.А. Data Mining: учебное пособие. – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

2. Алгоритмы интеллектуального анализа данных [Электронный ресурс]. – Режим доступа: https://msdn.microsoft.com/ru-ru/library/ms175595.aspx (дата обращения: 22.04.2015).