УДК 004.89

ОСОБЕННОСТИ ТЕХНОЛОГИИ *DATA MINING*

Д. В. Нечипорук

Донской государственный технический университет, Ростов-на-Дону, Российская Федерация

Rabbit.93@inbox.ru

Исследуются и анализируются особенности технологии *Data Mining*. Подробно описываются задачи, которые решает данная технология, перечисляются методы решения этих задач. Уделено отдельное внимание нечёткой логике, генетическим алгоритмам и нейронным сетям, продемонстрирован процесс решения задачи методами *Data Mining*.

Ключевые слова: *Data Mining*, анализ данных, нейронные сети, генетические алгоритмы, статистические методы.

UDC 004.89

DATA MINING TECHNOLOGY FEATURES

D. V. Nechiporuk

Don State Technical University, Rostov-on-Don, Russian Federation

Rabbit.93@inbox.ru

The article examines and analyzes the features of Data Mining technology. It describes in detail the tasks that this technology solves and methods for solving problems. Special attention in the article is paid to fuzzy logic, genetic algorithms and neural networks; it shows the process of solving the problem using Data Mining methods.

Keywords: Data Mining, data analysis, neural networks, genetic algorithms, statistical methods

Введение. В буквальном переводе с английского языка *Data Mining* — это «добыча данных». Но по существу, термин *Data Mining* означает сбор, анализ и изучение данных или информации. В 1989 году Григорий Пятецкий-Шапиро провёл семинар, на котором была поставлена такая задача: имеется очень крупная база данных, в которой находятся так называемые «скрытые знания», нужно разработать методы обнаружения этих скрытых знаний в большом объеме исходных данных. Эта задача и послужила появлению термина *Data Mining*, который использовался для обозначения методов обнаружения данных [1].

Основная часть. *Data mining* включает в себя следующее:

- методы прогнозирования, моделирования и классификации на основе применения искусственных нейронных сетей, деревьев решений, ассоциативной памяти, эволюционного программирования, нечёткой логики и генетических алгоритмов;
- статистические методы, такие как факторный анализ, дисперсионный анализ, дескриптивный анализ, корреляционный анализ, регрессионный анализ, анализ временных рядов, компонентный анализ.

Итак, суть технологии *Data Mining* состоит в поиске неочевидных, но полезных на практике закономерностей, которые скрыты в больших объемах данных [2]. В основе *Data Mining* лежит концепция шаблонов, которые представляют собой определённые закономерности. Им соответствуют задачи, которые решает технология *Data Mining* — классификация, прогнозирование, ас-

социация, кластеризация, обнаружение и анализ отклонений, анализ связей, оценка, визуализация и подведение итогов.

Все вышеуказанные задачи делятся на описательные и предсказательные. Описательные задачи направлены на улучшение понимания анализируемых данных, ключевой момент таких моделей — простота восприятия человеком. Предсказательные задачи решаются в два этапа: сначала на основе данных с известными результатами строится модель, а затем она используется для предсказания результатов на основании новых наборов данных.

Для решения каждой из вышеописанных задач существуют свои методы решения, например, для решения задачи классификации применяются байесовские сети, индукция деревьев решений, метод ближайшего соседа, а для решения задач прогнозирования применяются методы математической статистики и нейронные сети. Также существуют базовые методы решения задач *Data Mining*, которые используются в большом количестве задач.

К базовым методам *Data Mining* относят основанные на переборе алгоритмы. Простой перебор всех объектов требует $2*N$ операций, где N — это количество объектов. Очевидно, что при большом объеме данных применить простой перебор невозможно, поэтому для задач с высокой вычислительной сложности используют разного рода эвристики. Достоинства таких методов — это простота понимания и реализации, а недостатками является отсутствие формальной теории и сложности, связанные с исследованием и развитием.

Также к базовым методам *Data Mining* относятся элементы теории статистики. Это корреляционный, регрессионный, дисперсионный и другие виды статистического анализа данных. Основной недостаток таких методов — это усреднение значений, что приводит к потере информативности данных [3].

Методы решения разбиваются на две большие группы: «обучение с учителем» и «обучение без учителя». Методы «обучения с учителем» работают в два этапа. Строится модель анализируемых данных, называемая классификатор, затем классификатор подвергается обучению, то есть проверяется качество его работы, и, если оно неудовлетворительное, проводится дополнительное обучение классификатора. Так продолжается до тех пор, пока не будет достигнут требуемый уровень качества. «Обучение без учителя» включает задачи, которые выявляют описательные модели и закономерности.

В настоящее время в технологии *Data Mining* используются методы нечёткой логики, генетические алгоритмы и нейронные сети. Очень часто исходные данные по тем или иным причинам содержат неполную или неточную информацию. Недостоверность бывает физической, источником которой является внешняя среда, и лингвистической, которая возникает в результате словесного обобщения. Для обработки физических неопределенностей используются методы теории вероятностей и теория множеств, а для решения задач с лингвистической неопределенностью используются методы нечеткой логики. Нечеткая логика позволяет представить мышление человека. Человек не использует формальное моделирование на основе математических выражений, он пользуется нечетким естественным языком, и в процессе принятия решения он разделяет ситуацию на отдельные события, используя большое количество различных критериев. Именно таким образом работают методы нечеткой логики, позволяя оперировать со множеством частных правил вместо одного четкого обобщённого правила. Нейронные сети — это модели, которые основаны на биологической аналогии с человеческим мозгом. После прохождения этапа обучения на основе имеющихся данных они используются для решения различных задач анализа данных. В результа-

те обучения нейронная сеть находит закономерности в данных, однако в отличие от классических моделей, эти зависимости не могут быть записаны в явном виде. Нейронные сети могут выдавать прогноз очень высокого качества. Их главная особенность в том, что с их помощью можно аппроксимировать любую непрерывную функция [4].

Генетические алгоритмы можно отнести к числу универсальных методов решения задач различных типов. В области *Data Mining* — это поиск наиболее оптимальной модели и определение значимых параметров операционного базиса. Очень эффективным является интеграция генетических алгоритмов и нейронных сетей. Такой подход позволяет решить проблемы поиска оптимальных значений весов входов нейронов. Интеграция генетических алгоритмов и нечёткой логики дает более оптимизированную систему продукционных правил, которые используются для управления операторами генетических алгоритмов.

Рассмотрим процесс решения задачи методами *Data Mining*. Он включает в себя определенные этапы, которые проиллюстрированы на рис. 1. Процесс *Data Mining* может быть как успешным, так и неуспешным. Если процесс решения задачи был закончен неуспешно, возможно стоит попробовать решить задачу другими методами или изменить параметры модели.

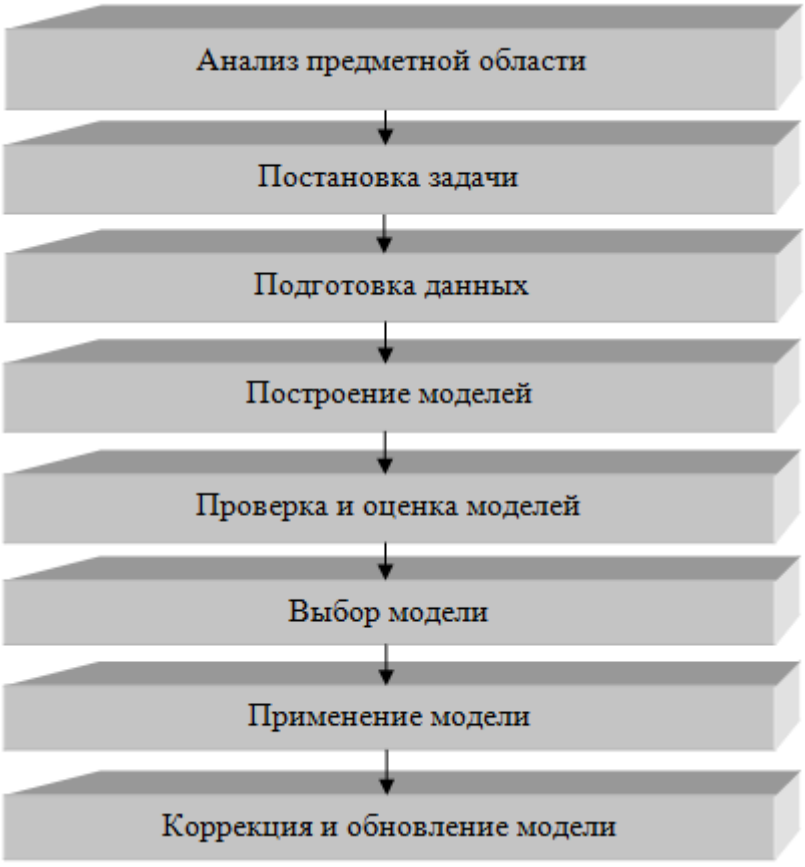

Рис. 1. Процесс *Data Mining*

Заключение. Итак, *Data Mining* — это процесс поддержки принятия решений, который основан на поиске в больших объемах данных скрытых закономерностей. У этой технологии есть свои достоинства и недостатки, но у нее, очевидно, есть перспективы развития. На данный момент самым серьезным недостатком является цена. Средства *Data Mining* относятся к очень дорогим программным инструментам и основные потребители — это крупные торговые предприятия, бан-

ки, страховые и финансовые компании. Однако, постепенная популяризация технологии должна привести к появлению более бюджетных программных средств, которыми сможет пользоваться каждый.

Библиографический список.

1. *Data mining* [Электронный ресурс] // Википедия. — Режим доступа : https: // ru.wikipedia.org/wiki/ Data_mining (дата обращения: 30.04.16).
2. Луньков, А. Д. Интеллектуальный анализ данных [Электронный ресурс] / А. Д. Луньков, А. В. Харламов // Саратовский национальный исследовательский университет. — Режим доступа : elibrary.sgu.ru›uch_lit/1141.pdf (дата обращения : 12.05.16).
3. Барсегян, А. А. Анализ данных и процессов / А. А. Барсегян. — Санкт-Петербург : БХВ-Петербург, 2009. — 512 с.
4. Искусственная нейронная сеть [Электронный ресурс] // Википедия. — Режим доступа : https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть (дата обращения: 04.05.16).