УДК 004.89

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Прокопец Татьяна Николаевна[1], к.э.н., доцент,
Синюк Татьяна Юрьевна[1], к.э.н., доцент,
Рыбалко Юлия Александровна[2], к.э.н., доцент
[1]Ростовский государственный экономический университет
[2]Полесский государственный университет
Tatyana Prokopetz[1], PhD, Rostov State University of Economics, hatani@mail.ru
Tatyana Sinyuk [1], PhD, Rostov State University of Economics, t_sinyuk@mail.ru
Yulia Rybalko [2], PhD, Polessky State University, rybalko.u@polessu.by

Аннотация. В статье представлены методы интеллектуального анализа данных, что позволило выявить их достоинства и недостатки с позиции более тщательного обзора литературных источников.

Ключевые слова: анализ, методы, данные, кластеризация, дерево принятия решений.

Интеллектуальный анализ данных с каждым годом становится все более актуальным направлением изучения во всех сферах человеческой деятельности: банковский, страховой, государственный сектора и другие.

Интеллектуальный анализ данных (ИАД или data mining) – это процесс обнаружения в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний (закономерностей), необходимых для принятия решений в различных сферах человеческой деятельности. Данное определение принято считать классическим, однако, на наш взгляд, оно содержит несколько неточностей:

– в определение этого понятия входят слова «анализ» и «знания», тогда как знания нужны для управления, т.е. достижения цели;

– для интерпретации доступны не только знания, но уже и информация;

– термин «ИАД» не подразумевает какого-либо одного метода анализа данных, но является собирательным и объединяет многие направления исследований и разработок.

Поэтому предлагается другое, более точное определение понятия. ИАД – это совокупность математических моделей, численных методов, программных средств и информационных технологий, обеспечивающих обнаружение в эмпирических данных доступной для интерпретации информации и синтез на основе этой информации ранее неизвестных, нетривиальных и практически полезных для достижения определенных целей знаний.

Исходя из сказанного видно, что целью интеллектуального анализа данных является поиск (обнаружение) в данных скрытых закономерностей (шаблонов информации). При этом накопленные сведения автоматически обобщаются до информации, которая может быть охарактеризована как знания.

В общем случае процесс ИАД состоит из трёх стадий:

1) выявление закономерностей (свободный поиск);

2) использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование);

3) анализ исключений, предназначенный для выявления и толкования аномалий в найденных закономерностях.

Иногда в явном виде выделяют промежуточную стадию проверки достоверности найденных закономерностей между их нахождением и использованием (стадия валидации).

Далее представлены основные **задачи** интеллектуального анализа данных.

1. Задача классификации заключается в том, что для каждого варианта определяется категория или класс, к которому он относится. Множество классов должно быть заранее известно и быть конечным и счетным.

2. Задача регрессии многим похожа на классификацию, особенностью является то, что в ходе ее решения производится поиск шаблонов для определения числового значения. В данном случае предсказываемый параметр – это число из непрерывного диапазона.

3. Задача прогнозирования новых значений на основании имеющихся значений числовой последовательности (или нескольких последовательностей, между значениями в которых наблюдается корреляция). При этом могут учитываться имеющиеся тенденции (тренды), сезонность, другие факторы.

4. Задача кластеризации заключается в делении множества объектов на кластеры схожих по параметрам. При этом, в отличие от классификации, число кластеров и их характеристики могут быть заранее неизвестны и определяться в ходе построения кластеров исходя из степени близости объединяемых объектов по совокупности параметров.

5. Задача определения взаимосвязей, также называемая задачей поиска ассоциативных правил, заключается в определении часто встречающихся наборов объектов среди множества подобных наборов [1].

Интенсивное применение интеллектуального анализа данных (ИАД) осуществляется благодаря наличию рабочих инструментов, реализующих разнообразные методы ИАД. По мнению некоторых экспертов, в ближайшее десятилетие интеллектуальный анализ данных и его ядро – Data Mining – станут наиболее перспективными направлениями разработки программного обеспечения.

Рассмотрим более подробно некоторые из методов интеллектуального анализа данных.

Дерево принятия решений. Дерево решений – это дерево, в котором каждой внутренней вершине поставлен в соответствие некоторый атрибут, каждая ветвь, выходящая из данной вершины, соответствует одному из возможных значений атрибута, а каждому листу дерева сопоставлен конкретный класс или набор вероятностей классов. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе. Дерево решений, как правило, лучше всего подходит для задач, в которых экземпляры представлены в виде пар атрибут-значение и целевая функция имеет дискретные значения.

Обычно деревья решений используются для реализации задач классификации.

Однако при использовании дерева решений может получится так, что не вся необходимая информация для построения модели может быть получена. Связано это с тем, что интерпретация результата зависит от качества имеющихся данных, т.е. поддеревья необходимо создавать, используя максимум возможной и накопленной информации.

Кластеризация. Задача кластеризации состоит в разбиении заданной выборки объектов (наблюдений) на подмножества (как правило, непересекающиеся), называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Данный метод применяется для моделирования и использует методы, которые могут обрабатывать реляционные данные.

Кластеризация данных включает в себя следующие этапы:

1. Выделение характеристик. Для начала необходимо выбрать свойства, которые характеризуют наши объекты, ими могут быть количественные характеристики (координаты, интервалы), качественные характеристики (цвет, статус и т.д.) Затем стоит попробовать уменьшить размерность пространства характеристических векторов, то есть выделить наиболее важные свойства объектов. Выделенные характеристики стоит нормализовать. Далее все объекты представляются в виде характеристических векторов. Однако при создании модели пользователя информация, используемая для создания кластеров, например, часто используемые пункты меню, особенности пользователя и т.д., не может быть представлена в виде численного вектора.

2. Определение метрики. Следующим этапом кластеризации является выбор метрики, по которой мы будем определять близость объектов.

Однако при использовании метода кластеризации возникают следующие проблемы:

1. каким образом определить понятие «расстояние»,

2. для использования алгоритма кластеризации необходимо заранее знать количество кластеров.

Относительно первой проблемы необходимо дать определение понятию расстояние. Однако при моделировании пользователя сделать это сложно из-за данных, необходимых для построения модели в рамках обзора литературы – это показатели, которые не выражаются в виде числа.

При решении второй проблемы предположим, что число кластеров известно заранее. Однако при моделировании невозможно заранее предсказать количество кластеров, которое будет использовано. Это означает, что необходимо разработать методику определения числа кластеров в процессе создания модели.

Таким образом, кластеризация может быть использована для создания групп литературных источников, имеющих одинаковые характеристики.

Нейронные сети. Ключевым элементом этой парадигмы является структура системы обработки информации. Она состоит из большого числа тесно взаимосвязанных элементов обработки – нейронов, работающих параллельно

Структуру нейронной сети можно представить следующим образом:

– Множество простых процессоров – нейронов.

– Структура связей – отражает детали конструкции сети, а именно то, какие элементы соединены и в каком направлении работают соединения, каков уровень значимости (т.е. вес) каждого соединения.

– Правило вычисления сигнала активности, позволяющий вычислить выходной сигнал по совокупности входных сигналов.

– Правило обучения, корректирующее связи [2].

В отличие от кластерного анализа нейронные сети не требуют наличие какого-либо показателя, что делает в свою очередь их полностью независимыми от приложения.

Нечеткая логика Нечеткая логика применяется для обработки данных с размытыми значениями истинности, которые могут быть представлены разнообразными лингвистическими переменными. Нечеткое представление знаний широко применяется в системах с логическими выводами (дедуктивными, индуктивными, абдуктивными) для решения задач классификации и прогнозирования.

Метод k-ближайших соседей (k-nearest neighbors) – это метод решения задач классификации и задач регрессии, основанный на поиске ближайших объектов с известными значения целевой переменной [3]. Он относит объекты к классу, которому принадлежит большинство из k его ближайших соседей в многомерном пространстве признаков. Число k – это количество соседних объектов в пространстве признаков, которые сравниваются с классифицируемым объектом.

Преимущества: алгоритм прост и легко реализуем; нет необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения; алгоритм универсален, его можно использовать для обоих типов задач: классификации и регрессии.

Недостатки: алгоритм работает значительно медленнее при увеличении объема выборки, предикторов или независимых переменных; из аргумента выше следуют большие вычислительные затраты во время выполнения; всегда нужно определять оптимальное значение k.

Каждый из перечисленных методов имеет свои сильные и слабые стороны, представляет информацию по-разному, разный по сложности и способам представления входных данных.

Таблица – Сравнение методов интеллектуального анализа данных

Методы	Сложность	Точность	Размер обучающих данных	Интерпретация
Дерево принятия решений	высокая	низкая	средний	низкая
Кластеризация	высокая/средняя	средняя	средний/ большой	низкая
Нейронные сети	высокая	высокая	большой	низкая
Нечеткая логика	средняя	низкая	не доступно	высокая
Метод k-ближайших соседей	высокая	очень низкая	большой	высокая/ нейтральная

Примечание – Таблица составлена автором по результатам собственных исследований.

Немаловажно, что методы интеллектуального анализа данных характеризуются определенными свойствами, которые могут быть определяющими при выборе одного их них. Можно сравнивать ИАД между собой, оценивая характеристики их свойств. Основные свойства и характеристики

методов интеллектуального анализа данных: точность, масштабируемость, интерпретируемость, проверяемость, трудоемкость, гибкость, быстрота и популярность.

В данной связи нами проведена сравнительная оценка методов интеллектуального анализа данных, которая позволит определить какой из них стоит применять к обзору литературных источников (табл.).

Исходя из проведенного исследования, можно сделать вывод о том, что каждый из перечисленных методов интеллектуального анализа данных имеет свои достоинства и недостатки. Оценка методов интеллектуального анализа в различных областях исследования позволила выявить наиболее подходящий в рамках обзора литературы. Это метод иерархической кластеризации, позволяющей разделять данные на разные группы на основе некоторых мер сходства.

Список использованных источников

1. Дядичев, В.В. Задачи и методы интеллектуального анализа данных / В.В. Дядичев, Е.В. Ромашка, Т.В. Голуб // Геополитика и экогеодинамика регионов. – Симферополь, 2015. – Т. 1 (11). – № 3. – С. 23-29.

2. Амаева, Л.А. Использование методов интеллектуального анализа данных для моделирования пользователя / Л.А. Амаева // Вестник Технологического университета. – Казань, 2015. – Т. 18. – № 1. – С. 320-322.

3. Метод ближайших соседей (kNN) [Электронный ресурс] // Fandom. – Режим доступа: https://learnmachinelearning.fandom.com/ru/wiki/Метод_ближайших_соседей_(kNN). – Дата доступа 15.03.2022.