Геополитика и экогеодинамика регионов.
Том 1 (11). Вып. 3. 2015 г. С. 23–29.

УДК 004.82

В. В. Дядичев [1]
Е. В. Ромашка [2]
Т. В. Голуб [3]

Задачи и методы интеллектуального анализа данных

[1]ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского», г. Симферополь
e-mail: mr.dyadichev@mail.ru
[2]ГОУ ВПО «Луганский государственный университет имени Владимира Даля», г. Луганск
e-mail: RomashkaLena171@gmail.com
[3]ГОУ ВПО «Луганский государственный университет имени Владимира Даля», г. Луганск
e-mail: Romashka@nxt.ru

Аннотация. *Необходимость систематизации знаний и прогнозирования событий, исходя из имеющихся данных, привела к созданию интеллектуального анализа данных. В статье выявлены задачи интеллектуального анализа данных. Приведены примеры его применения в различных сферах деятельности. Рассмотрены методы интеллектуального анализа данных. А также проведена сравнительная характеристика интеллектуального анализа данных. Сделаны выводы о преимуществах и недостатках различных методов интеллектуального анализа данных.*

Ключевые слова: *интеллектуальный анализ данных, методы, дерево решений, прогнозирование.*

Введение

Объемы информации, которыми приходится оперировать человеку, растут с огромной скоростью. В связи с этим, возникает необходимость систематизации данных. Однако и этого недостаточно. Необходимо автоматизировать анализ этих данных и прогнозировать возможные ситуации исходя из полученных данных. Возникает необходимость использования интеллектуального анализа данных.

Интеллектуальный анализ данных — это процесс обнаружения в сырых данных (row data) раннее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний (закономерностей), необходимых для принятия решений в различных сферах человеческой деятельности.

Материалы и методы

Целью интеллектуального анализа данных является обнаружение неявных закономерностей в наборах данных.

Задачи интеллектуального анализа данных

При проведении интеллектуального анализа данных происходит исследование множества вариантов. В большинстве случаев его можно представить в виде таблицы, каждая строка которой соответствует одному из вариантов, а в столбцах содержатся значения параметров, его характеризующих.

Зависимая переменная – параметр, значение которого рассматриваем как зависящее от других параметров (независимых переменных). Собственно эту зависимость и необходимо определить, используя методы интеллектуального анализа данных.

Ниже приведены основные задачи интеллектуального анализа данных.

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.

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.

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.

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.

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

Анализ последовательностей или сиквенциальный анализ иногда приводится как вариация предыдущей задачи, иногда выделяется отдельно. Целью, в данном случае, является обнаружение закономерностей в последовательностях событий. Подобная информация позволяет, например, предупредить сбой в работе информационной системы, получив сигнал о наступлении события, часто предшествующего сбою подобного типа. Другой пример применения – анализ последовательности переходов по страницам пользователей web-сайтов.

Анализ отклонений позволяет отыскать среди множества событий те, которые существенно отличаются от нормы. Отклонение может сигнализировать о каком-то необычном событии или, например, об ошибке ввода данных оператором [1].

В таблице 1 приведены примеры задач интеллектуального анализа данных из различных областей.

Основу методов интеллектуального анализа данных составляют всевозможные методы классификации, моделирования и прогнозирования. К методам интеллектуального анализа данных нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов). Такие методы, однако, предполагают некоторые априорные представления об анализируемых данных, что несколько расходится с целями интеллектуального анализа данных

(обнаружение ранее неизвестных нетривиальных и практически полезных знаний).

Таблица 1.

Примеры применения интеллектуального анализа данных

	Информационные технологии	Торговля	Финансовая сфера
Классификация			Оценка кредитоспособности
Регрессия			Оценка допустимого кредитного лимита
Прогнозирование		Прогнозирование продаж	Прогнозирование цен акции
Кластеризации		Сегментация клиентов	Сегментация клиентов
Определения взаимосвязей		Анализ потребительской корзины	
Анализ последователь-ностей	Анализ переходов по страницам web-сайта		
Анализ отклонений	Обнаружение вторжений в информационные системы		Выявление мошенничества с банковскими картами

Одно из важнейших назначений методов интеллектуального анализа данных состоит в наглядном представлении результатов вычислений, что позволяет использовать инструментарий интеллектуального анализа данных людьми, не имеющими специальной математической подготовки. В то же время, применение статистических методов анализа данных требует хорошего владения теорией вероятностей и математической статистикой [2].

Существует множество методов и алгоритмов интеллектуального анализа данных, таких как:

- искусственные нейронные сети
- дерево решений
- метод 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.

Основная часть аналитических методов, которые используются в технологии интеллектуального анализа данных – это известные математические методы и алгоритмы. Однако теперь появилась возможность их использования при решении тех или иных конкретных проблем, появившаяся благодаря новым возможностям технических и программных средств. Большинство методов интеллектуального анализа данных были разработаны в рамках теории искусственного интеллекта.

Метод представляет собой норму или правило, определенный путь, способ, прием решений задачи теоретического, практического, познавательного, управленческого характера [3].

Свойства методов интеллектуального анализа данных

Различные методы интеллектуального анализа данных характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств.

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

Масштабируемость – свойство вычислительной системы, которое обеспечивает предсказуемый рост системных характеристик, например, быстроты реакции, общей производительности и прочего, при добавлении к ней вычислительных ресурсов [4, 5, 6, 7].

Результаты и обсуждение

В таблице 2 приведена сравнительная характеристика некоторых распространенных методов. Оценка каждой из характеристик проведена следующими категориями, в порядке возрастания: чрезвычайно низкая, очень низкая, низкая/нейтральная, нейтральная/низкая, нейтральная, нейтральная/высокая, высокая, очень высокая.

Таблица 2.
Сравнительная характеристика методов интеллектуального анализа данных

Алгоритм	Точность	Масштаби-руемость	Интерпрети-руемость	Пригодность к исполнению
1	2	3	4	5
Линейная регрессия	Нейтральная	Высокая	Высокая/ нейтральная	Высокая
Нейронные сети	Высокая	Низкая	Низкая	Низкая
Методы визуализации	Высокая	Очень низкая	Высокая	Высокая

Продолжение таблицы 2

1	2	3	4	5
Деревья решений	Низкая	Высокая	Высокая	Высокая/ нейтральная
k-ближайшего соседа	Низкая	Очень низкая	Высокая/ нейтральная	Нейтральная
Алгоритм	Трудоемкость	Разносторон-ность	Скорость	Популярность
Линейная регрессия	Нейтральная	Нейтральная	Высокая	Низкая
Нейронные сети	Нейтральная	Низкая	Очень низкая	Низкая
Методы визуализации	Очень высокая	Низкая	Чрезвычайно низкая	Высокая/ нейтральная
Деревья решений	Высокая	Высокая	Высокая/ нейтральная	Высокая/ нейтральная
k-ближайшего соседа	Нейтральная/ низкая	Низкая	Высокая	Низкая

Выводы

Исходя из предоставленных данных, можно сделать вывод о том, что каждый из методов имеет свои достоинства и недостатки. Но ни один метод, какой бы не была его оценка с точки зрения присущих ему характеристик, не может обеспечить решение всего спектра задач интеллектуального анализа данных.

Литература

1. Ian H. Witten Data Mining: Practical Machine Learning Tools and Techniques [Text] / Ian H. Witten, Eibe Frank and Mark A. Hall. ; 3rd Edition. - Morgan Kaufmann. 2011. - P. 664.
2. Барсегян А. М. Методы и модели анализа данных: OLAP и Data Mining [Текст] / А.М. Барсегян, М.И. Куприянов, В.Ф. Степаненко, И.Н. Холод. - СПБ: БХВ-Петербург, 2004. - 336 с.
3. Сегаран Тоби Программируем коллективный разум [Текст] / Тоби Сегаран. - СПБ: Символ – Плюс, 2012. - 368 с.
4. Паклин Н.Б. Бизнес-аналитика: от данных к знаниям [Текст] / Н.Б.Паклин, В.И. Орешков. - СПБ: Питер, 2013. - 704 с.
5. Ромашка Е.В. Защита информации с применением электронной цифровой подписи [Текст] / Е.В.Ромашка, Н.В. Любомирский, А.В. Дядичев // Сборник трудов I научной конференции профессорско-преподавательского состава, аспирантов, студентов и молодых ученых «Дни науки КФУ им. В.И. Вернадского», г. Симферополь, 28-30 окт.2015. - Симферополь, 2015. - С. 66-72.
6. Рыбцев И.В. Вопросы анализа угроз и безопасности данных в корпоративных компьютерных сетях [Текст] / И.В. Рыбцев, А.В. Дядичев, А.В. Колесников //

Сборник трудов I научной конференции профессорско-преподавательского состава, аспирантов, студентов и молодых ученых «Дни науки КФУ им. В.И. Вернадского», г. Симферополь, 28-30 окт. 2015. - Симферополь, 2015. - С.48-53.

7. Стоянченко С.С. Целесообразность использования систем для сбора и анализа информации полученной с WEB-ресурса [Текст] / С.С. Стоянченко, А.В. Колесников, А.В. Дядичев // Сборник трудов I научной конференции профессорско-преподавательского состава, аспирантов, студентов и молодых ученых «Дни науки КФУ им. В.И. Вернадского», г.Симферополь, 28-30 окт.2015. - Симферополь, 2015. - С.111-117.

V. V. Dyadichev[1],
E.V. Romashka[2]
T.V. Golub[3]

Objectives and methods of data mining

[1]FSAEI HE «V. I. Vernadsky Crimean Federal University», Simferopol
e-mail: mr.dyadichev@mail.ru
[2]SEI HPE " Vladimir Dal Lugansk State University", Lugansk
e-mail: RomashkaLena171@gmail.com
[3]SEI HPE " Vladimir Dal Lugansk State University", Lugansk
e-mail: Romashka@nxt.ru

Abstract. The necessity of knowledge systematize and prediction of events, based on the available data, has led to the creation of data mining. The problem of data mining were highlighted in the article. Examples of its application in various fields of activity were painted. The methods of data mining were described. As well as a comparative characterization of data mining were defined. The conclusions about the advantages and disadvantages of various methods of data mining were done.

Keywords: data mining, methods, decision tree, prediction.

References

1. Ian H. Witten Data Mining: Practical Machine Learning Tools and Techniques [Text] / Ian H. Witten, Eibe Frank and Mark A. Hall. ; 3rd Edition. - Morgan Kaufmann. 2011. - P. 664.
2. Barsegyan A. M. Metodyi i modeli analiza dannyih: OLAP i Data Mining [Tekst] / A.M. Barsegyan, M.I. Kupriyanov, V.F. Stepanenko, I.N. Holod. -SPB: BHV-Peterburg, 2004. - 336 s.
3. Segaran Tobi Programmiruem kollektivnyiy razum [Tekst] / Tobi Segaran. SPB: Simvol – Plyus, 2012. - 368 s.
4. Paklin N.B. Biznes-analitika: ot dannyih k znaniyam [Tekst] / N.B.Paklin, V.I. Oreshkov. - SPB: Piter, 2013. - 704 s.
5. Romashka E.V. Zaschita informatsii s primeneniem elektronnoy tsifrovoy podpisi [Tekst] / E.V.Romashka, N.V. Lyubomirskiy, A.V. Dyadichev // Sbornik trudov I nauchnoy konferentsii professorsko-prepodavatelskogo sostava, aspirantov,

studentov i molodyih uchenyih «Dni nauki KFU im. V.I. Vernadskogo», g. Simferopol, 28-30 okt.2015. - Simferopol, 2015. - S. 66-72.

6. Ryibtsev I.V. Voprosyi analiza ugroz i bezopasnosti dannyih v korporativnyih kompyuternyih setyah [Tekst] / I.V. Ryibtsev, A.V. Dyadichev, A.V. Kolesnikov // Sbornik trudov I nauchnoy konferentsii professorsko-prepodavatelskogo sostava, aspirantov, studentov i molodyih uchenyih «Dni nauki KFU im. V.I. Vernadskogo», g. Simferopol, 28-30 okt. 2015. - Simferopol, 2015. - S.48-53.
7. Stoyanchenko S.S. Tselesoobraznost ispolzovaniya sistem dlya sbora i analiza informatsii poluchennoy s WEB-resursa [Tekst] / S.S. Stoyanchenko, A.V. Kolesnikov, A.V. Dyadichev // Sbornik trudov I nauchnoy konferentsii professorsko-prepodavatelskogo sostava, aspirantov, studentov i molodyih uchenyih «Dni nauki KFU im. V.I. Vernadskogo», g.Simferopol, 28-30 okt.2015. - Simferopol, 2015. - S.111-117.

Поступила в редакцию 29.09.2015г.