УДК 62-5, 681.5.015, 004.942

Применение методов анализа данных для определения наиболее популярных функций приложения по журналу действий пользователя

А.А. Личман, О.Ю. Чередникова
Донецкий национальный технический университет
E-mail: anton.lichman@yandex.ru

Аннотация
Рассмотрены методы решения задачи определения наиболее часто используемых функций приложения. Предложено использовать для этой цели нейронные сети. Выполнен анализ существующих нейронных сетей и особенностей их применения для различных целей. Предложен и реализован метод определения наиболее часто используемых функций приложения на основе нейронной сети Кохонена и прикладного пакета Deductor для предобработки данных. Это позволит разработчикам программных продуктов модернизировать уже готовые версии под новые потребности.

Введение

Одним из начальных этапов разработки или модернизации программного обеспечения (ПО) должно быть определение его функциональных потребностей. Для этого обычно анализируют текущую версию ПО или работу приложений, выполняющих похожие задачи. Для производственных целей часто существует многофункциональное ПО, которое выполняет сложные расчеты и приобретается на платной основе. Однако для многих конкретных задач предприятия достаточно небольшого набора функций. Например, часто используемое в геологии приложение «Micromine 2014», которое выполняет сложные расчеты, аналитику, часто применяют только в качестве визуализатора объемных моделей. В этом случае возможно рациональнее разработать собственное приложение, которое будет работать быстрее за счет меньших функциональных возможностей.

Для приложений, которые поддерживают ведение журнала действий пользователя, узнать требуемый пользователю функционал возможно, выполнив анализ журнала.

В настоящий момент актуальной является задача автоматизации анализа данных.

Исследование существующих решений анализа данных

Многие ведущие кампании заинтересованы в автоматизации анализа данных, в частности для выполнения анализа данных, получаемых от пользователей. Если объем данных небольшой, его возможно проанализировать вручную или программно с помощью офисных прикладных пакетов. Однако такие корпорации как Valve Steam и прочие пользуются нейронными сетями для анализа данных. Преимуществом нейронных сетей является скорость и качество их работы, а недостатком сложность реализации.

Еще с 2017 года Microsoft стали лидерами в Data mining за счет использования нейросетей для определения наиболее часто используемых пользователем функций и разработки на основе анализа новых версий операционной системы.

Алгоритмы с применением нейросетей, такие как метод ограниченного перебора приносят своим создателям коммерческую выгоду.

Компания WizSoft, например, разработала систему анализа данных WizWhy, основанную на алгоритме ограниченного перебора нейронных сетей, усовершенствовав этот метод при помощи алгоритма «Априори». Достоинством системы является простота в использовании и минимизация субъективных причин. Однако, главный ее недостаток - неспособность находить логические правила, содержащие более 6 элементарных событий [1].

Поэтому разработка методов анализа данных в настоящий момент остается актуальной и востребованной.

Постановка задачи

Технологии анализа данных (Data mining) применяют в различных отраслях человеческой деятельности

Целью применения Data Mining при анализе действий пользователя ПО является обнаружение наиболее частых действий, чтобы оптимально определить необходимый функционал разрабатываемого или модернизируемого ПО.

В работе предлагается решение следующих научных задач:

- исследование методов анализа данных;
- анализ типов нейросетей;

- реализация анализа журнала действий пользователя для определения наиболее частых действий на основе нейросети Кохонена.

Общие методы анализа данных

Помимо метода нейронных сетей следует выделить следующие методы анализа данных: деревья решений, генетические алгоритмы, нечеткая логика, алгоритмы ограниченного перебора, эволюционное программирование, системы рассуждения на основе аналогичных случаев, индукция правил, анализ с избирательным действием, логическая регрессия, алгоритмы определения ассоциаций и последовательностей, визуализация данных, комбинированные методы [2, 3, 6].

В технологии анализа данных (Data mining) большинство методов известны. Научной новизной является их адаптация под реализацию конкретных задач, благодаря развитию технологий последних лет.

Основная часть методов data mining была разработана в рамках теории искусственного интеллекта.

Метод нейронных сетей обычно используется для классификации, кластеризации, прогнозирования и распознавания образов. Для решения рассматриваемой в статье задачи нейросеть выполняет классификацию задач производства, основанную на анализе действий оператора, а также прогнозирования, чтобы делать приложение с заделом на возможные будущие задачи, или задатки под их выполнение [4].

Модель нейронной сети может быть следующих типов:

1) сети прямого распространения (backpropagation): одна из наиболее распространенных архитектур, в основном используется в таких областях, как прогнозирование и распознавание образов;

2) сети с обратной связью: такие, как дискретная модель Хопфилда, в основном используется для оптимизации вычислений и ассоциативной памяти;

3) самоорганизующиеся сети: включают модели адаптивной резонансной теории (ART) и модели Кохонена, в основном используется для кластерного анализа.

В настоящее время при анализе в data mining используются нейронные сети прямого распространения. Их недостатком является медленный темп обучения, а также высокий риск попасть в локальный минимум, из-за чего параметры обучения определить трудно.

Ввиду этих проблем многие перешли к методу объединения искусственных нейронных сетей с генетическими алгоритмами и достигли лучших результатов.

Одно из главных преимуществ нейронных сетей состоит в том, что они могут аппроксимировать любую непрерывную функцию, что позволяет исследователю не принимать заранее какие-либо гипотезы относительно модели. К существенным недостаткам нейронных сетей можно отнести тот факт, что окончательное решение зависит от начальных установок сети и его практически невозможно интерпретировать в традиционных аналитических терминах, что в целом не сильно мешает.

Процесс анализа данных, основанный на нейронной сети

Процесс анализа данных (data mining) может быть представлен тремя основными фазами:

- подготовка данных;
- анализ данных;
- выражение и интерпретация результатов (рис. 1).

Интеллектуальный анализ данных, основанный на нейронной сети, состоит из: подготовки данных, извлечения правил и оценки правил, то есть также трех этапов, как показано на рис. 2.

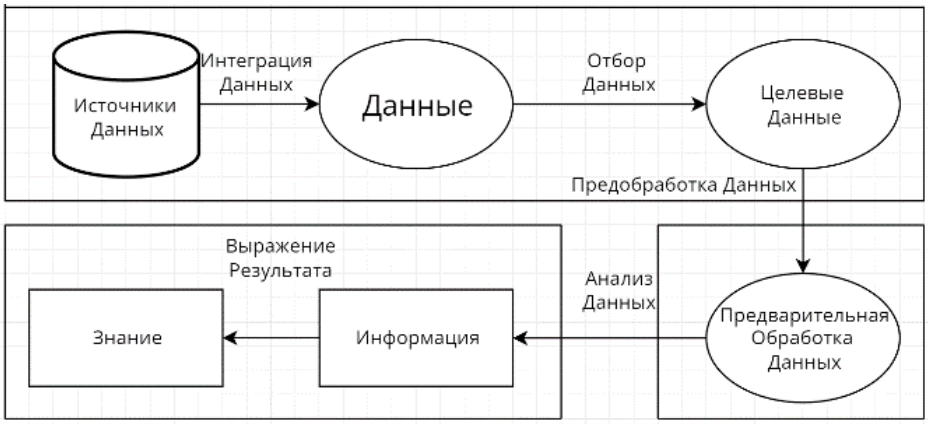

Рисунок 1 – Фазы процесса анализа данных

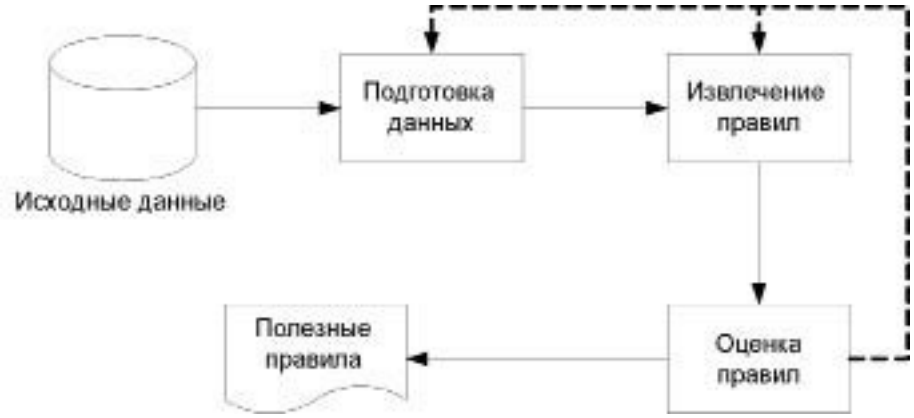

Рисунок 2 – Этапы интеллектуального анализа данных

Подготовка данных

Процесс подготовки данных должен определить и обработать добываемые данные, чтобы сделать их пригодными для конкретных методов интеллектуального анализа. Подготовка данных является первым важным шагом на пути интеллектуального анализа и играет в нем решающую роль. Как правило, подготовка данных включает в себя четыре процесса:

1. Очистка данных. Должна заполнить вакантные значения данных, устранить зашумленные данные и исправить несогласованность в данных.

2. Выбор данных. Должен определить расположение используемых в данном анализе данных.

3. Предварительная обработка данных. Является расширением процесса очистки данных, которые были выбраны.

4. Выражение данных. Должно преобразовать данные после предварительной обработки в форму, которая может быть принята по условию алгоритма анализа данных, основанного на нейронной сети.

Анализ данных, основанный на нейронной сети, может работать только с числовыми данными, из чего следует, что необходимо преобразовывать символьные данные в числовые. Простейший способ заключается в создании таблицы соответствий между символьными данными и числовыми.

Извлечение правил

Существует множество методов извлечения правил, среди которых наиболее часто используются LRE (Limited Relative Error) метод, метод черного ящика, метод извлечения нечетких правил, метод извлечения правил из рекурсивной сети, алгоритм извлечения правил двоичного входа и выхода (BIO-RE), алгоритм частичного извлечения правил (Partial-RE) и алгоритм полного извлечения правил (Full-RE).

Правила оценки

Несмотря на то, что цель правил оценки зависит от конкретного применения, в общем они могут быть оценены в соответствии со следующими задачами:

1) найти оптимальную последовательность извлечения правил; сделав это, получим лучшие результаты в ряде определенных данных;

2) проверить точность извлеченных правил;

3) определить количество знаний в нейронной сети, которые не были извлечены;

4) определить противоречия между извлеченными правилами и обученной нейронной сетью.

Анализ различных видов нейронных сетей

Существует множество алгоритмов анализа данных, основанных на нейронных сетях, проанализируем два наиболее популярных, основанных на самоорганизующихся нейронных сетях и на нечетких сетях.

Анализ данных, основанный на самоорганизующейся нейронной сети

Самоорганизационный процесс - процесс обучения без учителя. При таком обучении обучающее множество состоит из значений входных переменных, а в процессе обучения нет сравнивания выходов нейронов с желаемыми значениями. Можно сказать, что такая сеть учится понимать структуру данных.

Идея сети Кохонена принадлежит финскому ученому Тойво Кохонену. Принцип работы этих сетей заключается во введении в правило обучения нейрона информации о его расположении, то есть составляются карты размещения нейронов.

Самоорганизующиеся карты Кохонена используются для моделирования, прогнозирования, поиска закономерностей в больших массивах данных, выявления наборов независимых признаков и сжатия информации [4, 5, 6].

Анализ данных, основанный на нечеткой нейронной сети

В основе нечетких нейронных сетей лежит идея использования существующей выборки данных для определения параметров функций принадлежности, выводы делаются на основе аппарата нечеткой логики, а для нахождения параметров функций принадлежности используются алгоритмы обучения нейронных сетей. Такие системы могут использовать заранее известную информацию, обучаться, приобретать новые знания, прогнозировать временные ряды, выполнять классификацию образов. Но одним из главных достоинств является наглядность работы такой сети для пользователя.

Из таблицы 1 видно, что и сети Кохонена, и нечеткие нейронные сети имеют достоинства и недостатки.

Таблица 1 - Преимущества и недостатки популярных нейронных сетей в data mining

Тип нейросети	Область применения	Преимущества	Неодстатки
Сеть Конохена	Классификация, кластерный анализ, прогнозирование, сжатие данных	Устойчивость к зашумленным данным, неуправляемое обучение, быстрое обучение, возможность визуализации, возможность упрощения многомерной структуры	Эвристичность алгоритма обучения, предопределенность числа кластеров
Нечеткая нейронная сеть	Прогнозирование, классификация	Хорошая сходимость, быстрое обучение, интерпретируемость накопленных знаний, наглядность работы, легко определить размер сети, допустимость к зашумленным и неточным данным, способны аппроксимировать функции любой степени нелинейности, параллельные вычисления	Априорное определение компонентов

Основное отличие сетей Кохонена от других типов нейронных сетей состоит в наглядности и удобстве использования. Эти сети позволяют упростить многомерную структуру, их можно считать одним из методов проецирования многомерного пространства в пространство с более низкой размерностью. Другое принципиальное отличие сетей Кохонена от других моделей нейронных сетей - неуправляемое или неконтролируемое обучение, что позволяет задавать лишь значения входных переменных. Важнейшим преимуществом нейронечеткой сети является возможность построения одной сети для вычисления нескольких выходных значений по нескольким входным, а также способность к логическому описанию процессов и ручной корректировке функций принадлежности.

Однако нечеткие нейронные сети выгодно отличаются от других типов тем, что вобрали в себя все плюсы нечетких множеств. Таким образом, объединив нечеткие множества и нейронные сети, получили универсальные системы, компенсирующие недостатки нейронных сетей. Основным достоинством применения нейронных сетей является возможность решать различные неформализованные задачи. При этом можно очень просто моделировать различные ситуации, подавая на вход сети различные данные и оценивая выдаваемый сетью результат.

Из рассмотренных моделей анализа данных, основанных на нейронных сетях, можно сказать, что нейронные сети, системы нечеткой логики являются прогрессивным инструментом интеллектуального поиска и извлечения знаний, т. к. обладают способностью выявления значимых признаков и скрытых закономерностей в анализируемых показателях

Реализация анализа журнала действий пользователя для определения наиболее частых действий на основе нейросети Кохонена

Для решения конкретной задачи – анализа данных о действиях пользователей ПО, был использован алгоритм ограниченного перебора, с установлением ассоциаций. Этапы этого алгоритма показаны на рис.3.

Анализ действий пользователя сводится к задаче обнаружения знаний в базах данных, называемый KDD (Knowledge Discovery in Databases). Это процесс поиска полезных знаний в 'сырых данных', который включает в себя вопросы, позволяющие обнаруживать знания.

Этими знаниями могут быть правила, описывающие связи между свойствами данных (деревья решений), часто встречающиеся шаблоны (ассоциативные правила), а также результаты классификации (нейронные сети) и кластеризации данных (карты Кохонена) и т.д.

Для решения поставленной задачи будет использован пакет Deductor – полнофункциональный инструмент для Knowledge Discovery in Databases.

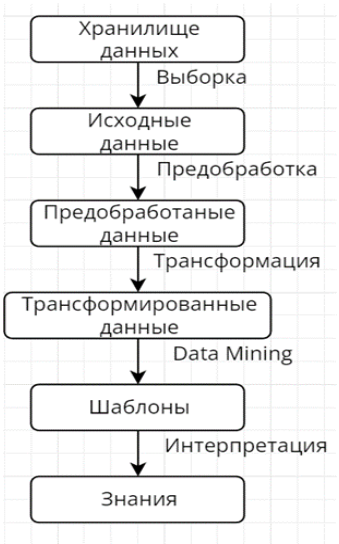

Рисунок 3 – Процесс получения знаний о действиях пользователя

Прежде всего должна быть выполнена подготовка исходного набора данных. В каждом приложении из состава пакета для этого предназначен специальный мастер подключения. Мастер позволяет импортировать данные из СУБД. Следующий шаг алгоритма - предобработка данных (удаление пиковых значений). Для выполнения этого шага в пакете существует приложение RawData Analyzer.

Далее необходимо выполнить трансформацию (нормализацию) данных. Многие приложения из состава пакета производят трансформацию данных автоматически. Например, для нейронных сетей, приложение само переводит числовые поля в нужный диапазон (нормализует), преобразует строковые, булевые и поля типа дата к числовым значениям.

После проведенной подготовки данных выполняется основной этап - Data Mining. В состав пакета включены приложения, реализующие популярные и эффективные методы DM. Neural Analyzer – нейронные сети, Tree Analyzer – деревья решений, Somap Analyzer – самоорганизующиеся карты Кохонена [7, 8, 9].

Для проверки алгоритма была рассмотрена задача анализа действий над паролями пользователя. Анализ выполнялся по шести критериям (количество изменений пароля, максимальный и минимальный период действия пароля, минимальная длина пароля и т.д.). На рис.4 показаны карты Кохонена по каждому критерию. Цветом отображена степень редкости того или иного события. За пределами двух линий находятся информационные шумы.

Кроме графической визуализации результатом алгоритма являются значения по каждому критерию, позволяющие выполнить анализ действий над паролями (табл.2).

Все приложения из состава пакета позволяют эффективно использовать полученные знания или модели на других данных [10].

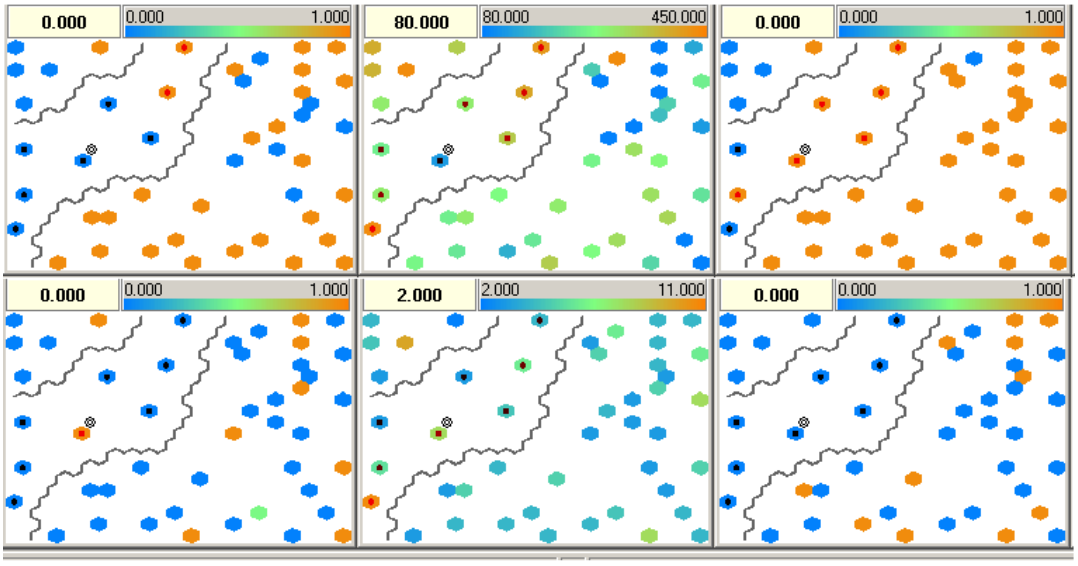

Рисунок 4 – Карты Кохонена

Таблица 2 – Результат анализа действий над паролями

Принудительное сохранение истории паролей	24 Пароля сохранено
Максимальное время действия пароля	42 дня
Минимальное время действия пароля	1 день
Минимальная длина пароля	7 символов
Соответствие требованиям сложности	Включено
Обратное шифрование в паролях	Выключено

Пакет Deductor удовлетворяет всем требованиям для успешного взаимодействия с экспертом, имеет развитый интерфейс, поддержку различных форматов хранения данных, интеграцию с офисными пакетами и т.д.

Заключение

Основные результаты работы следующие:

- использование нейронных сетей возможно для анализа действий пользователя, и имеет ряд преимуществ, но требует продуманного подхода к выбору типа нейросети;
- только средствами нейросетей на данный момент затруднительно проводить анализ данных пользователя, поэтому наиболее оптимальным вариантом является сочетание нейронных сетей, в частном случае сети Конохена, баз данных, и пакетов для предобработки;
- процесс получения знаний о действиях пользователя в условии наличия журнала записей не требует знаний в программировании, вполне можно выполнить все этапы процесса при помощи сторонних программ;
- данные о действиях пользователя не только помогут узнать на что конкретно делать упор при разработке приложения, но и позволять создателю ПО модернизировать уже готовые версии под новые потребности пользователя не нагромождая программу.

Литература

1. Дюк, В.А. Data Mining - интеллектуальный анализ данных // Информационные технологии: сайт. - URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html (дата обращения 01.11.2010)

2. Манжула, В.Г. Методы «мягких» вычислений для аналитической обработки информации в условиях неопределенности / В.Г. Манжула, С.А. Морозов, С.В. Федосеев // Фундаментальные исследования. - 2009. - № 4. - С. 75-76.

3. Назаров, А.В. Нейросетевые алгоритмы прогнозирования и оптимизации систем/ А. В. Назаров, А. И. Лоскутов - СПб.: Наука и Техника, 2003. - 384 с.

4. Чубукова, И. А. Data Mining. - М.: Изд-во «Интернет-университет информационных технологий - ИНТУИТ.ру», 2006. - 384 с.

5. Ярушкина, Н. Г. Основы теории нечетких и гибридных систем: учеб. пособие. - М.: Финансы и статистика, 2004. - 320 с.

6. Xianjun, Ni. Research of Data Mining Based on Neural Networks // World Academy of Science, Engineering and Technology. - 2008. - № 39. - P. 381-384.

7. Редько, В. Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики / В. Г. Редько. - М.: Ленанд, 2015. - 224 с.

8. Редько, В. Г. Подходы к моделированию мышления / В.Г. Редько. - М.: Ленанд, 2014. - 392 с.

9. Яхъяева, Г.Э. Нечеткие множества и нейронные сети: Учебное пособие / Г.Э. Яхъяева. - М.: БИНОМ. ЛЗ, ИНТУИТ.РУ, 2012. - 316 с.

10. Шеннон, К. Работы по теории информации и кибернетике. М.: Иностранная литература, 1963. — 832 с.

Личман А.А., Чередникова О.Ю. Применение методов анализа данных для определения наиболее популярных функций приложения по журналу действий пользователя. *Рассмотрены методы решения задачи определения наиболее часто используемых функций приложения. Предложено использовать для этой цели нейронные сети. Выполнен анализ существующих нейронных сетей и особенностей их применения для различных целей. Предложен и реализован метод определения наиболее часто используемых функций приложения на основе нейронной сети Кохонена и прикладного пакета Deductor для предобработки данных. Это позволит разработчикам программных продуктов модернизировать уже готовые версии под новые потребности.*

Ключевые слова: *data mining, нейронные сети, сеть Кохонена*

Lichman A.A., Cherednikova O. Yu. Application of data analysis methods to determine the most popular application functions based on the user activity log. *Methods of solving the problem of determining the most frequently used application functions are considered. It is proposed to use neural networks for this purpose. The analysis of existing neural networks and the features of their application for various purposes is carried out. A method for determining the most frequently used application functions based on the Kohonen neural network and the Deductor application package for data preprocessing is proposed and implemented. This will allow software developers to upgrade ready-made versions to meet new needs.*

Key-words*: data mining, neural networks, Kohonen network*

Статья поступила в редакцию 04.05.2023

Рекомендована к публикации профессором Мальчевой Р. В.