УДК 004.056

Я. А. Бекенева, С. И. Лебедев, И. И. Холод, Е. С. Новикова
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Классификация событий по составу независимых атрибутов

Предлагается метод разбиения наборов данных на группы в зависимости от состава атрибутов. Метод основан на использовании алгоритмов классификации данных и в качестве целевой переменной содержит некоторый набор атрибутов. Метод был опробован на реальных данных, полученных с предприятия. При описании экспериментов подробно рассмотрены предварительные преобразования данных, а также показано, что некоторые методы обработки могут быть не применимы к данным определенного формата. Тем не менее, эти методы могут показывать высокие результаты при разбиении данных, имеющих иной формат. Предлагаемый метод планируется использовать при анализе больших групп разнородных данных, получаемых от разнотипных средств контроля и описывающих разнотипные события. Применение метода будет способствовать получению групп данных с одинаковым или схожим составом атрибутов, благодаря чему может быть повышена точность их классификации при выявлении нарушений.

Группировка данных, атрибуты данных, данные от разнородных источников, дерево решений, интеллектуальный анализ данных

В настоящее время существует множество систем мониторинга, которые способны фиксировать события различного рода. К таким системам относят контрольно-пропускные пункты, системы видеонаблюдения, системы контроля доступа, а также разнообразные датчики, фиксирующие различные параметры (температуру, давление, влажность и т. п.). Любой производственный процесс характеризуется определенным набором событий. Каждая из таких систем фиксирует определенного рода инциденты, относящиеся к конкретным событиям.

Такие системы, как правило, генерируют наборы данных, характеризующихся определенным набором атрибутов и соответствующих им значений. При этом состав атрибутов зависит как от типа датчика, так и типа субъекта, инициировавшего инцидент. При этом одно и то же событие может быть описано с помощью нескольких записей, созданных разными системами мониторинга.

Как правило, задача выявления аномалий связана с анализом данных, построением шаблонов поведения, классификацией данных и т. д. Поступающие от разнородных источников сырые данные зачастую оказываются не пригодными для анализа.

Основной проблемой построения классификатора по информации, поступающей от разнородных средств контроля, является разный состав и тип атрибутов. Так, информация, поступающая от видеокамеры, может содержать сведения, выявленные методами распознавания образов (например, номер, цвет машины), а система контроля доступа может предоставлять информацию о сотруднике, приложившем пропуск. В связи с этим корректное построение классификатора по таким событиям невозможно, т. е. состав независимых атрибутов для разных подмножеств событий будет разный.

При совокупном анализе таких данных методами интеллектуального анализа может возникнуть проблема с выбором подходящего для такого набора данных метода. Кроме того, возможна ситуация, при которой для разных данных могут быть использованы разные методы анализа. В связи с этим одной из наиболее актуальных задач является адекватное разбиение наборов

разнородных данных на группы в зависимости от состава атрибутов для их дальнейшего анализа. Как правило, базы данных содержат большое количество различных записей, поэтому ручная группировка может оказаться невыполнимой.

В связи с этим можно сделать вывод о необходимости создания нового способа обработки данных для выявления нарушений.

Авторы предлагают решать задачу автоматического выявления нарушений методами Data Mining в несколько этапов:

1. Выделение групп событий.
2. Подготовка данных.
3. Обучение модели классификации для каждой группы.
4. Применение обученных моделей к новым данным.

Рассмотрим первый этап, связанный с особенностями выделения групп событий.

На этом этапе решается задача разбиения больших наборов данных на классы для дальнейшего применения различных классификаторов. Основной целью является разработка метода автоматической группировки данных в зависимости от состава атрибутов записей.

Задача группировки данных по атрибутам может решаться в различных сферах разными способами. Например, в исследованиях в области геолокации зачастую решаются задачи, связанные с кластеризацией объектов, в том числе по присущим им атрибутам [1], [2]. В медицине решается задача кластеризации данных, полученных в результате анализов, по набору атрибутов, при этом определенный набор атрибутов и присущих им значений у каждого кластера соответствует определенному набору диагнозов [3].

Авторы настоящей статьи решают иную задачу, требующую своего подхода. Основной проблемой является то, что разные средства мониторинга могут описывать события с использованием различных наборов атрибутов. Поэтому описание событий может различаться и зависеть от типа события и средств, которые фиксируют такие типы событий. При этом небольшая часть атрибутов является общей для всех возможных событий и средств, их фиксирующих, однако большая часть атрибутов остается вариативной. Для любого зафиксированного события в базе хранится запись, в которой каждая строка описывает свой отдельно взятый атрибут и соответствующее ему значение.

При совокупном анализе данных для получения максимально полного описания происходящих процессов необходим совместный анализ всех получаемых данных от всех возможных средств мониторинга. В результате объединения данных будет получен общий файл, содержащий множество записей от всех возможных средств мониторинга. Каждая строка такого файла будет соответствовать одной записи, полученной от какого-либо источника данных. Столбцы в таком файле будут описывать все возможные атрибуты, фиксируемые всеми имеющимися средствами фиксации инцидентов.

Основной сложностью такого объединения является получение большого количества пропущенных значений. Это связано с тем, что каждая система мониторинга фиксирует определенный набор параметров, который может лишь частично пересекаться для разных типов средств фиксации событий.

Как правило, наименования одних и тех же смысловых атрибутов в записях от разных источников данных могут различаться. Соответственно, данные, описывающие одни и те же параметры для разных строк, будут находиться в разных колонках, что приведет к еще большему увеличению пропущенных значений в совокупном файле. Таким образом, возникает задача выделения групп данных, которые имели бы одинаковый или максимально схожий состав атрибутов (рис. 1). Разделение данных на группы позволит получить выборки, содержащие минимальное количество пропущенных значений или не содержащее их вовсе.

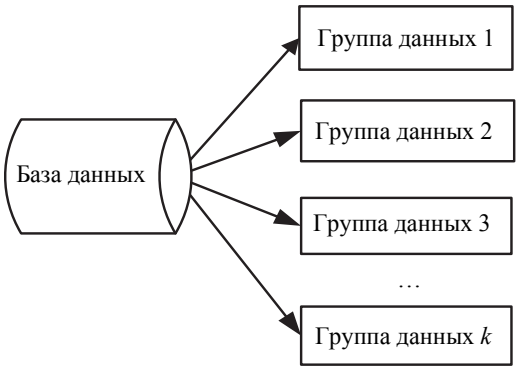

Рис. 1

В общем виде предлагается следующее выделение групп событий:

1. Чтение данных из таблицы атрибутов.
2. Выбор атрибутов, связанных с идентификатором события, именем и значением.
3. Если идентификатор события представлен в виде числа – преобразование идентификатора инцидента из числового значения в текстовое.

4. Конкатенация полей, указывающих на имя события, для одинаковых значений идентификатора события с ограничением не более трех значений в одну запись.

5. Объединение данных, полученных на предыдущем шаге.

6. Если выполнялся шаг 3 – обратное преобразование идентификатора события в числовое значение.

7. Чтение данных из таблицы событий.

8. Объединение двух таблиц по идентификатору события.

9. Задание целевого атрибута.

10. Выбор метода для разбиения данных на группы.

11. Разбиение данных на группы по составу атрибутов.

В результате будут получены группы данных с одинаковым составом атрибутов. Каждая из этих групп может быть проанализирована независимо от других с помощью различных методов анализа данных.

Для проверки работоспособности предложенного метода была проведена серия экспериментов с реальными данными, полученными с предприятия.

При проведении экспериментов использовалось оборудование со следующими характеристиками:

– сервера: процессор: Intel Xeon CPU E5-2667, 2.90GHz, 6 ядер; RAM: 64Gb DDR3; ОС:Windows 10 Pro (64-разрядная); Mysql: 5.7; Rapidminer: 8.001;

– данных:

1. Таблица incident, в которой представлены описания инцидентов, содержит 6 506 825 строк и 9 столбцов.

2. Таблица incident_attr, в которой представлены описания атрибутов в формате <имя, значение>, 15 522 224 строки и 4 столбца.

Исследуемые данные содержали записи, поступающие от разных источников (камер видеонаблюдения, датчиков контрольно-пропускных систем, измерительных устройств и т. д.) и описывающие события, инициируемые разными типами субъектов.

Все эксперименты проводились в среде RapidMiner [4].

На первом этапе экспериментов проводились общие предварительные преобразования данных. После чтения данных из таблицы incident_attributes выбирались атрибуты incident_id, name, value, затем производилось преобразование pivot и последующее объединение с таблицей incident. Полученная в результате таблица содержала записи обо всех зафиксированных событиях и всех возможных атрибутах событий. При этом имелось большое количество пропущенных значений, что затрудняло дальнейший анализ данных методами интеллектуального анализа.

На следующем этапе анализа осуществлялось разбиение данных на группы по составу атрибутов согласно методу, описанному выше. Для этой цели были опробованы алгоритмы классификации Single Rule Induction и Decision Tree.

Метод Single Rule Induction оказался не применим к имеющемуся формату данных, в связи с чем для решения задачи был использован метод Decision Tree. Авторами были опробованы различные параметры построения дерева решений, разные критерии расщепления, а также использованы разные корневые атрибуты для построения дерева. Кроме того, было проведено несколько экспериментов с разным составом атрибутов, используемых для классификации данных.

На рис. 2 представлена схема процесса, выполняемого в среде Rapidminer при проведении экспериментов.

Операторы на рис. 2 выполняют следующие функции:

Read incident_attributes, Read incident – чтение таблицы из базы данных;

Select attributes – выбор атрибутов (в данном случае идентификатор инцидента, имя, значение);

Numerical to Polynomial – преобразование данных (идентификатор инцидента incident_id из числового вида в строковый) – это необходимо для успешного выполнения оператора Loop Values;

Loop values – цикл по значениям incident_id. Для каждого уникального incident_id добавляются атрибуты «name». Например, при наличии несколь-

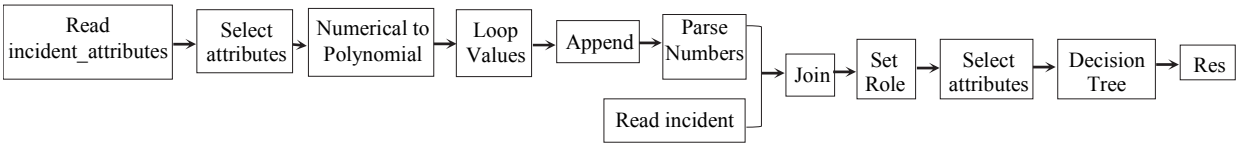

Рис. 2

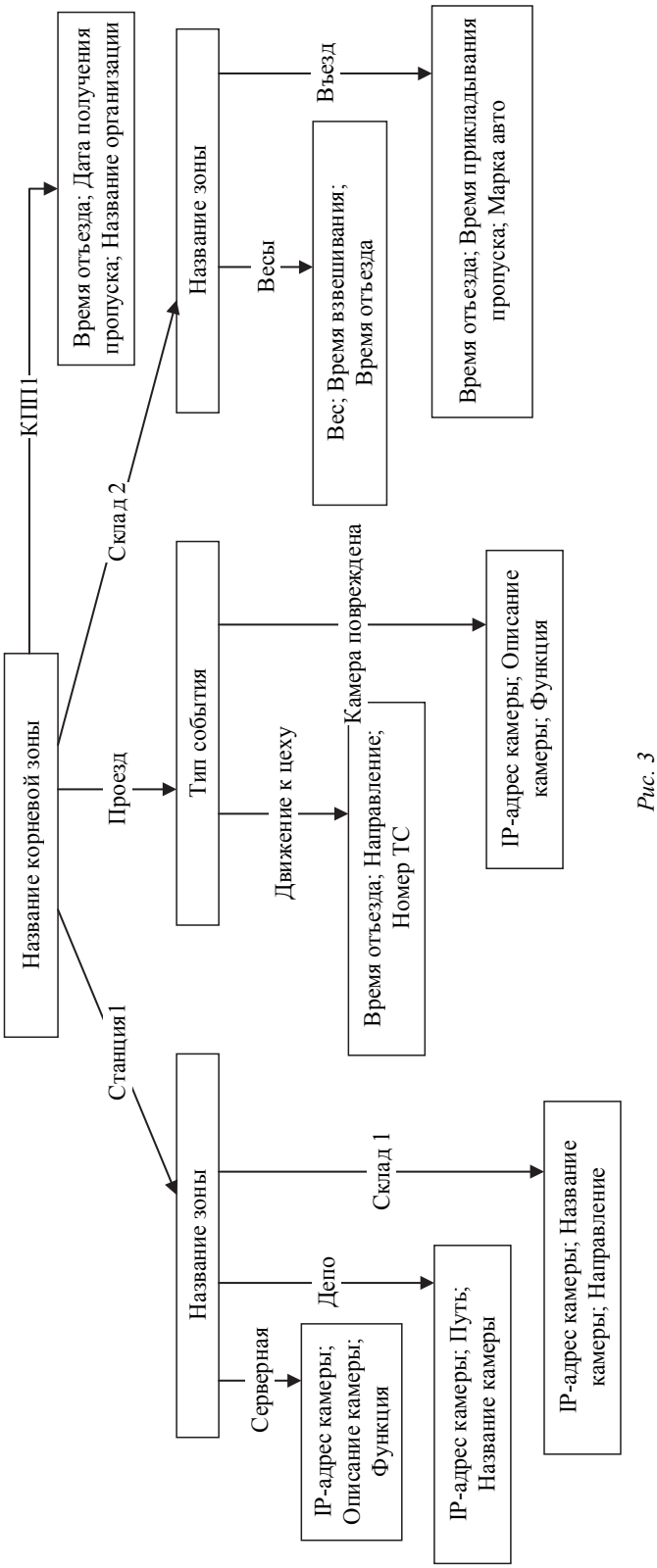
Название корневой зоны
КПП1
Время отъезда; Дата получения пропуска; Название организации
Проезд
Склад 2
Название зоны
Весы
Вес; Время взвешивания; Время отъезда
Въезд
Время отъезда; Время прикладывания пропуска; Марка авто
Тип события
Движение к цеху
Время отъезда; Направление; Номер ТС
Камера повреждена
IP-адрес камеры; Описание камеры; Функция
Станция1
Название зоны
Серверная
IP-адрес камеры; Описание камеры; Функция
Депо
IP-адрес камеры; Путь; Название камеры
Склад 1
IP-адрес камеры; Название камеры; Направление

Рис. 3

ких одинаковых incident_id с разными значениями name после применения данного оператора будет получена одна запись incident_id с объединенными данными в поле name (конкатенация строк);

Append – преобразование данных, полученных после выполнения оператора Loop Values из списка таблиц в одну таблицу;

Parse Numbers – обратное преобразование incident_id из строкового формата в числовой;

Join – объединение данных из двух таблиц;

Set Role – задание целевой переменной (в данном случае name);

Decision Tree – оператор для построения дерева решений.

При выполнении экспериментов были использованы различные параметры операторов. Каждое полученное дерево решений оценивалось при помощи оператора Cross Validation, позволяющего оценить его точность.

В серии экспериментов варьировались такие параметры, как критерий расщепления, использование обрезки pruning и использование предварительной обрезки preprunning. Основное отличие между обрезками pruning и preprunning заключается в том, что pruning применяется к дереву решений, построенному на всей выборке данных, а preprunning представляет собой предварительную обрезку выборки данных. При совместном использовании обоих типов обрезки сначала производится обрезка данных preprunning, по полученной выборке строится дерево решений, которое затем обрезается с помощью pruning.

В таблице представлены значения точности для каждого полученного дерева решений, построенного при различной комбинации параметров построения.

Критерий расщепления	Использование обрезки pruning	Использование предварительной обрезки preprunning	Точность решений, %
gain_ratio	–	–	92.19
Information gain	–	–	90.60
Gini index	–	–	90.60
gain_ratio	–	+	64
gain_ratio	+	–	90.60
gain_ratio	+	+	64

Как видно из таблицы, наиболее высокую точность показало дерево решений, к которому не применялись инструменты обрезки, т. е. наиболее полное дерево, построенное по полной выборке данных. При этом использовался критерий расщепления gain_ratio. Деревья решений, построенные с другими критериями расщепления, показали менее высокую точность. Использование обрезки уже построенного дерева несколько снижает его точность. Использование предварительной обрезки существенно снижает точность получаемого дерева. При этом, как показали эксперименты, применение обрезки к дереву, построенному на предварительно обрезанных данных, не повлияло на точность полученного в итоге дерева решений.

После серии экспериментов было выбрано дерево решений, представленное на рис. 3, точность которого составила 92.19 %. Таким образом, авторам удалось получить автоматическую группировку записей по составу атрибутов. Представленное на рис. 3 дерево решений было построено с использованием четырех переменных, связанных с названием корневых зон, зон, являющихся внутренними для корневых зон, типом события и переменной, которая являлась целевой и указывала на объединенные атрибуты записей. Для построения дерева использовался критерий расщепления gain_ratio, обрезки при этом не использовались.

Результат применения полученного дерева решений будет использован на следующих этапах анализа данных, рассмотренных в данном исследовании. Разработанный авторами метод может быть использован для группировки иных наборов данных, полученных как от реальных источников, так и в результате моделирования. В дальнейших работах будут представлены исследования по проведению классификации событий, сгруппированных с помощью предложенного метода. Предполагается, что произведенная предварительная группировка данных по составу атрибутов позволит более качественно обучить модели классификации и получить более высокую точность выявления возможных нарушений в потоке разнотипных событий.

Предложенный авторами метод позволяет получить наборы данных, одинаковых по составу атрибутов, а следовательно, минимизировать число пропущенных значений. Выделенные наборы данных в дальнейшем будут исследованы разными методами классификации с целью выявления возможных аномальных событий.

Работа выполнена при поддержке Министерства образования и науки Российской Федерации в рамках работы «Организация проведения научных исследований» (Задание № 2.6113.2017/6.7).

СПИСОК ЛИТЕРАТУРЫ

1. Колесенков А. Н. Современные подходы к обработке данных при построении геоинформационных систем экологического мониторинга // Изв. Тульского гос. ун-та. Техн. науки. 2016. № 9. С. 103–111.

2. The space syntax toolkit: Integrating depthmapX and exploratory spatial analysis workflows in QGIS / J. Gil, T. Varoudis, K. Karimi, A. Penn // SSS 2015–10th Intern. Space Syntax Symp. Space Syntax Laboratory, The Bartlett School of Architecture. London, UK: University College London, 2015. Vol. 10. P. 1–19.

3. Доан Д. Х., Крошилина С. В., Жулева С. Ю. Использование нечеткой кластеризации в анализе статистической нечеткой медицинской информации для формирования наборов вариантов течения болезни в системах поддержки принятия медицинских решений // Тр. Междунар. науч.-техн. конф. Воронеж: ФГБОУ ВО «Воронежский гос. техн. ун-т», 2017. Т. 1. С. 268.

4. Сайт RapidMiner: Data Science Platform. URL: https://rapidminer.com/ (дата обращения 01.06.2018).

Ya. A. Bekeneva, S. I. Lebedev, I. I. Kholod, E. S. Novikova
Saint Petersburg Electrotechnical University «LETI»

EVENT CLASSIFICATION DEPENDING ON THE SET OF INDEPENDENT ATTRIBUTES

Events at the production facility that can be recorded by monitoring devices could be not related to each other and carried out by different types of entities. In this case, the attribute sets may not coincide completely, except for the attributes that are common to all the fixed events. In this regard, large data sets describing a variety of different processes often consist of records that differ significantly in structure and composition of attributes. When aggregate analysis of such data by the methods of intellectual analysis, there may be a problem with the choice of a method suitable for such a set of data. In addition, the authors suggest that different analysis methods can be used for different data groups. In this paper, we propose a method for splitting the data sets into groups, depending on the composition of the attributes.

Data grouping, data attributes, data from heterogeneous sources, decision tree, data mining

УДК 006.72

А. А. Воевода, Д. О. Романников
Новосибирский государственный технический университет

Формирование структуры нейронной сети посредством декомпозиции исходной задачи на примере задачи управления роботом-манипулятором

Предлагается способ формирования структуры нейронной сети, основанный на декомпозиции исходной задачи, результатом которой является набор состояний, в которых может находиться система, и признаков смены состояний. Предлагается построить конечный автомат, в котором переходы и сам конечный автомат представлены составными частями нейронной сети, а каждому состоянию соответствует отдельная нейронная сеть, с помощью которой выполняется управление роботом-манипулятором в соответствующем состоянии. Построение нейронной сети состоит из трех этапов: 1) реализация части сети для определения признаков переходов между состояниями; 2) реализация части сети для определения состояния системы; 3) реализация нейронной сети для каждого состояния при формировании выходных сигналов и объединение выходных сигналов от разных состояний. Итоговая нейронная сеть позволяет получить большую наблюдаемость и возможность «отладки», а также в значительной мере упростить процесс обучения за счет использования более простых типов нейронных сетей и решаемых задач.

Нейронные сети, структура нейронной сети, обучение с подкреплением, автоматизация, конечный автомат, управление роботом-манипулятором

В настоящее время нейронные сети применяются для решения таких задач, как распознавание речи и изображений [1], генерация заголовков изображений [2] и др. В [3] предлагается метод обучения нейронных сетей с подкреплением (reinforcement learning), который на основе взаимо-