Метод преобразования данных от разнородных средств контроля для выявления нарушений

Я. А. Бекенева[1], С. И. Лебедев[2], И. И. Холод[3], А. В. Шоров[4], Е. С. Новикова[5]
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]yana.barc@mail.ru, [2]tylubz@gmail.com, [3]iiholod@mail.ru, [4]ashxz@mail.ru, [5]novikova.evgenia123@gmail.com

***Аннотация*. В работе предлагается метод обработки данных, получаемых от разнородных систем мониторинга. Сами по себе данные от таких систем лишь частично описывают событие, поэтому их может быть недостаточно для принятия решения при классификации событий. Набор данных от разнородных источников позволяет получить максимально полную картину о событиях, однако получаемые от разных источников данные могут значительно отличаться друг от друга. В данной работе предложено решение проблемы, связанной с группировкой данных о событиях и приведению их к виду, пригодному для дальнейшего анализа.**

Ключевые слова: события; обработка данных; преобразование данных; разнородные источники данных; классификация событий; выявление аномалий; выявление нарушений

I. Введение

Мониторинг производственных процессов и выявление в нем возможных нарушений является актуальной и важной задачей. Любой производственный процесс характеризуется определенным набором событий, каждое из которых может быть описано с помощью ряда атрибутов.

В настоящее время существует множество систем мониторинга, которые способны фиксировать различного рода события. К таким системам относятся, например, контрольно-пропускные пункты, камеры фото- и видеонаблюдения, системы контроля доступа, а также различные датчики, фиксирующие параметры микроклимата в рабочих и иных помещениях.

Каждая из таких систем фиксирует определенного рода инциденты, которые относятся к определенным событиям. При этом одно и то же событие может описываться с помощью нескольких записей, созданных разными системами мониторинга. Так, в работе [1] событие, связанное со входом отдельно взятого студента в аудиторию сопровождается двумя записями от систем мониторинга. Первая запись генерируется системой RFID и содержит данные о магнитной карточке, приложенной студентом на входе в аудиторию. Вторая запись генерируется камерой, фиксирующей лицо студента.

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации в рамках государственного задания «Организация научных исследований», задание #2.6113.2017/6.7

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 непригодными для анализа. В связи с этим можно выделить ряд проблем, которые в настоящее время требуют решения:

1. Базы данных, как правило, предназначены для хранения и не могут быть использованы для анализа данных в существующем виде.
2. Несколько записей, поступивших от разных источников, могут относиться к одному и тому же событию.
3. Данные, поступающие от разных типов источников, имеют разный формат и разный состав атрибутов.

В данной работе предложен метод преобразования данных от разнородных источников и приведения их к единому формату. Для проведения экспериментов была использована выборка данных, полученная с предприятия. Данные, преобразованные с помощью предложенного метода, были проанализированы средствами Data Mining.

II. Релевантные работы

В настоящее время существует большое количество работ, связанных с исследованиями процессов, которые могут описываться с помощью данных, полученных от разнородных источников. В некоторых случаях, как, например, в работе [2] не требуется преобразование данных и приведение их к единому формату. В данной работе основным источником данных является непрерывный сигнал GPS, с помощью которого описываются перемещения рейсовых автобусов. Сигналы SCATS используются в качестве вспомогательных.

Важную роль в выборе метода группировки и обработки данных играет сам формат данных. Например, в работе [3] авторы ставят перед собой задачу реализовать мониторинг производственного процесса в режиме реального времени, используя сигналы, полученные от разнородных датчиков. Авторы предлагают преобразование данных, полученных от датчиков, в

неопределенную систему линейных уравнений, которую, в свою очередь, предлагается решать с помощью жадного байесовского метода. Такой метод может быть оптимальным для решения определенной задачи и при этом быть совершенно неприемлемым для других задач, в частности, для обработки данных, исследуемых в настоящей работе.

Достаточно близкой по смыслу является работа [4], однако основной ее целью является сжатие данных с целью уменьшения их объема. Для решения этой задачи авторы предлагают пространственно-временную кластеризацию. Однако такой подход не учитывает субъекта, инициировавшего событие, а кроме того, может иметь большую погрешность при выделении отдельных событий или вовсе оказаться неприменим для данной цели.

III. Описание подхода

В работе [5] событие характеризуется таким набором параметров как $E = < type, sm_j, sb_k, sv_u, p_g g >$, где type – тип события; sm_j – объект наблюдения, зафиксировавший событие; sb_k – объект, который инициировал событие; sv_u – объект, над которым выполнено событие; $p_g g$ = 1..s – дополнительные параметры события. Процессы, происходящие на объекте, могут быть представлены в виде упорядоченных по времени множеств пар события и времени, в которое оно произошло: $P = \{< t, e_i >: e_i \in E, r \in R\}$, где t – время, в которое произошло событие; r – категория нарушения из множеств R .

Как правило, набор данных от разнородных источников имеет разный формат и разный набор атрибутов. Сами по себе такие данные от каждого отдельно взятого источника могут быть недостаточными для классификации событий и выявления возможных нарушений, однако совокупный набор записей от различного рода источников данных будет полным и информативным. Таким образом, для анализа событий и выявления среди них возможных нарушений необходимо решить следующие задачи:

1. Выполнить общие преобразования, позволяющие привести данные к виду, пригодному для дальнейшего анализа. Так, для методов Data Mining необходим набор в виде матрицы (таблицы), где каждая строка представляла бы собой описание события, а столбцы – атрибуты этого события.
2. Записи, поступившие от разных систем мониторинга и относящиеся к одному и тому же событию, должны быть сгруппированы. Для этого, в первую очередь, необходимо определить критерии, по которым следует выделять такие записи и объединять их.
3. Выделить группы событий, имеющих одинаковый или схожий состав атрибутов.
4. Выбрать алгоритмы классификации, применимые к исследуемому набору данных.

A. *Общие преобразования*

Любое событие e_i происходит в определенной точке пространства в определенный момент времени [5]. При этом событие может быть инициировано одним субъектом или группой субъектов. В рамках данной работы будут рассматриваться события, инициированные одним субъектом sb_k.

Таким образом, каждое событие имеет набор атрибутов A, относящихся к субъекту, инициировавшему событие, месту и времени совершения события. Такие атрибуты можно назвать общими и выделить их в группу *Incident* (рис 1).

Разные средства мониторинга могут описывать события с использованием разных наборов атрибутов. Поэтому описание событий может отличаться между собой и зависеть от типа события и средств, которые фиксируют такие типы событий. Вариативные атрибуты могут быть выделены в группу *Incident_attributes* (рис. 1). Для каждого зафиксированного события в базе хранится запись, в которой каждая строка описывает каждый отдельно взятый атрибут и соответствующее ему значение.

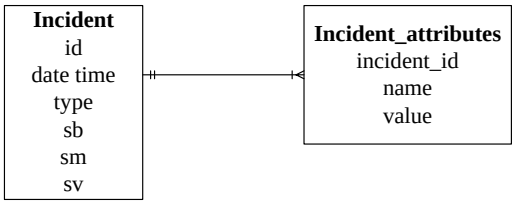

Рис. 1. Структура хранения данных об инцидентах

Если событие e_i может быть описано с помощью нескольких записей от разных систем мониторинга, то такие записи должны быть сформированы в одно и то же время источниками данных, расположенными в одном и том же месте, и указывать на один и тот же субъект, инициировавший событие.

Для комплексного анализа событий необходимо интегрировать данные, полученные от всех типов источников, в единую таблицу. Для баз данных, где данные хранятся в виде, представленном на рис. 1, необходимо выполнить следующие преобразования:

1. К таблицам из группы *Incident_attributes* применить преобразование pivot. Это позволит расположить атрибуты по столбцам и записать соответствующие им значения в виде одной строки.
2. Объединить таблицы *Incident* и *Incident_attributes* по *id* инцидента.

В результате 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 количество 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 отличаться. 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.

B. Объединение записей, описывающих одно и то же событие

Пусть некоторый субъект sb_k совершает действия, фиксируемые системами мониторинга sm_1, sm_2, sm_3. Каждая из этих систем генерирует записи с определенным набором параметров. Допустим, при данные, формируемые системой sm_1, описываются таким набором атрибутов как <time, zone, sb_name_cam>. Аналогично, данные, формируемые системой sm_2, описываются набором атрибутов <time, zonename, sb_name, weight, full>, а данные, формируемые системой sm_3 – <time, zone, sb_nm, dir>.

При объединении всех записей в единый файл будет получена запись следующего вида:

ТАБЛИЦА I ФОРМАТ ДАННЫХ ПОСЛЕ ОБЪЕДИНЕНИЯ ЗАПИСЕЙ ОТ РАЗНОРОДНЫХ ИСТОЧНИКОВ

time	zone	sb_name_cam	zonename	sb_name	weight	full	sb_nm	dir
X	X	X						
X			X	X	X	X		
X	X						X	X

В таблице показан вариант, когда в полученном файле имеется одновременно 3 атрибута, идентифицирующих субъект, инициировавший событие. Формат записи может как отличаться (один из атрибутов может указывать на субъект в виде слова, другой – указывать на идентификатор в виде номера и пр.), а может быть одинаковым для всех трех колонок. В некоторых случаях исходная строка может содержать и 2 атрибута, указывающих на субъект: например, цифровой идентификатор и соответствующая ему фамилия, номер транспортного средства и пр. Таким образом, на данном этапе необходимо сперва найти и объединить атрибуты, описывающие один и тот же смысловой параметр.

При получении таблицы подобного вида необходимо выделить все атрибуты, указывающие на субъект, и объединить их в один общий атрибут. Для этого можно как создать совершенно новый атрибут, идентифицирующий субъект, так и выбрать один из уже существующих. В приведенном примере будет создан новый атрибут value_sb. Аналогично должны быть проанализированы и преобразованы все атрибуты с общим смыслом. После преобразования будет получена таблица следующего вида:

ТАБЛИЦА II ФОРМАТ ДАННЫХ ПОСЛЕ ОБЪЕДИНЕНИЯ ОДИНАКОВЫХ ПО СМЫСЛУ АТРИБУТОВ

time	zone	value_sb	weight	full	dir
X	X	X			
X	X	X	X	X	
X	X	X			X

Очевидно, что таблица такого вида является более наглядной и удобной. Она не перегружена лишними параметрами, а количество пропущенных значений существенно меньше, чем в табл. 1. Тем не менее, в такой таблице еще имеется некоторое количество пропущенных значений, что по-прежнему создает неудобства при её анализе.

Как было сказано ранее, любое событие имеет три наиболее важных характеристики:

- время совершения события;
- место, где событие было совершено;
- субъект, который инициировал событие.

При анализе и группировке записей, которые лишь частично описывают одно и то же событие, важно следовать трем основным правилам:

- должны совпадать временные атрибуты события;
- должны совпадать пространственные атрибуты события;
- субъект должен быть одним и тем же.

При анализе отдельно взятых событий или отдельных типов событий важно понимать, из каких процедур, фиксируемых средствами мониторинга, состоит это событие. Необходимо определить последовательность этих процедур и временные задержки между ними. Например, при входе в офисное здание сотрудник организации прикладывает пропуск на входе, а через несколько секунд его лицо попадает в объектив камеры наблюдения. Следовательно, необходимо понимать, что временные атрибуты записей, относящихся к одному событию, могут отличаться на определенное значение, а не быть одинаковыми.

В некоторых случаях процесс может включать в себя события, совершаемые разными типами субъектов (например, автобусы и поезда), действия которых фиксируются разными средствами контроля. Параметры записей при этом могут существенно различаться для разных типов субъектов. В таких случаях количество атрибутов, идентифицирующих субъект, будет равно количеству типов субъектов.

В общем виде метод дальнейшей группировки данных может быть представлен следующим образом.

1. Сортировка данных по атрибуту, идентифицирующих выбранный тип субъекта.
2. Если количество типов субъектов больше одного, необходимо выполнить фильтрацию по выбранному атрибуту (значение не пустое или пустое). В общем случае на выходе фильтра будет

получено два набора данных. Первый набор данных будет относиться к событиям, инициированным субъектами выбранного типа. Второй набор данных будет относиться к событиям, инициированным субъектами остальных типов.

3. Если необходимо провести анализ данных для субъектов всех имеющихся типов, второй набор данных, полученный на шаге 2, необходимо подать на вход нового процесса сортировки.
4. Если шаг 3 выполняется, то шаги 1–2 необходимо выполнять до тех пор, пока на шаге 1 выходной набор данных не станет единственным.
5. Сортировка по времени для каждого отдельно взятого субъекта.
6. Объединение строк с одинаковым идентификатором субъекта при равенстве значений атрибутов, идентифицирующих зону, и разницы во времени, не превышающей допустимую задержку.
7. Удаление атрибутов, не содержащих полезной информации.
8. Запись данных в файл.

Шаги 5–8 выполняются для каждого набора данных, полученного на шаге 2. В результате будет получен набор файлов, описывающих различные события, инициированные разными типами субъектов. Количество таких файлов будет равняться количество типов субъектов. После выполнения шагов 5–7 каждый отдельно взятый набор данных, имевший вид, представленный в табл. 2, будет преобразован в следующий вид:

ТАБЛИЦА III ФОРМАТ ДАННЫХ ПОСЛЕ ПРЕОБРАЗОВАНИЯ

time	zone	value_sb	weight	full	dir
X	X	X	X	X	X

Полученные данные будут пригодны для дальнейшего анализа методами Data Mining.

IV. ЭКСПЕРИМЕНТЫ

Для проверки применимости алгоритмов Data Mining к данным, преобразованным с помощью предложенного авторами метода, в среде RapidMiner [6] были проведены эксперименты с выборкой данных, полученной с предприятия. Выборка представляла собой две таблицы из базы данных, при этом их структура хранения соответствовала структуре, представленной на рис. 1. Авторами было произведено преобразование pivot таблицы *incident_attributes* и последующее объединение с таблицей *incident*. В результате была получена общая таблица, состоящая из более чем тысячи строк, соответствующих инцидентам, и 74 столбцов, соответствующих атрибутам. В таблице были представлены данные, описывающие разные типы событий, инициированные разными типами субъектов.

После группировки атрибутов, имеющих одинаковое смысловое значение, и последующего выбора записей, описывающих действия определенного типа субъектов, авторами была получена таблица, состоящая из 600 строк и 39 столбцов. После объединения записей, относящихся к одному событию, была получена окончательная таблица размерностью 35 столбцов и 550 строк. Из данной таблицы были выделены записи об инцидентах, зафиксированных на складе предприятия, полученная выборка была подана на вход дерева решений (рис. 2).

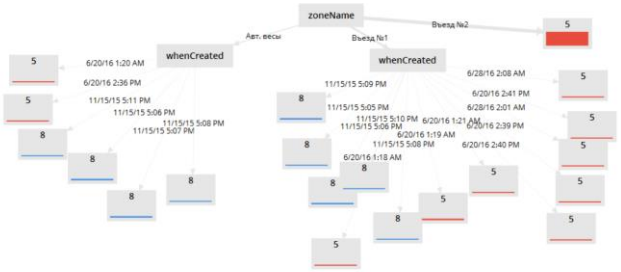

Рис. 2. Дерево решений для преобразованной выборки данных

Таким образом, предложенный авторами метод преобразования данных может быть использован для приведения разнородных данных к виду, пригодному для дальнейшего интеллектуального анализа.

V. ЗАКЛЮЧЕНИЕ

В работе представлен разработанный авторами метод преобразования данных от разнородных источников. С помощью данного метода выборка разнотипных данных была преобразована в формат, пригодный для дальнейшего анализа, что подтверждено успешно проведенным экспериментом с классификатором «Деревья решений». В дальнейшем авторы планируют исследовать данные различными классификаторами, провести их обучение и апробацию на тестовых выборках.

СПИСОК ЛИТЕРАТУРЫ

[1] Pss S., Bhaskar M. RFID and pose invariant face verification based automated classroom attendance system //Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on. IEEE, 2016. С. 1-6

[2] Artikis A., Weidlich M., Schnitzler F., Boutsis I., Liebig T., Piatkowski N., Gal, A. Heterogeneous Stream Processing and Crowdsourcing for Urban Traffic Management //EDBT. 2014. Т. 14. С. 712-723.

[3] Bastani K., Rao P. K., Kong Z. An online sparse estimation-based classification approach for real-time monitoring in advanced manufacturing processes from heterogeneous sensor data //IIE Transactions. 2016. Т. 48. №. 7. С. 579-598.

[4] Yang C., Zhang X., Zhong C., Liu C., Pei J., Ramamohanarao K., Chen J. A spatiotemporal compression based approach for efficient big data processing on cloud //Journal of Computer and System Sciences. 2014. Т. 80. №. 8. С. 1563-1583.

[5] Kholod I.I., Bekeneva Y.A., Novikova E.S., Shorov A.V. Intellectual model for violations detection in the business process //Young Researchers in Electrical and Electronic Engineering (EIConRus), 2018 IEEE Conference of Russian. IEEE, 2018. С. 313-317.

[6] RapidMiner: Data Science Platform. URL: https://rapidminer.com/