Акимушкин Василий Александрович,
Поздняков Сергей Николаевич

УДК 004.832.22, 378:004.9

ОБЗОР МЕТОДОВ EDUCATIONAL DATA MINING ДЛЯ АНАЛИЗА ПРОТОКОЛОВ ВЗАИМОДЕЙСТВИЯ ОБУЧАЕМОГО С «НАУЧНЫМИ ИГРАМИ»

Аннотация

Одной из тенденций развития электронного обучения является внедрение интерактивных обучающих систем. Взаимодействие обучаемого с интерактивной обучаемой системой порождает огромный массив данных, который может быть использован для корректировки учебного процесса. Для решения этой задачи могут использоваться методы Educational Data Mining. Educational Data Mining (EDM) является молодой междисциплинарной наукой, которая занимается разработкой методов для исследования данных, возникающих в образовательном контексте. Educational Data Mining использует как стандартные методы Data Mining, такие как кластеризация, классификация, регрессия, корреляция, визуализации и др., так и ряд специфичных, например, из области психометрики. В статье дается обзор методов Educational Data Mining, применительно к анализу потока данных, порождаемого при взаимодействии пользователя с инструментальными средами, лежащими в основе научных игр образовательного назначения.

Ключевые слова: Educational Data Mining, серьёзные (научные) игры, анализ протоколов, электронное обучение.

Термин Data Mining (DM) введён Григорием Пятецким-Шапиро и определяется как исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [6].

Основными методами DM являются классификация, регрессия, поиск ассоциативных правил и кластеризация [4]:

1. Классификация определяет класс объекта по его характеристикам. Фактически, происходит определение значения зависимой переменной объекта на основании других переменных, характеризующих данный объект. При этом, множество классов, к которым может быть отнесен объект, известно заранее. В случае задачи регрессии вместо конечного множества классов, используется множество действительных чисел.

2. При поиске ассоциативных правил целью является нахождение частых зависимостей (ассоциаций) между объектами или событиями. Найденные зависимости представляются в виде правил и могут быть использованы как для лучшего понимания природы анализируемых данных, так и для предсказания появления событий. Разновидностью задачи поиска ассоциативных правил является секвенциальный анализ, целью которого является установление отношения порядка между исследуемыми наборами.

3. Задача кластеризации заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных. В отличие от задачи классификации разбиение объектов по кластерам происходит одновременно с их формированием, заранее они не известны. Для определения сходства объектов вводится меры близости – расстояние.

Educational Data Mining (EDM) – применение методов DM для анализа данных, порождаемых образовательными процессами

с целью решения образовательных задач, таких как, адаптация курса обучения под конкретного обучаемого, улучшение понимания процесса обучения и т. д. [8]

Выделяют ряд важных особенностей, которые отличают применение DM в образовании от применения DM в других областях [1]:

1. *Цель.* Цель DM отличается в различных областях. В EDM выделяют две основные цели: прикладная – улучшение процесса обучения и определение направления для обучаемых и фундаментальная – более глубокое понимание процесса познания. Эти параметры сложно измерить количественно и они требуют собственного специального набора метрик.

2. *Данные.* В сфере образования огромное количество данных, из которых можно извлекать знания. Необходимо учитывать педагогические аспекты обучаемого и систем обучения при извлечении информации.

3. *Методы.* Образовательные данные и проблемы обладают рядом характеристик, которые требуют специальных методов обработки. Хотя некоторые методы DM могут быть применены напрямую, другие не могут и требуют адаптации. Более того, специфические методы могут быть использованы для решения специфических задач EMD.

Важной особенностью протоколов работы как данных, подлежащих анализу, является отношение порядка между событиями. Исходя из этого можно выделить две области DM, в рамках которых возможен анализ протоколов: секвенциальный анализ, который решает задачу поиска наиболее частых последовательностей событий, и анализ процессов (process mining), решающий задачу построения модели процесса на основе журнала событий.

СЕКВЕНЦИАЛЬНЫЙ АНАЛИЗ

Секвенциальный анализ работает с последовательностями происходящих событий. Последовательностью называется упорядоченное множество объектов. Для этого на множестве должно быть задано отношение порядка.

$$S = \{\ldots, i\_p, \ldots i\_q | i\_p < i\_q\}.$$

Транзакцией называется одна из множества всех анализируемых последовательностей. Транзакция T содержит последовательность S, если объекты, входящие в S, входят в T с сохранением отношения порядка. При этом в последовательности T между объектами из последовательности S могут быть другие элементы.

Поддержкой последовательности S называется отношение количества транзакций, в которое входит последовательность S, к общему количеству транзакций.

Задачей секвенциального анализа является поиск всех частых последовательностей, то есть тех, для которых уровень поддержки превышает некоторое минимальное значение.

Наиболее распространенными алгоритмами секвенциального анализа являются AprioriALL и GSP.

Трудностью применения секвенциального анализа к протоколам работы обучаемых является неизвестность «стратегии», используемой обучаемым. В отличие от «тактики», которую условно можно сопоставить техническим записям, «стратегию» трудно предсказать.

Схема анализа протоколов, предложенная в [9], состоит из трех этапов:

1. *Предварительная обработка* (preprocessing). Входными данными на этом шаге являются «сырые» протоколы, детально отражающие технические события и содержащие записи, отсортированные в порядке времени наступления событий. Как правило, исходные протоколы содержат много избыточной информации и анализировать из на поведенческом уровне невозможно. При помощи механизма разбора, работающего на основе библиотеки действий, исходные протоколы преобразуются в протоколы более высокоуровневых событий. Библиотека действий создается исследователем и позволяет выделять нужный семантический уровень для извлечения новых знаний.

2. *Выделение шаблонов* (pattern discovery). Целью данного шага является выделение значимых шаблонов на основе секвенциального анализа. Значимым считается

шаблон, который можно интерпретировать в контексте исследования. В зависимости от способа применения секвенциального анализа к данным, будет отличаться контекст. Например, можно использовать протоколы работы одного обучаемого в различные сессии или протоколы обучаемых в одну сессию из определенной группы.

3. *Анализ полученных шаблонов, проверка гипотез*.

Без предварительной обработки алгоритмы Process Mining извлекут огромное количество последовательностей, которые не будут нести какую-то полезную информацию.

PROCESS MINING

Применение методов Data Mining для анализа информации о реальных процессах, выполняемых системами, автоматизирующими бизнес-процессы, получило в литературе название Process Mining (технология построения формальных моделей экземпляров процессов по протоколам работы систем). Часто в литературе также встречается и понятие Workflow Mining (технология выявления часто встречающихся экземпляров процессов (шаблонов) из протоколов работы систем) [4]. Методы Process Mining применяются к протоколам работы информационных систем. В них отражается реальное выполнение бизнес-процессов через взаимодействие их исполнителей с информационными системами. Применение к ним методов Process Mining позволяет автоматически построить модели бизнес-процессов. Построенные таким образом модели бизнес-процессов отражают реальность и доступны для восприятия и анализа человеку.

К основным задачам PM относят:

- Построение модели процесса на основе протоколов работы.
- Проверку адекватности модели существующим протоколам работы.
- Улучшение существующей модели на основе протоколов работы.

Анализ процессов не сводится к выявлению пути, фокус может быть не на последовательности, а, например, на времени прохождения. Особенностью PM присущая процессам параллельность и точки выбора, игнорируемые большинством алгоритмов из data mining.

Существует ряд алгоритмов для извлечения процессов: эвристический (Heuristic miner), альфа-алгоритм и его модификации, нечеткий (Fuzzy Miner), генетические алгоритмы, многофазный (Multi-phase miner).

Выходная модель может быть представлена сетью Петри, событийной цепочкой процесса (EPC-диаграмма) и другими представлениями в зависимости от алгоритма [12].

Для оценки полученной модели процесса используют 4 конкурирующих между собой критерия [13]:

1. *Соответствие* (fitness) – возможность «проиграть» существующие протоколы на построенной модели.

2. *Простота* (simplicity) – модель не должны быть слишком нагруженной (иногда полученные модели имеют вид «спагетти»).

3. *Точность* (precision) – допустимое моделью поведение, которое не соответствует существующим протоколам, должно быть минимально.

4. *Обобщение* (generalization) – модель должна быть в состоянии воспроизвести будущее поведение процесса.

Для анализа средствами PM протоколы должны удовлетворять следующим требованиям [4]:

– все события, зафиксированные в протоколе, должны быть идентифицированы с экземплярами процессов;

– все события должны быть упорядочены по времени выполнения;

– разнотипные события должны различаться.

Существуют стандарты записи протоколов, поддерживаемые программными инструментами PM, например, ProM [9]: MXML [10] и XES [11]. Оба стандарта записи протокола являются расширением XML.

Корневым узлом MXML-документа является «WorkflowLog», соответствующий файлу протокола. Элемент WorkflowLog может содержать произвольное количество

элементов «Process» – процессов. Однократные выполнения этого процесса представлены экземплярами процесса – элементами «ProcessInstance». Каждый ProcessInstance соответствует однократному протеканию процесса. Конкретные события в рамках протекания процесса представляются элементами «AuditTrailEntry», обязательными дочерними элементами которых являются «WorkflowModelElement» – задача, которая была выполнена и «EventType» – тип события, который описывает стадию выполнения задачи. Кроме того, дополнительно могут указываться «Timestamp» (точная дата и время) и «Originator» (инициатор события, может быть человек или информационная система). Расширяемое поле Data может содержать произвольное количество атрибутов, которые являются парами.

В документе XES корневым элементов является «Log», в котором располагаются элементы-экземпляры процессов – «Trace». «Trace» содержит события в рамках конкретного протекания процесса – «Event». Все элементы могут содержать произвольное количество атрибутов различных типов. Формат XES является более современным.

Важным этапом при обработке протоколов является предварительная обработка. Как показано в [2] и [3], использование специальных способов предварительной обработки протоколов позволяет получить более информативные модели.

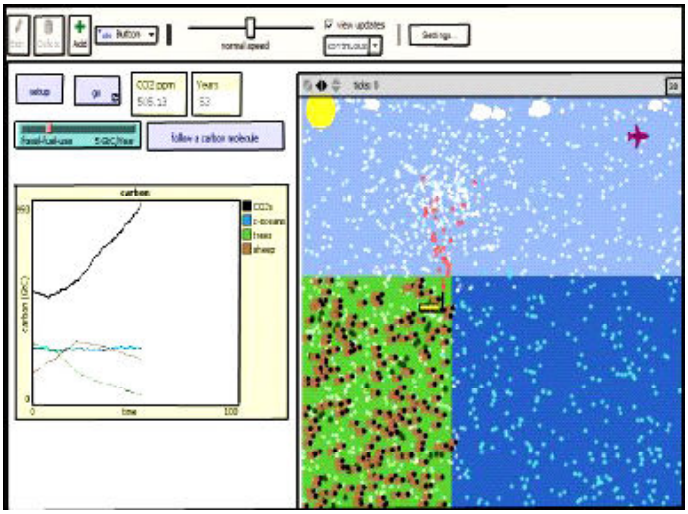

Рис. 1. Модель круговорота углерода в природе

В работе [3] рассматривается замещение исходных записей в протоколах более абстрактными сущностями. Предложены общие шаблоны и способы их выделения: циклические конструкции и подпроцессы. Преобразованный протокол будет состоять из более высокоуровневых сущностей – деятельностей («activities»).

Несколько измененный подход применялся в [2], здесь набор технических событиях сегментировался в пакеты, после чего использовалась классификация на основе эвристики для выделения деятельностей. Для анализа использовались протоколы работы, порожденные при взаимодействии обучаемых с моделью изменения климата в среде NetLogo [5]. NetLogo – свободно распространяемая программа для разработки и изучения агентно-ориентированных моделей. Модель круговорота углерода в данной среде представлена на рис. 1.

Обработка протоколов разбивалась на три этапа. На первых двух этапах производилось выделение последовательности действий из последовательности событий, на третьем – выделение последовательности деятельности из действий.

Каждый файл включал записи событий экземпляра процесса (например, нажатие кнопки click). На первом этапе протокол был преобразован в другой протокол, в котором было зафиксировано семь технических событий. В частности, нажатие кнопки «Go On/Off» было разделено на различные события: «Start» и «Stop», а изменение скорости не учитывало направление и считалось одним техническим событием. Далее, используя семантические события «Start», «Stop» и «Setup», были выделены непересекаемые последовательности событий.

На втором шаге каждой выделенной последовательности событий был присвоен идентификатор. Каждому событию, входящему в последовательность, соответствовала двоичная цифра: «1», если событие происходило один или более раз в последовательности и «0»,

если нет. Порядок событий при этом не учитывался. Таким образом, все выделенная последовательность рассматривалась как некоторый «мешок событий» («a bag of events»).

На последнем шаге определялись виды деятельности, используя три особенности: тип действия, тип события и продолжительность. Тип действия определялся расположением выделяющих последовательность действия разделителей «start» и «stop» и имел значения «pause» или «run». Тип события включал 4 эвристические категории: управление (включал нажатия запуска, паузы и сброса настроек), взаимодействие (настройка параметров симуляции), конфигурация (изменение параметров модели) и комбинированный тип, включающий первые три. В результаты было получено 9 видов деятельности.

В работе [2] выполнено построение диаграмм процессов на основе 3 логов: событий, действий и деятельностей. При этом показано, что использование больших семантических блоков и изучение их последовательности позволяет получить дополнительные сведения о том, как обучаемые взаимодействуют с компьютерной моделью.

SERIOUS GAMES

В последнее время большой интерес вызывают серьезные игры, с одной стороны, как активно растущий рынок, на который постоянно выходят новые разработки, с другой стороны, как область академических исследований, в которой увеличивается количество публикаций [14], [15]. Появляются методы создания серьезных игр, позволяющие получить конечный продукт без навыков разработки программного обеспечения [17].

Идея использования игр в целях, отличных от развлечений, была впервые сформулирована в [16]. Игры можно рассматривать как серьезные в том смысле, что в них заложена явная и тщательно продуманная образовательная цель, и они не предназначены в первую очередь для развлечения [16]. При этом образовательная цель не обязательно должна быть встроена в игру, а может быть связана с контекстом использования игры.

В настоящие время серьезные игры рассматриваются как потенциально ценный образовательный инструмент.

Обзор исследований в области оценки эффективности серьезных игр и оценки действий пользователя в них дан в [13].

Оценка обучения и тренировки требует системного подхода для определения достижений обучаемого и трудных для понимания областей. Подходы, основанные на использовании вспомогательных технологий, используются на протяжении многих лет для оценки успеваемости учащихся, благодаря потенциалу оптимизации процесса использования стандартизированных тестов, упрощению вычисления результатов обучения и составления отчетности. Большое множество существующих инструментов для проектирования тестов и анализа результатов тестирования можно разделить на три категории [13]:

- *Системы управления оцениванием* (Assessment management systems) – системы поддержки преподавателей для создания, администрирования, оценки и анализа тестов.
- *Инструменты оценивания ответов на естественных языках* (Tools for natural language answer assessment) – инструменты, позволяющие оценивать ответы, данные в свободной письменной форме.
- *Аудиторная система отклика* (Classroom response system) – интерактивные системы, позволяющие учителям моментально оценивать обучение в классе.

При всех плюсах подход, основанный на тестах, имеет ряд ограничений, связанных с оценкой решения сложных задач, взаимодействия, навыков рассуждения.

Альтернативой является оценка на основе игры (play-based), или встроенная в игру (in-game), которая, будучи лишенной недостатков тестов, может предоставить более содержательную и надежную информацию об обучении [13]. Традиционно под оцениванием на основе игры понимается анализ того, что делает ученик в процессе игры, для определения когнитивного развития. Этот подход, в отличие от тестового, не отвлекает внимание от игры, что потенциально мо-

жет понизить интерес к ней. Большим преимуществом цифровых игр является возможность отслеживать все действия пользователя [13].

Как отмечают в [16], серьезные игры (и игры в целом) могут и, как правило, содержат в игре проверку эффективности. Более конкретно, по мере прохождения игры, участники накапливают баллы и опыт, который позволяет перейти на следующие этапы и столкнуться с более сложными задачами.

По определению серьезные игры содержат модель оценки эффективности действия пользователя, так как очки должны коррелировать с решением содержательных задач. Отличительной особенность любой качественной серьезной игры должна быть соответствие игровых механизмов – оценки, уровни, бонусы – педагогическим целям. Игрок в итоге должен для обучения просто пытаться преодолеть игровые задачи.

В [13] выделяется два направления исследований в области оценки действий игрока: характеристика деятельности игрока и лучшая интеграция оценки в игру. Характеристика деятельности игрока включает характеристику игры (содержание, уровень сложности и т. д.) и профиль пользователя, содержащий индивидуальную информацию (перспективным представляется анализ нейрофизиологических сигналов). Важной особенностью профиля должна быть переносимость между различными играми. Лучшая интеграция оценки в играх, по существу, является задачей определения надлежащих механизмов и условий для их активации. Важно, чтобы эти механизмы были общими и модульными, что позволит их легко применять в различных играх.

ВЫВОДЫ

Как было показано, существует весьма богатый алгоритмический аппарат для анализа последовательных событий, генерируемых в процессе взаимодействия пользователя с интерактивными системами. Кроме того, существует ряд доступных программных средств, реализующих соответствующий аппарат, и которые можно использовать, преобразовав данные протоколов к стандартному виду.

Важной задачей является предварительная обработка протоколов, выделение событий более высокого уровня, чем технические события. Без предварительной обработки секвенциальный анализ будет генерировать большое количество последовательностей, а результатом анализа процесса будут неинформативные «спагетти-модели».

Process mining ориентирован на обработку большого количества протоколов протекания различных случаев (кейсов) одного и того же процесса. При обучении процесс для разных обучаемых будет различным. Даже при анализе протоколов обучения в рамках большой обучающей системы, многое «остается за кадром». Различные сессии одного пользователя также представляются разными процессами.

Литература

1. *Romero C., Ventura S.* Educational Data Mining: A Review of the State-of-the-Art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 2010. Vol. 40. Is. 6.

2. *Southavilay V. and others.* From “Events” to “Activities”: Creating Abstraction Techniques for Mining Students’ Model-Based Inquiry Processes / The 6th International Conference on Educational Data Mining, 2013.

3. *Bose R.P.J.C., Aalst W.* Abstractions in process mining: A taxonomy of patterns / 7th International Conference on Business Process Management, 2009.

4. *Баганесян А.А. и др.* Анализ данных и процессов 3-е издание. БХВ-Петербург, 2009.

5. NetLogo / http://ccl.northwestern.edu/netlogo/ (дата обращения 18.12.2013).

6. *Fayyad U., Piatetsky-shapiro G., Smyth P.* From Data Mining to Knowledge Discovery in Databases // American Association for Artificial Intelligence, 1996.

7. *Baker E. and others.* International Encyclopedia of Education (3rd edition). Elsevier, 2010

8. *Barnes T., Desmarais M., Romero C., Ventura S.* Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 2009.

9. *Romero C. and others.* Handbook of Educational Data Mining. CRC Press, 2011.
10. Стандарт записи протоколов MXML / http://www.processmining.org/WorkflowLog.xsd (дата обращения 18.12.2013).
11. Стандарт записи протоколов XES / http://www.xes-standard.org/ (дата обращения 18.12.2013).
12. *Aalst W.* Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer, 2011
13. *Bellotti F. and others.* Educational Data Mining: Assessment in and of Serious Games: An Overview. Advances in Human-Computer Interaction. Hindawi, 2013.
14. *Breuer J., Bente G.* Why so serious? On the Relation of Serious Games and Learning. Journal for Computer Game Culture. Eludamos, 2010.
15. *Abt C.C.* Serious Games. Viking Compass,1975.
16. *Becker K., Parker J.R.* The Guide to Computer Simulations and Games. John Wiley & Sons, 2011.
17. **Tang S., Hanneghan M.** A Model Driven Serious Games Development Approach for Game-based Learning / International Conference on Software Engineering Research and Practice, 2013.

REVIEW OF EDUCATIONAL DATA MINING METHODS AS APPLIED TO INTERACTION PROTOCOLS ANALYSIS IN «SCIENTIFIC GAMES»

Abstract

One of tendencies of development of electronic training is introduction of interactive training systems. Interaction of the trainee with interactive trained system generates a huge data file which can be used for correction of educational process. For the solution of this task the Educational Data mining methods can be used. Educational Data Mining (EDM) is young interdisciplinary science which is engaged in development of methods for research of the data arising in an educational context. Educational Data Mining uses both the standard Data mining methods, such as a clustering, classification, regression, correlation, visualization, etc., and a row specific, for example, from psychometrics area. The article contain review of the Educational Data mining methods for analysis of the data flow generated at interaction of the user with scientific games.

Keywords: Educational Data Mining, serious (academic) games, logs analysis, e-learning.

Акимушкин Василий Александрович,
магистр СПбГЭТУ «ЛЭТИ»,
аспирант математико-механическрго факультета СПбГУ,
vasiliy.akimushkin@gmail.com,

Поздняков Сергей Николаевич,
профессор кафедры ВМ-2
СПбГЭТУ «ЛЭТИ»
pozdnkov@gmail.com.