Выбор метода обнаружения аномалий в образовательных данных

А.С. Кузнецов, Е.Ю. Семенов

***Аннотация* — В статье рассматриваются подходы к выбору методов интеллектуального анализа в образовательных данных. Авторами обозначена проблема оценки эффективности алгоритмов обнаружения аномалий в данных в случае невозможности априорной классификации объектов. Предложена модель обоснованного выбора метода на основе статистической оценки выходных данных.**

***Ключевые слова* — интеллектуальный анализ данных, интеллектуальный анализ данных в образовании, обнаружение аномалий, OSVM, Random Forest, Elliptic Envelope.**

I. Введение

Общий вектор повышения уровня информатизации, а также сбора и накопления данных во всех сферах человеческой деятельности неизбежно требует увеличения доступности практических реализаций методов интеллектуального анализа данных (ИАД). Все это в последние годы привело к значительному расширению области их практического применения.

Использование методов ИАД легло в основу целых научных направлений, а потребность в исследованиях и специалистах в данной области в последнее десятилетие достигла огромных по меркам специализированной технической дисциплины значений.

К базовым задачам ИАД относятся [1]:

1. Кластеризация.
2. Классификация.
3. Регрессия.
4. Поиск ассоциативных правил.
5. Обобщение.
6. Обнаружение аномалий.

Обнаружение аномалий является одной из наиболее перспективных для изучения проблем, так как алгоритмы, решающие задачи в данной области могут использоваться не только как самостоятельные единицы, но и в качестве инструментов предобработки перед решением остальных базовых задач ИАД. Наиболее широко данные методы применяются в следующих областях:

1. Информационная безопасность (выявление мошеннических операций в банковской сфере и обнаружение угроз в компьютерных системах).
2. Медицина (диагностика заболеваний).
3. Производство (автоматизированный мониторинг технологических процессов и контроль качества).

Статья получена 27 февраля 2017.

А.С. Кузнецов, преподаватель, Орловский юридический институт МВД России им. В.В. Лукьянова, (e-mail: kuznetsov_as@ bk.ru).

Е.Ю. Семенов, старший преподаватель, Орловский юридический институт МВД России им. В.В. Лукьянова, (e-mail: john_neg@mail.ru).

II. Сбор и интеллектуальный анализ данных в образовании

Образовательная среда является большой областью, где целесообразно использование методов ИАД в различных формах. На сегодняшний день наибольшее развитие данное направление получило в странах и образовательных организациях, где активно внедряется и используются технологии электронного обучения. В исследованиях [2]-[5] подробно описываются различные сценарии применения методов ИАД в образовании.

В работах последних лет большое внимание отводится применению алгоритмов классификации и регрессии для решения задач прогнозирования успеваемости обучающихся, а также формирования индивидуальной образовательной траектории.

Правильная реализация сбора и ИАД в образовательных организациях даже при отсутствии интегрированного подхода к использованию электронного обучения и специализированных систем управления обучением (Learning Management System) могут значительно повысить прозрачность протекающих процессов.

Применение методов обнаружения аномалий для образовательных данных может позволить более эффективно решать задачу управления учебным процессом [6]. К примеру, на сегодняшний день функции контроля учебного процесса слабо автоматизированы и лежат в области экспертного анализа. Разработка систем поддержки принятия решений, основанных на интеллектуальном анализе многомерных наборов данных о ходе учебного процесса может повысить точность оценки текущего состояния объектов учебного процесса.

III. Задача выбора метода обнаружения аномалий для образовательных данных

Выбор подходящего метода обнаружения аномалий для образовательных данных является достаточно нетривиальной задачей в силу специфики анализируемых объектов. Необходимо отметить, что сам по себе процесс выявления нехарактерных для заданного массива данных объектов, то есть аномалий, является частным случаем бинарной классификации. В

данном случае выбранный метод должен промаркировать объекты массива, разделив их на две группы: «аномальные» и «нормальные».

Принятие решения об использовании конкретного метода обнаружения аномалий должно опираться на соответствующие научное обоснование. Стоит отметить, что эффективность работы конкретного алгоритма сильно зависит от характера анализируемых данных. По этой причине, зачастую, работы, посвященные оценке качества алгоритмов, применительны к определенной сфере и к определенным наборам данных.

A. Классическая оценка качества алгоритмов обнаружения аномалий

Классическим методом оценки качества алгоритмов бинарной классификации является построение так называемой ROC-кривой или кривой ошибок [7]. Она показывает отношение между долей верно классифицированных аномальных объектов от общего числа аномальных объектов, и долей ошибочно определенных объектов как аномальных от их общего числа, являющимися нормальными. Количественным показателем качества анализируемого алгоритма в данном методе является величина площади, ограниченной ROC-кривой и осью доли ложных положительных классификаций AUC (under ROC curve) (рисунок 1).

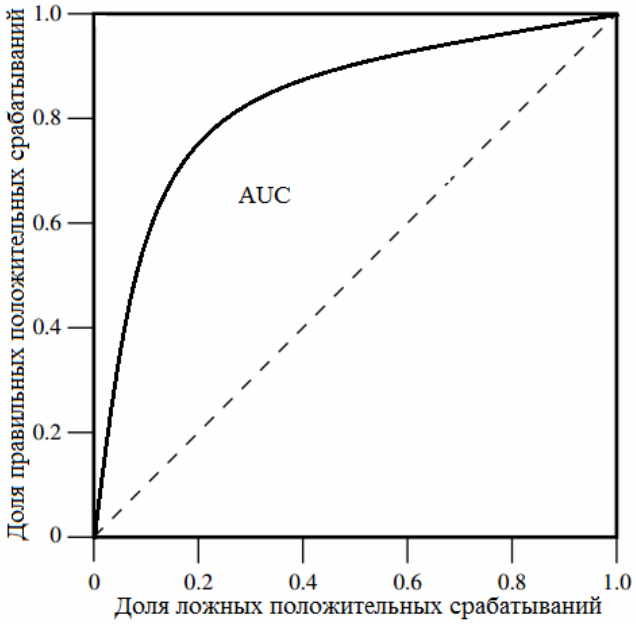

Рисунок 1. Пример кривой ошибок.

Данный метод позволяет достаточно полно оценить работу алгоритмов в большинстве случаев, однако он неприменим для оценки точности в случае анализа образовательных данных в силу специфики рассматриваемых объектов. Расчет цены ошибок первого и второго рода в данном случае невозможен из-за отсутствия возможности априорной проверки гипотезы об аномальности объектов.

Приведенный выше тезис легко пояснить на примере использования методов обнаружения аномалий для выявления мошеннических операций в банковской сфере, где факт мошенничества может быть однозначно установлен в случае маркировки события как аномального, так и в случае признания системой его нормальности.

Оценка человеческой деятельности с помощью формальных параметров является сложной задачей даже в случае выбора оптимальной комбинации таких параметров. Метрикой аномальности объектов в этом случае при верификации являются их параметры, а не фактическая принадлежность их самих к классу аномальных. Например, в случае оценки успеваемости низкие показатели в учебной деятельности или большое количество пропусков сами по себе являются критерием аномальности и сигналом к тому, что необходимо вмешательство для установления причин и коррекции ситуации.

B. Дополнительные условия отбора методов

Отсутствие однозначной априорной классификации не только затрудняет оценку эффективности методов ИАД, но и значительно сужает круг их выбора до группы алгоритмов «обучения без учителя», так как сформировать необходимый тренировочный массив данных для групп методов «обучение с учителем» и «обучение с частичным привлечением учителя» практически невозможно.

Помимо перечисленных выше особенностей выбора методов обнаружения аномалий для образовательных данных присутствуют ограничения, связанные с необходимостью практической реализации разработанной модели применения в виде законченной системы поддержки принятия решений. Эти ограничения заставляют отказаться от некоторого числа алгоритмов, обладающих высокой ригидностью в вопросе их программной интеграции в состав разрабатываемой системы из-за большого числа входных параметров.

Фактическим решением задачи об обнаружении аномалий при изложенных выше условиях является построение некоторой гиперплоскости в многомерном пространстве, являющейся решающей границей, где лежащие внутри области объекты будут считаться нормальными, а остальные аномальными. В качестве исходных параметров указывается предполагаемый процент аномальных объектов в заданном массиве.

IV. Оценка эффективности выбранных методов обнаружения аномалий

В качестве кандидатов для окончательного выбора были отобраны следующие методы обнаружения аномалий без учителя, соответствующие описанным выше требованиям:

1. Метод эллиптической огибающей (Elliptic Envelope), использующей ковариационную оценку с расстоянием Махалонобиса. Данный подход показывает неплохие результаты в классических оценках качества для массивов данных близких к гауссовскому распределению. Полное описание можно найти в исследовании [8].

2. Одноклассовый метод опорных векторов (OSVM). Сущность данного подхода заключается в

использовании модифицированного метода опорных векторов, применяемого для решения задач классификации. Подробное описание и необходимые математические выкладки приведены в исследовании [9].

3. Метод изолирующих деревьев (Isolation Forest). В отличии от многих методов обнаружения аномалий изначально разработан для этих целей. Основное отличие от других методов заключается в изолировании выбросов, а не в расчете профиля нормальных объектов. Концепция подхода, а также сравнение эффективности с методами ORCA и LOF приведены в работе [10].

A. Вычисление показателей эффективности выбранных алгоритмов

Предлагаемые этапы процесса оценки эффективности отобранных методов приведены на рисунке 2.

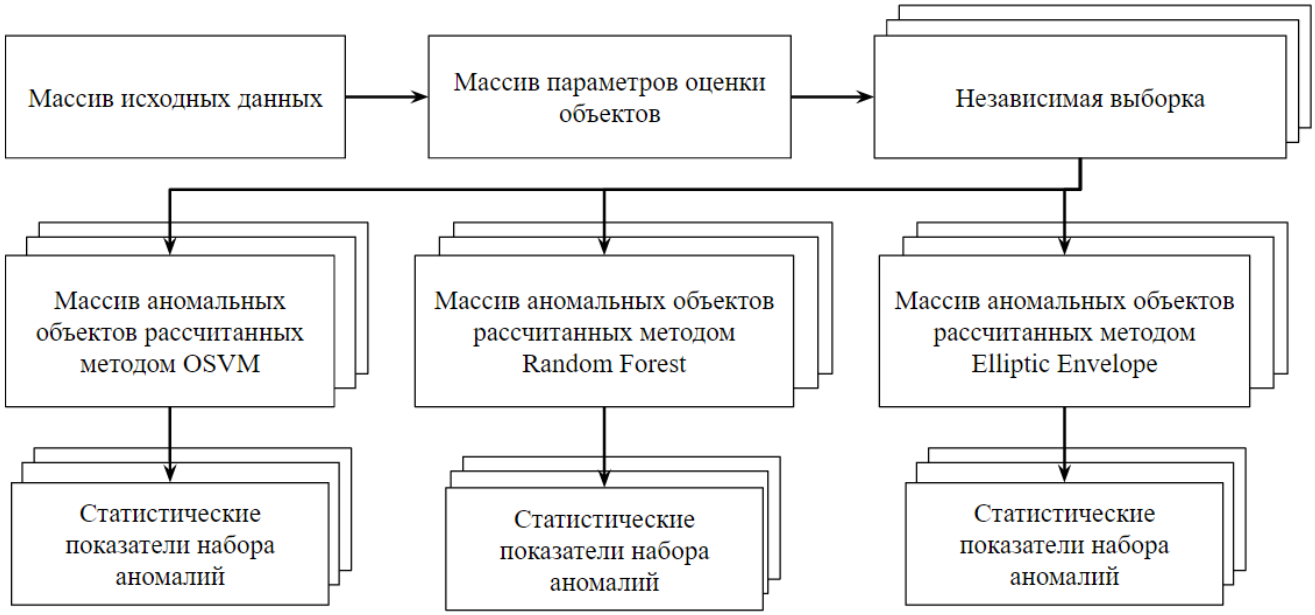

Рисунок 2. Этапы вычислений для оценки эффективности методов обнаружений аномалий.

Сравнение эффективности методов было выполнено на массиве данных полученным путем экспорта из базы данных локальной системы учета успеваемости.

На первом этапе производится формирование массива заданных параметров и его последовательное хронологическое разбиение на несколько независимых выборок путем запросов к полю даты события. Для данного случая было выбрано помесячное разделение одного учебного года.

Объектами оценки были выбраны обучающиеся. В качестве параметров объектов были выбраны четыре переменных, характеризующих их участие в учебном процессе:

1. Средний балл.
2. Число оценок.
3. Число неудовлетворительных оценок.
4. Число пропусков занятий.

На втором этапе производится необходимая предобработка – стандартизация, которая заключается в приведении исходного массива к набору сравнимых элементов из распределения с нулевым средним и среднеквадратическим отклонением, равным единице.

Затем, к полученным наборам данных последовательно применяются выбранные для анализа методы обнаружения аномалий.

В качестве общего параметра для всех алгоритмов было выбрано допущение о том, что 5% всех объектов являются аномальными.

Для метода OSVM в расчетах использовалось ядро радиальных базисных функций Гаусса, как наиболее универсальное для решения задач ИАД.

В качестве выходных данных алгоритмы должны показать заданный процент обучающихся, имеющих выдающиеся комбинации выбранных переменных. Важно отметить, что характер девиаций не всегда носит сугубо отрицательный характер: достаточно часто встречаются положительные аномалии с очень высоким средним баллом и малым количеством пропусков.

Заключительным этапом является вычисление статистических показателей вычисленных наборов аномальных значений для каждого из трех методов.

B. Вывод и интерпретация полученных результатов

Представляется, что вычисление среднеквадратического отклонения (СКО) для полученных наборов аномалий по всем одномерным срезам многомерной независимой выборки является объективной мерой экстремальности найденных значений. Математическое описание СКО представлено выражением:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad (1)$$

Исходя из математической интерпретации смысла СКО можно сделать вывод, что его наибольшее значение будет характеризовать набор аномальных значений метода как лучший среди вычисленных, а развертка по каждому из одномерных массивов, вычисленная по нескольким независимым выборкам, даст необходимую репрезентативность.

Количественную оценку СКО полученных для определения наиболее эффективного метода

обнаружения аномалий следует провести по каждой переменной отдельно ввиду их различной природы в массивах признаков объектов.

На рисунке 3 представлены графики конечных значений СКО одномерных наборов для нескольких независимых выборок для алгоритмов Isolation Forest, OSVM и Elliptic Envelope для каждой из выбранных переменных.

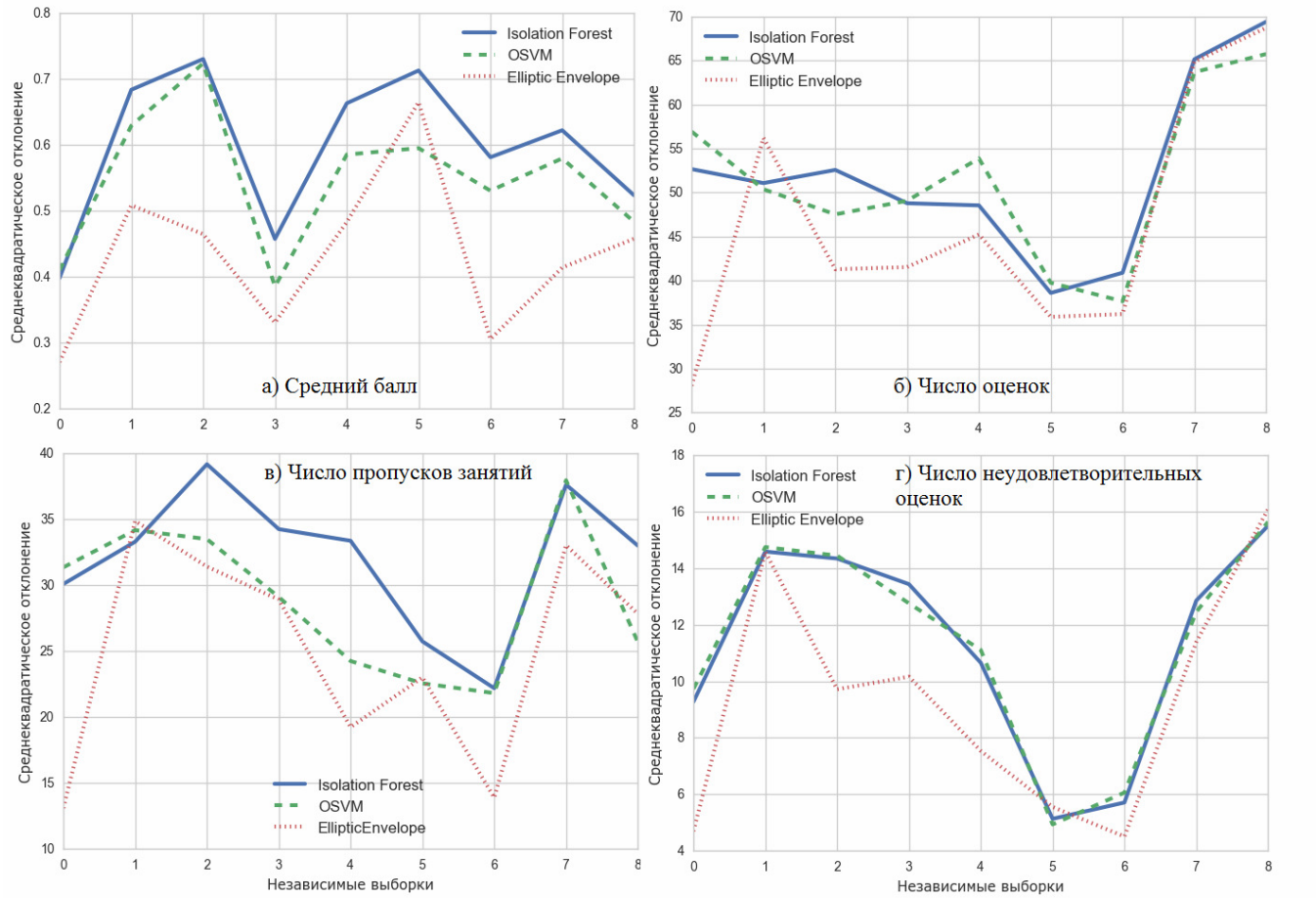

Рисунок 3. Результаты вычислений статистических характеристик массивов аномалий, полученных разными методами

Из представленных графиков можно сделать вывод о том, что алгоритм Isolation Forest, учитывая принятую гипотезу, в большинстве случаев оказался лучше конкурентов при анализе среднего балла и пропусков (рисунок 3-а, 3-в). Метод OSVM, являющийся достаточно универсальным и проверенным показал себя практически также хорошо. Алгоритм Elliptic Envelope показал худшие результаты по всем переменным.

Стоит отметить, что характер распределения значений переменных для аномальных объектов значительно повлиял на конечные расчеты, например, для числа неудовлетворительных оценок граница решающей функции для алгоритмов Isolation Forest и OSVM практически совпали, что отразилось на значениях статистических характеристик (рисунок 3-г).

В целом полученные результаты коррелируют со сравнительными тестами выбранных алгоритмов при использовании классического метода вычисления AUC.

V. Заключение и выводы

Использование методов ИАД позволяет решать сложные задачи в различных сферах человеческой жизни. Разработка подходов к применению таких методов в образовательных системах могут позволить повысить эффективность как системы управления учебным процессом, так и самих учебных курсов.

Системы, построенные на базе алгоритмов обнаружения аномалий в данных, успешно решают различные задачи в различных сферах человеческой деятельности. Их применение для анализа образовательных данных классических образовательных систем требует глубокого изучения. В статье удалось сформулировать особенности выбора таких методов с учетом невозможности однозначной классификации объектов.

Результаты проведенного исследования и предложенная модель выбора метода обнаружения аномалий на основе статистической оценки полученных массивов аномальных объектов могут помочь как исследователям, так и практикам при выборе подходящей основы для создания аналитических инструментов.

Библиография

[1] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.

[2] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1.1 (2009): 3-17.

[3] Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146.

[4] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40.6 (2010): 601-618.

[5] Baker, Ryan Shaun, and Paul Salvador Inventado. "Educational data mining and learning analytics." Learning analytics. Springer New York, 2014. 61-75.

[6] Кузнецов А.С., Семёнов Е.Ю. "Некоторые подходы к применению анализа данных в управлении учебным процессом" Информационные системы и технологии 6 (2016): 25-29.

[7] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[8] Rousseeuw, Peter J., and Katrien Van Driessen. "A fast algorithm for the minimum covariance determinant estimator." Technometrics 41.3 (1999): 212-223.

[9] Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." Statistics and computing 14.3 (2004): 199-222.

[10] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008.

The choice of anomaly detection method for educational data

A.S. Kuznetsov, E.Y. Semenov

Abstract — **This paper discusses the selection of data mining methods for educational data. The authors describe the efficiency estimation problem for anomaly detection methods in case of impossibility of objects a priori classification. In the paper proposed a verified model of choosing anomaly detection method based on the statistical evaluation of the output.**

Keywords — **data mining, educational data mining, anomaly detection, OSVM, Random Forest, Elliptic Envelope.**