УДК 004.052.42 DOI: 10.18698/0236-3933-2016-5-75-85

ВЫЯВЛЕНИЕ АНОМАЛИЙ ПРИ ПРОГНОЗНОМ АНАЛИЗЕ ДАННЫХ

В.И. Кузовлев
А.О. Орлов forewar@gmail.com

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Рассмотрены проблемы, возникающие при построении моделей в прогнозном анализе данных с учетом наличия в них аномальных выбросов. Обоснован выбор метода выявления аномалий и его применение в алгоритме построения прогнозной модели дерева решений. Описаны этапы работы этого алгоритма, методика поиска аномалий в данных. Приведено смысловое описание параметров настройки поиска и их принципиальное влияние на результат работы методики. Представлены результаты совмещения методики поиска аномалий с алгоритмом построения модели дерева решений, выраженные в повышении точности прогнозной модели за счет повышения устойчивости к выбросам в данных, а также в значительном повышении производительности анализа

Ключевые слова

Аномалии, выбросы в данных, прогнозный анализ, модель дерева решений

Поступила в редакцию 22.06.2015

Введение. В системах поддержки принятия решений (СППР) важное место занимают механизмы прогнозного анализа данных [1]. Прогнозный анализ данных является процессом формирования суждений о будущих фактах на основе обработки и анализа исходного набора статистических данных, называемого обучающим множеством, или генеральной совокупностью. Результат обучения — аналитическая модель, используемая в дальнейшем при формировании прогнозов. Серьезным препятствием при построении прогнозной модели может быть наличие шумов в исходных обучающих данных. Вызванные шумом искажения влияют на процесс построения прогнозной модели, а также на качество ее работы, выражающееся в точности распознавания объектов при прогнозировании. В конечном счете искажения в исходных данных снижают эффективность работы СППР, влияя на решения и управляющие оперативные воздействия, формируемые системой [2].

Задачей, которую ставят перед собой авторы настоящей работы, является исследование и разработка методик выявления аномалий в исходных данных, на которых строятся прогнозные модели. Широкий обзор существующих подходов к решению проблемы обнаружения аномалий приведен в работе [3], в которой существующие методы разбиты на несколько категорий по общему характеру. Методы каждой категории имеют достоинства и недостатки и, по мнению авторов, должны выбираться в зависимости от специфики предметной области отдельно взятой задачи.

Процесс построения СППР и, в частности, прогнозной модели анализа данных начинается с обучения модели на исходных данных, поэтому рассмотрим методы, способные работать на этапе обучения модели, а именно методы, основанные на широко известном подходе *k* ближайших соседей, когда объекты анализируют совместно с другими объектами, ближайшими к ним. Ключевая проблема при обнаружении аномалий — поиск расстояний между объектами данных, поскольку в системах принятия решений используют не только числовые данные, шкалы измерений которых часто заранее известны, но и категориальные данные, выраженные в вербальной форме, что затрудняет их сравнение. Обзор существующих критериев оценки расстояний между значениями категориальных атрибутов данных проведен в работе [4], также в этой работе выбран оптимальный критерий оценки расстояния. Проблемой указанного критерия является его зависимость от общего числа объектов данных. Это затрудняет расчет расстояний в динамических системах, объекты данных в которые могут попадать в процессе работы систем, а не только на начальном этапе формирования модели анализа.

Цель работы — анализ и описание методики обработки шума в данных и основанного на ней алгоритма построения дерева решений, позволяющего преодолеть следующие проблемы, имеющиеся в существующих алгоритмах построения моделей деревьев решений:

1) проблема наличия разнородных искажений в данных;
2) проблема выбора эффективной стратегии повышения качества данных.

Разработанный алгоритм должен обрабатывать искажения двух типов: аномальные значения атрибутов данных; отсутствующие значения. Для обработки аномальных значений необходимо использовать методы поиска аномалий в данных, для обработки отсутствующих значений — алгоритмы заполнения пропусков в данных. В настоящей работе показано, насколько успешно можно применять алгоритмы поиска выбросов в прогнозных моделях.

Научная новизна работы заключается в разработке подхода к решению задачи выявления аномалий на этапе построения модели принятия решений с помощью методики обработки шума в данных. В известных авторам настоящей статьи работах не представлено аналогичных или подобных подходов оптимизации построения модели решающего дерева.

Задача обработки исходных данных в целях обнаружения и коррекции шума имеет существенную актуальность, так как любой из описанных типов шума может влиять на процесс построения прогнозной модели, в особенности в областях, связанных с обеспечением безопасности человека [5, 6].

Задача выявления аномалий при построении прогнозной модели принятия решений. Определим использующиеся в настоящей работе понятия исходных данных и шума. Имеется исходное множество информационных объектов (объектов данных) $X = \{X_1, X_2, \ldots, X_n\}$ и множество атрибутов $A = \{A_1, A_2, \ldots, A_k\}$. Каждый объект является кортежем значений атрибутов $X_i = \{a_{i1}, a_{i2}, \ldots, a_{ik}\}$.

Шумом называют искаженные значения атрибутов объектов. Объект X_i полагают искаженным объектом, т. е. содержащим шум, если существует такой атрибут A_j, $j=\overline{1,k}$, значение a_{ij} которого является искаженным (содержащим шум). Рассмотрим шум двух типов: 1) отсутствие значений; 2) аномальные значения. Шум типа «отсутствие значения» обозначим как $a_{ij}=\text{null}$. Если некоторые объекты данных имеют пропуски в значениях каких-либо атрибутов, полагаем, что эти пропуски не несут физического смысла и маркируются как шум. Искажения типа «аномальные значения» могут иметь или не иметь физического смысла.

В настоящей статье рассмотрены аномалии в данных. Этот тип искажений представляет интерес в связи со сложностью их обнаружения по сравнению с пропусками или искажениями, которые легко обнаружить перебором словаря. Аномалии могут нести физический смысл и не являться фактической ошибкой в данных. Однако при построении модели в прогнозном анализе опираются на фундаментальное предположение о сохранении тренда: события или явления, имевшие место в прошлом, сохранят вероятность их появления в будущем. Поэтому аномалии или выбросы в данных рассматривают как искажения, подлежащие выявлению и очистке.

Объекты генеральной совокупности представляют собой экземпляры некоторых сущностей, обладающие одинаковым набором атрибутов. Значения этих атрибутов анализируют для выявления закономерностей всех объектов генеральной совокупности. Выбросами, или аномалиями называют такие объекты данных, которые не удовлетворяют параметрам, характерным для большинства других объектов генеральной совокупности. Поскольку каждый объект данных обладает рядом атрибутов, можно утверждать о степени схожести объектов, основываясь на сравнении всех значений соответствующих атрибутов этих объектов.

Большинство методов поиска выбросов в данных основаны на вычислении расстояний между объектами данных [3]. Метод поиска выбросов, основанный на методе расчете показателя локальной аномальности *LOF* [7], описан в работах [2, 8]. Одно из важных преимуществ метода — способность давать некоторую вероятностную оценку принадлежности каждого объекта данных к аномалиям. Это позволяет более гибко оценивать результат анализа, в отличие от методов, однозначно определяющих принадлежность объектов к аномалиям. В то же время, необходимы инструменты для управления указанным преимуществом метода расчета *LOF*, а именно требуется создание набора правил оценки результатов работы метода. Следует ввести некоторые дополнительные критерии, идентифицирующие выбросы.

Метод расчета *LOF* основан на известном методе *k* ближайших соседей, в связи с чем возникает задача выбора параметра *k*. Общие рекомендации по выбору параметра *k* приведены в работе [7], в которой предложено выбирать параметр *k* отдельно для каждой задачи с учетом специфики анализируемых данных, их количества, прогнозируемого числа возможных выбросов и т. д.

Одна из наиболее известных и эффективных моделей в прогнозном анализе — *дерево решений*. Эта модель относится к виду алгоритмов обучения с учителем, т. е. для построения модели используют некоторую выборку информационных объектов, называемую *обучающей выборкой*. Деревья решений организованы в виде иерархической структуры, состоящей из узлов принятия решений по оценке значений определенных переменных для прогнозирования результирующего значения [2]. Любое дерево решений определяет прогнозируемое значение, полученное в результате оценки некоторых входных атрибутов. Каждый уровень в дереве можно рассматривать как одно из решений. Узел дерева обеспечивает проверку условия, а каждое ребро обозначает один из возможных вариантов. Узлы принятия решений содержат критерии выбора, а ребра выражают взаимоисключающие результаты проверки соответствия этим критериям.

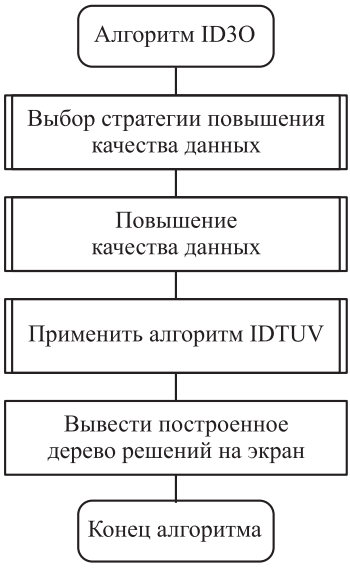

Рис. 1. Алгоритм ID3O

Метод ID3O построения модели решающего дерева. Алгоритм построения модели дерева решений ID3O (рис. 1) приведен в работе [2]. На первом этапе происходит выбор стратегии повышения качества данных в соответствии с показателями, предложенными в работе [9]. На втором этапе происходит повышение качества данных по этому алгоритму заполнения отсутствующих атрибутов данных, а также по рассмотренному в работе [8] алгоритму выявления аномалий. Далее строится дерево решений с помощью алгоритма IDTUV [10].

Алгоритм выявления аномалий, работающий на втором этапе метода ID3O в рамках процесса повышения качества данных, в свою очередь, проводит обработку выбросов в два этапа. На первом этапе выбросы в данных необходимо идентифицировать. Для идентификации аномалий применяют метод расчета *LOF*. На втором этапе обнаруженные объекты подлежат обработке.

Методика выявления и обработки аномалий. Применение этой методики в работе алгоритма построения прогнозных моделей обусловлено использованием алгоритма поиска аномалий LOF, который, как было отмечено выше, имеет преимущества по сравнению с аналогичными алгоритмами, но требует интерпретации результатов работы.

Методика основана на понятии ядра объектов обучающего множества [11]. Объекты анализа — объекты обучающего множества, которые представляют собой кортежи атрибутов данных. Атрибуты могут быть как дискретными, так и непрерывными, и в совокупности представляют собой кортеж, который рассматривают в рамках методики как единый объект анализа. Каждый атрибут

картежа является единицей данных, а весь кортеж — объектом, имеющим информационную значимость, или вес. Очевидно, что различные объекты анализа будут иметь разный информационный вес, уменьшающийся при появлении шума в атрибутах этих объектов.

Если представить объекты анализа сферическими телами, то можно принять частоту $f_n(a_i)$ появления значения a_i атрибута A_n в объектах генеральной совокупности как массу сферы. Чем чаще значение атрибута появляется среди объектов генеральной совокупности, тем «весомее» данное значение. Действительно, шум в данных, имеющий в большинстве своем хаотичный случайный характер, представляется как информационно более «легкий» объект.

Введем параметр ρ, характеризующий плотность объектов. Примем, что плотность всех объектов одинакова. Такое предположение правомерно, поскольку отсутствует априорная информация о вероятности возникновения шума в каких-либо конкретных атрибутах кортежа данных. Тогда, изменяя плотность ρ, можно регулировать объем тел и, соответственно, занимаемую ими площадь в общем информационном пространстве W, созданном множеством объектов генеральной совокупности.

Если пересечение объектов a_i, a_j в некотором пространстве W не пусто: $a_i \cap a_j \neq \varnothing$, то примем, что объекты принадлежат множеству C: $a_i \in C$ и $a_j \in C$. Множество C всех объектов, имеющих пересечения, называют ядром в пространстве W:

$$C = \left\{ a_1, a_2, \ldots, a_k \middle| \left(\bigcup_{i=1}^{k} \bigcup_{j=1}^{k} (a_i \cap a_j) \right) \neq \varnothing \right\}. \tag{1}$$

Методику выявления аномалий выполняют в три этапа. На первом этапе рассчитывают расстояния между всеми объектами анализа по формуле, предложенной в работе [12]:

$$\operatorname{dist}_{A_n}(a_i, a_j) = \sqrt{\frac{f_n(a_i) + f_n(a_j)}{f_n(a_i) f_n(a_j)}}, \tag{2}$$

где A_n — атрибут, принимающий значения $D(A_n) = \{a_1, \ldots, a_p\}$; $f_n(a_i)$ — величина, определяемая прямым подсчетом числа значений a_i атрибута A_n из объектов генеральной совокупности.

Вычисляют показатели локальной аномальности *LOF* для каждого объекта. На втором этапе происходит автоматический анализ среднего показателя $\overline{LOF}$ объектов ядра:

$$\overline{LOF} = \frac{\sum_{i=1}^{|C|} LOF(x_i)}{|C|}. \tag{3}$$

Здесь $|C|$ — мощность множества C, т. е. число объектов ядра. Если представить множество C на плоскости, то $S \le (C)$ — площадь фигуры C. Определяют отношение площади фигуры ядра к общей площади фигур объектов

$$S_{rel} = \frac{S(C)}{S(D(A_n))}. \quad (4)$$

Параметр плотности объектов ρ уменьшается с заданным шагом, который автоматически корректируют по мере продвижения процесса анализа. При уменьшении плотности площадь объектов увеличивается, новые объекты попадают в пересечения, становясь частью ядра. Затем находят средний показатель $\overline{LOF}$ по формуле (3) и отношение площадей по формуле (4). Плотность ρ уменьшается до тех пор, пока все объекты не попадут в ядро, т. е. станет справедливо равенство $S_{rel} = 1$.

На третьем этапе формируется зависимость среднего показателя локальной аномальности объектов ядра $\overline{LOF}(S_{rel})$ от отношения площадей фигуры ядра к общей площади объектов. Вся процедура повторяется несколько раз для разных значений параметра k, характеризующего число ближайших объектов при расчете показателя *LOF*.

Описанные выше этапы работы методики можно более формально записать в виде следующей последовательности шагов.

Шаг 1. Исходные данные представляют собой набор значений некоторого отдельно взятого категориального атрибута, являющийся подмножеством генеральной совокупности.

Шаг 2. По формулам (1), (2), (4) проводят анализ значений категориального атрибута. При этом начальная плотность должна быть задана из тех соображений, чтобы в момент начала анализа не существовало пересечений объектов (ядро было пустым). Далее плотность автоматически регулируется в процессе анализа.

Шаг 3. По результатам анализа данных строят зависимость среднего показателя $\overline{LOF}$ ядра от отношения площади ядра к суммарной площади всех объектов.

Шаг 4. Шаги 2, 3 повторяют несколько раз для разных значений параметра k в диапазоне $[1, p-1]$, где p — число уникальных значений рассматриваемого категориального атрибута. Таким образом, получают набор зависимостей среднего показателя $\overline{LOF}$ ядра от его относительной площади.

Шаг 5. В зависимости, соответствующей выбранному значению параметра k, определяют точку X начала возрастания функции. Выбросами считают точки, не вошедшие в ядро в точке X.

Экспериментальные исследования. Для экспериментов использованы наборы данных, полученных в Калифорнийском университете [11]. Построена зависимость среднего показателя локальной аномальности объектов ядра $\overline{LOF}$

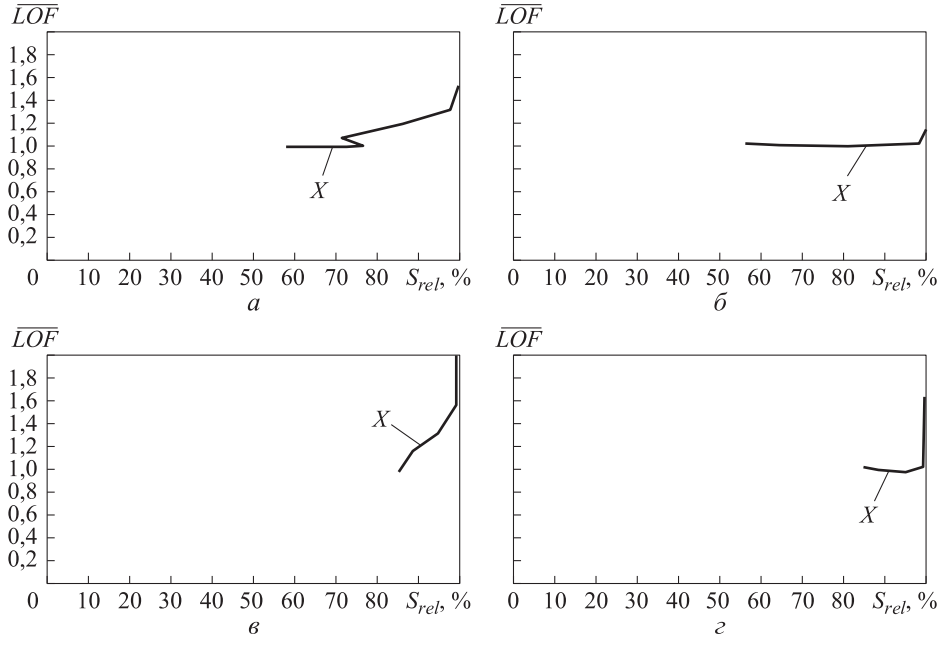

Рис. 2. Зависимость показателя локальной аномальности объектов ядра $\overline{LOF}(S_{rel})$ при k = 5 (*а*), 10 (*б*), 2 (*в*) и 4 (*г*)

от точек ядра к относительной площади фигуры ядра. Анализ проведен для различных значений параметра k. Пример зависимости приведен на рис. 2, дополнительные зависимости — в работе [11]. Результаты анализа атрибутов представлены в таблице.

Результаты анализа атрибутов

k	Показатель $\overline{LOF}$	Разброс точек ядра ΔLOF	Число выбросов
5	1,003	0,056	9
10	0,999	0,035	4
2	1,000	0,033	4
4	1,001	0,078	2

Для каждой зависимости эксперт определяет некоторую точку X, в которой начинается возрастание функции, показатель $\overline{LOF}$ ядра в точке X, а также разброс ΔLOF точек ядра. Точки, не вошедшие в ядро в точке X, считались выбросами при данном значении k. Результаты экспериментов подтверждают, что при увеличении параметра k зависимость среднего показателя локальной аномальности объектов ядра от относительной площади фигуры ядра становится более пологой, сигнал о появлении выбросов появляется позднее, т. е. большее число объектов попадает в ядро и меньше точек идентифицируются как выбросы. Таким образом, параметр k можно рассматривать как «регулятор» степени жесткости идентификации выбросов. Чем выше значение параметра k, тем «мягче» анализ и меньше объектов будут отнесены к выбросам.

Точки, входящие в ядро, имеют разброс показателя *LOF* в пределах одной десятой. При расширении границ ядра в него начинают попадать точки, являющиеся выбросами. В этот момент отношение среднего показателя $\overline{LOF}$ ядра к его относительной площади начинает увеличиваться, что расценивается построенной моделью как сигнал о попадании в ядро потенциального выброса. Полученные выводы позволяют интерпретировать значения показателя *LOF*, а также гибко выбирать значения параметра k на основе субъективных ожиданий эксперта средствами нечеткой логики [11].

После обнаружения и очистки аномалий построенная по алгоритму ID3O модель проверяется. Для проверки точности классификации используют подготовленную заранее тестовую выборку, объекты которой уже классифицированы экспертами. С помощью анализа сравнивают результаты классификации тестовой выборки, сформированной прогнозной моделью, с результатами классификации экспертов, которую полагают эталонной. Для оценки точности применяют критерий *ErrRatio*, называемый коэффициентом ошибки классификатора [9]. Этот критерий определяют как отношение числа неверно классифицированных объектов к общему числу объектов

$$ErrRatio = \frac{|X_f|}{|X|}. \qquad (5)$$

Здесь X — множество объектов в тестовой выборке; X_f — множество объектов, ошибочно классифицированных построенной моделью дерева решений.

Заключение. В результате применения методики выявления аномалий удалось совместить эффективный метод поиска выбросов в данных *LOF* с алгоритмом построения модели дерева решений ID3O. Это обеспечило последнему высокую устойчивость к искажениям в данных одновременно со значительным увеличением производительности системы при построении модели. Устойчивость к искажениям определена как снижение точности классификации при различных уровнях шума в данных, которое при использовании предложенной методики оказалось существенно меньше снижения точности при применении других алгоритмов [2]. Увеличение производительности составляет $p/2$ раз для каждого атрибута объекта данных, где p — число значений атрибута, проверяемого на аномальность [8].

Применение рассмотренной методики при построении прогнозных моделей позволяет эффективно обрабатывать искажения в данных и снижать влияние шума на результат работы систем поддержки принятия решений.

ЛИТЕРАТУРА

1. *Толочко С.И., Черненький В.М.* Анализ информационных систем и определение понятия информационная система поддержки оперативных решений // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2011. Спецвыпуск. С. 69–80.

2. *Кузовлев В.И., Орлов А.О.* Прогнозный анализ данных методом ID3O // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2012. № 10.
DOI: 10.7463/1012.0483286 URL: http://technomag.neicon.ru/doc/483286.html

3. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: A survey // ACM Computing Surveys. 2009. Vol. 41. No. 3. Article 15. 58 p.

4. *Boriah S., Chandola V., Kumar V.* Similarity measures for categorical data: A comparative evaluation // In Proceedings of the 8th SIAM International Conference on Data Mining, 2008.

5. *Черненький В.М., Гапанюк Ю.Е.* Методика идентификации пассажира по установочным данным // Инженерный журнал: наука и инновации. 2012. Вып. 3. DOI: 10.18698/2308-6033-2012-3-89 URL: http://engjournal.ru/catalog/it/biometric/89.html

6. *Толочко С.И., Черненький В.М., Спиридонов И.Н., Мартынов П.И.* Создание и внедрение автоматизированных систем паспортного контроля // Инженерный журнал: наука и инновации. 2012. Вып. 3. DOI: 10.18698/2308-6033-2012-3-94 URL: http://engjournal.ru/catalog/it/biometric/94.html

7. *Shubert E., Zimek A., Kriegel H.-P.* Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video and network outlier detection // Data Min. and Knowl. Disc. 2014. Vol. 28. Iss. 1. P. 190–237. DOI: 10.1007/s10618-012-0300-z

8. *Кузовлев В.И., Орлов А.О.* Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений // Наука и образование. МГТУ им. Н.Э. Баумана. Электрон. журн. 2012. № 9. DOI: 10.7463/0912.0483269 URL: http://technomag.neicon.ru/doc/483269.html

9. *Кузовлев В.И., Орлов А.О.* Вероятностный подход к оценке показателя достоверности элементов результатов профилирования // Инженерный журнал: наука и инновации. 2012. Вып. 3. DOI: 10.18698/2308-6033-2012-3-115
URL: http://engjournal.ru/catalog/it/hidden/115.html

10. *Достоверный* и правдоподобный вывод в интеллектуальных системах / В.Н. Вагин, Е.Ю. Головина, А.А. Загорянская, М.В. Фомина. М.: Физматлит, 2008. 712 с.

11. *Кузовлев В.И., Орлов А.О.* Методика выбора параметров и интерпретации результатов анализа выбросов в данных систем поддержки принятия решений // Инженерный журнал: наука и инновации. 2013. Вып. 11.
DOI: 10.18698/2308-6033-2013-11-1045
URL: http://engjournal.ru/catalog/it/hidden/1045.html

12. *Орлов А.О.* Проблема поиска расстояний между значениями категориальных атрибутов при обнаружении выбросов в данных // В мире научных открытий. 2012. № 8.1. С. 142–155.

Кузовлев Вячеслав Иванович — канд. техн. наук, доцент кафедры «Системы обработки информации и управления» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5).

Орлов Антон Олегович — канд. техн. наук, доцент кафедры «Системы обработки информации и управления» (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5).

Просьба ссылаться на эту статью следующим образом:

Кузовлев В.И., Орлов А.О. Выявление аномалий при прогнозном анализе данных // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2016. № 5. С. 75–85.
DOI: 10.18698/0236-3933-2016-5-75-85

ANOMALIES DETECTION IN PROGNOSTIC DATA ANALYSIS

V.I. Kuzovlev
A.O. Orlov forewar@gmail.com

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

Designing data models for prognostic purposes require anomalies detection method. This article describes the choice of the method and how it applies for the decision tree model algorithm. The authors not only describe the methods of data anomalies search, but also explain basic steps of the algorithm itself. The work analyzes search parameters and their major influence on the method application outcome. As a result of both anomalies detection methods and decision tree model algorithm design the accuracy of the prognostic model increases. It happens due to improved model robustness and also a significant performance improvement of the analysis

Keywords

Anomalies, outliers in data, prognostic analysis, decision tree model

REFERENCES

[1] Tolochko S.I., Chernen'kiy V.M. Information system analysis and the definition of a notion of information system of prompt decision support. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr., Spetsvyp.* [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng., Spec. Issue], 2011, pp. 69–80 (in Russ.).

[2] Kuzovlev V.I., Orlov A.O. Prognostic analysis of data by ID3O. *Nauka i obrazovanie. MGTU im. N.E. Baumana* [Science & Education of the Bauman MSTU. Electronic Journal], 2012, no. 10. DOI: 10.7463/1012.0483286 Available at: http://technomag.neicon.ru/en/doc/483286.html

[3] Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM Computing Surveys*, 2009, vol. 41, no. 3. Article 15. 58 p.

[4] Boriah S., Chandola V., Kumar V. Similarity measures for categorical data: A comparative evaluation. *In Proceedings of the 8th SIAM International Conference on Data Mining*, 2008.

[5] Chernen'kiy V.M., Gapanyuk Yu.E. The passenger identification technique using passenger name record data. *Jelektr. nauchno-tekh. izd. "Inzhenernyy zhurnal: nauka i innovacii"* [El. Sc.-Tech. Publ. "Eng. J.: Science and Innovation"], 2012, iss. 3.
DOI: 10.18698/2308-6033-2012-3-89 Available at: http://engjournal.ru/eng/catalog/it/biometric/89.html

[6] Tolochko S.I., Chernen'kiy V.M., Spiridonov I.N., Martynov P.I. Development and implementation of automated passport-control systems. *Jelektr. nauchno-tekh. izd. "Inzhenernyy*

zhurnal: nauka i innovacii" [El. Sc.-Tech. Publ. "Eng. J.: Science and Innovation"], 2012, iss. 3. DOI: 10.18698/2308-6033-2012-3-94 Available at: http://engjournal.ru/eng/catalog/it/biometric/94.html

[7] Shubert E., Zimek A., Kriegel H.-P. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video and network outlier detection. *Data Min. and Knowl. Disc.*, 2014, vol. 28, iss. 1, pp. 190–237. DOI: 10.1007/s10618-012-0300-z

[8] Kuzovlev V.I., Orlov A.O. Method of detecting anomalies in the source data at constructing a prognostic model of a decision tree in decision support systems. *Nauka i obrazovanie. MGTU im. N.E. Baumana* [Science & Education of the Bauman MSTU. Electronic Journal], 2012, no. 9. DOI: 10.7463/0912.0483269 Available at: http://technomag.neicon.ru/en/doc/483269.html

[9] Kuzovlev V.I., Orlov A.O. Probabilistic approach to estimating the validity factor of elements of profiling results. *Jelektr. nauchno-tekh. izd. "Inzhenernyy zhurnal: nauka i innovacii"* [El. Sc.-Tech. Publ. "Eng. J.: Science and Innovation"], 2012, iss. 3.
DOI: 10.18698/2308-6033-2012-3-115 Available at: http://engjournal.ru/eng/catalog/it/hidden/115.html

[10] Vagin V.N., Golovina E.Yu., Zagoryanskaya A.A., Fomina M.V. Dostovernyy i pravdopodobnyy vyvod v intellektual'nykh sistemakh [Credible and plausible inference in intelligent systems]. Moscow, Fizmatlit Publ., 2008. 712 p.

[11] Kuzovlev V.I., Orlov A.O. The method of parameters selection and data anomaly analysis results interpretation in decision support systems. *Jelektr. nauchno-tekh. izd. "Inzhenernyy zhurnal: nauka i innovacii"* [El. Sc.-Tech. Publ. "Eng. J.: Science and Innovation"], 2013, iss. 11. DOI: 10.18698/2308-6033-2013-11-1045 Available at: http://engjournal.ru/eng/catalog/it/hidden/1045.html

[12] Orlov A.O. The problem of search distances between values of categorical attributes detection emissions data. *V mire nauchnykh otkrytiy* [In the World of Scientific Discoveries], 2012, no. 8.1, pp. 142–155 (in Russ.).

Kuzovlev V.I. — Cand. Sci. (Eng.), Assoc. Professor of Information Processing and Control Systems Department, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, Moscow, 105005 Russian Federation).

Orlov A.O. — Cand. Sci. (Eng.), Assoc. Professor of Information Processing and Control Systems Department, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, Moscow, 105005 Russian Federation).

Please cite this article in English as:

Kuzovlev V.I., Orlov A.O. Anomalies Detection in Prognostic Data Analysis. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr.* [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng.], 2016, no. 5, pp. 75–85. DOI: 10.18698/0236-3933-2016-5-75-85