

УДК 519.2, 681.3

Выявление и анализ аномалий в статистических данных на примере посещаемости сайта

*Донцов В. В., студент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»*

*Научный руководитель: Просуков Е.А., доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
irudakov@bmstu.ru*

С появлением первых носителей информации, начался процесс непрерывного накопления знаний. С ростом информационных технологий, емкости электронных носителей, технологий передачи и обмена данными скорость притока информации только увеличилась.

Увеличение накопленного объема информации обуславливает рост актуальности её обработки. В связи с этим появилась новая область - интеллектуальный анализ данных.

Интеллектуальный анализ данных - собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Классы методов, используемые в этом направлении, включают в себя решение задач классификации, кластеризации, прогнозирования и другие.

Одним из типов задач является определение отклонений. Решение позволяет обнаружить и проанализировать данные, наиболее отличающиеся от общей совокупности – аномалии.

Аномалиями могут быть данные, которые резко отличаются от общей выборки. Ошибки измерения и необычная природа входных данных могут являться причиной отклонений.

Методы решения данного класса задач могут применяться во многих областях, требующих оперативного выявления отклонений от какого-либо типичного сценария получения данных. Например, к таким областям может относиться безопасность клиента банка (резкое увеличение сумм транзакций может говорить о захвате данных клиента

злоумышленником). Постоянный рост поступающей информации, увеличение уровня безопасности и контроля обуславливает потребность в решении данного класса задач.

Также к таким областям относится контроль электронного ресурса в сети. Если у пользователя есть интернет-ресурс и информация о посещаемости этого ресурса, то резкое падение показателей может свидетельствовать об отказе оборудования или других технических неполадках.

В статье кратко изложены результаты работы, целью которой являлись поиск и разработка метода, способного на основе статистики посещаемости сайта за предыдущие периоды времени, выявлять и проводить анализ аномалий в текущих данных.

В ходе работы был проведён анализ с целью определения правил для формирования релевантной выборки.

В статье рассмотрено совместное использование методов выявления аномалий, основанных на вычислении фактора локального отклонения и среднеквадратичного отклонения.

Фактор локального отклонения (local outlier factor) – метод, позволяющий оценить степень принадлежности рассматриваемого объекта к группе ближайших по отношению к нему объектов. Данный метод был получен синтезом алгоритмов DBSCAN и OPTICS и направлен на выявление значений, отклоняющихся от ближайшей совокупности объектов.

Основная идея алгоритма состоит в сравнении локальной плотности объекта с локальными плотностями его соседей (рис. 1).

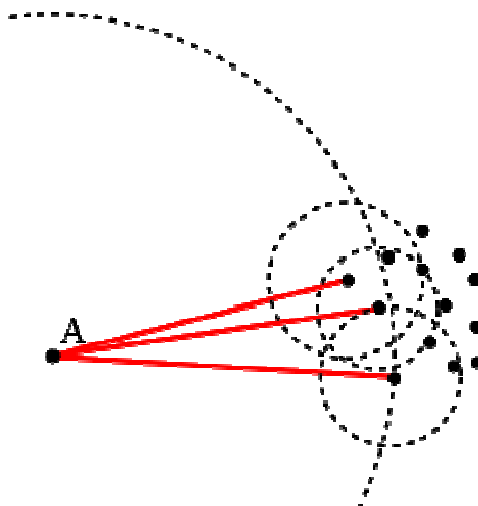


Рис. 1. точка А имеет меньшую плотность, чем её соседи

Локальная плотность объекта рассчитывается на основе k ближайших соседей и соседей, которые попадают в область, ограниченную выбранными соседями. Сравнивая

плотности объектов с плотностями их соседей можно выделить регионы со сравнительно одинаковой плотностью и объекты, чья плотность меньше плотности их соседей. Такие объекты будут являться аномалиями.

Метод позволяет получить значение, доступное для интерпретации и анализа, но обладает следующими ограничениями:

- зависимость результата от числа k ближайших соседей: увеличение способствует повышению точности результата, однако увеличивает время вычисления локальной плотности объектов;
- необходимость калибровки допустимого предела, превышение которого будет строго соответствовать аномальному значению;
- необходимость регулирования выборки путём устранения ранее найденных аномалий.

Последнее ограничение можно обойти путём увеличения числа k ближайших соседей. При этом локальная плотность аномалий будет уменьшена.

Получение фактора локального отклонения выполняется достаточно просто в несколько шагов:

$$\begin{aligned}
 k\text{-distance}(A) &= d(A, B) : B \in N_k(A), \forall C \in N_k(A) \ d(A, B) \geq d(A, C), \\
 \text{reachability-distance}_k(A, B) &= \max\{k\text{-distance}(B), d(A, B)\}, \\
 \text{lrd}(A) &:= 1 / \left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right), \\
 \text{LOF}_k(A) &:= \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A),
 \end{aligned}$$

где d – расстояние между объектами, k -distance - расстояние, в пределах которого лежат k соседей, reachability-distance - расстояние достижимости, lrd – локальная плотность объекта, $N_k(A)$ – множество соседей объекта A .

При этом если $\text{LOF} \approx 1$, то объект сравним с ближайшими соседями. Если $\text{LOF} \gg 1$, то объект – аномалия.

Среднеквадратичное отклонение - показатель рассеивания значений случайной величины относительно её математического ожидания. Со среднеквадратичным отклонением нормального распределения связано правило трёх сигм. Согласно этому правилу приблизительно с 0,9973 вероятностью значение нормально распределённой

случайной величины лежит в интервале $(\bar{x} - 3\sigma, \bar{x} + 3\sigma)$, где σ - среднее квадратичное отклонение, $\bar{x}$ – математическое ожидание заданной выборки. Использование данного правила совместно с предположением о том, что закон распределения данных о посещаемости сайта близок к нормальному закону распределения, может точно выделить данные, которые являются аномалиями.

Данные, которые будут лежать за пределами диапазона трёх сигм, с наибольшей вероятностью окажутся аномалиями. Так же, можно корректировать интервал в соответствии с определённой степенью покрытия. Для оправданности использования метода закон распределения выборки должен быть близок к нормальному закону распределения.

Среднее квадратичное отклонение вычисляется по формуле:

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

где $\bar{x}$ – математическое ожидание заданной выборки.

Важным этапом в определении аномалий является формирование соответствующей выборки. В ходе работы было выявлено, что пользователи, составляющие целевую аудиторию сайта, посещают сайт с некоторой закономерностью. Такая закономерность может обуславливаться различными факторами – среднестатистический период рабочего дня, временная направленность сайта (ночные чаты, утренние новости и т.д.), время публикации новостей или событий, ключевыми интересами общества в определённые промежутки времени и др.

Таким образом, для посещаемости определённого сайта можно выделить условную закономерность в характере этих данных. На рисунке 2 можно проследить изменение уровня посещаемости на протяжении недели. Периоды пиковой активности приходятся на дневное время. Небольшое падение характерно в период с 17 до 19 вечера, предположительно в конец рабочего дня.

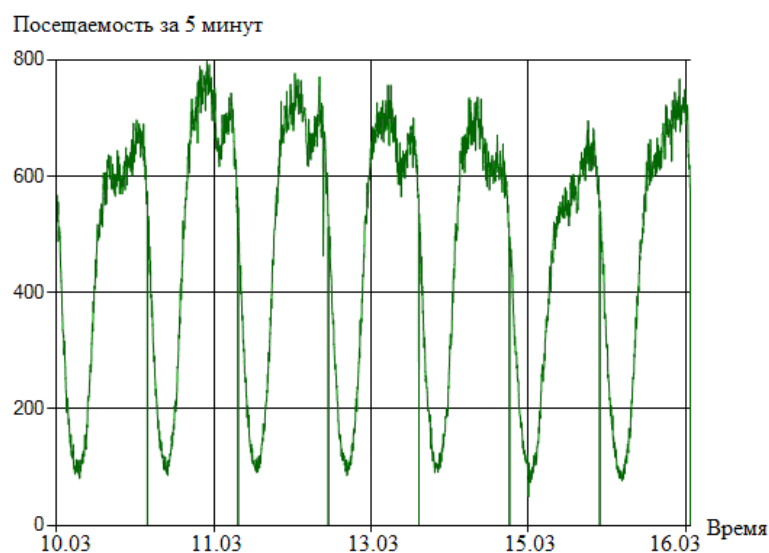


Рис. 2. Посещаемость сайта <http://forum.ixbt.com/> за неделю
(с 2014-03-10 по 2014-03-16).

В выходные уровень посещаемости закономерно ниже, чем в будние дни. Данная тенденция прослеживается на протяжении нескольких недель. Также для любого буднего дня прослеживается закономерность спада в период окончания рабочего дня. На рисунке 3 можно проследить, что характер поведения данных в целом однороден для характерных дней недели.

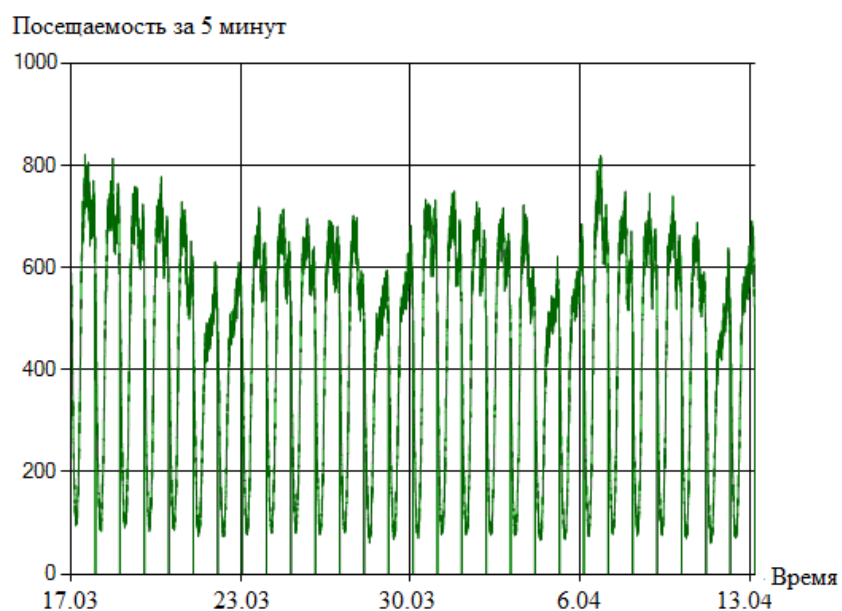


Рис. 3. Посещаемость сайта <http://forum.ixbt.com/> с 2014-03-17 по 2014-04-13.

Основным недостатком использования среднеквадратичного отклонения является то, что анализируемые данные не обязательно будут подчиняться нормальному закону распределения. В данной работе используется приближение о том, что данные являются нормально распределёнными. При этом, ожидается, что метод будет давать хорошие результаты при анализе скорости изменения данных, т.к. эти данные более приближены к нормальным (рис. 4).

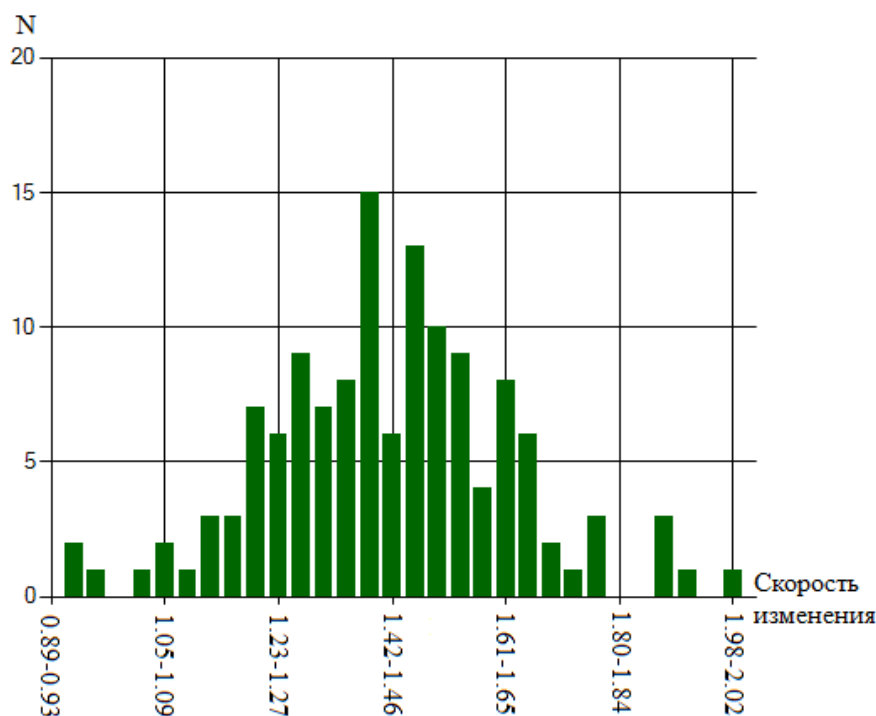


Рис. 4. Распределение данных о скорости изменения посещаемости сайта <http://forum.ixbt.com/> с 11:00 до 12:00 каждую среду с 2014-02-03 по 2014-04-23

Для тестирования методов была проведена серия экспериментов. Алгоритм LOF, выявляющий аномалии на основе фактора локального отклонения использует 10 соседей и пороговое значение равное 1,5. Алгоритм 3G, основанный на среднеквадратичном отклонении, использует интервал $(\bar{x} - 3\sigma, \bar{x} + 3\sigma)$. В экспериментах использовались выборки данных с 11:00 до 12:00 каждую среду с 2014-02-03 по 2014-04-23 для данных с различной среднесуточной посещаемостью. Использовались данные посещаемости и данные скорости изменения посещаемости. В расчёт были взяты случайные сайты представляющие малую среднесуточную посещаемость (меньше 1000), среднюю (от 1000 до 500 000), крупную (от 500 000). Используемые сайты:

— «Каталог программиста» <http://articles.org.ru/> (среднесуточная посещаемость 500 посетителей).

— «Конференция iXBT.com: всё о железе и не только» <http://forum.ixbt.com/> (среднесуточная посещаемость 65 000 посетителей).

— «Мой Мир@Mail.Ru» <http://my.mail.ru> (среднесуточная посещаемость 3 000 000 посетителей).

Данные о посещаемости этих сайтов находятся в открытом доступе и были скачены с сайта <http://top.mail.ru>. В данном тестировании алгоритмы применялись для трёх случаев:

- В случае отсутствия возмущений – без аномалий.
- В случае внесения возмущений, отклоняющих значение не более чем на 10% - слабых аномалий.
- В случае внесения возмущений, отклоняющих значение более чем на 10% - сильных аномалий.

Результаты экспериментов приведены в таблице 1.

Таблица 1

Результаты тестирования алгоритмов LOF и 3G. Обозначения: НН – не найдены, НВ – найдены все, ЛА – ложная аномалия, ПС – по скорости, ПД – по данным

Алгоритмы	LOF			3G		
Посещаемость	Малая	Средняя	Крупная	Малая	Средняя	Крупная
Без аномалий	НН, 3ЛА ПС	НН	НН	НН, 2ЛА ПС	НН	НН
Слабые аномалии	НВ, 5ЛА ПС	НВ	НВ ПС, 50% ПД	НВ	НВ ПД, 50% ПС	НН ПД, 50% ПС
Сильные аномалии	НВ, 5ЛА ПС	НВ	НВ ПС, 50% ПД	НВ	НВ	НВ ПС, 50% ПД

Из результатов эксперимента можно увидеть, что при использовании алгоритма 3G практически не возникает ложных аномалий, однако общий процент найденных аномалий уменьшается при увеличении посещаемости ресурса. Также из результатов следует, что алгоритм LOF является более чувствительным к слабым аномалиям. Ложные срабатывания определяются сравнительно большой разницей значений для сайтов с малой посещаемостью. Для таких сайтов рекомендуется резко завышать пороговое значение алгоритма LOF.

При использовании варианта совместной работы данные считаются аномальными, если оба алгоритма посчитают их таковыми.

Таблица 2

Результаты тестирования комбинации алгоритмов. Обозначения: НН – не найдены, НВ – найдены все, ЛА – ложная аномалия, ПС – по скорости, ПД – по данным

Алгоритм	Комбинированный		
Посещаемость	Малая	Средняя	Крупная
Без аномалий	НН, 1ЛА ПС	НН	НН
Слабые аномалии	НВ ПД, НН ПС, 1 ЛА ПС	НВ: 50% ПС, 50% ПД	НВ ПС, НН ПД
Сильные аномалии	НВ ПД, НВ ПС	НВ: 50% ПС, НВ ПД	НВ ПС, НН ПД

Эффективность совместного использования методов определяется уменьшением количества выявленных ложных аномалий. Исходя из полученных результатов, чем больше среднесуточная посещаемость электронного ресурса, тем эффективнее работает анализ по данным о скорости посещаемости. Для улучшения результативности поиска аномалий необходимо формировать выборку из значений скорости изменения данных. При использовании комбинированного варианта работы наиболее предпочтительны именно эти данные. При этом анализ способен выявить аномалии в случае действительно критичных отклонений.

Наиболее статистически значимая выборка получается путём выбора данных в определённые дни недели и временные диапазоны в эти дни. Для статистически данных другой области так же рекомендуется формирование выборки на основе выявленных закономерностей.

Для наибольшего покрытия выявленных аномалий необходимо применять комплекс методов. Целесообразнее проводить анализ на сайтах со средней или крупной среднесуточной посещаемостью. Алгоритм LOF порождает ложные аномалии. Алгоритм 3G может оказаться недостаточно точным. Для эффективного поиска отклонений необходимо использовать метод 3G для всей выборки и по аномалиям данной выборки рассчитывать значения LOF. Такой подход позволяет сократить количество совершаемых вычислений. Параметры методов могут быть откалиброваны под нужды пользователя.

Список литературы

1. Черняк Л. Большие Данные - новая теория и практика. Режим доступа: <http://www.osp.ru/os/2011/10/13010990/> (дата обращения 27.11.2014).
2. Outlier Detection Techniques. The 2010 SIAM International Conference on Data Mining. Available at: <https://www.siam.org/meetings/sdm10/tutorial3.pdf>, accessed 25.11.2014.
3. Breunig M.M., Kriegel H.P., Sander J. LOF: Identifying Density-based Local Outliers // International Conference on Management of Data: proceedings. Dallas: ACM SIGMOD, 2000. P. 93–104. DOI:10.1145/335191.335388.