электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51036. ISSN 2307-0595

Повышение качества данных с использованием методики поиска аномалий на примере Портала открытых данных правительства Москвы

 # 08, август 2014

Кузовлев В. И., Орлов А. О.

УДК: 004.052.42

Россия, МГТУ им. Баумана
forewar@gmail.com

1. Введение

В системах поддержки принятия решений (далее СППР) исследуются статистические данные для формирования тех или иных аналитических моделей. Наличие искажений (или шума) в данных оказывает влияние на результат бизнес-процесса и процесса в СППР: происходит анализ искаженных данных, в результате могут вырабатываться неверные и неэффективные решения и организационные воздействия. Для решения данной проблемы при построении модели необходимо использовать механизм, способный обрабатывать искаженные данные таким образом, чтобы они оказывали минимальное воздействие на результат работы системы [1]. На основе метода выявления аномалий в исходных данных в [1] предложена методика подготовки, проведения и интерпретации результатов анализа [2]. В данной статье рассмотрено применение методики [2] при анализе открытых статистических данных. Целью данной статьи является демонстрация эффекта от применения методик анализа данных в рамках системы поддержки принятия решений для анализа набора данных открытого доступа. В статье приведен анализ открытого набора данных, задачей которого является информирование целевой аудитории по наличию и доступности инфраструктурных спортивных объектов на территории города Москвы. Анализ заключается в выявлении степени информативности атрибутов набора данных с учетом применения методики выявления аномалий [1].

2. Анализ информативности данных

Созданная в рамках исследований система поддержки принятия решений использует прогнозную модель дерева принятия решений [5, 6]. Деревья решений организованы в виде иерархической структуры, состоящей из узлов принятия решений по оценке значений определенных переменных для прогнозирования результирующего значения. Данная модель относится к виду алгоритмов обучения с учителем, то есть для построения модели исполь-

зуется некоторая выборка информационных объектов, называемая обучающей выборкой. Суть модели дерева принятия решений заключается в построении такого решающего дерева, каждый узел которого на каждом шаге построения модели является наиболее информативным атрибутом среди всех еще не рассмотренных атрибутов. Мера информативности определяется как количество передаваемой информации (энтропия). Если имеется n равновероятных значений атрибута, то вероятность p каждого из них равна $1/n$ и информация, сообщаемая значением атрибута, равна $-\log_2 p = \log_2 n$. Если P – это дискретное распределение $P = (p_1, p_2, \dots, p_n)$, то энтропия P вычисляется следующим образом:

$$I(P) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i \qquad (1)$$

Формула (1) называется энтропией Шеннона или информационной энтропией. Если обучающее множество S разбито на попарно непересекающиеся классы *C1, C2, ..., Ck*, то информация, необходимая для идентификации класса отдельного примера S_i, равна $Info(S_i) = I(P)$, где P – дискретное распределение вероятностей появления соответствующего примера из обучающего множества, сопоставленное набору классов *C1, C2, ..., Ck:* $P = (\frac{|C1|}{|S|}, \frac{|C2|}{|S|}, \dots, \frac{|Ck|}{|S|})$. Разбив множество примеров на основе значений некоторого атрибута A на подмножества *S1, S2, ..., Sn*, можно вычислить $Info(S_i)$ как взвешенное среднее информации, необходимой для идентификации класса примера в каждом подмножестве:

$$Info(A, S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} \cdot Info(Si) \qquad (2)$$

Величина $Gain(A, S) = Info(S) - Info(A, S)$ показывает количество информации, сообщаемое атрибутом A. Наборы открытых данных проанализированы путем вычисления количества информации, сообщаемого всеми атрибутами наборов.

3. Методика поиска аномалий в статистических данных

Поиск и устранение аномалий в исследованных наборах данных применялся с целью повышения эффекта от использования открытого набора данных для конечного пользователя.

Важной частью методики поиска аномалий является метод расчета показателя локальной аномальности LOF [3]. Одним из важных преимуществ метода LOF является его способность к расчету степени аномальности каждого объекта данных, позволяющая более гибко оценивать результат анализа в отличие от методов, однозначно определяющих принадлежность объектов к аномалиям. К основным недостаткам метода LOF можно отнести необходимость предварительного выбора параметров, в частности параметра k, определяющего количество ближайших соседей, анализируемых моделью [7]. Кроме того, поскольку метод дает числовую оценку степени аномальности объектов, необходимо вводить некоторые дополнительные критерии, идентифицирующие выбросы. В методике [2] учтены данные осо-

бенности метода LOF, предложено использование механизмов теории нечетких множеств. Этапом дефаззификации при этом может являться переход от значения степени аномальности объекта данных к решению о принадлежности его к выбросам.

Методика поиска аномалий состоит из трех основных этапов. На первом этапе рассчитываются расстояния между всеми объектами анализа. Расчет расстояний производится по формуле инверсной гравитации (3), предложенной в [4].

$$dist_{A_n}(x_i, x_j) = \sqrt[2]{\frac{f_n(x_i) + f_n(x_j)}{f_n(x_i) \cdot f_n(x_j)}}, \quad (3)$$

где A_n – категориальный атрибут, принимающий значения $D(A_n) = \{x_1, \dots, x_p\}$; $f_n(x)$ – количество объектов генеральной совокупности, атрибут A_n которых принимает значение x.

Также вычисляются показатели локальной аномальности LOF для каждого объекта.

В [2] введено понятие ядра, которое заключается в следующем. Объекты анализа представляются как сферические тела с частотой $f_n(x)$ появления значения x атрибута A_n среди объектов генеральной совокупности. Эта частота является массой сферы. Если считать плотность ρ всех объектов одинаковой, тогда, изменяя ρ, можно регулировать объем тел и, соответственно, занимаемую ими площадь.

Если пересечение объектов x_i, x_j в некотором пространстве W: $x_i \cap x_j \neq \emptyset$, тогда $x_i \in C$ и $x_j \in C$. Множество C всех объектов, имеющих пересечения, называется ядром в пространстве W.

$$C = \{x_1, x_2, \dots, x_k \mid (\bigcup_{i=1}^{k} \bigcup_{j=1}^{k} (x_i \cap x_j)) \neq \emptyset\}, \quad (4)$$

На втором этапе происходит автоматический анализ среднего показателя LOF среди объектов ядра, а также отношения площади фигуры ядра к общей площади фигур объектов:

$$\overline{LOF} = \frac{\sum_{i=1}^{|C|} LOF(x_i)}{|C|} \quad (5)$$

$$S_{rel} = \frac{S(C)}{S(D(A_n))} \quad (6)$$

Параметр плотности объектов ρ уменьшается с заданным шагом, который автоматически корректируется по мере продвижения процесса анализа. При уменьшении плотности площадь объектов увеличивается, новые объекты попадают в пересечения, становясь частью ядра. Снова рассчитывается средний показатель LOF по формуле (5) и отношение площадей по формуле (6). Плотность ρ уменьшается до тех пор, пока все объекты не попадут в ядро, то есть станет справедливо равенство $S_{rel} = 1$.

На третьем этапе формируется график $\overline{LOF}(S_{rel})$ зависимости среднего показателя локальной аномальности объектов ядра от отношения площадей фигуры ядра к общей

площади объектов. Вся процедура повторяется несколько раз для разных значений параметра k, характеризующего количество ближайших объектов при расчете показателя LOF.

4. Применение методики на открытых данных

Использовались наборы Портала открытых данных правительства Москвы (http://data.mos.ru/datasets) из категории «Зимние наборы данных». Наборы содержат данные о доступных для посещения зимних спортивных площадках города Москвы. В таблице 1 приведен формат анализируемых данных.

Таблица 1 Формат данных анализа

Атрибут данных	Тип атрибута	Уникальных значений
Тип объекта	Категориальный	78
Округ	Категориальный	10
Наличие раздевалки	Булев	2
Освещение	Категориальный	5
Доступность	Категориальный	2

Выбранные наборы данных обладают большим количеством категориальных атрибутов. Они представляют больший интерес по сравнению с числовыми, так как требуют дополнительных процедур расчета расстояний между объектами в силу того обстоятельства, что категориальные значения не принадлежат заранее каким-либо шкалам. Для расчета расстояний между значениями категориального атрибута использовалась формула инверсной гравитации (3).

Необходимо пояснить атрибуты с малым количеством принимаемых значений. Булев атрибут «Наличие раздевалки» принимает значения «да/нет». Категориальный атрибут «Доступность» принимает значения «Платно/Бесплатно». Категориальный атрибут «Территория» принимает значения «Открытый/Крытый».

Объединенный набор данных содержит 1500 записей. Набор был разбит на 5 примерно равных пересекающихся обучающих наборов данных. По полученным наборам данных были построены прогнозные модели дерева решений. Для построения моделей дерева решений использовался метод ID3O [5]. Алгоритм ID3O имеет высокую устойчивость к искажениям в данных по сравнению с подобными алгоритмами [8] благодаря механизмам поиска и устранения аномалий в данных, а также заполнения пропущенных значений атрибутов. Построение моделей происходило в два этапа. На первом этапе строилось дерево решений без предварительного поиска и очистки данных от аномалий. Оценивалась информативность атрибутов данных на каждом шаге построения модели. На втором этапе дерево решений для каждого набора данных строилось заново после применения методики поиска аномалий. Снова оценивалась информативность атрибутов данных. Эффект от применения методики рассчитывался как разница между информативностью данных до и после удаления аномалий.

В таблице 2 приведены расчеты информативности всех атрибутов на первом шаге построения модели дерева решений для решающего атрибута «Доступность».

Таблица 2 Информативность атрибутов на первом шаге построения модели дерева решений

	Атрибут			
	Тип объекта	**Округ**	**Раздевалка**	**Освещение**
Исходные данные	0.152	0.093	0.052	0.011
После удаления аномалий	0.078	0.074	0.038	0.009

Видно, что корнем дерева выбран атрибут «Тип объекта», как обладающий наибольшей информативностью. Данный выбор достаточно очевиден, поскольку атрибут «Тип объекта» обладает намного большим количеством уникальных значений по сравнению с другими атрибутами, как показано в таблице 1. При этом значение остальных атрибутов набора данных существенно обесценивается. Поиск нужных объектов для пользователя сводится в большей степени к выбору из единственного атрибута «Тип объекта», который обладает большим количеством вариантов, и СППР не может эффективно выполнять возложенные на нее функции.

После применения методики поиска и удаления аномалий набор данных снова был использован для построения модели дерева решений. Как и в первом случае, наиболее информативным атрибутом является атрибут «Тип объекта», который и признается корнем дерева. Однако значение информативности этого атрибута существенно приблизилось к информативности других атрибутов, что повышает их значимость при построении модели. В этом случае система поддержки принятия решений может предложить пользователю классификацию объектов с возможностью выбора из нескольких значимых атрибутов.

5. Заключение

Анализ открытых данных, содержащих информацию об инфраструктурных зимних спортивных объектах города Москвы, показал достаточно низкое качество исходных данных и необходимость в применении дополнительных процедур по очистке и повышению качества информации. Наблюдается существенный перевес информативности отдельных атрибутов по сравнению с остальными, что затрудняет эффективную классификацию данных для представления их в широкий доступ в виде справочной системы с возможностью поиска. Результаты применения методики выявления аномалий в данных показали, что очистка информации с применением данной методики позволяет повысить значимость всех атрибутов данных за счет понижения излишней информативности, вызванной отдельными аномальными значениями в данных.

Список литературы

1. 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. № 09. DOI: http://dx.doi.org/10.7463/0912.0483269

2. Кузовлев В. И., Орлов А. О. Методика выбора параметров и интерпретации результатов анализа выбросов в данных систем поддержки принятия решений // Вестник МГТУ им. Баумана. Сер. «Приборостроение». М. 2013. 10 с.
3. Breunig M., Kriegel H.-P., T. Ng R., Sander J. LOF: Identifying Density-Based Local Outliers // Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press. P. 93-104.
4. Орлов А. О. Проблема поиска расстояний между значениями категориальных атрибутов при обнаружении выбросов в данных // В мире научных открытий. Красноярск, 2012. № 8.1. С. 142-155.
5. Кузовлев В. И., Орлов А. О. Прогнозный анализ данных методом ID3O // Наука и образование. МГТУ им. Н. Э. Баумана. Электрон. журн. 2012. № 10. DOI: http://dx.doi.org/10.7463/1012.0483286
6. Utgoff P. E. Incremental induction on Decision Trees // Machine Learning. 1989. V. 4. P. 161-186.
7. Boriah S., Chandola V., Kumar V. Similarity measures for categorical data: A comparative evaluation // Proceedings of the 8th SIAM International Conference on Data Mining. Atlanta, 2008. P. 253-254.
8. Вагин В. Н. и др. Достоверный и правдоподобный вывод в интеллектуальных системах / под ред. В. Н. Вагина, Д. А. Поспелова. 2-е изд., испр. и доп. М.: ФИЗМАТЛИТ, 2008. 712 с.