Метод выявления аномалий в исходных данных при построении прогнозной модели решающего дерева в системах поддержки принятия решений

09, сентябрь 2012

DOI: 10.7463/0912.0483269

Кузовлев В. И., Орлов А. О.

УДК 004.052.42

Россия, МГТУ им. Н.Э. Баумана

forewar@gmail.com

Введение

Системы поддержки принятия управленческих решений (далее СППУР) принадлежат к типу систем поддержки принятия решений (далее СППР). СППУР применяются для анализа обширных наборов данных, отражающих различные аспекты бизнес-процессов в организациях, и формирования рекомендаций по выработке на основе проведенного анализа управляющих воздействий, направленных на решение практических задач. Механизмы анализа данных в СППУР основаны на принципах прогнозного анализа. Суть прогнозного анализа заключается в формировании суждений о будущих фактах на основе анализа исторических данных. Основой прогнозного анализа данных являются методы машинного обучения, в частности механизмы распознавания образов, где в свою очередь широко применяются алгоритмы классификации.

Наличие искажений (или шума) в данных оказывает влияние на результат работы механизма прогнозирования в СППР: происходит анализ искаженных данных, в результате могут вырабатываться неверные и неэффективные решения и организационные воздействия. Для решения данной проблемы при построении прогнозной модели необходимо использовать механизм, способный обрабатывать искаженные данные таким образом, чтобы они оказывали минимальное воздействие на результат работы системы.

Постановка задачи исследования

Статья посвящена разработке алгоритма поиска и обработки аномалий в данных для применения в прогножном анализе в системах поддержки принятия управленческих

решений. Поскольку в качестве прогнозной модели используется модель дерева решений, разрабатываемый алгоритм обработки шума в данных должен иметь следующие преимущества перед существующими алгоритмами построения модели дерева решений:

1. Алгоритм должен обрабатывать аномалии в данных.
2. Необходимо разработать новый метод поиска аномалий в данных, имеющий преимущества перед существующими методами.

Для решения поставленной задачи проведен обзор существующих алгоритмов построения дерева решений, а также существующих методов поиска аномалий в данных.

Существующие методы построения деревьев решений

Рассматривается прогнозная классифицирующая модель дерева решений. 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 используется некоторая выборка информационных объектов, называемая обучающей выборкой.

Существует достаточно много алгоритмов, реализующих принципы модели деревьев решений [1-4]: ID3, C4.5, ДРЕВ, ID5R, Reduce.

Алгоритм ID3 (итеративный дихотомайзер) предложен Джоном Куинланом [3]. Он строит решающее дерево, на каждом уровне которого выбирается атрибут, имеющий наибольшую информационную значимость, которая определяется через понятие энтропии, то есть количества передаваемой данным атрибутов информации.

Алгоритм C4.5 предложен также Куинланом, он расширяет возможности ID3, имея возможность работать как с дискретными так и с непрерывными атрибутами. Также алгоритм C4.5 в отличие от ID3 определяет отсутствующие значения атрибутов и игнорирует их, не используя в построении дерева.

Алгоритм IDTUV разработан В. Н. Вагиным [1]. Он использует алгоритмы ID3 и C4.5 совместно, при этом исправляет отсутствующие значения атрибутов, позволяя учитывать соответствующие объекты при построении дерева.

Однако, ни один из рассмотренных алгоритмов не обрабатывает шум типа «аномальные значения».

Существующие методы поиска аномальных значений в данных

В отличие от атрибутов с пустыми значениями, для успешной работы с аномальными значениями (выбросами) атрибутов необходим специальный механизм идентификации аномалий. В [5] исследуются различные методы обнаружения аномалий в

данных. Выделяется подход, в основе которого используется широко известный метод «k ближайших соседей». Данный подход применяется в методе LOF (Local Outlier Factor), описанном и подробно обсуждающимся в [6]. Данный метод основан на оценке плотности расположения объектов, проверяющихся на выбросы. Объекты, лежащие в областях наиболее низкой плотности, считаются выбросами. Преимущество метода LOF перед другими методами, работающими с плотностью расположения объектов, заключается в том, что в LOF рассматривается так называемая «локальная плотность». Таким образом, LOF успешно распознает выбросы в ситуациях, когда в выборке присутствуют объекты разных классов, не являющиеся аномалиями.

Рисунок 1. Механизм LOF. Случай областей с разной плотностью

На рисунке 1 показан пример, когда объекты обучающей выборки принадлежат двум классам C_1 и C_2 . Объекты в двух классах обладают разной плотностью. Точки o_1 и o_2 являются аномалиями. Благодаря вычислению локальной плотности классов LOF успешно распознает оба выброса. Методы, основанные на вычислении средней плотности всех объектов, в большинстве случаев обнаруживают выброс o_1 , но пропускают выброс o_2 , что подтверждается в [5, 6].

Метод выявления аномалий в исходных данных

Используемые в данной статье понятия исходных данных и шума определяются следующим образом. Имеется исходное множество информационных объектов (объектов данных)

$$X = \{X_1, X_2, \dots, X_n\}.$$

Множество атрибутов описывается следующим образом:

$$A = \{A_1, A_2, \dots, A_k\}$$

Множество значений некоторого категориального атрибута A_i равно

$$D(A_i) = \{a_1, \dots, a_p\}$$

Каждый объект является кортежем значений атрибутов

$$X_i = \{a_{i1}, a_{i2}, \dots, a_{ik}\}.$$

Объект X_i является искаженным объектом, то есть содержит шум, если существует такой атрибут A_j , $j = \overline{1, k}$, значение a_{ij} которого является искаженным, то есть содержит шум. Таким образом, шумом называются искаженные значения атрибутов объектов. В данной работе рассматривается шум двух типов: отсутствие значений или аномальные значения. Шум типа «отсутствие значения» обозначается как $a_{ij} = null$. Если некоторые объекты данных имеют пропуски в значениях каких-либо атрибутов, считается, что данные пропуски не несут физического смысла и маркируются как шум. Искажения типа «аномальные значения» могут иметь или не иметь физического смысла. Например, значение атрибута «температура по Цельсию» равное 1200 имеет физический смысл, однако является аномальным значением в контексте измерения температуры человеческого тела. В другом случае, некоторое качественное значение атрибута, написанное с опечаткой, также является аномальным среди множества значений данного атрибута, не содержащих опечаток. Такое значение не несет физического смысла. Любой из описанных типов шума может оказывать влияние на процесс построения прогнозной модели, поэтому задача обработки исходных данных с целью обнаружения и коррекции шума имеет существенную актуальность. Стоит отметить при этом, что в отличие от шума типа «отсутствие значений» шум типа «аномальные значения» требует дополнительных процедур идентификации.

Механизмы, основанные на вычислении расстояний между объектами, нуждаются в метриках, позволяющих эти расстояния найти. Шум может возникать в количественных и в качественных атрибутах объектов. Как для количественных, так и для многих качественных атрибутов легко установить меры сравнения и расчета расстояния между объектами. Основной проблемой являются шкалы для вычисления расстояния между категориальными атрибутами. Категориальными называются качественные атрибуты, значения которых не принадлежат какой-либо интервальной или порядковой шкале [7].

В [8] предложена формула для вычисления расстояний между значениями категориального атрибута. Значение $f_j(x)$ равно количеству объектов, атрибут A_j которых принимает значение x :

$$f_j(x) = \sum_{i=1}^n \lambda_i. \quad (1)$$

Здесь $\lambda_i = \begin{cases} 1, & \text{если } a_{ij} = x; \\ 0, & \text{иначе.} \end{cases}$

Пусть имеется некоторый категориальный атрибут A_n , принимающий значения $D(A_n) = \{a_1, \dots, a_p\}$. Тогда расстояние между значениями x_i и x_j ($i, j \leq p$) при условии, что $x_i \neq x_j$, обозначается как $dist_{A_n}(x_i, x_j)$ и вычисляется следующим образом [8]:

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

То есть расстояние между значениями категориального атрибута рассчитывается как корень из отношения суммы количеств появления соответствующих значений атрибута к их произведению.

В [8] показано, что точность классификации прогнозной модели при использовании предложенной формулы не уступает существующим формулам, исследованным в [9], при этом получен существенный выигрыш в производительности при добавлении новых объектов в множество исходных данных. Выигрыш в производительности составляет $\frac{p}{2}$ раз для каждого атрибута объекта данных, где p – это количество значений атрибута, проверяемого на аномальность.

Обработка выбросов данных происходит в два этапа. На первом этапе выбросы в данных необходимо идентифицировать. Для идентификации аномалий применяется механизм LOF. На втором этапе обнаруженные объекты подлежат обработке. Шум в значениях атрибутов может носить как характер ошибки, так и иметь физический смысл. Например, атрибут «цвет» у некоторого объекта может принимать значение «красный» и определяться как шум. В то же время этот же атрибут «цвет» у другого объекта может иметь значение «зеленый» и тоже определяться как шум. Однако, в первом случае, шум носит характер опечатки и подлежит корректировке, а во втором случае значение является достоверным, тогда решение о корректировке должно приниматься на основании дополнительных внешних знаний. В том случае, когда принимается решение о сохранении значения шума, необходимо исключить данный атрибут рассматриваемого объекта из дальнейших расчетов, в которых он может оказать неожиданное влияние на

оценку других объектов. Иными словами, такому атрибуту должен быть присвоен нулевой вес в дальнейших расчетах.

На основе механизма LOF и предложенной формулы расчета расстояний между значениями категориального атрибута разработан метод поиска и обработки аномалий (выбросов) в объектах данных, представленный в таблице 1.

Таблица 1. Алгоритм обработки выбросов в данных

Алгоритм обработки выбросов в данных

Вход: множество объектов $X = \{X_1, X_2, \dots, X_n\}$, содержащих выбросы в данных; атрибут $A_j, j = \overline{1, k}$, по которому проводится проверка; параметр *MinPts*, определяющий количество ближайших объектов проверки; параметр *Restore* = {0, 1}, определяющий действие над найденными аномалиями.

Выход: множество объектов, содержащих аномалии $X' = \{X'_1, \dots, X'_m\}, m \leq n$; множество весов значения атрибута A_j объектов с аномалиями $V = \{V_1, \dots, V_m\}$, причем $V_i = \{0, 1\}$.

Начало алгоритма.

Шаг 1. Положить $X' = \{\emptyset\}$ пустое множество. Сформировать множество значений атрибута $D(A_j) = \{a_1, \dots, a_k\}, 1 < k \leq n$.

Шаг 2. Для каждого a_i из $D(A_j)$ вычислить $LOF_{MinPts}(a_i)$.

Шаг 3. Вычислить $\overline{LOF}$ как среднее арифметическое полученных ранее $LOF_{MinPts}(a_i)$.

Шаг 4. Добавить в X' объекты из X такие, что $\forall X_z \in X: LOF_{MinPts}(X_z(A_j)) > \overline{LOF}$, где $X_z(A_j)$ – значение атрибута A_j объекта X_z .

Шаг 5. Если значение параметра *Restore* равно 1, значит значения шума подлежат корректировке. Перейти к шагу 6. Если значение *Restore* равно 0, перейти к шагу 8.

Шаг 6. Положить значения атрибута A_j для всех объектов из X' равными пустому значению $\forall X_z \in X: X_z(A_j) = null$.

Шаг 7. Положить веса $V = 1$ для всех объектов из X' . Перейти к шагу 9.

Шаг 8. Положить веса $V = 0$ для всех объектов из X' .

Шаг 9. Конец.

Конец алгоритма.

Алгоритм выполняется для каждого атрибута, в результате формируются множества объектов, содержащих аномалии в значениях соответствующих атрибутов. В зависимости от стратегии корректировки шума, определяющейся входным параметром *Restore*, значения-выбросы либо обнуляются, либо остаются прежними. В случае обнуления соответствующему атрибуту объекта присваивается вес равный единице, что позволяет учитывать влияние данного атрибута в дальнейших расчетах. Задача поиска значения в этом случае сводится к задаче заполнения пропусков в данных. В случае, если аномальное значение атрибута остается неизменным, данному атрибуту текущего объекта присваивается нулевой вес, что позволяет не учитывать аномальное значение в дальнейших расчетах. Такой подход позволит корректно оценивать базовые показатели достоверности обработки информации [10].

Заключение

Описана проблема влияния искажений в исходных данных на результат работы систем поддержки принятия решений. Рассмотрены существующие алгоритмы построения прогнозной модели решающего дерева, обнаружено отсутствие механизмов для работы с данными, содержащими шум типа «аномальные значения». Рассмотрены существующие методы обнаружения аномалий в данных, выделен метод оценки показателя локальной аномальности LOF. Предложена формула расчета расстояний между значениями категориального атрибута. Предложенная формула позволит вычислять показатель локальной аномальности для объектов данных, содержащих как числовые, так и категориальные атрибуты. На основе результатов исследований разработан метод обнаружения и обработки аномалий в данных. Применение данного метода на начальном этапе построения прогнозной модели позволит своевременно обрабатывать искажения в данных и снизить влияние шума на результат работы систем поддержки принятия решений.

Список литературы

1. Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В. 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.
2. Minger J. An Empirical Comparison of Pruning Methods for Decision Tree Induction // Machine Learning. 1989. Vol. 4, no. 2. P. 227-243. DOI: 10.1023/A:1022604100933
3. Quinlan J.R. Induction of Decision Trees // Machine Learning. 1986. Vol. 1, no. 1. P. 81-106. DOI: 10.1023/A:1022643204877

4. Utgoff P.E. Incremental induction on Decision Trees // Machine Learning. 1989. Vol. 4, no. 2. P. 161-186. DOI: 10.1023/A:1022699900025
5. Chandola V., Banerjee A., Kumar V. Anomaly detection: A Survey // ACM Computing Surveys. 2009. Vol. 41, no. 3. Article 15. DOI: [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882)
6. 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.
7. Орлов А.И. Эконометрика: учебник. М.: Экзамен, 2002. 576 с.
8. Орлов А.О. Проблема поиска расстояний между значениями категориальных атрибутов при обнаружении выбросов в данных // В мире научных открытий. 2012. №8.1 (32). С. 142-155.
9. 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, GA, USA: SIAM, 2008. P. 243–254.
10. Кузовлев В.И., Липкин Д.И. Определение базовых показателей достоверности обработки информации проектных решений АСОИУ. М., 2001. 12 с. Деп. в ВИНТИ № 1094-B2001.

Method of detecting anomalies in the source data at constructing a prognostic model of a decision tree in decision support systems

09, September 2012

DOI: 10.7463/0912.0483269

Kuzovlev V.I., Orlov A.O.

Russia, Bauman Moscow State Technical University
forewar@gmail.com

The article describes methods of working with noise in the source data at constructing models of data analysis in decision support systems. The paper consists of five parts. The introduction describes the problem of presence of distortions in the input data when decision support systems are in operation. The authors set the task of research and development of methods of analyzing distortions in the data in the decision tree model. The second part of the paper is a survey of existing algorithms of decision trees and methods of work with corrupted data in them. Also the authors reviewed the literature on existing methods for anomaly search in data. The third part describes a method for estimating a local anomaly; the authors propose an extension of this method by using a new formula for calculating the distance between the values of a categorical attribute. In the fourth part the authors propose a method for detection of anomalies in the source data at constructing a predictive model of the decision tree in decision support systems. In the conclusion, the results of the study are listed.

Publications with keywords: [decision support systems \(DSS\)](#), [decision tree model](#), [data anomalies](#), [local occurrence frequency method](#)

Publications with words: [decision support systems \(DSS\)](#), [decision tree model](#), [data anomalies](#), [local occurrence frequency method](#)

References

1. Vagin V.N., Golovina E.Iu., Zagorianskaia A.A., Fomina M.V. *Dostoverniy i pravdopodobnyy vyvod v intellektual'nykh sistemakh* [Reliable and plausible inference in intelligent systems]. Moscow, Fizmatlit, 2008. 712 p.
2. Minger J. An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning*, 1989, vol. 4, no. 2, pp. 227-243. DOI: 10.1023/A:1022604100933
3. Quinlan J.R. Induction of Decision Trees. *Machine Learning*, 1986, vol. 1, no. 1, pp. 81-106. DOI: 10.1023/A:1022643204877

4. Utgoff P.E. Incremental induction on Decision Trees. *Machine Learning*, 1989, vol. 4, no. 2, pp. 161-186. DOI: 10.1023/A:1022699900025
5. Chandola V., Banerjee A., Kumar V. Anomaly detection: A Survey. *ACM Computing Surveys*, 2009, vol. 41, no. 3, article 15. DOI: [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882)
6. 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, pp. 93-104.
7. Orlov A.I. *Ekonometrika* [Econometrics]. Moscow, Ekzamen, 2002. 576 p.
8. Orlov A.O. Problema poiska rasstoianii mezhdu znacheniiami kategorial'nykh atributov pri obnaruzhenii vybrosov v dannyykh [The problem of search distances between values of categorical attributes detection emissions data]. *V mire nauchnykh otkrytii* [In the World of Scientific Discoveries], 2012, no. 8.1 (32), pp. 142-155.
9. 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, GA, USA, SIAM, 2008, pp. 243-254.
10. Kuzovlev V.I., Lipkin D.I. *Opredelenie bazovykh pokazatelei dostovernosti obrabotki informatsii proektnykh reshenii ASOIU* [The definition of basic confidence factors of processing of information of designed solutions of the computer-aided systems of information processing and control]. Moscow, 2001. 12 p. Dep. VINITI no. 1094-V2001.