НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н. Э. БАУМАНА
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408
электронный научно-технический журнал

Прогнозный анализ данных методом ID3O

 # 10, октябрь 2012
DOI: 10.7463/1012.0483286
Кузовлев В. И., Орлов А. О.
УДК 004.052.42

Россия, МГТУ им. Н.Э. Баумана
forewar@gmail.com

Введение

Прогнозный анализ данных применяется в системах поддержки принятия решений. Суть прогнозного анализа заключается в формировании суждений о будущих фактах на основе анализа некоторого набора исходных данных. Такой набор данных называется обучающим множеством, а процесс анализа исходных данных называется обучением с учителем. В процессе обучения строится прогнозная модель, которая далее используется для формирования прогнозов. В данной статье рассматривается модель решающего дерева (decision tree model) [1-3]. Деревья решений организованы в виде иерархической структуры, состоящей из узлов принятия решений по оценке значений определенных переменных для прогнозирования результирующего значения. Любое дерево решений выводит прогнозируемое значение, полученное в результате оценки некоторых входных атрибутов. Каждый уровень в дереве может рассматриваться как одно из решений; узел принятия решений 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.

Существует достаточно много алгоритмов, реализующих принципы модели деревьев решений [1-3]: ID3, C4.5, ДРЕВ, ID5R, Reduce. Все эти алгоритмы построения дерева решений не предполагают возможности наличия искажений в данных. Отсутствующие или искаженные данные оказывают существенное влияние на конечный вид построенного дерева решений и, как следствие, на результат работы модели.

В реальных системах поддержки принятия решений формирование прогнозной модели подвержено влиянию искажений в данных, по которым строится прогнозная

модель. В [4] обсуждаются такие типы искажений в данных, как «отсутствие значений» и «аномальные значения». Рассматривается влияние искажений этих типов на процесс формирования прогнозной модели при использовании различных алгоритмов построения дерева решений.

Зачастую искажения в исходных данных не могут быть исправлены автоматически с приемлемым уровнем точности. В таком случае требуется ручная проверка и исправление таких данных. Процесс ручной обработки сталкивается с проблемой ограниченности временных и материальных ресурсов.

Постановка задачи исследования

Целью данной работы является разработка методики обработки шума в данных и основанного на ней алгоритма построения дерева решений, решающего следующие проблемы, не учитывающиеся в существующих алгоритмах построения 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 и отсутствующие значения. Для обработки аномальных значений необходимо использовать методы поиска аномалий в данных. Для обработки отсутствующих значений необходимо использовать алгоритмы заполнения пропусков в данных. Для решения проблемы выбора эффективной стратегии повышения качества данных необходим некоторый показатель качества данных, который позволит ранжировать объекты данных для достижения максимальной эффективности ручной обработки данных.

Методика обработки шума в данных

Обработка объектов данных с целью устранения шума может проводиться в автоматическом режиме или вручную.

Ручная обработка объектов данных необходима в тех случаях, когда автоматический подбор значений методом ближайших соседей представляется некорректным с точки зрения смыслового наполнения информационного элемента. Зачастую различные атрибуты объектов данных имеют разные значения для реальных процессов, то есть разным уровнем затрат от искажения этих атрибутов. Поэтому имеет значение понятие стоимости восстановления значений отдельных атрибутов.

Выбор стратегии повышения качества данных важен в реальных условиях ограниченности материальных ресурсов. Необходимо использовать такой механизм выбора стратегии повышения качества данных, который позволит максимизировать эффект от повышения качества данных в условиях ограниченности ресурсов. В [6-8] предлагаются критерии для оценки качества информационных элементов при ручной обработке данных. Появляется возможность обеспечить такую методику обработки, когда в первую очередь будут обрабатываться те объекты, значимость которых определены как наивысшая. Таким образом, обеспечивается интенсивность повышения качества данных в условиях параллельной эксплуатации системы [7].

При автоматической обработке значения атрибутов объектов, являющиеся шумом, должны быть скорректированы. Одним из наиболее эффективных и популярных методов коррекции таких данных является метод «ближайших соседей» [1]. Этот метод основан на предположении о том, что если объекты схожи по некоторому подмножеству их атрибутов, то эти же объекты, вероятно, будут схожи и по другому подмножеству атрибутов. Таким образом, для объекта, содержащего шум в значениях некоторого атрибута, находятся наиболее близкие объекты, которые не содержат шума в данном атрибуте. После этого шум корректируется на основании значений соответствующих атрибутов ближайших объектов. Для расчета близости объектов используется некоторая заданная заранее метрика. Обработка выбросов данных происходит в два этапа. На первом этапе выбросы в данных необходимо идентифицировать. Для идентификации аномалий применяется метод выявления аномалий, предложенный авторами в [4]. Этот метод основан на механизме LOF [5]. На втором этапе обнаруженные объекты подлежат обработке.

Для идентификации шума, характеризующегося пропусками в данных, не требуется дополнительных процедур идентификации. Объекты, имеющие пустые значения атрибутов, помечаются как объекты с шумом. Для заполнения пропущенного значения атрибута некоторого объекта проводится поиск наиболее близких объектов, после чего пропущенное значение восстанавливается на основе значений найденных объектов.

Для нахождения ближайших объектов необходимо ввести метрику, позволяющую рассчитывать расстояния между объектами. Для расчета расстояний между объектами предлагается формула, основанная на широко известной формуле расчета расстояния Хэмминга. Если имеются два объекта X_i, X_j из множества объектов X, тогда расстояние между объектами определяется по формуле

$$D(X_i, X_j) = \sum_{p=1}^{k} d(x_{ip}, x_{jp}) \cdot V(x_{ip}) \cdot V(x_{jp}), \quad (1)$$

где k – количество атрибутов объектов, $d(x_{ip}, x_{jp})$ – расстояние между значениями p-го атрибута объектов, $V(\cdot)$ – вес соответствующего значения атрибута объекта, $V(x_{ip}) = 1$ в том случае, если значение атрибута x_{ip} не помечено как аномальное, в противном случае $V(x_{ip}) = 0$.

Расстояние $d(x_{ip}, x_{jp})$ для непрерывных признаков [6] равно

$$d(x_{ip}, x_{jp}) = \frac{|x_{ip} - x_{jp}|}{|a_{pmax} - a_{pmin}|}, \quad (2)$$

где $|a_{pmax} - a_{pmin}|$ – разница между максимальным и минимальным значениями p-го атрибута среди всех объектов.

Если одно из значений непрерывного атрибута неизвестно ($x_{jp} = null$), то расстояние определяется по известному значению [6] формулой

$$\mathrm{d}(\mathrm{x_{ip}}, \mathrm{x_{jp}}) = \frac{\max\left((\mathrm{x_{ip}} - \mathrm{a_{pmax}}), (\mathrm{x_{ip}} - \mathrm{a_{pmin}})\right)}{|\mathrm{a_{pmax}} - \mathrm{a_{pmin}}|}. \quad (3)$$

Если оба значения неизвестны, расстояние максимально и равно единице $d(x_{ip}, x_{jp}) = 1$.

Для дискретных атрибутов расстояние вычисляется при условии, что $x_{ip} \neq x_{jp}$:

$$d(x_{ip}, x_{jp}) = dist_{A_p}(x_i, x_j) \cdot V(x_{ip}) \cdot V(x_{jp}). \quad (4)$$

Если $x_{ip} = x_{jp}$, тогда $d(x_{ip}, x_{jp}) = 0$. Если одно из значений дискретных атрибутов неизвестно ($x_j = null$), тогда расстояние равно

$$d(x_{ip}, x_{jp}) = dist_{A_p}(x_i, x_{maxp}), \quad (5)$$

где x_{maxp}– наиболее частое значение p-го атрибута, то есть $\mathrm{f_A}(\mathrm{x_{maxp}}) = max(f_A(x_z))$, где $z = \overline{1, k}$, где k – количество различных значений атрибута

A_p. Если оба значения дискретных атрибутов неизвестны, тогда $d(x_{ip}, x_{jp}) = 1{,}5$, что является округленным в большую сторону максимальным расстоянием.

Формула расчета $dist_{A_n}(x_i, x_j)$ предложена и обоснована авторами в [4]:

$$dist_{A_n}(x_i, x_j) = \sqrt[2]{\frac{f_n(x_i) + f_n(x_j)}{f_n(x_i) \cdot f_n(x_j)}}. \quad (6)$$

Найденные наиболее близкие объекты участвуют в формировании нового значения атрибута. В случае количественного атрибута, пустое значение заполняется средним арифметическим соответствующих значений атрибутов ближайших объектов:

$$x'_{ip} = \sum_{j=1}^{h} \frac{x_{jp}}{h}. \quad (7)$$

Здесь h – количество наиболее близких объектов.

В случае качественного атрибута значение выбирается как наиболее часто встречающееся среди соответствующих значений атрибутов наиболее близких объектов.

В таблице 1 приводится алгоритм заполнения отсутствующих значений атрибутов данных, основанный на вычислении расстояния Хэмминга между объектами данных с учетом возможного появления аномальных значений атрибутов данных.

Таблица 1. Алгоритм заполнения отсутствующих значений атрибутов данных

Алгоритм заполнения отсутствующих значений атрибутов данных

***Вход*:** множество объектов $X = \{X_1, X_2, \dots, X_n\}$, содержащих пустые значения и выбросы в данных; множество объектов, содержащих аномалии $X' = \{X'_1, \dots, X'_m\}$, где $m \le n$; множество весов значения атрибута A_j объектов с аномалиями $V = \{V_1, \dots, V_m\}$, причем $V_i = \{0, 1\}$; параметр K, определяющий количество искомых ближайших объектов.

***Выход*:** множество объектов $X'' = \{X''_1, X''_2, \dots, X''_n\}$, не содержащих пустых значений атрибутов.

Начало алгоритма.

Шаг 1. Начало цикла по X. Выбрать следующий объект $X_i \in X$.

Шаг 2. Если X_i не содержит пустых значений атрибутов, то добавить этот объект в выходное множество $X'' = X'' \cup \{X_i\}$ и перейти к шагу 1. Иначе перейти к шагу 3.

Шаг 3. Сформировать множество $D(X_i) = \{D(X_i, X_1), \dots, D(X_i, X_{i-1}), D(X_i, X_{i+1}), \dots, D(X_i, X_n)\}$ расстояний от объекта X_i до остальных объектов из X.

Шаг 4. Из множества $D(X_i)$ выбрать K наиболее близких объектов к X_i и сформировать из них множество $D_K = \{X_{k1}, \dots, X_{kK}\}$.

Шаг 5. Заполнить пустые значения атрибутов объекта X_i на основании значений атрибутов объектов из D_K.

Шаг 6. Добавить объект в выходное множество $X'' = X'' \cup \{X_i\}$.

Шаг 7. Конец цикла по X.

Конец алгоритма.

Алгоритм ID3O

В [1] предлагается алгоритм IDTUV, использующий алгоритмы построения дерева решений ID3 и C4.5 совместно с алгоритмом ВОССТАНОВЛЕНИЕ, позволяющим восстанавливать пропущенные данные.

Для шума типа «аномальные значения» в [4] предлагается метод выявления аномалий в исходных данных.

На основе исследованных методов и подходов разработан алгоритм ID3O, представленный на рисунке 1. Данный алгоритм использует предложенную методику обработки шума в данных и позволяет строить модель дерева решений с учетом шума в данных, а также ограниченности ресурсов при ручной обработке.

На первом этапе происходит выбор стратегии повышения качества данных в соответствии с показателями, предложенными в [8]. На втором этапе происходит повышение качества данных по предложенному алгоритму заполнения отсутствующих атрибутов данных, а также по предложенному в [4] алгоритму выявления аномалий. Далее строится дерево решений по алгоритму IDTUV [1].

Рисунок 1. Алгоритм ID3O

Необходимость проверки прогнозной модели на различных наборах исходных данных обоснована тем фактом, что разные наборы данных содержат разное количество числовых и категориальных атрибутов, а также различно количество уникальных значений этих атрибутов. Таким образом, проверка на различных наборах исходных данных позволит составить наиболее полную картину работы сравниваемых алгоритмов.

Для проверки точности классификации используется подготовленная заранее тестовая выборка, объекты которой уже классифицированы экспертами. В результате анализа сравниваются результаты классификации тестовой выборки построенной прогнозной моделью с результатами классификации экспертов, которая признается эталонной. Для оценки точности используется критерий $ErrRatio$, называемый коэффициентом ошибки классификатора [1]. Данный критерий определяется как отношение числа неверно классифицированных объектов к общему числу объектов:

$$ErrRatio = \frac{|X_f|}{|X|}. \quad (8)$$

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

Для проверки точности распознавания использовались общедоступные наборы данных из коллекции Калифорнийского университета.

1) Данные задач Монахов [10].

2) Данные для распознавания типов цветков Ириса. Одна из наиболее популярных выборок данных для проверки моделей классификации [9].

3) Данные для распознавания флагов стран [8].

4) Данные о сердечной недостаточности проекта Statlog [11].

На первом этапе проводились эксперименты над различными наборами данных, не содержащих шум.

В таблице 2 представлены данные экспериментов над тестовыми наборами данных, не содержащих шума. Алгоритмы IDTUV и ID3O выдают результат, аналогичный результатам алгоритмов ID3 и C4.5 в случаях, когда в наборе данных содержатся только категориальные или только числовые атрибуты соответственно. В среднем по всем наборам данных алгоритмы IDTUV и ID3O показали лучшую точность классификации по сравнению с ID3 и C4.5, что подтверждает результаты, полученные в [1].

Таблица 2. Точность классификации на данных без шума

	ID3	**C4.5**	**IDTUV**	**ID3O**
Monks	94,4	90,3	94,4	94,4
Iris	0	92,7	92,7	92,7
Flags	47,8	73,9	69,6	69,6
StatlogHeart	0	77,3	83,3	83,3
Среднее	**35,55%**	**83,5%**	**85%**	**85%**

На втором этапе в наборы данных вносился шум типов «отсутствие значений атрибутов» и «аномальные значения атрибутов». На втором этапе сравнивались результаты работы алгоритмов IDTUV и ID3O, поскольку они обрабатывают и корректируют шум в данных. Результаты представлены в таблице 3.

Таблица 3. Точность классификации на данных с шумом

Шум 5%		
	IDTUV	**ID3O**
Monks	93,6	94,4
Iris	90,3	92,7
Flags	68,5	69,6
StatlogHeart	77,7	83,2
Среднее	**82,53%**	**84,98%**

Шум 10%		
	IDTUV	**ID3O**
Monks	92,8	94,1
Iris	88,8	92,6
Flags	66,3	69,1
StatlogHeart	76,3	81,8
Среднее	**81,05%**	**84,4%**

Шум 20%		
	IDTUV	**ID3O**
Monks	90,7	92,8
Iris	87,9	92,2
Flags	64,2	65,2
StatlogHeart	74,2	78,1
Среднее	**79,25%**	**82,08%**

Заключение

По итогам экспериментов стало ясно, что предложенный в данной работе алгоритм ID3O имеет высокую устойчивость к искажениям в данных благодаря механизмам поиска и устранения аномалий в данных, а также заполнения пропущенных значений атрибутов. Так, при уровне шума в 5 % средний результат по всем наборам данных всего на 0,02 % ниже результата при отсутствии шума. При шуме в 10 % и 20 % снижение точности классификации составило в среднем 0,6 % и 2,92 % соответственно, что существенно меньше, чем снижение точности классификации других рассмотренных алгоритмов при аналогичных уровнях шума.

Таким образом, предложенный алгоритм ID3O показал способность к эффективной работе с данными, содержащими шум.

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. Quinlan J.R. Induction of Decision Trees // Machine Learning. 1986. Vol. 1, no. 1. P. 81-106. DOI: 10.1023/A:1022643204877
3. Utgoff P.E. Incremental induction on Decision Trees // Machine Learning. 1989. Vol. 4, no. 2. P. 161-186. DOI: 10.1023/A:1022699900025
4. 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: http://dx.doi.org/10.7463/0912.0483269
5. 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.
6. Кузовлев В.И., Липкин Д.И. Определение базовых показателей достоверности обработки информации проектных решений АСОИУ. М., 2001. 12 с. Деп. в ВИНИТИ № 1094-В2001.
7. Кузовлев В.И., Орлов А.О. Учет взаимосвязей между объектами результатов профилирования // Наука и образование. МГТУ им. Н. Э. Баумана. Электрон. журн. 2012. № 08. Режим доступа: http://engbul.bmstu.ru/doc/482766.html (дата обращения 03.09.2012).
8. Кузовлев В.И., Орлов А.О. Вероятностный подход к оценке показателя достоверности элементов результатов профилирования // Вестник МГТУ им. Н.Э. Баумана, 2012. Режим доступа: http://vestnik.bmstu.ru/catalog/it/hidden/115.html (дата обращения 21.11.2012).
9. UCI Machine Learning Repository: Flags Data Set. Режим доступа: http://archive.ics.uci.edu/ml/datasets/Flags (дата обращения 02.09.2012).
10. UCI Machine Learning Repository: Iris Data Set. Режим доступа: http://archive.ics.uci.edu/ml/datasets/Iris (дата обращения 02.09.2012).
11. UCI Machine Learning Repository: Monk's Problems Data Set. Режим доступа: http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems (Дата обращения: 02.09.2012)
12. UCI Machine Learning Repository: Statlog Heart Data Set. Режим доступа:. http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 (Дата обращения: 02.09.2012)

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU
SCIENCE and EDUCATION
EL № FS77 - 48211. №0421200025. ISSN 1994-0408
electronic scientific and technical journal

Prognostic analysis of data by ID3O

 # 10, October 2012
DOI: 10.7463/1012.0483286

Kuzovlev V.I., Orlov A.O.

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

The article deals with prognostic analysis, namely construction of a prognostic model of the decision tree. The introduction describes principles of the decision tree model and singles out significant problems in algorithms for constructing a decision tree - in particular, the problem of building a decision tree in the presence of noise in the data. The second part describes methods of automatic and manual processing of noise in the data, indicates the problem of limitation of material resources and time resources at manual processing. The third part describes known methods of noise processing in the data and building a decision tree model. On the basis of the research, the authors propose the ID3O algorithm for building a decision tree model in the presence of noise in the raw data and with limited resources for processing and improving data quality. The conclusion presents the results of the proposed algorithm in comparison with existing methods of building a prognostic model of the decision tree.

Publications with keywords: decision tree model, prognostic data analysis, noise in data, the ID3O method
Publications with words: decision tree model, prognostic data analysis, noise in data, the ID3O method

References

1. Vagin V.N., Golovina E.Iu., Zagorianskaia A.A., Fomina M.V. *Dostovernyi i pravdopodobnyi vyvod v intellektual'nykh sistemakh* [Reliable and plausible inference in intelligence systems]. Moscow, Fizmatlit, 2008. 712 p.
2. Quinlan J.R. Induction of Decision Trees. *Machine Learning*, 1986, vol. 1, no. 1, pp. 81-106. DOI: 10.1023/A:1022643204877
3. Utgoff P.E. Incremental induction on Decision Trees. *Machine Learning*, 1989, vol. 4, no. 2, pp. 161-186. DOI: 10.1023/A:1022699900025

4. Kuzovlev V.I., Orlov A.O. Metod vyiavleniia anomalii v iskhodnykh dannykh pri postroenii prognoznoi modeli reshaiushchego dereva v sistemakh podderzhki priniatiia reshenii [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 and Education of the Bauman MSTU], 2012, no. 9. DOI: http://dx.doi.org/10.7463/0912.0483269
5. 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.
6. 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.
7. Kuzovlev V.I., Orlov A.O. Uchet vzaimosviazei mezhdu ob"ektami rezul'tatov profilirovaniia [Accounting interconnections between objects profiling results]. *Nauka i obrazovanie MGTU im. N.E. Baumana* [Science and Education of the Bauman MSTU], 2012, no. 8. Available at: http://engbul.bmstu.ru/doc/482766.html , accessed 03.09.2012.
8. Kuzovlev V.I., Orlov A.O. Veroiatnostnyi podkhod k otsenke pokazatelia dostovernosti elementov rezul'tatov profilirovaniia [A probabilistic approach to the evaluation of the indicator of reliability of elements of profiling results]. *Vestnik MGTU im. N.E. Baumana* [Herald of the Bauman MSTU], 2012. Available at: http://vestnik.bmstu.ru/catalog/it/hidden/115.html , accessed 21.11.2012.
9. *UCI Machine Learning Repository: Flags Data Set.* Available at: http://archive.ics.uci.edu/ml/datasets/Flags , accessed 02.09.2012.
10. *UCI Machine Learning Repository: Iris Data Set.* Available at: http://archive.ics.uci.edu/ml/datasets/Iris , accessed 02.09.2012.
11. *UCI Machine Learning Repository: Monk's Problems Data Set.* Available at: http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems , accessed 02.09.2012.
12. *UCI Machine Learning Repository: Statlog Heart Data Set.* Available at: http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 , accessed 02.09.2012.