УДК 519.23

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ НАБЛЮДЕНИЙ ПРИ КЛАССИФИКАЦИИ ОБЪЕКТОВ

© 2014 В.А. Алексеева, Ю.С. Донцова, В.Н. Клячкин

Ульяновский государственный технический университет

Поступила в редакцию 28.05.2014

Рассматривается задача классификации объектов, при которой часть исходной информации утрачена, и ее необходимо восстановить. Исследуется эффективность различных методов восстановления пропущенных наблюдений. Проводится сравнительный анализ эффективности 4 методов восстановления нечисловых данных при классификации объектов.

Ключевые слова: восстановление, наблюдение, классификация, объект, моделирование

Рассматривается задача классификации объектов, в которой каждый объект характеризуется *m*-мерным вектором признаков $(X_1 ... X_m)$. Предположим, что некоторые наблюдаемые значения данных признаков, которые имеют нечисловую природу, были утрачены (X_i=*N*) в силу определенных обстоятельств (см. табл. 1). Следовательно, возникает задача восстановления нечисловых данных.

Таблица 1. Выборка данных с пропущенными значениями для *n* объектов

	X_1	X_2	X_3	X_4	…	X_m
1	x_{11}	x_{12}	*N*	x_{14}	…	x_{1m}
2	x_{21}	*N*	x_{23}	x_{24}	…	x_{2m}
3	x_{31}	x_{32}	x_{33}	x_{34}	…	*N*
…	…	…	…	…	…	…
n	x_{n1}	*N*	*N*	*N*	…	x_{nm}

Проблема восстановления пропущенных данных исследуемых объектов возникает при решении многих практических задач. В матрицах исходных наблюдений по разным причинам (неисправность измерительного прибора, грубая ошибка при подготовке данных, удаление резко выделяющихся наблюдений и т.д.) могут появляться пропуски отдельных элементов или каких-то частей выборки. Исключать по причине потери данных из дальнейшего исследования весь объект (строку, в которой обнаружены пропуски) или признак (столбец, в котором обнаружены пропуски) нецелесообразно. Неполная априорная информация объектов, как правило, усложняет процесс построения и дальнейшего применения различных математических моделей, а также может привести к неадекватным результатам. В связи с этим возникает задача поиска наилучшего метода восстановления пропущенных наблюдений по некоторому критерию качества. Выбор критерия восстановления стертых данных производится исходя из характера последующей обработки данных и в зависимости от окончательных целей исследования [1].

В зависимости от решаемой проблемы исследователю может потребоваться либо оценить некоторые параметры при наличии пропущенных значений, либо оценить сами пропущенные значения, либо то и другое вместе. Две последние задачи требуют больше исходных допущений, чем задача оценки параметров. Методы их решения основаны на использовании некоторой избыточной информации, которая возникает вследствие связи между признаками.

В настоящее время наиболее распространенными методами по восстановлению пропущенных данных являются такие, как заполнение пропусков средними значениями, метод ближайших соседей, регрессионный метод, метод максимального правдоподобия и EM-алгоритм, алгоритм ZET, алгоритм ZetBraid, метод Бартлетта, Resampling, эволюционный метод и другие. Однако перечисленные методы работают с данными объектов, значения которых представлены в числовой форме [2-3]. Для решения сформулированной задачи (с учетом нечисловой природы наблюдений) воспользуемся следующими методами [4-5].

Метод 1: замена пропущенных значений на моду. Как правило, мода представляет собой значение на множестве наблюдений, которое

Алексеева Венера Арифзяновна, кандидат технических наук, доцент кафедры «Прикладная математика и информатика». E-mail: v.fashutdinova@mail.ru

Донцова Юлия Сергеевна, аспирантка

Клячкин Владимир Николаевич, доктор технических наук, профессор кафедры «Прикладная математика и информатика». E-mail: v_kl@mail.ru

встречается наиболее часто. Поскольку наблюдаемые значения объекта носят случайный характер, определим по имеющимся данным дискретное распределение пропущенного параметра (см. табл. 2) и затем во всех записях, где он отсутствует, проставим его моду. Этот способ хорошо применять, когда отсутствующих данных относительно мало.

Таблица 2. Дискретное распределение пропущенного параметра по всей выборки

N	$N=X_1$	$N=X_2$	$N=X_3$	$N=X_4$	…	$N=X_m$
P	p_1	p_2	p_3	p_4	…	p_m

здесь $p_1,\ldots, p_m-$ вероятности, причем $\sum_{i=1}^{m} p_i(N = X_i) = 1.$

Метод 2: замена пропущенного значения на моду, но с использованием условного распределения по присутствующим параметрам. Данный метод (см. табл. 3) в отличие от первого метода требует больше вычислений.

Таблица 3. Условное распределение пропущенного параметра для первого объекта

	X_1	X_2	X_3	X_4	…	X_m
X_1	p_{11}	p_{12}	p_{13}	p_{14}	…	p_{1m}
X_2	p_{21}	p_{22}	p_{23}	p_{24}	…	p_{2m}
X_3	p_{31}	p_{32}	p_{33}	p_{34}	…	p_{3m}
X_4	p_{41}	p_{42}	p_{43}	p_{44}	…	p_{4m}
…	…	…	…	…	…	
X_m	p_{m1}	p_{m2}	p_{m3}	p_{m4}	…	p_{mm}

здесь $p_{ij}-$ вероятности переходов, причем для всех i: $\sum_{j=1}^{m} p_{ij} = 1.$

Метод 3: моделирование пропущенных данных с использованием дискретного распределения пропущенного параметра, но теперь в каждом случае (для каждого объекта) производится случайный эксперимент с использованием этого распределения, и на место отсутствующего значения записывается исход этого эксперимента.

Метод 4: моделирование пропущенных данных с использованием условного распределения по присутствующим параметрам, но теперь в каждом случае (для каждого объекта) производится случайный эксперимент с использованием этого условного распределения, и на место отсутствующего значения записывается исход этого эксперимента.

Для верификации описанных выше методов рассмотрим выборку, состоящую из 1000 объектов, каждый из которых характеризуется 11 признаками [6]. Из данной выборки сформируем тестовую выборку, на которой случайным образом смоделируем различные типы пропусков (в начале выборки, в конце выборки, в середине выборки, несколько пропусков подряд и т.д.), к которым применимы методы восстановления пропущенных значений. Результаты применения этих методов представлены в виде таблиц. Заметим, что замена пропущенного значения на моду с использованием условного распределения по присутствующим параметрам (метод 2) в рассматриваемой выборке верно восстановила менее половины пропущенных значений.

Таблица 4. Результаты восстановления пропущенных значений с помощью метода 1

№ эксп.	**Верно восстановленные значения**	**Ошибочно восстановленные значения**
1	87%	13%
2	86%	14%
3	86%	14%
4	85%	15%

Таблица 5. Результаты восстановления пропущенных значений с помощью метода 3

№ эксп.	**Верно восстановленные значения**	**Ошибочно восстановленные значения**
1	92%	8%
2	93%	7%
3	90%	10%
4	92%	8%

Таблица 6. Результаты восстановления пропущенных значений с помощью метода 4

№ эксп.	**Верно восстановленные значения**	**Ошибочно восстановленные значения**
1	92%	8%
2	93%	7%
3	90%	10%
4	92%	8%

Представленные в табл. 4-6 результаты исследования позволяют сделать вывод о том, что третий и четвертый метод с использованием распределения параметра (простого и условного) для каждого объекта имеют более высокую точность восстановления пропущенных значений – не менее 90%. Следовательно, применение данных методов позволит сохранить в исходной выборке больше полезной информации, которая необходима для адекватного построения

математических моделей при решении практических задач классификации объектов.

Рассмотренный пример позволяет предложить следующий алгоритм восстановления пропущенных значений:

1. Из заданной (исходной) выборки отбирается подмножество данных, не имеющих пропущенных значений.
2. На этом подмножестве моделируются различные типы пропусков, характерных для исходной выборки.
3. Смоделированные пропущенные данные восстанавливаются с использованием различных методов.
4. По доле верно восстановленных значений выбирается наилучший метод.
5. Этот метод используется для восстановления реально пропущенных (а не смоделированных) данных в исходной выборке.

Работа выполнена в рамках задания Минобрнауки России №2014/232.

СПИСОК ЛИТЕРАТУРЫ:

1. *Айвазян, С.А.* Прикладная статистика. Основы моделирования и первичная обработка данных / *С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин*. – М.: Финансы и статистика, 1983. 471 с.
2. *Злоба, Е.* Статистические методы восстановления пропущенных данных / *Е. Злоба, И. Яцкие* // Computer Modelling & New Technologies. 2002. Vol. 6, № 1. P. 51-61.
3. *Снитюк, В.Е.* Эволюционный метод восстановления пропусков в данных // Сборник трудов междунар. конф. «Интеллектуальный анализ информации». – Киев: 2006. С. 262-271.
4. *Шепелева, М.В.* Модели кредитного и поведенческого скоринга. http://www.masters.donntu.edu.ua/2006/kita/shepeleva/library/metod%20scoring.pdf
5. *Chen, G.G.* Bound and collapse bayesian reject inference when data are missing not at random / *G.G. Chen, T. Astebro* // Mathematical Approaches to Credit Risk Management. Conference Proceedings, Banff International Research Station for Mathematical Innovation and Discovery. 2003. 205 p.
6. *Клячкин, В.Н.* Сравнительный анализ точности нелинейных моделей при прогнозировании состояния системы на основе марковской цепи / *В.Н. Клячкин, Ю.С. Донцова* // Известия Самарского научного центра Российской академии наук. 2013. Т. 15. № 4(4). С. 924-927.

MISSING OBSERVATIONS RECOVERY DURING OBJECTS CLASSIFICATION

© 2014 V.A. Alekseeva, Y.S. Dontsova, V.N. Klyachkin

Ulyanovsk State Technical University

It is considered the problem of objects classification where part of initial information is lost and it is necessary to recover it. It is studied efficiency of different missing observations recovery methods. It is performed a comparative analysis of four non-numeric observations recovery methods efficiency at objects classification.

Key words: *recovery, observation, classification, object, modeling*

Venera Alekseeva, Candidate of Technical Sciences, Associate Professor at the Department «Applied Mathematics and Computer Science». E-mail: v.fashutdinova@mail.ru
Yuliya Dontsova, Post-graduate Student
Vladimir Klyachkin, Doctor of Technical Sciences, Professor at the Department «Applied Mathematics and Computer Science». E-mail: v_kl@mail.ru