OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.8

АЛГОРИТМ ПРЕДОБРАБОТКИ И ВОССТАНОВЛЕНИЯ АНКЕТНЫХ ДАННЫХ

Сибирев И.В., Афанасьева Т.В.

Ульяновский государственный технический университет,
г. Ульяновск, Россия

ivan.sibirev@yandex.ru

tv.afanasjeva@gmail.com

В статье предлагается алгоритм, восстанавливающий недостающие анкетные данные, на основе методов нечёткой кластеризации.
Ключевые слова: предобработка анкетных данных; восстановление данных; нечёткая кластеризация.

Введение

В настоящее время актуальными проблемами являются проблемы обработки данных, в частности, результатов анкетирования, данных, снятых с множества датчиков, результатов целевого запроса и т.д. Данные могут восприниматься как «кадр», характеризующий текущее состояние изучаемой системы, процессы и тенденции изучаемого объекта или явления.

В последние годы возникло множество фирм, берущихся за обработку анкетных данных: составление анкет, проведение анкетирования, перевод анкет в электронный вариант, повторное проведение анкетирования (если это необходимо), обработку анкетных данных вручную или при помощи программных средств. Некоторые фирмы берутся за разработку программного обеспечения для обработки анкетных данных.

Проблемой этапа предобработки анкетных данных является неполнота при заполнении анкет или недоступность отдельных данных.

Существуют два основных подхода при работе с недостающими данными. Первый подход – восстановление данных из сторонних источников информации. Это может быть повторное анкетирование, задание уточняющих вопросов, восстановление информации с использованием внешних баз данных. В IT мире в подавляющем большинстве случаев это реализуется с помощью многообразных систем контроля версий или резервного копирования, как в масштабах отдельного рабочего места, так и в масштабах отдельного сервера или DATA-центра.

Второй подход – редукция (удаление) данных. Если часть анкетных данных отсутствует, при этом невозможно провести повторное анкетирование или задать уточняющие вопросы, тогда в дело вступает редукция. Удаляются те вопросы и те анкеты, в которых не заполнены («повреждёны») данные. Часто удаляют наиболее незаполненные вопросы и анкеты. Позиции, оставшиеся пустыми, заполняют, например, самыми типичными ответами по каждому вопросу для данного анкетирования, либо используют статистические модели восстановления данных. Обзор статистических моделей восстановления данных в кластерном анализе приведен в работе [Busse и др, 2005]. Редукция и грубое заполнение пустых мест значительно ухудшают качество анализируемых материалов.

Цель некоторых исследований, обрабатывающих большие объёмы данных, – не в точной конкретизации каждого отдельного параметра, но – в описании общей картины, состояний и тенденций, в том числе, в терминах нечёткой логики.

Нами предлагается алгоритм восстановления данных на стадии предобработки, позволяющий избежать избыточной редукции. Данный алгоритм относится к неточным, но он позволяет заполнить недостающие данные с большей точностью, чем заполнение типичным представителем или средним арифметическим, приводит к меньшим искажениям общей картины при обработке данных методами многомерного анализа (кластеризация, факторный анализ и т.д.) .

Данный алгоритм основан на методах кластерного анализа (см.[Дюран и др, 1977], [Жамбю, 1988], [Ким и др., 1989], [Райзин и др, 1980], [Малышев и др, 1991], [Мандель, 1988] и др.). Мы разбиваем анкетные данные на несколько

кластеров. Затем с помощью нечёткой кластеризации (см. [Малышев и др, 1991])получаем коэффициенты принадлежности каждой отдельной анкеты к каждому кластеру. В дальнейшем, опираясь на таблицу принадлежности, уточняем незаполненные анкетные данные.

1. Алгоритм восстановления данных

Опишем алгоритм и проиллюстрируем его на примере вычислительного эксперимента.

Шаг 0. Входные данные – числовые или символ N, который обозначает незаполненный ответ на вопрос.

Таблица 1– Исходные данные

Параметры	П1	П2	П3	П4	П5	П6
Анкета 1	1	65	3	N	14	5
Анкета 2	N	23	72	N	7	26
Анкета 3	1	23	43	N	N	52
Анкета 4	124	45	N	N	57	N
Анкета 5	N	23	72	N	23	35
Анкета 6	59	56	43	N	45	12
Анкета 7	N	N	N	N	N	N

Шаг 1. Редукция. Удаляем полностью незаполненные параметры и анкеты.

Таблица 2. Данные после редукции

Параметры	П1	П2	П3	П5	П6
Анкета 1	1	65	3	14	5
Анкета 2	**N**	23	72	7	26
Анкета 3	1	23	43	**N**	52
Анкета 4	124	45	**N**	57	**N**
Анкета 5	**N**	23	72	23	35
Анкета 6	59	56	43	45	12

Шаг 2. Временно заполняем все незаполненные ответы средними значениями по параметру.

Таблица 3. Первичное заполнение недостающих данных

Параметры	П1	П2	П3	П5	П6
Анкета 1	1	65	3	14	5
Анкета 2	**46,25**	23	72	7	26
Анкета 3	1	23	43	**29,2**	52
Анкета 4	124	45	**46,6**	57	**26**
Анкета 5	**46,25**	23	72	23	35
Анкета 6	59	56	43	45	12

Шаг 3. Кластеризация анкетных данных. Может проводиться одним из методов кластерного анализа. Мы будем использовать центроидный метод [].

Шаг 3.1.Подбор количества кластеров.

Постараемся подобрать количество кластеров так, чтобы не было кластеров, для которых какой-либо параметр изначально был полностью неизвестным.

Ниже представлены разбиения на 2, 3, 4, 5 кластеров:
x5: C1:A1; C2:A2; C3:A4; C4:A6; C5: A3, A3; - нет;
x4: C1:A4; C2:A6; C3:A3, A5; C4:A1, A2; - нет;
x3: C1:A3, A5; C2:A1, A2; C3:A4, A6; - да;
x2: C1:A4, A6; C2:A1, A2, A3, A5; - да;
где x2, …, x5 – разбиения на 2, …, 5 кластеров, C1, C2, … – первый, второй и т.д. кластеры, A1, A2, … – анкета номер 1,2 … , C2:A1, A2, A3, A5; -означает что анкеты 1,2,3,5 попали в кластер 2.

После каждого разбиения выписано «да» или «нет», что означает отсутствие или присутствие в данном разбиении кластеров с полностью неизвестным параметром, соответственно.

Выберем разбиение, в котором количество кластеров было бы наибольшим при отсутствии кластеров с полностью неизвестным параметром. То есть – выбираем разбиение со словом «да», где наибольшее количество кластеров. Этим условиям в нашем примере соответствует разбиение на три кластера.

Шаг 3.2. Кластеризация центроидным методом на выбранное количество кластеров.

Таблица 4. Кластеризация центроидным методом на 3 кластера

C1	**П1**	**П2**	**П3**	**П5**	**П6**
Анкета 3	1	23	43	**29,2**	52
Анкета 5	**46,25**	23	72	23	35
C2					
Анкета 1	1	65	3	14	5
Анкета 2	**46,25**	23	72	7	26
C3					
Анкета 4	124	45	**46,6**	57	**26**
Анкета 6	59	56	43	45	12

C1, C2, C3- названия кластеров. П1, П2, П3, П4, П5 – названия параметров.

Шаг 4. Кластеризуем FCM-методом FCM-метод [Малышев и др, 1991] – метод нечёткой кластеризации, применяется в паре с другим методом кластеризации, в нашем случае с центроидным методом. Результатом нечёткой кластеризации является таблица коэффициентов принадлежности анкет к кластерам.

Таблица 5. FCM кластеризация

FCM	C1	C2	C3
Анкета 1	0.927	0.024	0.049
Анкета 2	0.047	0.041	0.913
Анкета 3	0.385	0.098	0.517
Анкета 4	0.018	0.95	0.032
Анкета 5	0.018	0.019	0.963
Анкета 6	0.249	0.383	0.372

Шаг 5. Уточнение временно заполненных анкетных данных.

Шаг 5.1. Для каждого кластера (из пункта 3.2) подсчитываем среднее значение каждого параметра по этому кластеру.

Таблица 6. Среднее значение каждого параметра по каждому кластеру

С1	**П1**	**П2**	**П3**	**П5**	**П6**
Анкета 3	1	23	43	**29,2**	52
Анкета 5	**46,25**	23	72	23	35
Среднее значение	23,63	23	57,5	26,1	43,5
С2					
Анкета 1	1	65	3	14	5
Анкета 2	**46,25**	23	72	7	26
Среднее значение	23,625	44	37,5	10,5	15,5
С3					
Анкета 4	124	45	**46,6**	57	**26**
Анкета 6	59	56	43	45	12
Среднее значение	91,5	50,5	44,8	51	19

Новое значение временно заполненных параметров анкеты вычислим как сумму произведений средних значений этого параметра для каждого кластера (из таблицы 6) на коэффициенты принадлежности данной анкеты кластеру (из таблицы 5). Результаты вычислений приведены в таблице 7.

Таблица 7. Результаты уточнения неизвестных параметров на первой итерации

Параметры	П1	П2	П3	П5	П6
Анкета 1	1	65	3	14	5
Анкета 2	**85,62**	23	72	7	26
Анкета 3	1	23	43	**37,44**	52
Анкета 4	124	45	**38,09**	57	**16,12**
Анкета 5	**88,99**	23	72	23	35
Анкета 6	59	56	43	45	12

Первая итерация закончена, переходим к шагу 3.

В нашем примере, спустя несколько итераций, были получены следующие анкетные данные.

Таблица 8. Результаты уточнения неизвестных параметров на четвёртой итерации.

Параметры	П1	П2	П3	П5	П6
Анкета 1	1	65	3	14	5
Анкета 2	**52.93**	23	72	7	26
Анкета 3	1	23	43	**41.13**	52
Анкета 4	124	45	**47.96**	57	**23.72**
Анкета 5	**53.87**	23	72	23	35
Анкета 6	59	56	43	45	12

Построим таблицу значений восстанавливаемых параметров по окончании каждой итерации.

Таблица 9. Значения неизвестных параметров после каждой итерации.

Итерация	А2П1	А3П5	А4П3	А4П6	А5П1
0	46.25	29.2	46.6	26	46.25
1	85.62	37.44	38.09	16.12	88.99
2	58.81	41.18	47.75	22.77	58.69
3	44.93	40.81	47.95	23.65	43.69
4	52.93	41.13	47.96	23.72	53.87

Здесь А4П6 – шестой параметр четвёртой анкеты.

Рисунок 1 иллюстрирует сходимость значений восстанавливаемых параметров при возрастании номера итерации. На горизонтальной оси – число итераций, на вертикальной оси – значения параметров. Каждая ломаная на рисунке соответствует значениям одного параметра на разных итерациях.

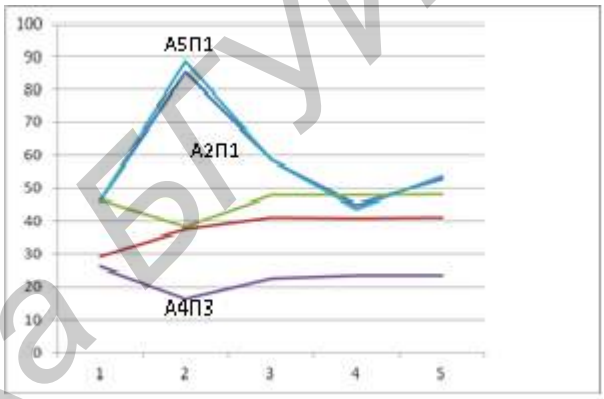

Рисунок 1. Генерируемые данные на первых нескольких итерациях

Наблюдается стремление значений параметров к некоторым предельным значениям при увеличении числа итераций.

2. Вычислительный эксперимент

Для изучения свойств алгоритма требуется апробация на реальных данных. Коллективом ученых Ульяновского государственного технического университета (Н.Г. Ярушкина, Т.В. Афанасьева, О.В. Шиняева, К.В. Святов, Л.М. Валкин, Д.А. Ефремов, К.Г. Калимуллин и др.) в рамках проекта «Исследование ИТ-кластера Ульяновской области». проведено анкетирование руководителей ИТ-предприятий г. Ульяновска. Собраны данные по 87 предприятиям по 39 вопросам анкеты [Ярушкина и др, 2013] . Нами проводились обработка и кластерный анализ этих данных [Афанасьева и др., 2014].

В результате редукции, с заданным порогом заполнения, количество обрабатываемых анкет сократилось с 87 до 49, количество обрабатываемых параметров сократилось с 39 до 33, то есть лишь 48% из потенциально заполняемых позиций подверглось дальнейшей обработке и анализу. Остальные 52% потенциально заполняемых позиций «оказались за бортом», причём из них были

заполнены 5% от общего числа позиций. Иными словами, в редуцированных данных было 10% заполненных позиций от числа удалённых.

С использованием алгоритма, описанного в этой статье, можно было не редуцировать данные вообще, т. к. не нашлось полностью незаполненных параметров или анкет. При этом классификационное значение мало заполненных параметров при кластеризации ничтожно.

Заключение

Данный алгоритм предназначен для восстановления данных на стадии предобработки, в частности, анкетных данных. В настоящий момент область применимости – вещественные данные, но в дальнейшем она может быть расширена за счёт некоторых модификаций.

Алгоритм не генерирует новую информацию, а распространяет известную информацию на незаполненные места анкет. Чем больше известной информации, тем информативнее будет результат.

Алгоритм не способен восстановить аномальные значения, но способен восстановить данные с учётом средних тенденций в этом и других кластерах. Данный алгоритм – более тонкий инструмент, чем «заливание» всех пустующих полей анкет средним арифметическим.

В текущей версии каждая итерация данного алгоритма может потребовать значительных вычислительных затрат. Спасает ситуацию быстрая сходимость алгоритма, что делает общее время вычислений приемлемым для практического использования. У этого алгоритма широкие возможности оптимизации быстродействия.

При использовании данного алгоритма нам удастся:
– сэкономить время и средства на повторном анкетировании;
– «спасти» значительную часть заполненных данных, которые были бы удалены при обычной редукции;
– восстановить незаполненные данные с точностью, позволяющей дальнейшую обработку методами многомерного анализа.

Данный алгоритм может быть использован на этапе предобработки данных как для исследования с использованием пакетного режима обработки данных, так и в исследованиях с использованием KDD систем.

Библиографический список

[Афанасьева и др., 2014] Программа «Сегментация и кластеризация рынка IT» / Т.В. Афанасьева, И.В. Сибирев // Инновации в науке. Сб.ст. по материалам XXIX междунар. науч.– практ. конф. №1. – Новосибирск : СибАК, 2014. – С. 46-53

[Дюран и др, 1977] Кластерный анализ / Б. Дюран, П. Оделл. – М.: Статистика, – 1977. – 128 с.

[Жамбю, 1988] Жамбю, М. Иерархический кластер-анализ и соответствия. Пер. с фр. / М. Жамбю– М.: Финансы и статистика, 1988. – 342 с.

[Ким и др., 1989] Факторный, дискриминантный и кластерный анализ: Пер с англ./ Дж. О. Ким. [и др.]; – М.: Финансы и статистика, 1989. – 215с.

[Райзин и др, 1980] Классификация и кластер. /Под ред. Дж. Вэн. Райзина. – М. : Мир, 1980, – 390 с.

[Малышев и др, 1991] Нечеткие модели для экспертных систем в САПР / Н.Г. Малышев и др . – М.: Энергоиздат, 1991. – 136 с.

[Мандель, 1988] Мандель, И Д. Кластерный анализ / И.Д. Мандель– М.: Финансы и статистика. 1988. – 176с.

[Ярушкина и др, 2013] Исследование ИТ-кластера Ульяновской области / Н. Г. Ярушкина [и др.]; – Ульяновск : УлГТУ, 2013. – 137 с.

[Busse и др, 2005] Cluster Analysis of Heterogeneous Rank Data. / L.M. Busse, P. Orbanz, J.M. – Zurich: Buhmann Institute of Computational Science, ETH Zurich, 8092

[Давыдов, 2015] Knowledge Discovery and Data Mining в системной социологии. [Режим доступа 2015] http://www.isras.ru/Davydov_Knowledge.html

ALGORITHM FOR PREPROCESSING AND RECOVERY OF QUESTIONNAIRES DATA

Sibirev I.V., Afanasyeva T.V.

Ulyanovsk State Technical University, Ulyanovsk, Russia

ivan.sibirev@yandex.ru;

tv.afanasjeva@gmail.com

The article proposes an algorithm for recovery missing questionnaires data. This algorithm bases on fuzzy clustering methods.

Introduction

The actual problem is preprocessing of data and recovery of missing questionnaires for allowing work by methods of multidimensional analysis based on fuzzy logic.

Main Part

We present algorithm for recovery missing questionnaires data in the preprocessing stage, avoiding excessive reduction.

This algorithm is based on the methods of cluster analysis. We divide data into several clusters. Then using fuzzy clustering we obtain the coefficients of membership of each questionnaire to each cluster. We based on the table of coefficients membership for upgrade questionnaires data.

The algorithm is illustrated by the example of the computational experiment.

We consider example where the results of the algorithm and the classical reduction are compared. The data are compared on the basis of a real experiment for clustering the IT-companies of the Ulyanovsk region.

Conclusion

This algorithm allows: to save time and resources on re-survey; to «save» main data part that would be reduced; to restore missing questionnaires data with precision, allowing work by methods of multidimensional analysis.