УДК 519-712.2

А.Н. Губский, С.А. Стенин, В.М. Корчинский

МЕТОД ВОССТАНОВЛЕНИЯ БИНАРНЫХ ДАННЫХ С ПРОПУСКАМИ

Аннотация. Предложен метод восстановления бинарных данных, основанный на максимизации критерия частоты совпадений данных в однородных группах источников информации.Предлагаемый метод является модификацией известных методов заполнения, где пропуски не имеют критического значения, однако в отличие от них он существенно повышает достоверность восстановления генеральной совокупности бинарных данных за счет анализа групповых свойств источников информации.

Ключевые слова: бинарные данные, источники информации, таблица опроса, критерий частоты совпадений.

Введение. С проблемой обработки пропусков в данных приходится сталкиваться в самых разнообразных приложениях статистического анализа [1-4]. Исследователи, как правило, стремятся как можно быстрее избавиться от пропусков с тем, чтобы впоследствии провести обработку “полных” данных стандартными средствами, мало задумываясь над тем, что такой подход может привести к сильному различию статистических выводов, сделанных при наличии пропусков в данных и при их отсутствии. Самыми распространенными приемами анализа данных с пропусками являются исключение некомпетентных наблюдений (т. е. содержащих хотя бы один пропуск) и традиционные методы заполнения пропусков. Эти методы в общем случае имеют малую эффективность, ведут, как правило, к смещенности и несостоятельности[5], к нарушению уровней значимости критериев и другим искажениям статистических выводов, не обладают устойчивостью к распределению пропусков.

Постановка задачи. Пусть данные формируются в виде прямоугольных таблиц . Строкам (столбцам) таблицы данных соответствуют различные источники информации. Тогда столбцы (строки) представляют собой исследуемые переменные (признаки, баллы, рейтинги

 ISSN 1562-9945

и т. д.). Элементами таблицы являются действительные числа(числовые характеристики продукции и товаров), бинарные числа (1,0) или (+, -), дискретные числа (например, при ранжировании критериев оценки качества).

Предполагается, что в таблице часть значений переменных отсутствует. Они могут отсутствовать по техническим причинам вследствие сбоя оборудования, либо мнение части источников информации не может оказать предпочтения одному критерию перед другим.

Задача автоматизированного восстановления данных с пропусками заключается в построении такого алгоритма восстановления данных, который на закономерностях поведения отдельных групп источников информации автоматически восстанавливает пропущенные в бинарной таблице данные с достаточной степенью состоятельности и достоверности.

Обзор существующих решений. Существуют четыре основные группы методов обработки данных с пропусками:

1. Методы исключения. При отсутствии некоторых переменных объекта мониторинга они удаляются из генеральной совокупности и оставшиеся данные обрабатываются. Эти методы легко реализуются и могут быть удовлетворительны при малом числе пропусков и большой генеральной совокупности данных. Однако иногда приводят к большим смещениям и не всегда бывают эффективными.

2. Методы заполнения.Пропуски заполняются и полученные данные обрабатываются обычными методами.Как правило, используются следующие процедуры: заполнение с выборочным подбором, когда подставляются значения переменных других объектов выборки; заполнение средними, когда подставляются средние присутствующих значений; заполнение с помощью регрессии, когда пропущенные значения оцениваются с помощью регрессии на присутствующие для анализируемого объекта переменные. Эти методы также не всегда эффективны, поэтому на практике в эти методы при решении конкретных прикладных задач, следует вводить модификации.

3. Методы взвешивания.Рандомизированные выводы по данным выборочных исследований с пропусками обычно построены на весах плана, обратно пропорциональных вероятности выбора. Взвешивание связано с заполнением средними

$$\frac{\sum p_i^{-1} x_i}{\sum p_i^{-1}}$$

где суммы берутся по извлеченным объектам. Методы взвешивания измеряют веса, чтобы учесть отсутствие значений. Подробно методы взвешивания описаны в [6].

4. Методы, основанные на моделировании.Методы основываются на построении модели порождения пропусков. Выводы получают с помощью функции правдоподобия, построенной при условии справедливости этой модели, с оцениванием параметров методами типа максимального правдоподобия. Преимущество этих методов состоит в том, что они являются гибкими, позволяют отказаться от методов, разработанных для частных случаев пропусков, и работать с неполными данными различного рода выборок на основе общего подхода максимизации функции правдоподобия.

Выбор того или иного метода зависит от характера данных и степени наличия пропусков в их генеральной совокупности.В частности, для обработки данных в виде бинарных таблиц, в которых элементы таблиц принимают значения 1 или 0 (“+” или “-”)и в которых есть пропуски,ниже предлагается метод, являющийся модификацией известных методов заполнения [7].

Метод восстановлениябинарных данных. Суть данного метода заключается в отыскании из анализа генеральной совокупности данных однородных групп источников информации и определении принадлежности источника информации, с которым временно была утеряна связь (наличие “пропуска”), к одной из выделенных групп. Далее используются традиционные процедуры методов заполнения. Покажем реализацию данного метода на следующем примере.

Пусть мы получили информацию от 10 источников с целью оценить покупательский спрос некоторой продукции или товара по 6 переменным признакам (критериям). Данные мониторинга сведены в таблицу 1. Здесь “+” – положительный ответ по данному признаку, “-” – отрицательный ответ, “*” – неопределенный ответ (безразличие), т. е. можно считать и “+” и “-”, “?” – утерянный по техническим причинам ответ (“пропуск”).

 ISSN 1562-9945

Таблица 1

Результаты опроса

	j	Субъекты опроса									
	i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Признаки товара	**1**	*	+	+	+	*	-	-	+	-	-
	2	*	+	+	+	*	-	*	+	+	+
	3	*	-	-	-			*	-	*	*
	4	*	+	+	+	*	-	*	+	+	-
	5	*	-	-	-	*	-	*	-	-	-
	6	*	?	+	+	*	-	?	-	-	-

Как видно из таблицы, для источников информации i=2,7 ответ на вопрос об оценке признака i=6 утерян, т. е. есть пропуски.

Сформируем для каждого из этих источников информации группы с одинаковыми оценками, давая значения пропуску “+” и “-”, при этом обозначим количество источников информации в этих группах через r, а количество одинаковых признаков S

Для субъекта i = 2

для “+”: {3,4,2}, r = 3, S = 4

для “-”: {2,8}, r = 2, S = 3

Для субъекта i = 7

для “+”: {1,7}, r = 2, S = 6

для“-”: {1,7,9,10}, r = 4, S = 6

Введем обобщенный показатель количества совпавших оценок в каждой группе(критерий частоты совпадений)

$$N = r.S. \qquad (1)$$

Отсюда для источника информации i = 2

для “+”: $N = 12$;

для “-”: $N = 6; N_{\max} = 12;$

для источника информации i = 7

для “+”: $N = 12$;

для “-”: $N = 24; \quad N_{\max} = 24;$

Отсюда, можем поставить на место пропуска для i = 2 - “+”, для i = 7 - “-”, т. к. именно для этих значений $N = N_{\max}$.

Далее, суммируя количество положительных и отрицательных ответов по каждой строке можем сделать вывод какой признак (или характеристику) товара необходимо улучшить для повышения покупательского спроса.

Заключение. Предлагаемый в статье метод является модификацией известных методов заполнения, где пропуски не имеют критического значения,однако в отличие от них онсущественно повышает достоверность восстановления генеральной совокупности данных за счет анализа групповых свойств источников информации.В случае критического значения пропусков целесообразно использовать более строгие методы, например, методы моделирования с использованием функций максимального правдоподобия или байесовских стратегий [8].

ЛИТЕРАТУРА

1. Орлов А.И. Теория принятия решений. Учебное пособие. – М.: Издво “Март”, 2004.-656с.
2. Миркин Б.Г. Проблема группового выбора – М.: Наука, 1974.-263с.
3. Венда В.Ф. Перспективы развития психологической теории обучения операторов// Психологический журнал. 1980-т.4.№ 1.-С.48-63
4. Бродецкий Г.Л. Экономико-математические методы и модели в лингвистике. Потоки событий и системы обслуживания. Учебное пособие. – М.: Изд. Центр ”Академия”, 2011г. – 272 с.
5. Феллер В. Введение в теорию вероятностей и ее приложения М.: Мир, 1967. т. 2.- 275 с.
6. Р.Дж. А. Литл, Д.Б. Рубин Статистический анализ данных с пропусками. – М.: Финансы и статистика. 1981-336 с.
7. Кохрен У. Методы выборочного исследования. М.: Статистика.-1976. 440с.
8. Андерсон Т. Введение в многомерный статистический анализ – М.: Физматгиз. 1963.500с.

 ISSN 1562-9945