Методы устранения неполноты данных при онлайн-мониторинге рынка труда

В.С. Гиоргашвили[1], М.А. Бакаев[1]

[1]Новосибирский государственный технический университет, пр. К. Маркса 20, Новосибирск, Россия, 630073

Аннотация. Проблема неполноты данных достаточно актуальна, в том числе при сборе онлайн-данных. Возможными причинами неполноты данных могут быть: ошибки и изменения на площадках-источниках данных, сбои и ошибки в работе инструментов, осуществляющих сбор данных, и т.д. В результате чего при анализе данных имеем неполный массив. Поскольку для осуществления анализа наличие пропусков в данных обычно нежелательно, то возникает выбор: отбросить неполные данные или заполнить недостающие значения. Второе, как правило, является предпочтительным решением, однако важно выбрать подходящий метод устранения пропусков в данных, не приводящий к искажению результатов. В данной статье представлен обзор современных методов устранения неполноты данных. В задаче заполнения пропусков в онлайн данных по рынку труда был использован метод k-средних.

1. Введение

Ситуация, когда необходимо обработать пропуски в массивах данных, может возникнуть при проведении разнообразных социологических, экономических и статистических исследований. Так, при работе с собранными данными по рынку труда нами была отмечена неполнота данных о значении средней заработной платы для отдельных сфер деятельности и регионов. То есть система сбора данных в отдельно взятые периоды не собирала информацию по рынку труда с онлайн-площадок, и, следовательно, база данных системы не пополнялась. Вероятно, это связано со сбоем в работе программы, или отсутствием в эти периоды публикуемых вакансий и резюме на сайтах поиска работы. Данная система для сбора, структурирования и анализа онлайн-данных (применяемая для мониторинга рынка труда в сибирских регионах [1, с. 18]) не имела встроенных механизмов для коррекции этой проблемы, в связи с чем, возникла необходимость подбора метода для устранения неполноты данных.

2. Обзор современных методов устранения неполноты данных

На сегодняшний день существует множество методов, позволяющих устранить неполноту данных, но каждый из них имеет свои преимущества и недостатки. К наиболее распространенным методам можно отнести следующие:

1) Исключение из таблицы строк с пропусками. Метод применяется в случае с таблицей большой размерности и при незначительном количестве пропусков. В противном случае такой метод приводит к смещению оценок, потому как строки с пропущенными значениями содержат новую информацию, необходимую для анализа. Главным недостатком данного метода является потеря информации при изъятии неполных данных.

2) Заполнение пропусков средними по столбцу значениями. Применение данного метода целесообразно только в том случае, когда пропуски в данных по переменным случайны и сам

механизм пропусков несущественен. Недостатками такого метода являются вносимые изменения в распределения данных и уменьшение дисперсии.

3) Метод ближайших соседей. Суть метода состоит в поиске строк таблицы, которые являются ближайшими по определенному критерию к строке с пропуском. Для его заполнения, значения переменной (в установленном столбце) в соседних строках усредняются с конкретными весовыми коэффициентами, которые обратно пропорциональны расстоянию к строке, в которой есть пропуск. Такой метод точнее предыдущего, но он практически неприменим в случае большого количества пропусков, т.к. опирается на существование связей между строками в таблице.

4) Метод регрессии. По имеющимся данным осуществляется построение уравнения множественной линейной регрессии, и вычисляются пропущенные значения переменных. Метод нельзя применить в случае, когда количество пропусков в строке больше одного, поскольку это приводит к множеству решений, и вместе с тем его точность является невысокой, поскольку в реальных задачах зависимости нелинейные [3, с. 52].

5) Метод максимального правдоподобия и EM-алгоритм. Метод, требующий проверки гипотез о распределении значений переменных. Применение такого метода затруднительно, если количество пропущенных значений переменной велико. Особенностью метода является построение модели порождения пропусков с дальнейшим получением выводов на основании функции правдоподобия, которая строится при условии справедливости данной модели, с оцениванием параметров методами максимального правдоподобия [4, с. 497].

6) Метод k-средних. При использовании данного метода, так же как в случае ближайших соседей, предполагается, что близкие по одним признакам строки должны быть близки и по другим признакам. Однако отличие метода состоит в том, что здесь осуществляется поиск не ближайших соседей для каждой строки с пропущенными значениями, а используется информация о центре кластера, куда попала конкретная строка с пропуском. Для разбиения на кластеры необходима начальная инициализация пропущенных значений [2, с. 282].

При использовании этого метода выполняется инициализация пропущенных значений с помощью замены средним значением по признаку, кластеризация производится методом k-средних. Пропущенные значения заменяются на соответствующие им значения центра кластера, в который попала каждая строка с пропуском. Этот алгоритм выполняется в течение нескольких итераций до сходимости или по достижению максимального заданного числа итераций.

3. Применение метода k-средних для заполнения пропусков в онлайн-данных по рынку труда

Основной причиной неполноты данных при применении системы онлайн-мониторинга рынка труда (подробное описание см. в [1, с. 17]) являлась изменчивость источников данных – площадок, служащих для размещения объявлений о вакансиях и резюме.

Для устранения неполноты имеющихся данных по рынку труда наиболее подходящим методом был признан метод k-средних. При использовании данного метода объекты объединяются в кластеры так, что в один кластер попадут максимально схожие объекты, а объекты различных классов будут максимально отличаться друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, характеризующих объекты.

Неполные данные о значениях средней заработной платы по вакансиям и резюме и их количестве рассчитывались на основе имеющихся данных (использовался пакет Statistica). Для расчета были взяты данные по Красноярскому краю за 2 полугодие 2014 года (таблица 1). Для этого региона отсутствовали значения для «Работы на дому» и «Временной работы».

Для начала восстанавливались данные по вакансиям. С помощью функции «Иерархическая классификация» было определено количество кластеров. Объектами в данном случае были выбраны наблюдения (строки) – сферы деятельности. Исходя из визуального представления результатов, было выявлено, что сферы образуют 3 естественных кластера.

Согласно методу k-средних, вычисления начинались с k случайно выбранных наблюдений (k=3), которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами. После изменения состава кластера вычисляется новый центр тяжести, чаще всего, как вектор средних значений по каждому параметру. Алгоритм продолжается до тех пор, пока состав кластеров не перестанет меняться. В результате данные по вакансиям разбились по кластерам так, как показано на рисунках 1-3.

Для того чтобы заполнить отсутствующие строки, было использовано среднее значение по второму кластеру (рисунок 4), в который попали сферы «Работа на дому» и «Временная работа». Таким образом, можно предположить, что в среднем за неделю публиковалось 11,9 вакансий в сферах «Работа на дому» и «Временная работа», а средняя заработная плата по этим сферам составила 13 353,41 руб./месяц.

Аналогичным образом заполнялись отсутствующие данные по резюме. В исходных данных по резюме отсутствовали значения средней заработной платы для «Продажи услуг», «Пищевой промышленности», «Непищевой промышленности», «Работы на дому» и «Временной работы» и информация о среднем количестве публикуемых за неделю резюме для «Продажи услуг», «Работы на дому» и «Временной работы».

Было выявлено, что сферы образуют также 3 кластера. Таким образом, в ходе всех действий были получены следующие результаты:

- в среднем за неделю публиковалось 6 вакансий в сферах «Продажа услуг», «Работа на дому» и «Временная работа»;
- средняя заработная плата в сферах «Продажа услуг», «Пищевая промышленность», «Работа на дому» и «Временная работы» составила 16 982,39 руб./месяц;
- средняя заработная плата в «Непищевой промышленности» составила 23 576,71 руб./месяц.

Таблица 1. Данные о вакансиях по Красноярскому краю за 2 полугодие 2014 года.

Сфера	Год	Вакансии	
		В среднем за неделю	Средняя зарплата, руб/мес
Страхование	2014 (II полугодие)	5,66	34616,82
Спорт, красота, здоровье	2014 (II полугодие)	15,17	23591,75
Рабочие професии	2014 (II полугодие)	59,98	35609,63
Прочее	2014 (II полугодие)	156,00	31890,10
Государственная служба	2014 (II полугодие)	2,69	25154,41
Торговля розничная	2014 (II полугодие)	60,08	23720,51
Торговля оптовая	2014 (II полугодие)	38,76	31820,46
Рестораны, кафе, общепит	2014 (II полугодие)	31,46	18941,07
Транспорт, автобизнес	2014 (II полугодие)	52,76	30285,18
Промышленность непищевая	2014 (II полугодие)	35,40	48676,99
Высший менеджмент	2014 (II полугодие)	23,06	38745,59
Логистика, склад, закупки	2014 (II полугодие)	32,92	28672,02
Строительство, архитектура	2014 (II полугодие)	61,94	40217,64
Бухгалтерия, финансы, банки	2014 (II полугодие)	73,74	31062,54
ИТ и Интернет	2014 (II полугодие)	38,14	33712,51
Маркетинг, реклама, PR	2014 (II полугодие)	29,41	22496,18
Сфера услуг	2014 (II полугодие)	12,80	16211,91
Охрана и безопасность	2014 (II полугодие)	17,00	25527,58
Медицина и формация	2014 (II полугодие)	40,08	30494,74
Персонал офиса, АХО	2014 (II полугодие)	26,10	16647,34

Недвижимость	2014 (II полугодие)	0,29	33250,00
Юриспруденция	2014 (II полугодие)	8,04	35462,30
Образование, наука, языки	2014 (II полугодие)	22,34	17797,61
Продажа услуг	2014 (II полугодие)	84,71	38298,39
Работа для студентов	2014 (II полугодие)	15,37	23154,49
Персонал для дома	2014 (II полугодие)	1,62	17626,38
Промышленность пищевая	2014 (II полугодие)	6,71	20432,76
Телекоммуникация и связь	2014 (II полугодие)	3,37	25364,44
Полиграфия, издательства, СМИ	2014 (II полугодие)	3,14	24694,44
ТЭК, энергетика, добыча сырья	2014 (II полугодие)	6,55	46544,90
Работа дома	2014 (II полугодие)	0,00	0,00
Туризм, гостиничное дело	2014 (II полугодие)	12,58	26065,35
Временная работа	2014 (II полугодие)	0,00	0,00
Кадровые службы,HR	2014 (II полугодие)	9,55	30401,16
Сельское хозяйство	2014 (II полугодие)	0,86	19602,97
Дизайн, творческие профессии	2014 (II полугодие)	8,66	27623,10
Всего за неделю/средняя ЗП:	2014 (II полугодие)	997,47	30914,96

Спорт, красота, здоровье	2270,628
Прочее	3598,287
Гос. служба	1165,817
Торговля розничная	2179,641
Торговля оптовая	3547,981
Транспорт, автобизнес	2462,415
Логистика, склад, закупки	1321,695
Бухгалтерия, финансы, банки	3012,191
Маркетинг, реклама, PR	3045,284
Охрана, безопасность	901,823
Медицина и фармация	2610,559
Работа для студентов	2579,815
Промышленность пищевая	4504,377
Телекоммуникации и связь	1017,116
Полиграфия, издательства, СМИ	1491,020
Туризм, гостиничное дело	521,683
Кадровые службы, HR	2544,432
Дизайн, творческие профессии	580,234

Рисунок 1. Первый кластер, содержит 18 наблюдений.

Рестораны, кафе, общепит	3951,096
Сфера услуг	2021,265
Персонал офиса, АХО	2329,182
Образование, наука, языки	3142,533
Персонал для дома	3021,455
Работа дома	9442,290
Временная работа	9442,290
Сельское хозяйство	4419,113

Рисунок 2. Второй кластер, содержит 8 наблюдений.

Страхование	2755,417
Рабочие профессии	2053,423
Промышленность непищевая	7186,689
Высший менеджмент	164,261
Строительство, архитектура	1205,207
ИТ и Интернет	3394,799
Недвижимость	3721,909
Юриспруденция	2157,577
Продажа услуг	156,527
ТЭК, энергетика, добыча сырья	5679,103

Рисунок 3. Третий кластер, содержит 10 наблюдений.

перемен.	Среднее	Стандарт отклон.	Дисперс.
В среднем за неделю	11,90	13,119	172
Средняя зарпата, руб./месяц	13353,41	8314,671	69133750

Рисунок 4. Среднее количество и заработная плата для кластера.

4. Заключение

В статье был представлен обзор наиболее распространенных методов восстановления пропусков в данных. Выбор метода устранения неполноты данных может зависеть от типов признаков, в которых имеются пропуски, от количества объектов, имеющих пропущенные значения, и от причины их возникновения. В каждом случае необходим индивидуальный подбор метода обработки пропущенных значений. В рамках данной статьи, для восстановления пропусков был использован на практике метод k-средних, реализуемый в пакете Statistica. В результате кластеризации удалось рассчитать значения для отсутствующих данных и заполнить все строки исходной таблицы по рынку труда для Красноярского края.

Применение эффективных методов позволило, с одной стороны, полностью устранить проблему отсутствия данных, а с другой – повысить точность прогнозируемых значений показателя. Основная идея при поиске решений заключалась в предположении о возможности спрогнозировать недостающую информацию, имея частичные данные в заданном периоде.

5. Благодарности

Исследование выполнено при финансовой поддержке РФФИ (РГНФ) в рамках научного проекта № 17-32-01087 _a2.

6. Литература

[1] Bakaev, M. Prospects and challenges in online data mining: experiences of three-year labour market monitoring project/ M. Bakaev, T. Avdeenko // Lecture Notes in Computer Science. Data Mining and Big Data. – 2016. – Vol. 9714. – P. 15-23.

[2] MacQueen, J. Some methods for classification and analysis of multivariate observations / J. MacQueen // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967. – Vol. 1. – P. 281-297.

[3] Злоба, Е. Статистические методы восстановления пропущенных данных / Е.Злоба, И. Яцкив // Computer Modelling & New Technologies. – 2002. – Vol. 6(1). – P. 51-61.

[4] Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – М.: Издательский дом Вильямс, 2008. – 1103 с.

Methods for rebuilding incomplete data in online labor market monitoring

V.S. Giorgashvili[1], M.A. Bakaev[1]

[1]Novosibirsk State Technical University, K. Marx Ave. 20, Novosibirsk, Russia, 630073

Abstract. The problem of incomplete data is quite relevant, including when collecting online data. Possible reasons for incompleteness can be: errors and changes at the sites-the sources of data, failures and errors in the instruments for collecting data, etc. With the result that at the stage of data analysis have an incomplete array. Because the analysis is the presence of missing data is usually undesirable, there is the choice to discard incomplete data or fill in missing values. Second, as a rule, is the preferred solution, however, it is important to choose a suitable method of eliminating the missing data, not distorting the results. This paper presents a review of modern methods of elimination of incompleteness of the data and describes the application of the method of k-means to fill the gaps in the online data on the labor market.

Keywords: data quality, missing data, web- scraping, labor market, k-mean.