Глава I. Проблемы обработки и итерационного моделирования неполных данных с помощью многообразий малой размерности

Введение

В первом параграфе дан обзор литературы по существующим методам обработки данных с пропусками. В частности, рассматриваются методы регрессионного анализа, методы максимального правдоподобия, эмпирические методы. Также в рассмотрение включаются и нейросетевые методы заполнения пропусков в таблицах.

Во втором параграфе приводится обзор существующей теории главных кривых.

В третьем параграфе приводятся теоретические и методологические основы моделирования неполных данных с помощью многообразий малой размерности. При этом вначале рассматриваются эмпирические таблицы и задачи эмпирического предсказания, требования к методам обработки таблиц эмпирических данных, а затем, в четвертом параграфе, описываются теоретические основы самого метода моделирования неполных данных с помощью многообразий малой размерности. Рассматриваются различные варианты применимости метода. В итоге ставятся задачи, решению которых посвящен разработанный метод.

I.1. Методы обработки данных с пропусками

В рамках классической теории статистического прогнозирования [41, 42, 49] известно, что оптимальным статистически достоверным прогнозом отсутствующего значения является его условное математическое ожидание – регрессия. Однако весь аппарат математической статистики основан на предположении о том, что функции распределения нормальны или близки к нормальным. Поэтому при восстановлении функции условного математического ожидания требуется проверять гипотезу о распределении эмпирических данных по нормальному закону или использовать аппарат непараметрической статистики, восстанавливающий оценки плотностей распределения вероятностей.

В докомпьютерное время (до 1960 г.), начиная со статьи Уилкса [86], исследования по проблеме заполнения пропусков носили в основном теоретический характер и касались большей частью оценок максимального правдоподобия (МП–оценки) по некомплектным выборкам. На практике же использовались примитивные способы борьбы с пробелами – вычеркивание некомплектных строк или столбцов, замена пробелов средними по столбцу, использование в вычислениях только комплектных пар и т.п. Полный обзор этих и многих других методов до 1966 г. можно найти в [55], некоторые из них приведены в [25].

С распространением ЭВМ были предложены более сложные машинные алгоритмы, основанные на методе наименьших квадратов: регрессионный метод [57, 85], метод главных компонент [64], пошаговая регрессия [62], метод многомерной линейной экстраполяции [50], метод прогностических переменных [36]. Учитывая тот факт, что оценки первых двух моментов полностью определяют оценки регрессии, многие авторы сосредоточились на проблеме оценивания ковариационной матрицы по данным с отсутствующими значениями [61, 63, 71]. Одновременно выяснилась и некоторая ограниченность методов, основанных на методе наименьших квадратов. Так, Уилкинсон [41, стр. 167] указывал, что если пробелы имеются только в отклике, то при совместном оценивании пробелов и коэффициентов регрессии метод наименьших квадратов требует вычеркивать все строки с пробелами. Это приводит к неполному использованию информации, содержащейся в данных.

Со второй половины 70-х годов особых успехов добилось направление, связанное с МП-оценками, особенно в рамках нормальных распределений. Появились практические алгоритмы, вычисляющие МП-оценки пробелов, например [56, 67, 84]. В работе [59] предложена мощная вычислительная процедура – ЕМ-алгоритм для решения общей задачи оценивания параметров в условиях некомплектной выборки. К настоящему времени эти методы интенсивно развиваются, созданы эффективные робастные варианты ЕМ-алгоритма [80]. Возобладала тенденция поиска для всех классических статистических методов аналогов, способных работать с некомплектными данными, не заполняя пробелов [70, 80, 82]. Более полный обзор теории и практики содержится в монографиях [60, 78].

Достоверное статистическое оценивание должно давать для отсутствующих данных их условные математические ожидания (условия – известные значения других признаков) и характеристики разброса – доверительные интервалы. Это, однако, требует либо непомерно большого объема известных данных, либо очень сильных предположений о виде функций распределения. Приходится вместо ***статистически достоверных*** уравнений регрессии использовать ***правдоподобные эмпирические*** методы заполнения пропусков.

Так, один из известных эмпирических методов – алгоритм “ZET” [39, 40] – основан на изучении “похожести” объектов. Из предположения избыточности строится "предсказывающая подтаблица", состоящая из наиболее связанной с интересующим нас неизвестным элементом информации. Далее из принципа локальной линейности прогнозируется оценка неизвестного значения.

В [27, 28, 29] приводится пример прогнозирования эмпирических (измеряемых) данных по теоретическим данным (например, месту объекта в какой-либо естественной классификации) – прогнозируются высшие потенциалы ионизации атомов на основе атомных номеров. Не так давно был создан новый эмпирический метод – метод транспонированной регрессии [30, 31, 32, 65], близкий по идее к полуэмпирическому методу [27–29].

Основной проблемой в задаче транспонированной регрессии является нахождение для каждого объекта наилучшей функции из данного класса

функций и наилучшей опорной группы объектов, по которым строится эта функция.

I.2. ГЛАВНЫЕ КРИВЫЕ

Впервые понятие главной кривой появилось в [68] в 1989 году. В этой работе главные кривые были определены как "self-consistent" гладкие кривые, которые проходят через середину *d*-мерного вероятностного распределения или облака данных. Self-consistency означает, что каждая точка кривой есть среднее всех точек данных, которые проецируются в эту точку кривой.

В качестве связи между главными компонентами и главными кривыми показано, что если прямая линия есть self-consistent, тогда она является главной компонентой. Также показано, что при некоторых условиях главные кривые представляют собой критические точки расстояния от наблюдения. Основываясь на этом свойстве, авторы разработали алгоритм для нахождения главных кривых и для распределений, и для множеств данных.

Определение главной кривой, данное в [73], имеет то преимущество, что главные кривые существуют всегда, если распределение имеет конечные вторые моменты. Новое определение также делает возможным проведение теоретического анализа обучения главных кривых на основе учебных данных. Основываясь на этом, определен алгоритм построения ломаных, который успешно сравнивается с предыдущими методами.

В [83] дано альтернативное определение главной кривой, основанное на смешанной модели. Оценка выполнена с использованием метода максимального правдоподобия.

В [58] на основе обобщения полной дисперсии вводится понятие главной точки, а потом рекурсивно определяются главные кривые и поверхности.

I.3. ТАБЛИЦЫ ЭМПИРИЧЕСКИХ ДАННЫХ

В данной работе исследуется классическая задача обнаружения эмпирических закономерностей [37, 38], рассмотрению которой посвящен широкий спектр работ. Объектом исследования в них является таблица эмпирических данных, в которой систематизированы сведения о результатах измерений некоторых свойств изучаемых объектов. Эта таблица используется для предсказания значения некоторого выделенного исследователем свойства (или множества свойств) объекта при отсутствии информации о структуре объекта и его внутренних взаимосвязях. Для этого используют в основном способ рассуждений по аналогии, при этом в таблице должны содержаться объекты с известными значениями выделенного свойства. Обычно строки эмпирической таблицы соответствуют множеству объектов, а столбцы – множеству свойств или признаков, отсюда другие названия: таблица “объект-свойство”, “объект-признак”. В результате анализа данных в таблице исследователь получает некоторые эмпирические закономерности, которые используются для прогнозирования значений. Методы обнаружения

закономерностей в таблицах эмпирических данных, в свою очередь, являются широким полем для исследований [26, 37, 38, 43]. При этом в каждом отдельном случае понятие закономерности конкретизируется.

Задачи эмпирического предсказания

Итак, под эмпирической таблицей понимается таблица, элементы которой есть результаты измерений ряда признаков у подмножества объектов A, выбранных из некоторого множества Г. Множество Г задается в зависимости от целей исследования. При этом считается, что исследователь имеет правило, по которому он может определить, принадлежит или не принадлежит множеству Г произвольный рассматриваемый объект. Любому объекту $a \in \Gamma$ можно сопоставить вектор $\mathbf{x}=(x_1,\ldots, x_j,\ldots, x_n)$ и значение x_0 в пространстве признаков $X_1,\ldots,X_j,\ldots,X_n$; X_0. Признак X_0 выделен в качестве целевого признака. Для каждого признака X_j определена область его значений D_j (j=1,…,n; 0) и указан тип шкалы, в которой он измерен [48]. Важно различать группы шкал, предназначенных для измерения признаков следующих трех видов: количественных (шкалы интервалов, отношений и абсолютная); порядковых (шкалы порядка, частичного порядка, рангов, баллов); номинальных (шкала наименований). Пусть выбрано некоторое множество объектов $A=\{a^1,\ldots,a^i,\ldots,a^N\}$, $A \subseteq \Gamma$. Множеству A соответствует таблица $\mathbf{X}=\{x_j^i\}$ (i=1,…,N, j=1,…,n; 0). Таблица **X** используется для решения пяти типов задач эмпирического предсказания [43], которые перечислены ниже.

1. ***Распознавание образов*** (предсказание значения целевого признака x_0 для любого объекта $a \in \Gamma$ по его описанию **x**; в этом случае признак X_0 замерен в шкале наименований).

2. ***Предсказание значения целевого признака*** x_0 ***для объекта*** $a \in \Gamma$ ***по его описанию x***. Признак X_0 – порядковый или количественный.

3. ***Упорядочивание объектов по их перспективности с точки зрения некоторого критерия*** (предсказание порядка на объектах некоторого подмножества A' ($A' \neq A$, $A' \subset \Gamma$).

4. ***Автоматическая группировка объектов***. В данном случае значения признака X_0 для подмножества объектов A не заданы. Необходимо эти значения определить, используя свойство “похожести” объектов по их описанию.

5. ***Динамическое прогнозирование значения целевого признака*** x_0 ***объекта*** $a \in \Gamma$, ***использующее временные измерения значений признаков*** $X_1,\ldots,X_n$ (анализ временных рядов). В качестве примера задачи динамического прогнозирования можно привести задачу ранней диагностики заболеваний на основе профилактических осмотров пациентов.

При решении данных задач используется следующая эмпирическая гипотеза: считается, что при выборе объектов подмножества A из множества Г не делается предпочтения одного объекта другому: объекты подмножества A выбираются из Г случайным образом, т. е. рассматривается статистическая постановка задачи. Известные методы обработки эмпирических таблиц строят решающее правило, максимизирующее качество предсказания на объектах

подмножества *A*. Если допустить, что указанная гипотеза не верна, то всегда можно так подобрать объекты подмножества *A* из Г, что это правило будет плохо работать на остальных объектах множества Г. Поэтому большинство методов обработки таблиц в явном или неявном виде используют эту гипотезу.

Очевидно, что все вышеприведенные задачи могут рассматриваться как единая задача заполнения пропусков в таблице. В начале этой главы были рассмотрены существующие методы решения этой задачи. В дальнейшем основное внимание будем уделять задаче восстановления пропущенного количественного или порядкового признака (второго типа).

Требования к методам обработки таблиц эмпирических данных

Рассмотрим основные особенности указанных в предыдущем разделе задач для случая изучения сложных объектов.

1. Задачи приходится решать в условиях высокой априорной неопределенности, когда практически ничего неизвестно о виде функций распределения вероятностей в пространстве признаков. Всякое “сильное” предположение (например, о нормальности распределения, некоррелированности признаков и т.д.) ставит вопрос об адекватности предлагаемого вида действительному. То же самое можно сказать и о предположении об унимодальности функций распределения. Поэтому методы решения задач должны быть универсальными, т.е. ориентированными на достаточно слабые ограничения на вид распределений.
2. При изучении сложных объектов возникают большие трудности при задании исходной системы признаков для их описания. Поэтому в признаковом пространстве может быть много “дублирующих” и “шумящих” признаков. В результате проблема выбора наиболее информативной подсистемы признаков приобретает важное значение, поскольку уменьшение числа признаков часто улучшает качество решения (и сокращает экономические и временные затраты на измерения или сбор информации).
3. Для описания объектов используются признаки, измеренные в разных шкалах и, возможно, разнотипные.
4. В связи со сложностью измерения некоторых параметров, отказом датчиков и т.д. в таблице могут отсутствовать некоторые значения исходных признаков и даже целевых у некоторых объектов.

В связи с этим методы решения задач обработки экспериментальных данных должны удовлетворять следующим требованиям:

1) должны работать при наличии пропусков в таблице;
2) работать даже в случае, если число измеренных признаков превышает число объектов, и число объектов достаточно мало;
3) должны обеспечивать возможность обработки разнотипных экспериментальных данных (без сведения всех признаков к одной шкале) и инвариантность к допустимым преобразованиям шкал признаков;
4) должна обеспечиваться достаточно высокая вычислительная эффективность.

И, дополнительно:

5) должен использоваться класс решающих функций, имеющий малую меру сложности;
6) должны обеспечиваться наглядность и легкая интерпретируемость полученных решающих правил.

I.4. ИТЕРАЦИОННОЕ МОДЕЛИРОВАНИЕ НЕПОЛНЫХ ДАННЫХ С ПОМОЩЬЮ МНОГООБРАЗИЙ МАЛОЙ РАЗМЕРНОСТИ

Всем перечисленным в предыдущем параграфе требованиям к методам решения задач обработки экспериментальных данных удовлетворяет метод моделирования данных с пробелами многообразиями малой размерности.

Итак, пусть задана таблица данных, строки которой соответствуют объектам, а столбцы – признакам. Пусть, далее, часть информации в таблице отсутствует – есть пробелы. Основная возникающая в связи с этим задача – правдоподобно заполнить существующие пропуски. Ей сопутствуют еще одна задача – произвести "ремонт" таблицы: выделить данные, имеющие неправдоподобные значения, и исправить их. Кроме того, по таблице, как правило, полезно построить правило вычислений для заполнения пробелов в данных о новых объектах (по мере их поступления) и ремонта этих новых данные. Построение такого правил вычисления предполагает, что данные о новых объектах связаны между собой теми же соотношениями, что и в исходной таблице.

Следует особенно подчеркнуть, что в этих проблемах невозможно говорить ни об истинных значениях данных, ни даже о статистической доказательности, но только о правдоподобии. Особую трудность (и в то же время – притягательность) описанные задачи имеют в тех случаях, когда плотность пробелов высока, расположены они нерегулярно, а данных немного, например, число строк примерно таково же, как и число столбцов.

Обычные алгоритмы регрессии состоят в построении эмпирических зависимостей одних данных от других. Этот подход здесь неприменим. Если расположение пробелов нерегулярно, то фактически требуется построение зависимостей неизвестных данных от известных для всех возможных их положений в таблице. Это означало бы построение 2^{n-1} зависимостей, где n – число признаков. Только в этом случае можно будет восстанавливать любой неизвестный набор данных, если хоть что-то известно. В связи с этим приходится использовать метод моделирования данных многообразиями малой размерности.

Суть метода моделирования данных такова. Вектор данных x с k пробелами представляется как k-мерное линейное многообразие L_x, параллельное k координатным осям, которые соответствуют пропущенным данным. При наличии априорных ограничений на пропущенные значения место L_x занимает прямоугольный параллелепипед $P_x \subset L_x$. Ищется многообразие M заданной малой размерности (чаще всего – кривая), наилучшим образом приближающее данные и удовлетворяющее некоторым требованиям

регулярности. Для комплектных векторов данных точность приближения определяется как обычное расстояние от точки до множества (нижняя грань расстояний до точек множества). Для неполных данных вместо него используется нижняя грань расстояний между точками M и L_x (или, соответственно, P_x). Из данных вычитаются ближайшие к ним точки многообразия M – получается остаток – и процесс повторяется, пока остатки не приблизятся в достаточной степени к нулю. Близость линейного многообразия L_x или параллелепипеда P_x к нулю означает, что мало расстояния от нуля до ближайшей к нему точки L_x (соответственно, P_x). Дальнейшая конкретизация метода состоит в указании того, как строится многообразие M.

Идея же моделирования данных с помощью многообразий малой размерности возникла давно. Самая известная, давняя и очень практичная ее реализация для данных без пробелов – это классический метод главных компонент. Он состоит в том, что данные моделируются с помощью их ортогональных проекций на "главные компоненты" – собственные векторы корреляционной матрицы, которым соответствуют наибольшие собственные значения. Другая алгебраическая интерпретация метода главных компонент – сингулярное разложение таблицы данных. Как правило, для достаточно точного представления данных требуется сравнительно немного главных компонент и размерность сокращается иногда в десятки раз.

Обобщение первой главной компоненты на нелинейный случай ("главная кривая") было предложено в 1988 г. [68, 75, 76]. Известны также обобщения классического метода главных компонент на данные с пробелами.

В работе описан метод построения системы моделей для некомплектных данных. В простейшем случае эти модели являются обобщением классического (линейного) метода главных компонент на данные с пробелами. Далее следует квазилинейный метод, надстраиваемый над линейным и использующий его результаты. Наконец, с помощью формализма самоорганизующихся кривых строится существенно нелинейный метод.

Для каждого метода приводится соответствующая механическая интерпретация, которая показывает сходства методов и их последовательное развитие.

В результате, построенная технология моделирования данных с пробелами многообразиями (линейными и нелинейными) малой размерности в общем случае представляется более эффективной по сравнению с обычными уравнениями регрессии.

Разрабатываемый алгоритм заполнения пробелов в отличии от многих других алгоритмов, предназначенных для той же цели, не требуют их предварительного априорного заполнения данных. Однако, что вполне естественно, он требует предварительной нормировки данных ("обезразмеривания") данных – перехода в каждом столбце таблицы к "естественной" единице измерения. Следует заметить, что в задаче обработки данных с пробелами невозможно перейти к однородной задаче центрированием

данных.

А что касается расположения пробелов в данных, то приведенный алгоритм применим в том случае, когда матрица данных не может быть приведена перестановкой строк и столбцов к следующему блочно-диагональному виду:

$$A=\begin{bmatrix} A_1 & @ & \dots & @ \\ @ & A_2 & \dots & @ \\ \dots & \dots & \dots & \dots \\ @ & @ & \dots & A_n \end{bmatrix},$$

где @ – прямоугольные матрицы с неизвестными элементами. Для таких таблиц связь между различными блоками A_i установить невозможно, а поэтому и невозможно решать задачу восстановления пропущенных данных по известным.

Постановка задачи

Пусть задана прямоугольная таблица $A=(a_{ij})$, клетки которой заполнены действительными числами или значком @, означающим отсутствие данных.

Требуется построить модели, которые позволяли бы решать следующие три задачи, связанные с восстановлением пропущенных данных:

1) правдоподобно заполнить имеющиеся пробелы в данных;

2) отремонтировать данные, т.е. исправить их значения таким образом, чтобы наилучшим образом работали построенные модели;

3) построить по имеющейся таблице вычислитель, который бы заполнял пробелы в данных и ремонтировал бы их по мере по мере поступления (в предположении, что данные в поступающей на вход строке связаны теми же соотношениями, что и в исходной таблице).

Первый возникающий вопрос: *как (в какой метрике) оценивать ошибку модели?* Выбор меры ошибки необходим и для построения моделей и для их тестирования. С точки зрения простоты вычислений наиболее привлекателен метод наименьших квадратов (МНК). Ошибка в нем вычисляется как сумма квадратов уклонений по всем известным данным (среднеквадратичная ошибка Mean Square Error – MSE). Однако и здесь имеется произвол, связанный с выбором масштабов, то есть с нормировкой данных.

В классическом методе главных компонент обычно производится нормировка исходных данных на единичную дисперсию. После такой нормировки первая главная компонента определяется как такое направление (вектор), что ортогональные проекции данных на него имеют максимальную дисперсию. Она соответствует главной оси эллипсоида рассеяния.

Однако нормировка на единичную дисперсию не всегда соответствует сути дела. Кроме среднего квадратичного уклонения σ данной величины на роль естественного масштаба претендуют также точность ее измерения и, что особенно важно, допуск на ее изменение.

Понятие "допуск" происходит из технических приложений и означает тот произвол в значении величины, который может быть допущен без ущерба для решения значимых задач. Допуск определяется индивидуальным пользователем или особыми соглашениями о стандартных допусках. Именно величина допуска, скорее всего, может быть наилучшим естественным масштабом измерения. Следует, однако, помнить, что эта величина определяется не только таблицей данных, но еще и теми задачами, которые с ее помощью будут решаться.

Исходно по построению главных компонент столько же, сколько исходных признаков – просто совершается переход к новой системе координат. Однако нет необходимости вычислять все главные компоненты и, тем более, сохранять их все в модели. Достаточно оставить несколько из них. Если из p данных отобрано m главных компонент ($m<p$), то приходим к так называемой m-факторной модели.

Всюду далее предполагаем, что данные нормированы приемлемым образом (например, на соответствующие допуски) и оцениваем ошибки по методу наименьших квадратов.