МАТЕМАТИЧЕСКИЕ МЕТОДЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 51-74, 004.93

ПОДХОД К ЗАПОЛНЕНИЮ ПРОПУСКОВ В ОБУЧАЮЩИХ ВЫБОРКАХ ДЛЯ КОМПЬЮТЕРНОГО КОНСТРУИРОВАНИЯ НЕОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

В.А. Дударев, доцент

кафедра Информационных технологий МИТХТ им. М.В. Ломоносова, Москва, 119571 Россия

e-mail: vic_dudarev@mail.ru

В статье предлагается методика заполнения пропусков в значениях свойств химических элементов в обучающих выборках для компьютерного конструирования неорганических соединений, основанная на построении линейной регрессии с учетом Периодического закона.

***Ключевые слова**: компьютерное конструирование неорганических соединений, линейная регрессия, интерполяция свойств химических элементов.*

Введение

Формальная постановка задачи компьютерного конструирования неорганических соединений может быть дана следующим образом. Пусть каждое неорганическое соединение описано вектором $x = (x_1^{(1)}, x_2^{(1)}, ..., x_M^{(1)}, x_1^{(2)}, x_2^{(2)}, ..., x_M^{(2)}, ..., x_1^{(L)}, x_2^{(L)}, ..., x_M^{(L)})$ где L – количество химических элементов в соединении, а M – количество параметров химических элементов, выбранных для описания. Каждое соединение также характеризуется принадлежностью к определенному классу: $a(x) \in \{1,2,...,K\}$, где K – количество классов. Таким образом, обучающая выборка состоит из N объектов: $S = \{x_i, i = 1,..., N\}$. Обозначим подмножество объектов обучающей выборки из класса a_j, $j = 1,..., K$ как $S_{aj} = \{x : a(x) = a_j\}$. Цель обучения – построить классифицирующие правила, которые позволяют не только отличить объекты разных классов, но и обладают прогностической способностью образовывать новые комбинации химических элементов, которые не использовались для обучения, и относить их к одному из K классов. Таким образом, осуществляется переход к классической задаче распознавания образов по прецедентам. Особенностью предметной области является только формирование признакового описания, имеющего составную структуру: набор свойств элементов (компонентов неорганического вещества) повторяется L раз.

Одной из основных сложностей при использовании программных комплексов распознавания образов для компьютерного конструирования неорганических соединений является наличие пропусков в значениях свойств химических элементов, на основе которых формируются выборки для обучения. В зависимости от алгоритмов распознавания пропуски (отсутствия значений) могут не только исказить правильность обучения и, следовательно, распознавания, но и вызвать отказ в обучении, вынуждающий исследователя полностью отказаться от использования признака, содержащего пропуски в значениях.

Существует два варианта действий: исследовать матрицу, не заполняя пропуски, или заполнить тем или иным образом пропуски и анализировать полученную заполненную матрицу. Первый вариант более строг, однако практически всегда заставляет отказаться либо от рассмотрения значительной части фактических данных, либо от применения ряда мощных методов, «воспринимающих» только полные матрицы. Второй вариант «работает», то есть не оказывает заметного искажающего влияния на результаты анализа, когда число пропусков в данных мало (до 20 %). Рассмотрим основные приемы обработки данных, содержащих пропуски.

Методы обработки пропусков в обучающих выборках

Метод исключения неполных векторов

Метод исключения неполных (некомплектных) векторов (casewise deletion) [1] состоит в том, что все векторы (строки или столбцы матрицы), содержащие пропуски, исключают из рассмотрения и в дальнейшем анализируют новую, редуцированную матрицу данных. Когда выборка содержит достаточное число комплектных объектов, такой подход следует признать наиболее целесообразным.

Однако на практике распространена ситуация, когда наличие даже небольшого числа случайно (или, как минимум, без явной закономерности) распределенных пропусков – при формально большой размерности данных – приводит к резкому уменьшению числа комплектных наблюдений. Так, например, если предположить, что пропуски распределены независимо по закону Бернулли, то в случае наличия 5% пропусков для матрицы с числом столбцов $m = 10$ ожидаемая доля комплектных наблюдений составит $0.95^{10} \sim 0.6$ (60%), то есть редуцированная матрица будет содержать 0.6/0.95 = 0.63 (63%) данных, присутствующих в исходной матрице.

Метод заполнения средними значениями

Метод заполнения пропусков безусловными средними значениями (mean substitution) является одним из самых простых и известных ме-

тодов заполнения пропусков. При этом пропуск заменяется средним по столбцу матрицы. При применении этого метода происходит смещение (уменьшение) дисперсии переменных, что приводит и к смещенным оценкам элементов ковариационной и корреляционной матриц. Коэффициенты ковариации оказываются занижены, а корреляции – завышены [2]. Насколько приемлемо полученное смещение для дальнейшего анализа, решается в каждом конкретном случае; в целом, метод пригоден при малом числе пропусков в данных.

Метод заполнения условными средними значениями

Метод заполнения пропусков условными средними значениями (метод Бака, imputation by regression) [2]. Для двух переменных матрицы, заметно коррелирующих между собой – m_1 и m_2, можно построить регрессионное уравнение зависимости одной переменной от другой: $m_2 = am_1 + b$ по наблюдениям, известным для обоих переменных, и оценить недостающие значения m_2 с помощью полученного регрессионного уравнения по имеющимся значениям m_1. Данные, заполненные по методу Бака, обеспечивают разумные оценки средних, в частности, если приемлемо предположение о нормальности наблюдений. Ковариационная матрица по заполненным методом Бака данным занижает величину дисперсий и ковариаций (а корреляционная матрица завышает величину корреляций), хотя и не так сильно, как при подстановке безусловных средних.

Метод заполнения выборочными значениями

Метод заполнения пропусков выборочными значениями (hot deck imputation) [2]. Существуют методы заполнения пропусков, основанные на использовании расстояния до обоих объектов (в некоторой метрике между парами объектов), которое определяется по значениям признаков. Считается, что если два объекта близки в пространстве измеренных признаков, то из этого следует и их близость по неизмеренным признакам. Метрика и пороговое значение расстояния, определяющее близость объектов, вводятся в зависимости от условий конкретной задачи – шкал, в которых признаки измерены, количества пропусков и т.д.

Например, пусть требуется оценить значение пропущенного признака m_j, то есть оценить элемент m_{ij} матрицы. Для этого формируется подматрица с измеренными значениями признака, из которой далее выделяется группа наиболее близких объектов в пространстве измеренных у этого объекта признаков. Затем неизвестное значение m_{ij} заменяется средним по выделенной однородной группе объектов значением признака m_{ij} или случайным значением из этой группы.

Метод максимального правдоподобия

Оценка пропусков методом максимального правдоподобия (Expectation-Maximization, EM algorithm) [1]. EM-алгоритм – общий итеративный алгоритм для задач оценивания методом максимального правдоподобия и не только для заполнения отсутствующих данных – например, он используется для оценивания компонент дисперсии, итеративно взвешиваемых оценок наименьших квадратов и т.д. Достоинством EM-алгоритма является его надежная сходимость, недостатком – то, что скорость сходимости может быть очень низкой, если пропущено много данных. Как и любой другой метод оптимизации, данный алгоритм «локален», то есть процесс оптимизации сходится к локальному минимуму.

Методика заполнения пропусков с учетом специфики предметной области

Как показывает практика, заполнение пропусков в обучающей выборке возможно осуществлять с использованием некоторых приближений, полученных с учетом специфики предметной области, в нашем случае – неорганического материаловедения. Из Периодического закона Д.И. Менделеева известно, что свойства химических элементов находятся в периодической зависимости от атомного номера.

Для решения проблемы заполнения пропусков нами предложен следующий комбинированный подход (рис. 1). Во-первых, из обучающей выборки удаляются признаки, имеющие больше *20%* (задается экспертом) пропусков, так как их информативность небольшая, но в случае заполнения пропусков они могут помешать правильному обучению системы. Заполнение же оставшихся пропусков предлагается реализовать по следующему алгоритму с учетом особенностей предметной области и Периодического закона.

Пусть неизвестно свойство *y* химического *x*, тогда проводится поиск химических элементов, из той же группы периодической системы, у которых искомое свойство известно. Указанные химические элементы, согласно Периодическому закону должны быть «близкими» по набору значений свойств к тому, у которого требуется заполнить пропуск. То есть получаем парную или простейшую линейную регрессию:

$$y_t = a + bx_t + \varepsilon_t,$$

где x_t – вектор атомных номеров, y_t – вектор значений свойств.

Далее полученная линейная регрессия решается, например, методом наименьших квадратов (МНК) [3]. После проверки адекватности регрессионной модели, например, по критерию

Фишера, принимается решение об использовании значения свойства, вычисленного с использованием найденной модели.

Если модель получается неадекватной (в силу возможных «выбросов» значений свойств внутри группы), то предлагается следующая схема вычисления недостающего значения. Находятся два «ближайших» элемента из той же группы с учетом их атомного номера, при этом относительное «расстояние» по атомным номерам между элементами не должно превышать заданного экспертом значения. Для найденных элементов (x_1, y_1) и (x_2, y_2) проводится линейная интерполяция, т.е. в результате неизвестное значение y вычисляется по формуле

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x - x_1)$$

Если подходящие элементы для линейной интерполяции не найдены, то неизвестное значение свойства элемента заменяется на среднее арифметическое значение этого признака у объектов с равным классообразующим признаком или этот признак исключается из обучающей выборки [4].

Заключение

Предложенная методика заполнения пропусков в свойствах химических элементов является комбинированным способом вычисления, основанным, прежде всего, на методе Бака, применяемом с учетом предметной области. По сравнению с методом, используемым в информационно-аналитической системе [4], основанным на методе вычисления безусловных средних, обеспечивается меньшее занижение дисперсии и увеличение корреляции, что дает в результате более качественную информацию для анализа алгоритмами распознавания образов, используемыми при компьютерном конструировании неорганических соединений.

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 12-07-00142 и 14-07-31032.

ЛИТЕРАТУРА:

1. Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками: пер. с англ. М.: Финансы и статистика, 1990. 336 с.

2. Zloba E., Yatskiv I. Statistical methods of reproducing of missed data // Computer Modelling & New Technologies. 2002. V. 6. № 1. P. 51–61.

3. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. 2-е изд. М.: Физматгиз, 1962. 337 с.

4. Столяренко А.В. Интегрированная информационно-аналитическая система для прогнозирования свойств неорганических соединений : автореф. дис. … канд. техн. наук. М.: Моск. гос. ин-т электроники и математики, 2008. 24 с.

APPROACH TO REPRODUCING OF MISSED DATA IN LEARNING SAMPLES FOR COMPUTER-AIDED INORGANIC COMPOUNDS DESIGN

V.A. Dudarev[@]

**M.V. Lomonosov Moscow State University of Fine Chemical Technologies, Moscow, 119571 Russia*

[@] Corresponding author e-mail: vic_dudarev@mail.ru

The article is devoted to questions of accumulated data usage to find out regularities by means of pattern recognition methods that allow predicting formation of not synthesized substances and estimating its properties. The formal task of computer-aided inorganic compounds design is stated. An approach to reproduce of missing data in learning samples for computer-aided inorganic compounds design is proposed. It is based on combination of linear regression and interpolation taking into consideration the problem domain – inorganic chemistry. The approach is more powerful than methods of missed data reproduction used currently in information-analytical system for inorganic compounds design running at IMET RAS.

Keywords: *computer-aided inorganic compounds design, linear regression, chemical elements property values interpolation.*