УДК 004.89:004.4

ПРИМЕНЕНИЕ КОМБИНИРОВАННЫХ ЛИНЕЙНЫХ МЕТОДОВ ДЛЯ ЗАПОЛНЕНИЯ ПРОБЕЛОВ В ТАБЛИЦАХ ДАННЫХ

Таскин А.С., Неволина С.С.
Научный руководитель – профессор Миркес Е.М.
Сибирский федеральный университет

При обработке числовых таблиц данных часто оказывается, что числовые величины в них принимают лишь несколько значений или групп (классов) значений. Тогда проблему заполнения пробелов в таких данных можно переформулировать как задачу классификации.

Под классификацией будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Классификация - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или объект.

Модель данных (распознаватель) строится на основе тех объектов выборки, которые не содержат пробелов.

Решение задачи классификации имеет большое значение в области медицинской диагностики. Объекты выборки – это пациенты. Признаки (свойства) характеризуют результаты обследований, симптомы заболевания и применявшиеся методы лечения. Примеры бинарных признаков: пол, наличие головной боли, слабости. Порядковый признак – тяжесть состояния (удовлетворительное, средней тяжести, тяжёлое, крайне тяжёлое). Количественные признаки – возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата. Признаковое описание пациента является, по сути дела, формализованной историей болезни.

На практике часто применяют комбинированные методы, сочетающие в себе достоинства и нейтрализующие недостатки базовых методов.

В данной работе предлагается для заполнения пробелов в таблицах данных использовать метод линейной регрессии, а для повышения точности результатов применить предварительную кластеризацию данных.

Линейная регрессия – метод, позволяющий аппроксимировать зависимость между несколькими входными и одной выходной переменной. Модель линейной регрессии описывается гиперплоскостью. Коэффициенты уравнения линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости вдоль оси целевого (вычисляемого) признака.

Линейная регрессия является одним из самых популярных линейных методов заполнения пробелов в данных.

Широкое применение линейной регрессии обусловлено тем, что большое количество реальных процессов в экономике и бизнесе можно с достаточной точностью описать линейными моделями.

Скользящий контроль – процедура контроля универсальности метода, его способности обобщать. При проведении скользящего контроля из обучающего множества поочередно исключается одна строка, затем на оставшихся проводится обучение (в

данном случае сроится гиперплоскость регрессии), а тестирование происходит на исключенной строке.

Ошибка классификации – количество «промахов» при определении классов значений. Считается, что класс значения некоторого признака определен верно, если предсказанное значение ближе к истинному значению, чем к любому другому значению из множества возможных дискретных значений этого признака.

Кластеризация – объединение объектов или наблюдений в непересекающиеся группы, называемые кластерами, на основе близости значений их атрибутов (признаков).

Чаще всего при анализе данных они представляются как облако точек в n-мерном пространстве (n – количество признаков), а их координаты задаются значениями соответствующего признака. При таком подходе естественной является кластеризация на основе взаимного расположения точек-объектов в пространстве. При этом кластеры формируются на основе сгущений точек. Однако в данной работе предлагается принципиально другая процедура выделения кластеров.

Для выделения кластеров необходимо, в первую очередь, выбрать признаки и по их значениям разделить выборку на кластеры. В один кластер попадут объекты с одинаковым значением данного признака. Такие признаки должны обладать следующими свойствами:

1) дискретность значений;
2) небольшое количество значений.

Эти требования возникают из соображений о том, что кластеров, полученных по значениям одного признака, не должно быть слишком много. Иначе количество объектов, им принадлежащих, будет весьма мало. Этот фактор может негативно сказаться на точности прогноза по такому «малому» кластеру.

После того, как для текущего объекта определены кластеры, в которые он входит, требуется найти «лучший» из них – тот, по которому далее будет проводиться регрессионный анализ.

Для решения задачи поиска признака, кластеризация по которому наиболее эффективна, каждый из классов подвергался процедуре скользящего контроля, и фиксировались полученные результаты ошибки классификации. Далее проводилось ранжирование классов по возрастанию их «качества». Под понятием «качества» подразумевается оценка того, насколько хорошо признаки объектов выражаются через их другие признаки. В данной работе критерием определения «качества» класса является величина средней ошибки классификации, приходящейся на объект класса.

Тестирование метода осуществлялось с использованием двух различных наборов данных. Первый набор данных – весьма популярная тестовая «задача о президентах». Она использовалась для получения общего представления о способности предлагаемого метода заполнять пробелы в данных. Второй набор данных – медицинская база – представляет собой выборку реальных (естественных) данных большой размерности. Она отображает истории болезней пациентов, перенесших инфаркт миокарда. Рассмотрим подробнее наборы данных, используемые для тестирования.

Медицинская база

Смысловые значения признаков представлены в табл. 1.

Таблица 1. Смысловые значения признаков

Признак (столбец)	Расшифровка
1	Возраст
2	Пол
с поля 3 по поле 55	Поля отражают течение инфаркта миокарда, лабораторные изменения, ЭКГ-измерения, проводимое лечение (кроме тромболитической терапии)
56	Летальный исход
57	Группа лечения пациентов тромболитическими препаратами: 2 – АКТИЛИЗЕ (дорогостоящий препарат) 1 – СТРЕПТОКИНАЗА (недорогой препарат) 0 – пациенты не получившие этот вид лечения

Так как для тестирования методов требуется обучающая выборка без пробелов, было принято решение минимизировать количество пробелов в таблице. Для этого были удалены признаки (столбцы), имеющие менее половины заполненных значений. Затем были удалены объекты (строки), содержащие хотя бы одно незаполненное значение. Таким образом, была сформирована таблица, которая использовалась для тестирования методов и их модификаций.

В итоговой таблице 247 строк, 57 столбцов, 14079 значений и нет пробелов (пропусков).

Задача о президентах.

Таблица имеет 31 строку и 13 столбцов. В таблицу занесена информация о выборах президентов США с 1864 по 1980 годы. Каждый объект (строка таблицы) описывает соответствующую предвыборную экономическую или общественно-политическую обстановку.

Признаки таблицы представлены в бинарной форме и представляют собой подтверждение или опровержение соответствующих утверждений об общественно- политической ситуации и характеристиках кандидатов на момент выборов (табл. 2).

Таблица 2. Смысловые значения признаков «задачи о президентах»

№ признака	Утверждение
1	правящая партия была у власти более 1 срока
2	правящая партия получила больше 50% на прошлых выборах
3	в год выборов была активна третья партия
4	была серьезная конкуренция при выдвижении от правящей партии
5	кандидат от правящей партии был президентом в год выборов
6	год выборов был временем спада или депрессии
7	рост среднего нац. валового продукта на душу населения > 2.1%
8	произвел правящий президент существенные изменения в политике
9	во время правления были существенные социальные волнения
10	администрация правящей партии виновна в серьезной ошибке/скандале
11	кандидат правящей партии - национальный герой
12	кандидат оппозиционной партии - национальный герой
13	победила правящая партия

Последнее утверждение отражает результат выборов. Победу может одержать либо правящая партия, либо оппозиционная.

В таблице данных задачи о президентах подтверждению утверждения соответствует значение «1», а отрицанию - «0». Таблица данных рассматриваемой задачи представлена в табл. 2.

В ходе численного эксперимента проводилось сравнение рассматриваемого метода линейной регрессии с предварительной кластеризацией и «классической» линейной регрессией. Сравнение осуществлялось по значениям ошибок классификации, полученных проведением скользящего контроля (табл. 3).

Таблица 3. Сравнительные результаты применения методов (скользящий контроль)

Выборка	линейная регрессия		линейная регрессия с кластеризацией	
	значение ошибки	в %	значение ошибки	в %
Задача о президентах	158	39,2	119	29,6
Медицинская база	1613	11,3	1735	12,1

На «тестовой» задаче о президентах предложенный комбинированный метод показал значительное улучшение результатов по сравнению с «базовым» методом линейной регрессии. Однако по результатам тестирования на таблице естественных данных большой размерности он незначительно уступил линейной регрессии.

Предложен комбинированный метод заполнения пробелов в таблицах данных, основанный на кластеризации по значениям признаков выборки и последующим применении линейной регрессии к сформированным кластерам.

Проведена апробация метода на различных таблицах данных. Получены неоднозначные результаты по точности восстановления данных предложенного метода по сравнению с «классической» линейной регрессией. Существует интуитивное предположение о необходимости более глубокой предварительной обработки при работе с таблицами естественных данных для реализации полного потенциала предложенного метода.