УДК 004.4

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ НАБЛЮДЕНИЙ

Ильина Е.А., студент гр. ПИб-121, IV курс
Научный руководитель: Пимонов А.Г., д.т.н., профессор
Кузбасский государственный технический университет
имени Т.Ф. Горбачева,
г. Кемерово

Проблема пропущенных наблюдений весьма актуальна во многих сферах нашей жизни, например, в экономике. Причин, по которым может возникнуть неполнота данных, достаточно много. В качестве таковых могут выступать следующие: сокрытие данных, невнимательность и т. д. В результате мы имеем неполный массив наблюдений. Данную проблему исследователи решают по-разному. Некоторые просто исключают из рассмотрения наблюдения с пропущенными данными. Другие подходят к решению проблемы пропущенных данных более рационально. Они стремятся на этапе первичной обработки заполнить пропуски в уже имеющихся данных, для того чтобы восстановить исходную зависимость. В настоящее время существует множество методов восстановления пропущенных наблюдений.

Нами ведется работа по созданию программного комплекса, в котором будут реализованы несколько алгоритмов для восстановления пропущенных наблюдений. В результате предварительного анализа были выбраны как универсальные алгоритмы, подходящие для любых массивов данных с пропусками, так и специальные алгоритмы, которые предназначены для восстановления данных в особых массивах.

Ядро системы составляют шесть программно реализованных алгоритмов [1]: 1) эволюционный; 2) исключения некомплектных строк; 3) заполнения средним значением; 4) Resampling; 5) Zetbrain; 6) Em.

Эволюционный метод восстановления пропусков в данных основывается на композиции нейронной сети [2] и генетического алгоритма [3]. То есть входные данные для обучения нейронной сети имеют пропуски значений, и необходимо решить задачу параметрической оптимизации с помощью генетического алгоритма. Разработанный эволюционный метод имеет ряд преимуществ. Так, его использование не требует выполнения ограничений на исходную информацию, связанных с линейностью модели, распределением параметров и т. д. Таблица исходных данных может иметь произвольную размерность и структуру пропусков [4].

Метод исключения некомплектных строк и метод заполнения средним значением применяется при большой размерности таблицы и незначительном количестве пропусков. В других случаях метод ведет к смещению оценки выборки, поскольку строки с пропущенными значениями содержат новую ин-

формацию, необходимую для анализа.

Resampling-метод применяется для решения задачи заполнения пропусков в неполных данных, когда значения для заполнения пропущенных элементов выбираются случайным образом из исходного множества данных. Значение для замены пропуска можно выбрать двумя способами: с вращением (когда ранее выбранное значение может участвовать в замене еще раз) и без вращения. После этого на всем массиве строится регрессионная модель, позволяющая предсказать значения для пробелов. Преимуществом данного метода является то, что информация, которая содержится в исходном массиве, используется более полно, а итерационный характер алгоритма позволяет получить более точный прогноз [5].

EM-алгоритм предназначен для работы с большими объемами данных. Его название происходит от слов «expectation-maximization», что переводится как «ожидание-максимизация». Это связано с тем, что каждая итерация содержит два шага – вычисление математических ожиданий и максимизацию [6].

Все вышеописанные алгоритмы реализованы в составе программного комплекс, для разработки которого использовались средства MS Visual Studio 2012. В качестве языка программирования был выбран язык C#. Программный комплекс предусматривает работу с массивами данных, хранящимися в текстовом файле (`*.txt`), либо в рабочей книге MS Excel (`*.xlsx`).

В перспективе предполагается развитие программного комплекса за счет добавления возможности дополнительного анализа полученных результатов.

Список литературы:

1. Круглов, В.В. Методы восстановления пропусков в массивах данных / В.В. Круглов, И.В. Абраменкова // Программные продукты и системы. – 2005. – №2. – С. 59-63.

2. Дороганов, В.С. Методы статистического анализа и нейросетевые технологии для прогнозирования показателей качества металлургического кокса / В.С. Дороганов, А.Г. Пимонов // Вестник Кемеровского государственного университета. – 2014. – №4, Т. 3. – С. 123-129.

3. Дороганов, В.С. Модифицированная сеть Ворда и гибридный метод обучения для прогноза показателей качества металлургического кокса / В.С. Дороганов, А.Г. Пимонов // Вестник Кузбасского государственного технического университета. – 2015. – №3. – С. 141-148.

4. Снитюк, В.Е. Прогнозирование. Модели, методы, алгоритмы: учебное пособие / В.Е. Снитюк. – Киев: Маклаут, 2008. – 364 с.

5. EM-масштабируемый алгоритм кластеризации [Электронный ресурс]. – Режим доступа: https://basegroup.ru/community/articles/em, свободный (дата обращения 28.03.2016).

6. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999.