УДК 004.42

Пимонов Александр Григорьевич, д. т. н., профессор
(КузГТУ, г. Кемерово; ИЭОПП СО РАН, г. Новосибирск)
Глебова Екатерина Андреевна, магистрант
(КузГТУ, г. Кемерово)
Глебов Вадим Витальевич, магистрант
(КузГТУ, г. Кемерово)
Alexander G. Pimonov, Doctor of Technical Science, Professor
(KuzSTU, Kemerovo; IEIE SB RAS, Novosibirsk)
Ekaterina A. Glebova, Master's Degree student
(KuzSTU, Kemerovo)
Vadim V. Glebov, Master's Degree student
(KuzSTU, Kemerovo)

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ВОССТАНОВЛЕНИЯ УТЕРЯННЫХ ДАННЫХ
SOFTWARE TO RESTORE LOST DATA

Проблема пропущенных данных весьма актуальна во многих сферах человеческой деятельности, 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 для дальнейшего анализа мы имеем неполный массив наблюдений. Эту проблему разные исследователи решают по-разному. Некоторые просто исключают из рассмотрения наблюдения с пропущенными данными. Другие подходят к решению проблемы пропусков более рационально. Они стремятся на этапе первичной обработки заполнить пропуски в уже имеющихся данных для восстановления исходной зависимости. Процедуру восстановления утерянных данных принято называть импутированием. В настоящее время существует множество методов и алгоритмов восстановления пропущенных наблюдений. В результате предварительного анализа [1, 2] для программной реализации нами были выбраны как универсальные алгоритмы, подходящие для любых массивов данных с пропусками, так и специальные алгоритмы, которые предназначены для восстановления данных в особых массивах: 1) исключение некомплектных строк; 2) заполнение средним значением; 3) эволюционный метод [3]; 4) Resampling-метод [3]; 5) Zetbraid-алгоритм [3]; 6) EM-алгоритм [3, 4].

Перспективным для дальнейшего эффективного использования является эволюционный метод восстановления пропусков в данных. Он основывается на композиции нейронной сети и генетического алгоритма. 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. Но требует дальнейших исследований эффективность использования нейронных сетей с неитеративными алгоритмами обучения [5] и выяснение влияния распределения значений факторов на точность восстановления пропусков.

Для создания программного обеспечения была использована интегрированная среда разработки Visual Studio 2012, исходный код написан на объектно-ориентированном языке программирования C# 5.0. Графический интерфейс пользователя разработан в Windows Forms. Программный комплекс предусматривает работу с массивами данных, хранящимися либо в текстовом файле (`*.txt`), либо в рабочей книге MS Excel (`*.xlsx`). При запуске программного комплекса на экране монитора появляется главное окно приложения (рис. 1), на котором расположены две вкладки. Первая вкладка (рис. 1) предназначена непосредственно для восстановления данных.

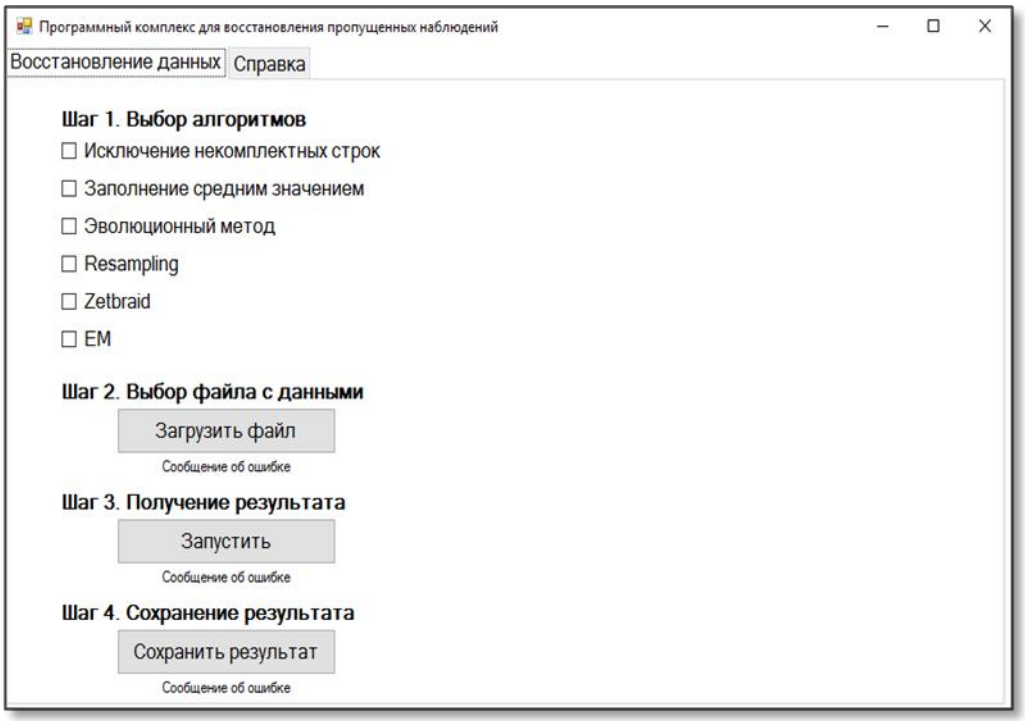

Рис. 1. Главное окно программного комплекса

`На первом шаге` пользователь выбирает те алгоритмы, с помощью которых будет происходить восстановление утерянных данных. Пользователь может выбрать как один алгоритм, так и все шесть.

`На втором шаге` выполняется загрузка файла с данными.

`На третьем шаге` происходит проверка данных в загружаемом файле и получение результата, если файл содержал данные допустимого формата.

На четвертом, завершающем шаге после окончания восстановления пропущенных наблюдений нужно сохранить результат. Файл с результатами будет иметь название того алгоритма, с помощью которого восстанавливались данные. В этом же файле содержится отчет о проделанной работе, а именно: 1) сколько значений было в файле; 2) сколько было пропусков; 3) сколько данных было восстановлено. В случае если пользователь отметил несколько алгоритмов, то будет создана общая папка для всех файлов с результатами.

В результате проведенного исследования было разработано программное обеспечение для восстановления утерянных данных. Ядро программного комплекса составляют шесть алгоритмов для восстановления пропущенных наблюдений: 1) эволюционный; 2) исключения некомплектных строк; 3) заполнения средним значением; 4) Resampling; 5) Zetbraid; 6) Em-алгоритм. Разработанное программное обеспечение для решения задачи импутирования может быть использовано в статистическом анализе экономических данных, где большие массивы с пропущенными наблюдениями представляют серьезную проблему. Возможно дальнейшее расширение функциональных возможностей программного комплекса за счет разработки интернет-сервиса для восстановления пропущенных наблюдений [6].

Список публикаций

1. Ильина, Е.А. Алгоритмы восстановления пропущенных наблюдений // Информационно-телекоммуникационные системы и технологии (ИТСиТ-2015): Материалы Всероссийской научно-практической конференции. – Кемерово, 2015.
2. Ильина, Е.А. Анализ, разработка и программная реализация алгоритмов решения задачи импутирования // Сборник лучших статей VIII Всероссийской, 61 научно-практической конференции молодых ученых «Россия молодая». – Кемерово: КузГТУ. – 2016. – С. 172-175.
3. Снитюк, В.Е. Прогнозирование. Модели, методы, алгоритмы. – Киев: Маклаут, 2008. – 364 с.
4. EM-масштабируемый алгоритм кластеризации [Электронный ресурс]. – Режим доступа: https://basegroup.ru/community/articles/em, свободный (дата обращения 10.10.2016).
5. 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.
6. Тайлакова, А.А. Web-сервис для поиска оптимальной конструкции нежестких дорожных одежд / А.А. Тайлакова, А.Г. Пимонов // Вестник Кузбасского государственного технического университета. – 2015. – № 6. – С. 176-181.