**Исследование точностных характеристик разработанного алгоритма восстановления данных на основе нечёткой кластеризации.**

[MAPE](http://en.wikipedia.org/wiki/Mean_absolute_percentage_error), [MAE](http://en.wikipedia.org/wiki/Mean_absolute_error), [MSE](http://en.wikipedia.org/wiki/Mean_squared_error), [ME](http://en.wikipedia.org/wiki/Mean), [RMSE](http://en.wikipedia.org/wiki/Root-mean-square_deviation), [E](http://en.wikipedia.org/wiki/Mean), [SD](http://en.wikipedia.org/wiki/Standard_deviation) обычно используются для оценки качества прогнозирования временных рядов. Также, они применимы для оценки точности восстановления пропущенных данных в числовом массиве.

Рассмотрим способы оценки результатов восстановления. [MAPE](http://en.wikipedia.org/wiki/Mean_absolute_percentage_error" \o "mean absolute percentage error) – средняя абсолютная ошибка в процентах:

,

где – фактическое значение рассматриваемой величины, – восстановленное значение рассматриваемой величины. N – количество восстановленных значений.

[MAE](http://en.wikipedia.org/wiki/Mean_absolute_error" \o "mean absolute error) – средняя абсолютная ошибка.

[MSE](http://en.wikipedia.org/wiki/Mean_squared_error) – среднеквадратичная ошибка

[ME](http://en.wikipedia.org/wiki/Mean) – средняя ошибка

[RMSE](http://en.wikipedia.org/wiki/Root-mean-square_deviation) – квадратный корень из среднеквадратичной ошибки.

[E](http://en.wikipedia.org/wiki/Mean) – ошибка.

[SD](http://en.wikipedia.org/wiki/Standard_deviation) – стандартное отклонение

Проводился эксперимент, цель которого – сравнить оценки восстановления данных предлагаемым нами методом с результатами заполнения пробелов таблицы средними арифметическими значениями по столбцам таблицы, именно такой способ наиболее популярен при предобработке таблиц с пропусками.

Входные данные эксперимента – числовая таблица размерности 50х10, заполненная искусственно сгенерированными числами с выраженным разделением на 5 кластеров. Таблица сгенерирована с помощью программы «Генератор входных данных для кластеризации» [Ссылка будет(Петербург 2017)]. Размах генерируемых значений от 0 до 1000.

В серии экспериментов из одной и той же таблицы произвольным образом вырезались 10 %, 20 %, …, 90 % данных. Данные восстанавливались тремя способами: как среднее арифметическое из заполненных данных по столбцу; предлагаемым нами алгоритмом восстановления данных; улучшенной модификацией нашего алгоритма. Каждый раз вычислялись все перечисленные выше оценки. Каждый эксперимент, например, с вырезанием 10 %, повторялся 6 раз. По 6 повторам эксперимента вычислялось среднее арифметическое значение оценок.

Таблица 1 содержит оценки [RMSE](http://en.wikipedia.org/wiki/Root-mean-square_deviation) для различных процентов вырезанных значений из всего массива данных.

Рис. 1. Таблица зависимости значений оценки [RMSE](http://en.wikipedia.org/wiki/Root-mean-square_deviation) от процента вырезанных значений из всего массива данных.

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE | С использованием алгоритма  восстановления пропущенных значений  (10 итераций) | С использованием модификации алгоритма  восстановления пропущенных значений  (50 итераций+ критерий остановки индивидуальный для каждого значений) | Заполнение пробелов средним арифметическим |
| 10% | 102.2526 | 98.85909 | 202.1201 |
| 20% | 112.6816 | 104.324 | 200.3675 |
| 30% | 166.193 | 148.1165 | 200.2874 |
| 40% | 193.088 | 180.4051 | 204.1945 |
| 50% | 199.0949 | 190.0077 | 207.9323 |
| 60% | 193.088 | 180.4051 | 203.0428 |
| 70% | 166.193 | 148.1165 | 200.2874 |
| 80% | 112.6816 | 104.324 | 207.9323 |
| 90% | 102.2526 | 98.85909 | 202.1201 |

Рисунок 1. Сравнение значений RMSE (ось ох – проценты вырезанных значений, от 10% до 90 %, ось оy – средние значения RMSE по различным экспериментам). V1, V2 – старая и новая версия нашего алгоритма, V3 – заполнение средними арифметическими.

Вывод. Применение алгоритма восстановления данных оправдано. Оно даёт результат, лучший, чем заполение пропусков средним арифметическим. При 10, 20, 30, 70, 80 , 90 наблюдается сильный выигрыш.

Литература.

[Хенрик Книберг] Kanban и Scrum: выжимаем максимум // Хенрик Книберг, Маттиас Скарин

[Cибирев, 2016 (1)] Cибирев И.В. Алгоритм предобработки и восстановления анкетных данных / И.В. Сибирев, Т.В. Афанасьева // Открытые семантические технологии проектирования интеллектуальных Систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2016) : материалы VI междунар. науч.-техн. конф. / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2016. – С. 271-271.

[Сибирев И.В, 2016 (2)] Сибирев И.В. Анализ эффективности алгоритма восстановления анкетных данных / И.В. Сибирев, Т.В. Афанасьева // Электронное обучение в непрерывном образовании 2016. III Международная научно-практическая конференция: сборник научных трудов. – Ульяновск : УлГТУ, 2016. – **С.367-373.**

[Сибирев, 2016 (3)] Сибирев И.В. Быстродействие программной реализации алгоритма «восстановление данных» / И.В. Сибирев // Информатика и вычислительная техника 2016. – Ульяновск : УлГТУ, 2016. –**С. 228-232.**

[Сибирев, 2016 (4)] Сибирев И.В. Предобработка данных в интеллектуальном анализе на основе восстановления пропущенных анкетных значений / Пятнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2016. Труды конференции. В 3-х томах. Т 1. – Смоленск Универсум, 2016. – С**. 378-386**