**2.Исследование временных и точностных характеристик разработанного алгоритма кластеризации, в условиях наличия пропущенных значений.**

Рассмотрим некоторую организационно-техническую систему, содержащую в своём составе узлы ОТС – «команды разработчиков». Описываемая ОТС обычно содержит в том или ином виде следующие роли: Product Owner – сотрудник, отвечающий за продукт и распределяющий приоритеты задач; команда, которая работает над задачами; Scrum, Kanban Master и т.п. – сотрудник, занимающийся настройкой и администрированием этой ОТС, в том числе и её процесса гибкой разработки программного обеспечения.

Есть основные целевые функции, согласно которым производится настройка и оптимизация ОТС: нагрузка или производительность; время выполнения задач или длительность цикла их выполнения; качество, уровень дефективности; предсказуемость или степень управляемости процесса.

Важно учитывать, с каким подходом гибкой разработки программного обеспечения работает данная организационно-техническая система, так как в разных подходах четыре основных целевых функции задаются по-разному. Все подходы гибкой разработки объединяет то, что для роли Master рекомендуется экспериментировать с настройками ОТС. Среди них: количество людей в команде, наложение или снятие ограничений на отдельные должностные инструкции в зависимости от ситуации, степень планирования, длина итерации обратной связи и т.д.

Важно отслеживать ответную реакцию процессов ОТС после изменения настроек. Предлагается возвращать старые значения параметров ОТС в случае ухудшения значения целевой функции или продолжать экспериментировать. Именно это рекомендуется в книге [Х. Книберг].

В процессе работы ОТС в рамках гибких методологий в момент окончания итерации осуществляется обратная связь. Происходит передача вектора параметров, снятых с ОТС, администратору (роль Master). По результатам нескольких итераций можно построить векторы возвращаемых параметров и судить о динамике (временные ряды) ОТС.

Далее все составные части организационно-технической системы будем называть элементами ОТС.

Иногда возникает ситуация, когда к концу итерации некоторые из параметров элементами ОТС не наблюдаемы. Нет возможности снять значения параметров (признаков) некоторых элементов и при помощи обратной связи их передать. Для решения этой проблемы может быть использован предлагаемый нами алгоритм восстановления значений параметров элементов ОТС.

Основная идея алгоритма – итеративное восстановление пропущенных числовых значений параметров обрабатываемых объектов как линейной комбинации средних значений параметра по кластерам с весовыми коэффициентами принадлежности объектов к кластерам. Весовые коэффициенты получены при помощи алгоритма нечёткой FCM-кластеризации. Алгоритм итеративен и восстанавливает пропущенные числовые значения с учётом всех заполненных числовых ячеек массива данных. Критерий остановки алгоритма – 10 итераций. Более подробно алгоритм описана в статьях [Cибирев, 2016 (1)], [Сибирев И.В, 2016 (2)], [Сибирев, 2016 (3)], [Сибирев, 2016 (4)]. Написана программа на языке #C, которая апробирована на массиве данных 103 элемента ОТС при 73 наблюдаемых числовых признаках.

2.1) Анализ быстродействия алгоритма в зависимости от производительности ПК.

В качестве входных данных использовались 103 элемента ОТС, и 73 числовых признака (параметры), которыми они были охарактеризованы.

Примерное время вычислений на персональной машине с 2-4 ядрами процессора сравнимо с двумя, тремя минутами (табл.1). На более слабых персональных машинах с меньшим числом ядер процессора время колеблется от двух до десяти минут.

Табл.1. Быстродействие алгоритма на разных машинах.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Машина № | Процессор | ОЗУ ГБ | Тип системы | V4 мин.:c. |
| 1 | Pentium(R) Dual-Core CPU E5200 – 2.50GHz; | 4 | 32 разряда | 2:26; 4:14;  2:36; 2:37 |
| 2 | Intel(R) Celeron(R)  CPU E3200 – 2.40 GHz; | 2 | 32 разряда | 4:23; 6:16;  10:6; 8:35 |
| 3 | Intel(R) Core(TM) i7-2630QM CPU – 2.00 GHz; | 8 | 64 разряда | 2:11; 2:7;  2:34; 1:56 |

Был проведен эксперимент по проверке наличия зависимости между процентом вырезанных (восстанавливаемых) значений и временем вычислений. Результаты приведены на рисунке 1 в виде графика, где вдоль абсцисс (OX) отложен процент вырезанных (восстанавливаемых) данных, вдоль ординат (OY) отложено время вычислений. Эксперимент показал, что нет явной зависимости между процентом вырезанных (восстанавливаемых) данных и временем вычислений.

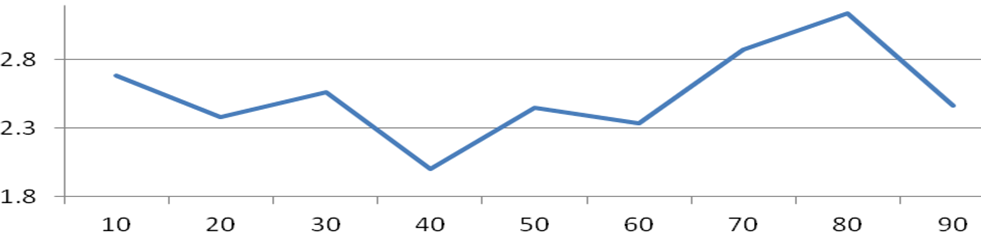


Рис. 1. Зависимость времени вычислений алгоритма (мин.) от процента вырезанных из массива данных.

Был проведён эксперимент, для проверки наличия зависимости между размерами массива обрабатываемых данных и временем вычислений. Для эксперимента использовались произвольно выбранные элементы ОТС, в количестве от 3 до 103, и все их признаки (параметры). Результаты эксперимента представлены на рисунке №2 в виде графика, где вдоль абсцисс (OX) отложено количество элементов ОТС, вдоль ординат (OY) отложено время вычислений. Подтверждена гипотеза о том, что чем больше элементов ОТС, тем дольше будет проходить восстановление попущенных значений.

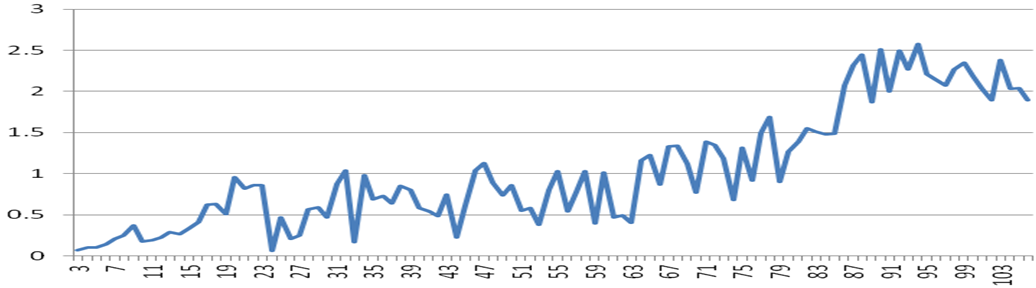


Рис. 2. Зависимость между количеством анкет и временем работы алгоритма.

Для следующей части эксперимента использовались произвольно выбранные параметры элементов ОТС, в количестве от 3 до 73, использовались все элементы ОТС. Результаты эксперимента представлены на рисунке 3 в виде графика, где вдоль абсцисс (OX) отложено количество элементов ОТС, вдоль ординат (OY) отложено время вычислений. Подтверждена гипотеза о том, что чем больше параметров рассматриваются, тем дольше будет проходить восстановление попущенных значений.

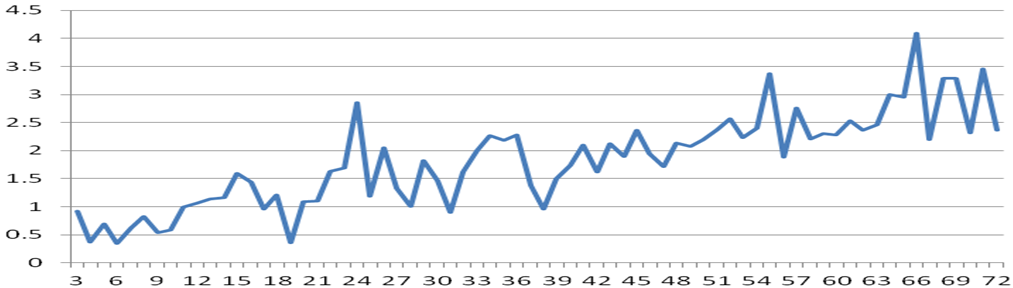


Рис. 3. Зависимость между количеством параметров и временем работы алгоритма.

В обоих экспериментах время вычислений не превосходит четырёх минут.

2.1) Анализ точности алгоритма

Анализ экспериментальных данных

В процессе эксперимента произвольным образом удалялись 10 %, 20 %,…, 90 % данных в одном столбце матрицы – для каждого параметра в отдельности, использовался алгоритм восстановления данных. Для каждого столбца находилось евклидово расстояние между столбцом восстановленных данных и столбцом реальных данных.

В разных столбцах таблицы значения были из промежутков от 0 до 100, от 0 до 10, 0 или 1. При сравнении результатов эксперимента для разных столбцов мы нормировали данные, переводя значения в столбцах в промежуток от 0 до 1.

На рис. 4. представлены графики зависимостей, где по горизонтальной оси отложены проценты вырезанных значений, по вертикальной оси – евклидовы отклонения между столбцами восстановленных и реальных нормированных данных. Для каждого исследованного столбца – свой график.

При возрастании числа вырезанных из столбцов элементов евклидово отклонение между реальной и восстановленной выборками растет. Найдем среднеарифметическое значение по всему пучку представленных на рис. 4 кривых. На рис. 5 представлен график зависимости среднеарифметического значения евклидовых отклонений по разным параметрам от процента вырезанных значений. Зависимость на рис. 5 – линейная возрастающая функция. При восстановлении 90% данных (95 элементов) суммарная ошибка – около 4 ед.

Возрастание отклонения происходит за счет накопления числа ошибок: при удалении n элементов, каждый из них даст ошибку восстановления, чем больше n, тем больше евклидово расстояние между реальным и восстановленным столбцами. Производная функции рис. 2 показывает, что один удаленный процент данных восстанавливается со средней ошибкой 0,03-0,04 (в долях от 1).

Ошибки зависят от информативности удаленной части столбца. Например, при изучении одного из параметров элементов ОТС при случайном вырезании 10% данных были удалены почти все значения, соответствующие наименьшим показателям от 0 до 18 (из 100%). В результате, данные были восстановлены несколько завышенными однотипными значениями от 10% до 30%, что дало увеличение ошибки при 10% вырезанных данных.

|  |  |
| --- | --- |
|  |  |
| Рис. 4. Графики зависимостей между процентом удаленных значений отдельного параметра и евклидовым расстоянием между столбцами восстановленных и реальных данных | Рис. 5. График зависимости между процентом удаленных значений и среднеарифметическим евклидовых отклонений для разных столбцов. |

Рассмотрим поведение алгоритма восстановления на разных типах данных. В таблицу 2 внесены некоторые из исследуемых параметров с данными булевого типа. Восстановленные данные получались из промежутка от 0 до 1 и характеризовали нечеткое тяготение результата к нулю или единице. Мы округлили их до целого числа и сравнили с реальными данными.

Таблица 2. Восстановление данных булевого типа. Процент верно восстановленных данных от числа восстанавливаемых

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Исследуемые параметры** | **Процент удаленных элементов столбца** | | | | | | | | |
| **10%** | **20%** | **30%** | **40%** | **50%** | **60%** | **70%** | **80%** | **90%** |
| Параметр № 1 | 80 | 71 | 71 | 74 | 79 | 75 | 81 | 80 | 80 |
| Параметр № 2 | 60 | 62 | 65 | 64 | 64 | 57 | 66 | 50 | 67 |
| Параметр № 3 | 80 | 76 | 84 | 83 | 74 | 84 | 49 | 80 | 83 |
| Параметр № 4 | 60 | 57 | 68 | 71 | 53 | 57 | 47 | 67 | 66 |
| Параметр № 5 | 70 | 71 | 61 | 64 | 58 | 67 | 61 | 58 | 41 |
| Параметр № 6 | 60 | 71 | 90 | 76 | 72 | 65 | 68 | 63 | 31 |
| Параметр № 24.1 | 100 | 100 | 97 | 98 | 98 | 98 | 100 | 99 | 99 |
| Параметр № 24.2 | 50 | 43 | 42 | 48 | 57 | 54 | 49 | 52 | 52 |
| Параметр № 24.3 | 67 | 67 | 55 | 57 | 57 | 60 | 45 | 36 | 62 |
| Параметр № 24.4 | 90 | 86 | 77 | 88 | 83 | 89 | 87 | 68 | 88 |
| Параметр № 24.5 | 90 | 100 | 97 | 98 | 100 | 98 | 99 | 99 | 99 |
| Параметр № 26.1 | 80 | 90 | 77 | 86 | 85 | 87 | 86 | 86 | 87 |
| Параметр № 26.2 | 80 | 57 | 77 | 81 | 71 | 49 | 76 | 81 | 87 |
| Параметр № 26.3 | 80 | 76 | 81 | 48 | 85 | 78 | 84 | 77 | 70 |

Процент верно восстановленных данных от числа восстанавливаемых значительно выше вероятности 0,5, которая получилась бы при случайном выпадении нуля или единицы. Таблица 1 показывает, что процент верно восстановленных данных примерно одинаков при 10 % - 90 % восстановления, он не зависит от количества вырезанных элементов.

Встречаются аномалии. При случайном вырезании наиболее информативной группы значений всё заполняется однотипными значениями, при этом наблюдаем низкий процент восстановленных данных. Например, параметры 3, 4 при 70%, параметры 5, 6 при 90 % вырезанных данных дали низкий процент восстановления.

Высокий процент восстановления дали столбцы, заполненные сравнительно однородными данными, например, параметры 24.1,4,5, 26.1. При восстановлении не булевых данных из некоторого промежутка таким подходом можно оценить лишь среднюю погрешность восстановления по столбцу.

[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, но для оценки результатов восстановления недостающих значений параметров (признаков) элементов ОТС они не применимы.

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. Данный способ не подходит для нашего случая, так как фактическое значение часто близко или равно нулю. Мы получаем некорректные значения [MAPE](http://en.wikipedia.org/wiki/Mean_absolute_percentage_error).

[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

Оценки [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) в контексте задачи восстановления недостающих значений параметров (признаков) элементов ОТС не применимы, так как параметры обладают разным размахом значений, разной степенью зашумлённости значений, разной степенью детерминированности значений. Получение общей оценки качества для всех восстанавливаемых значений возможно, но в таком контексте мало информативно. Приведём графики зависимости общего значения [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) по всем восстанавливаемым параметрам от процента вырезанных (недостающих) значений от общего массива данных. Рисунок 6, график зависимости значений оценки [MAE](http://en.wikipedia.org/wiki/Mean_absolute_error) (ось ординат) от процента вырезанных значений (ось абсцисс) из всего массива данных. Зависимость не подтверждена.

Рис. 6. График зависимости значений оценки [MAE](http://en.wikipedia.org/wiki/Mean_absolute_error) от процента вырезанных значений из всего массива данных.

Рисунок 7, график зависимости значений оценки [MAPE](http://en.wikipedia.org/wiki/Mean_squared_error) (ось ординат) от процента вырезанных значений (ось абсцисс) из всего массива данных. Зависимость не подтверждена.

Рис 7. График зависимости значений оценки [MAPE](http://en.wikipedia.org/wiki/Mean_squared_error) от процента вырезанных значений из всего массива данных.

Рисунок8, график зависимости значений оценки [ME](http://en.wikipedia.org/wiki/Mean_squared_error) (ось ординат) от процента вырезанных значений (ось абсцисс) из всего массива данных. Зависимость слабая. И объяснима особенностями формулы. Скорее этот график отражает зависимость между процентом вырезанных значений и количеством вырезанных значений, что также делает его малоинформативным.

Рис. 8. График зависимости значений оценки [ME](http://en.wikipedia.org/wiki/Mean_squared_error) от процента вырезанных значений из всего массива данных.

Рисунок 9, график зависимости значений оценки [MSE](http://en.wikipedia.org/wiki/Mean_squared_error) (ось ординат) от процента вырезанных значений (ось абсцисс) из всего массива данных. Зависимость не подтверждена.

Рис. 9. График зависимости значений оценки [MSE](http://en.wikipedia.org/wiki/Mean_squared_error) от процента вырезанных значений из всего массива данных. Зависимость не подтверждена.

По той же причине мы не можем привести графики [RMSE](http://en.wikipedia.org/wiki/Root-mean-square_deviation), [SD](http://en.wikipedia.org/wiki/Standard_deviation). Значения этих оценок оказались колеблющимися вокруг константы.

Значения оценок [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) получаются следующего характера, например, 1Е+14, т.е. . Такие большие значения оценок получаются по причине деления на число близкое к нулю, по причине накопления большого числа ошибок восстановленных данных с большим размахом, при агрегации ошибок восстановления большого числа недостающих значений в массивах большой размерности. Вывод. Данные оценки в рассматриваемом методе не применимы. Они не информативны из-за разного размаха параметров и масштаба данных столбцов.

Для заключительной оценки результатов восстановления значений параметров (признаков) предлагаем использовать приведённую погрешность, которая характеризует точность каждого отдельного восстанавливаемого значения в долях от размаха параметра.

Алгоритм за один раз восстанавливает множество пропущенных значений параметра, необходим способ их наглядного представления результатов оценки точности большого числа восстановленных данных. Приведём таблицу 3, в которой каждой сторожке соответствует интервал приведённой погрешности, каждому столбцу соответствует эксперимент, в котором произвольным образом вырезался определенный процент значений от общего объёма данных, в каждой ячейке – процент данных, попавший в этот интервал приведённой погрешности при данном эксперименте.

Таблица 3. Результаты эксперимента по изучению зависимости приведённой погрешности от процента вырезанных значений от общего объёма данных.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Процент вырезанных значений от общего объёма данных | | | | | | | | |
|  |  | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
| Интервал приведённой погрешности в процентах от размаха | 0-10 | 27,61 | 25,58 | 20 | 19,44 | 14,71 | 28,68 | 21,97 | 14,28 | 15,59 |
| 10-20 | 13,96 | 13,92 | 19,03 | 18,19 | 17,2 | 13,65 | 11,54 | 11,02 | 13,31 |
| 20-30 | 6,58 | 9,17 | 12,69 | 14,15 | 12,49 | 7,24 | 10,7 | 16,85 | 17,65 |
| 30-40 | 6,9 | 9,92 | 13,44 | 12,46 | 15,78 | 6,09 | 11,07 | 15,42 | 9,75 |
| 40-50 | 10,75 | 10,58 | 12,1 | 11,17 | 14,2 | 6,69 | 10,5 | 14,12 | 13,54 |
| 50-60 | 9,95 | 6,58 | 8,6 | 8,59 | 9,81 | 5,91 | 11,14 | 11,74 | 12,93 |
| 60-70 | 5,78 | 5 | 7,47 | 7,54 | 8,29 | 6,58 | 8,01 | 9,31 | 7,02 |
| 70-80 | 4,98 | 5,08 | 2,96 | 4,68 | 3,97 | 5,91 | 5,31 | 4,32 | 5,45 |
| 80-90 | 3,85 | 5,75 | 2,63 | 2,34 | 2,61 | 9,32 | 4,65 | 1,89 | 3,6 |
| 90-100 | 9,63 | 8,42 | 1,08 | 1,45 | 0,94 | 9,92 | 5,11 | 1,04 | 1,16 |

Наилучший исход для алгоритма восстановления параметров элементов ОТС достигается тогда, когда наибольшая часть данных восстановлены с наименьшей погрешностью. Для иллюстрации результатов эксперимента в таком контексте приводим таблицу 4.

Таблица 4. Накопленный процент восстановленных данных в заданных интервалах приведённой погрешности при различных процентах вырезанных значений от общего объёма данных.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Процент вырезанных значений от общего объёма данных | | | | | | | | |
|  |  | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
| Интервал приведённой погрешности в процентах от размаха | 0-10 | 27,61 | 25,58 | 20,00 | 19,44 | 14,71 | 28,68 | 21,97 | 14,28 | 15,59 |
| 0-20 | 41,57 | 39,50 | 39,03 | 37,62 | 31,91 | 42,33 | 33,51 | 25,31 | 28,90 |
| **0-30** | **48,15** | **48,67** | **51,72** | **51,77** | **44,40** | **49,57** | **44,21** | **42,16** | **46,56** |
| 0-40 | 55,06 | 58,58 | 65,16 | 64,23 | 60,18 | 55,65 | 55,28 | 57,58 | 56,31 |
| 0-50 | 65,81 | 69,17 | 77,26 | 75,40 | 74,38 | 62,35 | 65,78 | 71,70 | 69,84 |
| 0-60 | 75,76 | 75,75 | 85,86 | 83,99 | 84,19 | 68,26 | 76,92 | 83,44 | 82,77 |
| 0-70 | 81,54 | 80,75 | 93,33 | 91,53 | 92,48 | 74,84 | 84,93 | 92,75 | 89,79 |
| 0-80 | 86,52 | 85,83 | 96,29 | 96,21 | 96,45 | 80,76 | 90,24 | 97,07 | 95,24 |
| 0-90 | 90,37 | 91,58 | 98,92 | 98,55 | 99,06 | 90,08 | 94,89 | 98,96 | 98,84 |
| 0-100 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |

Таблица 4 показывает, что алгоритм работает примерно с одинаковой точностью восстановления как при 10%, так и при 90% вырезанных данных. Существующие алгоритмы восстановления данных основаны на статистических распределениях, имеют жёсткие ограничения по проценту восстанавливаемых данных и зависят от вида статистического распределения. Данный алгоритм основан на методах нечёткой кластеризации, таким образом для него эти ограничения сняты.

Рис. 10. График зависимости значений оценки [MSE](http://en.wikipedia.org/wiki/Mean_squared_error) от процента вырезанных значений из всего массива данных. Зависимость не подтверждена.

На рисунке 10 приведём графики накопленного процента восстановленных данных в зависимости от интервалов приведённой погрешности для каждого из экспериментов по восстановлению при различных процентах вырезанных значений. Графики близки друг к другу, что говорит о том, что для данного алгоритма качество восстановления пропущенных значений не зависит от процента вырезанных значений или зависимость незначительна.

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**