**Исследование алгоритмов сокращение размерности пространства признаков в задаче анализа клиентских данных**

Н.В. Попов1, Н.В. Размочаева2, Д.М. Клионский3

Факультет компьютерных технологий и информатики

Санкт-Петербургский государственный электротехнический

университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

1nikita.popov14916@gmail.com, 2nvrazmochaeva@etu.ru, 3dmklionsky@etu.ru

*Аннотация* — В данной работе была поставлена цель исследовать и определить самые точные и быстрые алгоритмы сокращения размерности пространства признаков в задаче анализа клиентских данных, так как большие базы данных могут содержать зашумляющую или дублирующую информацию, от которой следует избавиться для улучшения качества обработки данных. Для этого был проведен обзор существующих решений отбора и выделения признаков и выделены сильные и старые стороны. Результаты данной работы показали, что самыми точными и быстрыми алгоритмами являются одномерный отбор, использующий в качестве статического критерия Хи-квадрат и метод библиотеки scikit-learn SelectFromModel, принимающий оценки параметров модели логистической регрессии.

Ключевые слова: отбор данных, анализ клиентских данных, размерность пространства признаков, машинное обучение

# **Введение**

На сегодняшний день данные, описывающую ту или иную предметную область, часто содержат в себе большое количество различных признаков, которые определяют свойства каких-либо процессов или предметов. При этом эти наборы данных могут достигать колоссальных размеров, благодаря чему работа с данными превращается в очень ресурсозатратные и долгие процессы. Сокращение размерности пространства признаков влечет за собой уменьшение используемой памяти и избавление от зашумленной и дублирующей информации. Для достижения данной цели используются различные алгоритмы машинного обучения, каждый из которых имеет свою индивидуальную степень эффективности и обоснованности для своей области применения. Клиентские базы не являются исключением, они часто содержат в себе большие массивы информации, часть из которых является избыточной. Поэтому определение способов сокращения пространства признаков для более точного анализа клиентских данных – проблема, актуальность которой только возрастает с течением время. Объектом исследования являются алгоритмы сокращения размерности пространства признаков, а предметом – различные характеристики алгоритмов такие, как принцип работы, точность, скорость и наглядность результата. В статье решаются задачи:

* Обзор существующих алгоритмов сокращения пространства признаков.
* Формирование критериев для определения оптимальных алгоритмов.
* Обзор рассматриваемых в статье алгоритмов.
* Исследование характеристик рассмотренных алгоритмов.
* Сравнение эффективности алгоритмов согласно выявленным критериям.

# **Сравнение аналогов**

Для того, чтобы определить наилучшие алгоритмы, необходимо провести обзор на современные решения уменьшения количества параметров. При поиске существующих аналогов решения задачи сокращения размерности пространства признаков, было акцентировано внимание на работы во всех сферах деятельности, связанных с признаками, в том числе медицина, разработка оборудования и т.д. Были рассмотрены как чисто теоретические выкладки, так и выполненные на практике решения.

## Научная статья "Методы снижения размерности пространства статистических данных" [1].

Рассмотрено несколько способов снижения размерности признаков:

* Метод главных компонент
* Факторный анализ.
* Многомерное шкалирование.
* Алгоритмы снижения размерности в автоматизированном системно-когнитивном анализе.

Приведено теоретическое описание алгоритмов и методы их оценивания. Рассмотренные алгоритмы выделяют признаки, а не сокращают, поэтому их применение в работе с данными о клиентах находится под вопросом.

## Научная статья "Снижение размерности признакового пространства в задачах идентификации излучающих объектов по данным радиомониторинга с использованием искусственных нейронных сетей" [2].

Разбор метода анализа главных компонент и использования автоассоциативных нейронных сетей. Анализ возможностей искусственных нейронных сетей показал, что автоассоциативные нейронных сети позволяют минимизировать признаковое пространство без значительной потери информации в работе с радиосигналами.

## Научная статья "Алгоритмы и методы снижения пространства диагностических признаков" [3].

Проводится сравнительный анализ методов снижения размерности пространства диагностических признаков в лечебной практике таких, как:

* Кластерный анализ.
* Метод главных компонент.
* Метод экстремальной группировки параметров.
* Многомерное шкалирование.

Приведено теоретическое описание методов. Несмотря на исследование алгоритмов в сфере медицины в данной статье, эти алгоритмы могут быть применены и для других областей.

## Научная статья "Алгоритм GRAD для выбора информативного подпространства признаков" [4].

Проводится разбор алгоритмов AlDel, Grad и алгоритм комбинированного типа DX сокращения пространства признаков. Были проведены экспериментальные выкладки, проверяющие эффективность методов. Заключен вывод, что алгоритм GRAD, основанный на AdDel-методе наиболее целесообразен.

## Научная статья "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning" [5].

В статье описываются широко используемые техники отбора и выделения признаков.

Даны описания методов отбора признаков: Correlation Coefficien, BW-ration, PAV, mRmR, I-RELIEF, CMIM, TERACT, Genetic Algorithm, SVM-REF.

Даны описания методов выделения признаков: Independent Component Analysis, Principle Component Analysis, Nonlinear Principle Components Analysis, Probabilistic Principle Component Analysis, Kernel Principle Component Analysis.

Был сделан вывод, что для таких задач, как выявление болезней, более удобным являются методы выявления признаков, в то время, как для избавления от зашумляющих и дублирующих признаков предназначены алгоритмы сокращения.

# **Критерии сравнения аналогов:**

* Возможность настройки алгоритмов.

Работа алгоритмов на данных различных размеров может иметь разную эффективность, поэтому необходимо иметь возможность настройки параметров алгоритма.

* Количество рассмотренных алгоритмов.

Большое число алгоритмов позволяет определить, как можно больше эффективных сильных сторон для сравнения.

* Тип алгоритма сокращения.

Алгоритмы могут не только сокращать количество признаков, но и преобразовывать их в новые, из-за чего данные могут потерять свою наглядность для аналитиков. Поэтому следует различать методы сокращения и выделения.

Сравнение по критериям представлено в табл. 1.

TАБЛИЦА I. Сравнение по критериям

| **№ аналога** | **1** | **2** | **3** | **4** | **5** |
| --- | --- | --- | --- | --- | --- |
| Возможность настройки алгоритмов | + | + | + | - | + |
| Количество алгоритмов | 4 | 2 | 4 | 3 | 14 |
| Тип алгоритма отбора | выд. | выд. | выд. | сокр. | сокр. и выд. |

Выводы по итогам сравнения: в вышеразобранных статьях подробно описано большинство современных способов сокращения размерности пространства признаков, но также стоит отметить, что практическая реализация алгоритмов отсутствует, поэтому количественных критериев сравнения определить не удалось. В настоящее время разработаны алгоритмы для отдельных областей таких, как медицина или исследование физических явлений, но эти прикладные программы не покрывают всей области решения проблемы большого количества параметров в различных сферах производства и науки, в том числе и клиентских баз данных.

Алгоритмы первой, второй и третьей статьи являются методами выделения признаков, поэтому их использование в анализе данных о клиентах может не соответствовать условиям. Наша цель – определить характеристики алгоритма, работающего для клиентских данных, поэтому вторая и третья статья, базирующиеся на исследовании радиосигналов и медицинских диагнозов, менее актуальны в нашей работе.

Выбор метода решения

По итогам сравнения существующих аналогов были сформированы требования для одного метода или нескольких методов, которые бы могли привести данные о клиентах в пригодный для дальнейшего анализа вид – то есть сократить все зашумляющие и дублирующие признаки. При этом разработанные методы не должны зависеть от количества признаков или размера анализируемых данных, а должны иметь возможность настраивания параметров алгоритма для конкретных данных. Алгоритмы должны иметь высокую точность результата, высокую скорость работы и частично позаимствованные сильные стороны уже существующих алгоритмов сокращения. При этом конкретно для задачи анализа клиентских данных акцент должен быть сделан на методах отбора признаков, а не выделения.

# **Описание метода решения**

Описание рассматриваемых алгоритмов сокращения пространства признаков.

Для сокращения количества признаков в работе используются такие методы, как сокращение признаков с низкой вариативностью, одномерный отбор и метод главных компонент. Также рассматриваются методы отбора признаков такие, как функция SelectFromModel библиотеки scikit-learn и алгоритм рекурсивного исключения признаков. Эти методы используют результаты работы моделей-оценщиков, которые представлены в данной работе логистической регрессией, случайным лесом и сверхслучайным лесом.

## Сокращение признаков с низкой вариативностью.

Удаляются все объекты, дисперсия которых не соответствует некоторому порогу. По умолчанию метод удаляет все объекты с нулевой дисперсией, т. е. объекты, имеющие одинаковое значение во всех выборках.

## Одномерный отбор.

Данный метод отбирает признаки, имеющие наиболее выраженную взаимосвязь с целевой переменной с помощью статистического критерия Хи-квадрат. Формула Хи-квадрата приведена на (1):

|  |  |
| --- | --- |
|  | (1) |

## Метод главных компонент (PCA).

Метод позволяет сократить размерность пространства признаков данных с помощью преобразования на основе линейной алгебры путем проецирования данных в пространство меньшей размерности. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. Данный алгоритм является единственным среди представленных алгоритмом обучения без учителя, то есть для его работы не нужен целевой параметр.

## Метод SelectFromModel.

Метод принимает на вход модель-оценщик, которая предоставляет значения для каждого из параметров, которые сравниваются с параметром функции "threshold". Из данных удаляются признаки, значения которых ниже порогового значения.

## Метод рекурсивного исключения признаков.

Данный алгоритм также известен, как Recursive Feature Elimination (RFE). Алгоритм получает на вход модель-оценщик, предоставляющий значения для каждого из параметров. На основе этих значений производится поочередное удаление признаков из первоначального набора. Рассматривая все меньшие наборы признаков, алгоритм сохраняет только наиболее важные до тех пор, пока в наборе не останется определенное параметром "n\_features\_to\_select" количество признаков.

## Логистическая регрессия.

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится зависимая переменная, принимающая значения 0 и 1 и множество независимых переменных, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

В модели логистической регрессии имеются веса признаков, которые можно считать за важность параметров в задаче сокращения размерности пространства этих признаков. Чем ближе к нулю вес, тем меньшее влияние имеет этот признак на решение.

## Случайный лес.

Случайный лес – модель, состоящая из множества деревьев решений. Дерево принятия решений – средство поддержки принятия решений. Структура дерева представляет собой "листья" и "ветки". На "ветках" дерева решения записаны атрибуты, от которых зависит целевая функция. В "листьях" записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Классифицирование происходит путем спуска по дереву до листа и выдачи соответствующего значения.

В каждом узле дерево решений ищет такое значение определённого параметра, которое приведёт к максимальному уменьшению загрязнения (коэффициента) Джини. Загрязнение Джини – вероятность неверной маркировки в узле случайно выбранного образца. Коэффициент Джини узла равен разнице единицы и суммы отношений класса к общему количеству образцов, возведённых в квадрат, для каждого из множества классов.

Случайный лес – модель, в которой для построения множества деревьев выбирается случайная выборка образцов из набора данных. При разделении узлов выбираются случайные наборы параметров. Значимость параметра в случайном лесу – это суммарное уменьшение загрязнения Джини во всех узлах, использующих этот параметр для разделения. Именно это значение является важностью признака, по которому можно судить о его необходимости в наборе данных.

## Сверхслучайные деревья.

В сверхслучайных деревьях, как и в случайных лесах, используется случайное подмножество возможных признаков, но вместо поиска наиболее оптимальных порогов, пороговые значения произвольно выбираются для каждого возможного признака, и наилучший из этих случайно генерируемых порогов выбирается, как лучшее правило для разделения узла.

# **Исследование характеристик рассмотренных алгоритмов**

Для сравнения описанных выше методов были проведены замеры скорости и точности работы этих алгоритмов. В качестве языка программирования был выбран python, так как для него разработано большое число библиотек, которые широко используются в машинном обучении, в том числе библиотека scikit-learn, содержащая некоторые из алгоритмов отбора данных.

В качестве данных были использованы три набора данных [6-8], содержащие от 17 до 21 признаков и различного размера кортежи: от 2838 до 41188. При этом часть данных была исключена из наборов после предварительной обработки, так как содержала искаженную информацию или пробелы в данных кортежей. Так как все представленные алгоритмы кроме сокращения признаков с низкой вариативностью и PCA являются алгоритмами обучения с учителем, был отобран признак, определенный как целевой.

Каждый набор данных, полученный в качестве результата работы одного из алгоритмов применяется для построения модели, основанной на логической регрессии. Далее производится расчет целевого параметра тестовых данных (20% от общего числа набора) и сравнивается с оригинальным. Точность алгоритма сокращения признаков определяется отношением правильных прогнозов ко всем.

## Сокращение признаков с низкой вариативностью.

В библиотеке scikit-learn метод реализован функцией sklearn.feature\_selection.VarianceThreshold (версия v0.22) [7]. Результаты применения метода на используемом наборе данных представлены в табл. 2. Первый результат в табл. 2 характеризует прогноз без применения данного метода.

TАБЛИЦА II. Результаты применения метода VarianceThreshold

|  |  |  |  |
| --- | --- | --- | --- |
| **threshold** | **Количество исключенных признаков** | **Точность** | **Время работы** |
| – | – | 0,809606 | – |
| 0 | 0 | 0,809606 | 0,003509 |
| 0,005 | 1 | 0,808733 | 0,003040 |

Как показали результаты, точность модели только ухудшилась с использованием метода.

## Одномерный отбор.

Для реализации метода была использована функция sklearn.feature\_selection.SelectKBest (версия v0.22) [7], с параметром score\_func = chi2. Точность для различных наборов данных после проведения сокращения с помощью данного алгоритма представлена на рис. 1. Время работы метода составляет – 0,006986. В наивысших точках (при количестве признаков 11, 12 и 14) кривая имеет значение 0,810480.

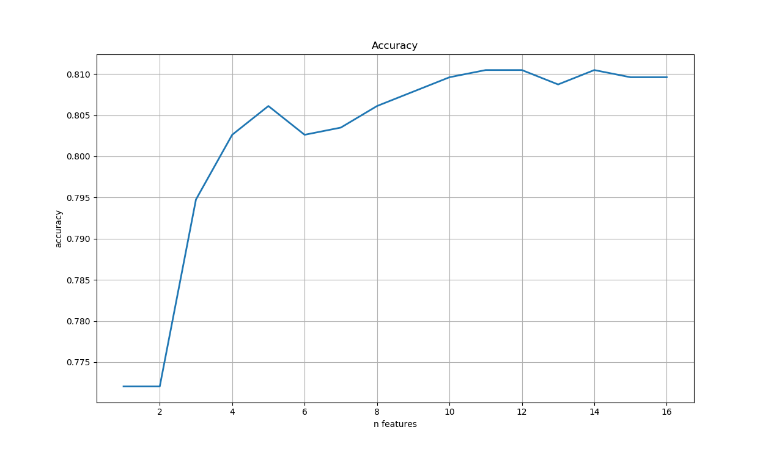


Рисунок 1 – Точность одномерного отбора для различного числа признаков

## Метод главных компонент.

В библиотеке scikit-learn метод реализован функцией sklearn.decomposition.PCA (версия v0.22) [7]. Зафиксированные результаты работы представлены на рис. 2. Скорость работы алгоритма составляет 0,011030 сек.

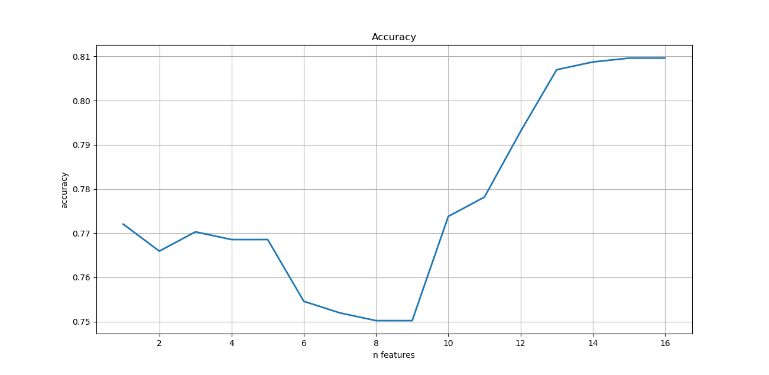


Рисунок 2 – Точность МГК для различного числа признаков

## Методы SelectFromModel и рекурсивного исключения признаков.

В библиотеке scikit-learn алгоритмы имеет реализацию в виде функций sklearn.feature\_selection. SelectFromModel (версия v0.22) [7] и sklearn.feature\_selection.RFE (версия v0.22) [7] соответственно.

Модели логистической регрессии, случайного леса и сверхслучайных деревьев представлены функциями sklearn.linear\_model.LogisticRegression (версия v0.22) [7], sklearn.ensemble.RandomForestClassifier (версия v0.22) [7] и sklearn.ensemble.ExtraTreesClassifier (версия v0.22) [7] соответственно. Результаты эффективности работы алгоритмов представлены в табл. 3 и на рис. 3-4.

TАБЛИЦА III. Результаты эффективности работы алгоритмов SelectModel и RFE

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Метод** | **SelectFromModel** | | | **RFE** | | |
| Модель | LR | RFC | ETC | LR | RFC | ETC |
| Скорость | 0,057502 | 0,237182 | 0,167122 | 0,487762 | 3,387273 | 2,251238 |
| Точность | 0,810480 | 0,809606 | 0,809606 | 0,810480 | 0,809606 | 0,809606 |

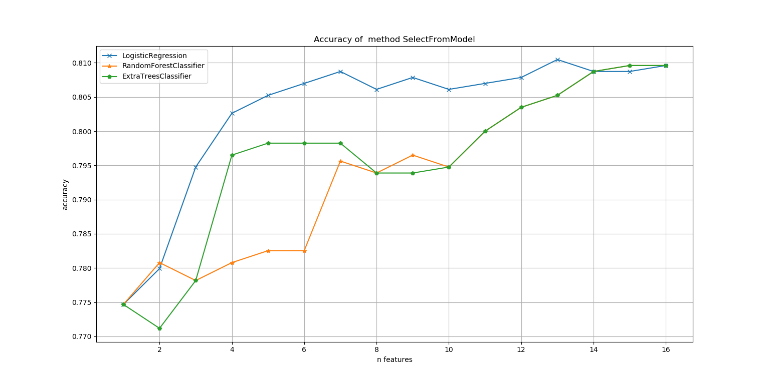


Рисунок 3 – Точность метода SelectFromModel для различного числа признаков

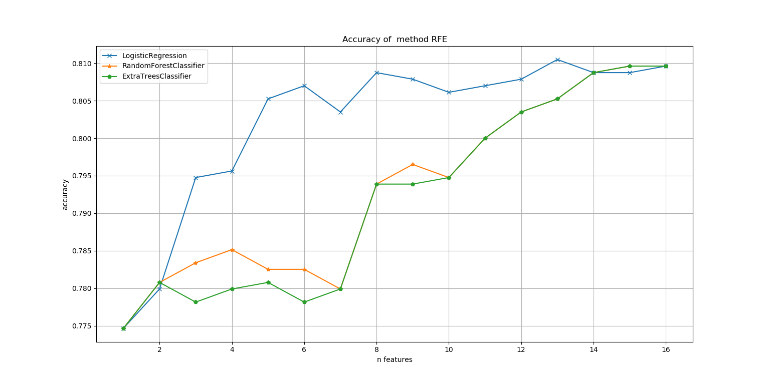


Рисунок 4 – Точность RFE для различного числа признаков

Модели логистической регрессии и случайного леса для методов SelectFromModel и RFE практически совпали, поэтому для сравнения этих алгоритмов был выведен график, на котором методы используют модель сверхслучайных деревьев, предоставленный на рис. 5.

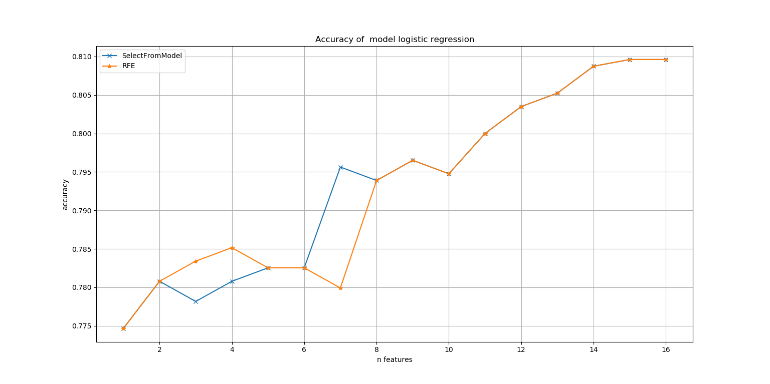


Рисунок 5 – Точность модели сверхслучайных деревьев для различного числа признаков

# **Сравнение эффективности алгоритмов.**

Применение метода сокращения признаков с низкой вариативностью привело к уменьшению точности решения. Это произошло по причине высоковариативности используемых данных. Поэтому данный метод наиболее эффективен для более объемных данных, в которых имеется несколько десятков признаков.

Одномерный отбор показал наиболее высокий уровень точности среди рассмотренных алгоритмов – 0,810480.

Метод главных компонент показал высокую точность только при уменьшении пространства признаков на 1-3 признака, тогда как при дальнейшем сокращении точность падает.

В алгоритмах SelectFromModel и рекурсивного исключения признаков среди моделей наивысшую точность с незначительным перевесом показала логистическая регрессия. Для первого метода хуже всего показала себя модель случайного леса, для второго – сверхслучайных деревьев.

Если сравнивать эти два алгоритма, то результаты практически идентичны, исключая случай применения классификатора сверхслучайных деревьев, где средняя точность метода SelectFromModel оказалась больше (на 0,6%)

Меньшее время работы показали алгоритмы сокращения признаков с низкой вариативностью, одномерный отбор, метод главных компонент и SelectFromModel, реализующий оценку логистической регрессии – меньше 0,1 сек. Самым продолжительным оказался алгоритм RFE, использующий модель случайного леса.

# **Выводы**

В ходе работы было установлено, что наибольшую точность (0,810480) сокращения размерности пространства признаков имеет одномерный отбор, использующий в качестве статического критерия Хи-квадрат. Скорость данного алгоритма высокая по сравнению с другими рассмотренными алгоритмами. Менее точный результат показали методы рекурсивного исключения признаков и SelectFromModel, использующих модели-оценщики логистической регрессии, случайного леса и сверхслучайных деревьев. Для первой и второй модели средняя точность алгоритмов оказались приблизительно одинаковой, но для сверхслучайных деревьев метод SelectFromModel оказался точнее на 0,6%. Также среди моделей оценщиков наибольшую точность показала логистическая регрессия (средняя точность на 1,2% больше случайного леса и на 1% – сверхслучайных деревьев). Скорость данных алгоритмов относительно других рассмотренных оказалась низкой, но это можно регулировать количеством оценщиков в моделях. Сокращение признаков с низкой вариативностью и метод главных компонент показали высокую скорость работы, но низкую точность. Также стоит отметить, что все рассмотренные алгоритмы предоставляют возможность регулировки количества сохраняемых параметров, а в случае моделей-оценщиков еще и количество оценщиков для настройки точности, и скорости работы алгоритма.

В качестве перспективного вектора развития можно определить исследование других современных алгоритмов предобработки и анализа данных, а также усовершенствование рассмотренных в статье: комбинирование методов, использование нейросетей.

Список литературы

1. Орлов Александр Иванович, Луценко Евгений Вениаминович Методы снижения размерности пространства статистических данных // Научный журнал КубГАУ - Scientific Journal of KubSAU. 2016. №119. URL: https://cyberleninka.ru/article/n/metody-snizheniya-razmernosti-prostranstva-statisticheskih-dannyh (дата обращения: 23.12.2019).
2. Аджемов С. С., Терешонок М. В., Чиров Д. С. Снижение размерности признакового пространства в задачах идентификации излучающих объектов по данным радиомониторинга с использованием искусственных нейронных сетей // T-Comm. 2008. №6. URL: https://cyberleninka.ru/article/n/snizhenie-razmernosti-priznakovogo-prostranstva-v-zadachah-identifikatsii-izluchayuschih-obektov-po-dannym-radiomonitoringa-s (дата обращения: 23.12.2019).
3. Максюта Н. В., Поворознюк А. И. Лгоритмы и методы снижения пространства диагностических признаков // Вестник НТУ ХПИ. 2005. №46. URL: https://cyberleninka.ru/article/n/lgoritmy-i-metody-snizheniya-prostranstva-diagnosticheskih-priznakov (дата обращения: 23.12.2019).
4. Загоруйко Н. Г., Кутненко О. А. Алгоритм GRAD для выбора информативного подпространства признаков //URL: http://math. nsc. ru/~ wwwzag/7. GRAD. pdf (дата обращения: 12.01. 2016). – 2005.
5. Khalid S., Khalil T., Nasreen S. A survey of feature selection and feature extraction techniques in machine learning //2014 Science and Information Conference. – IEEE, 2014. – С. 372-378.
6. Bank Marketing //URL: https://www.kaggle.com/henriqueyamahata/bank-marketing-classification-roc-f1-recall/data (дата обращения: 23.12.2019).
7. Библиотека sklearn //URL: https://scikit-learn.org (дата обращения: 23.12.2019).

**The Study of Dimension Reduction Algorithms of Feature Space in Task of Analyzing Client's Data**

Nikita V. Popov1, Natalya V. Razmochaeva2, Dmitry M. Klionskiy3

Department of Software Engineering and Computer Applications

Saint Petersburg Electrotechnical University "LETI"

1nvpopov@stud.eltech.ru, 2nvrazmochaeva@etu.ru, 3dmklionsky@etu.ru

In this paper, the goal was to investigate and determine the most accurate and fast algorithms for reducing the dimension of the feature space in the task of analyzing client data, because large databases may contain noisy or duplicate information that should be eliminated to improve the quality of data processing. To do this, a review of existing solutions for selection and extract features was conducted, and strengths and old sides were highlighted. The results of this work have shown that the most accurate and fast algorithms are one-dimensional selection, which uses Chi-square as a static criterion, and the method of the scikit-learn SelectFromModel library, which accepts estimates of the parameters of the logistic regression model.

Главным недостатком текста видится перевод англоязычных терминов. Например, термин dimensionality reduction, которому посвящена статья, стоит переводить как "снижение размерности", а не "сокращение".

Точно не повредило бы описание того, что такое клиентские данные, чем они отличаются от физических и медицинских, почему именно для них задача актуальна.

Проблемы с оформлением графиков:   
--дублирующиеся подписи  
--подписи, не оформленные как \caption{} для графиков  
--графики не векторизованные, плохо читаемые  
--много пустого пространства как следствие неправильного размещения графиков

Классика жанра: цветные графики надо делать такими, чтобы при чёрно-белой печати они оставались читаемыми.  
3. Мы пишем по-русски, в русском языке десятичный разделитель — запятая.  
4. Сущности кода наподобие `sklearn.linear\_model.LogisticRegression` имело бы смысл выделить соответствующим шрифтом.   
6. Не вдаваясь в подробности, почему это так, обращу, всё же, внимание на то, что в нижнем правом углу первой страницы стоит отметка "978-1-7281-5761-0/20/$31.00 ©2020 IEEE"

Нет детального описания экспериментов. Какие 20% от выборки брались в качестве тестового набора? Делалась ли кросс-валидация?  
2. Описанные методы были просто применены, никак не объясняется падение качества методов в определенных точках. Нет описания параметров, важных для методов. Например, говорится, что метод главных компонент дает хороший эффект только при снижении размерности с 17 признаков до 13-14, однако никак не вычислялась доля остаточной дисперсии.