# АЛГОРИТМЫ ВЫДЕЛЕНИЯ ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ПО ОЦЕНКЕ ТЕХНИЧЕСКОГО СОСТОЯНИЯ ОБЪЕКТА

**Голев А.В., Огородников О.В.**

*Институт проблем управления им. В.А. Трапезникова,*
*Россия, г. Москва, ул. Профсоюзная, д.65*
oiw23@mail.ru, o.v.ogorodnikov@gmail.com

*Аннотация: В докладе рассмотрены алгоритмы выделения информативных признаков в задаче интеллектуального анализа данных на примере оценки и прогноза технического состояния (задача ранней диагностики) электродвигателя беспилотного летательного аппарата, где анализируемое множество значений контролируемых параметров относится к определенному классу или уровню опасности.*

Ключевые слова: интеллектуальный анализ данных, выделение информативных признаков, классификация, диагностика, электродвигатель, беспилотный летательный аппарат, состояние технического объекта.

## Введение

Актуальность разработки систем контроля технического состояния (СКТС) электродвигателя обусловлена развитием технологий, связанных с созданием перспективных беспилотных электрических летательных аппаратов (ЛА) [1]. Внедрение высокоэффективных СКТС позволит обеспечить высокие показатели отказоустойчивости и снизить эксплуатационные затраты.

Задача разработки алгоритмов оценки состояния технических объектов на практике решается с использованием концептуальных подходов «Белого ящика», «Черного ящика» или их комбинирования [2].

При использовании подхода «Белого ящика» ТО рассматривают как динамический объект, для которого полностью известна его внутренняя структура, набор взаимодействующих элементов, характер связей между ними, внешние и внутренние условия и их влияние. В ходе анализа выполняется моделирование приведенного динамического объекта, изменение параметров которого описывается системой дифференциальных уравнений.

Многие динамические объекты имеют сложную структуру и состоят из значительного количества взаимодействующих элементов. При построении моделей таких объектов необходимо учитывать значительное количество внутренних и внешних факторов. Это часто обуславливает невысокую адекватность моделей, построенных на базе подхода «Белый ящик». К примеру, поведение даже простого динамического объекта, функционирующего в нештатном или аварийном режиме, не будет соответствовать построенной модели, т.к. набор возможных неисправностей объекта широк и не может быть предусмотрен в модели заранее. В противном случае, для описания нештатных ситуаций необходимо сильно усложнять модель.

На практике, поведение большинства объектов плохо поддается формализации. Поэтому подход к построению алгоритмов оценки состояния технических объектов, основанный только на аналитическом моделировании, применяется для сравнительно простых объектов с очевидными свойствами. Аналитические модели динамического объекта не способны приспосабливаться к изменившимся условиям функционирования, т.е. не способны обучаться и в конечном счете могут стать неадекватными.

При использовании концептуального подхода «Черный ящик» к построению алгоритмов оценки состояния ТО необходимость в точной информации о структуре ТО, как динамического объекта, отсутствует. В рамках данного подхода считается, что о функциях и реакциях ТО можно судить по выходным параметрам, наблюдаемым как результат на внешнее воздействие.

Для упрощения задачи поиска и формализации закономерностей в данных о состоянии ТО в них выделяют подмножества прецедентов и интерпретируют их как классы, которые соответствую подмножествам конкретных состояний ТО. Для решения задачи построения моделей классификации применяется дискриминантный анализ, например, линейный дискриминант Фишера, логистическая регрессия и т.д.

Большой интерес представляют различные методы ИАД обеспечивающие высокую эффективность в задачах поиска трудно формализуемых закономерностей в данных. В настоящий момент имеется значительное число современных методов ИАД, пригодных для решения задач оценки состояния ТО. Самые популярные из них это: нейронные сети [3], деревья решений [4], метод рассуждения на основе аналогичных случаев (метод k-ближайших соседей) [5], метод опорных

векторов [6]. Представленные методы и их различные модификации имеют одно общее свойство – используют информацию о ранее возникающих состояниях ТО (прецедентах) для поиска и формализации закономерностей в его поведении, однако различаются между собой по принципу формирования конечного результата.

Для повышения эффективности применения методов ИАД нередко производится предварительная обработка исходных данных, позволяющая выделить наиболее информативные признаки (ИП) для решения задачи классификации (снизить размерность пространства признаков).

Многие исследователи считают этап выделения информативных признаков одним из самых важных и сложных этапов решения задач классификации. В связи с этим становится актуальным вопрос выбора методов и алгоритмов выделения ИП, обеспечивающих наибольшую эффективность методов ИАД.

Задача выделения информативных признаков (снижения размерности пространства признаков) заключается в уменьшении количества исходных признаков (входных параметров), поступающих на вход алгоритмов или моделей классификации, в которых формализованы закономерности, отражающие взаимосвязи между параметрами и техническим состоянием ТО. Особенно значительное внимание уделяется этой проблеме теорией машинного обучения в прикладных задачах, решение которых основано на методах ИАД, так как часто поиск и формализация закономерностей могут быть затруднены из-за избыточности информации (проблема «проклятия размерности») и шума, что может приводить к переобучению (overfitting). Во многих случаях снижение размерности пространства признаков позволяет увеличить точность и упростить модели/алгоритмы классификации, создаваемые с применением методов ИАД.

Обычно при уменьшении размерности пространства признаков исходное множество признаков подвергается процедурам отбора или проекции (преобразования) признаков.

Под отбором признаков понимается формирование подмножества значимых признаков для модели/алгоритма классификации. Применение «оберточных» методов (wrapper method) [7] является наиболее универсальным подходом к отбору признаков. Данные методы обеспечивают оптимизацию подмножества параметров моделей/алгоритмов классификации на основе информации, полученной в результате итерационного выполнения полноценного поиска и формализации закономерностей с использованием методов ИАД. Главной особенностью «оберточных» методов является то, что на каждой итерации оптимизационного алгоритма создаются модели/алгоритмы, которые в дальнейшем могут непосредственно применяться для решения прикладной задачи. Таким образом для каждого подмножества признаков, сгенерированного «оберточным» методом, выполняется обучение нейронной сети, строится дерево решений, растущая пирамидальная сеть и т.д.

Оптимизация подмножества информативных признаков, может сочетаться с оптимизацией параметров применяемого метода ИАД (выбор архитектуры нейронной сети – количество слоев, нейронов, вид активационной функции; глубины дерева решения; параметров метода k-ближайших соседей – метрика сходства объектов, количество соседей).

Общие алгоритмы оптимизации, применяющиеся для решения комбинаторных задач формирования оптимальных подмножеств на основе заданного исходного множества, являются основой для «оберточных» методов. Для создания «оберточных» методов решения задачи отбора информативных признаков с использованием методов ИАД применяют следующие алгоритмы оптимизации: генетические алгоритмы [8], метод рассеяния (scatter search) [9], метод роя частиц [10], метод отжига [11] и т.д.

При значительном количестве исходных признаков, объеме анализируемых данных использование «оберточных» методов для отбора информативных признаков может оказаться неприемлемым из-за существенных вычислительных затрат на построение и тестирование множества моделей/алгоритмов на каждой итерации оптимизации.

Анализ мировой научной литературы показал актуальность применения методов ИАД в интеграции с методами выделения информативных признаков. Использование оберточных методов, методов фильтрации и проекции для выделения информативных признаков позволит повысить эффективность создания и применения алгоритмов оценки и прогноза технического состояния электродвигателей летательных аппаратов. В качестве методов фильтрации выбраны поиск корреляционных зависимостей, критерий Information gain, статистический критерий хи-квадрат-тест. В качестве методов проекции предлагается использовать спектральный анализ, позволяющий учитывать особенности функционирования электродвигателя, метод главных компонент и ядерный метод главных компонент. В качестве «оберточных» методов выбраны метод полного перебора для малого количества исходных признаков (<6) и генетические алгоритмы при большей размерности.

В результате решения задачи ранней диагностики с использованием методов ИАД выполняется оценка и прогноз технического состояния электродвигателя ЛА, при которых анализируемое множество значений контролируемых параметров относится к определенному классу или уровню опасности в зависимости от возможных последствий сложившейся ситуации. При этом наиболее важным является прогноз технического состояния электродвигателя. Возникновение в процессе эксплуатации нештатной или аварийной ситуации может свидетельствовать о случившемся выходе из строя электродвигателя, поэтому получение выводов о текущем состоянии контролируемого объекта является недостаточным. Основной задачей анализа информации при диагностике электродвигателя является определение негативных тенденций изменения технического состояния электродвигателя заранее.

Возможность предотвращения нештатных и аварийных ситуаций, возникающих в процессе эксплуатации электродвигателя, в основном определяется способностью алгоритмов, в которых реализованы методы анализа информации о контролируемых параметрах, точно идентифицировать и прогнозировать техническое состояние электродвигателя. Точность оценки и прогноза технического состояния электродвигателя во многом определяется возможностями алгоритмов выделения информативных признаков, которые применяются при построении алгоритмов оценки технического состояния электродвигателя с использованием методов ИАД.

Задача построения алгоритма оценки технического состояния электродвигателя с использованием методов ИАД может быть формализована, как задача классификации следующим образом. На основе известного конечного множества описаний состояния электродвигателя (базы прецедентов) $\{(\bar{x}_1, \bar{y}_1), \dots, (\bar{x}_k, \bar{y}_k)\}$, $\bar{x}_i \in X$, $\bar{y}_i \in Y$, i=1..k, требуется построить алгоритм оценки технического состояния электродвигателя $\Phi: X \to Y$, позволяющий для произвольного вектора $\bar{x} \in X$ получить вектор $\bar{y} \in Y$, где $X$ – множество векторов значений контролируемых параметров электродвигателя (признаков), $Y$ – множество меток классов, отражающих агрегированные состояния электродвигателя, в отношении которых могут приниматься одинаковые решения.

В частном случае задача классификации может быть сведена к задаче поиска аномальных данных – обнаружения (детектирования) неисправностей. Тогда $Y$ – множество меток классов, отражающих нормальное и нештатное состояние электродвигателя.

В общем случае $X = R^{n \times q}$, $Y = R^m$, где n – количество контролируемых параметров электродвигателя; q – количество последовательных моментов времени, в которые рассматриваются значения контролируемых параметров электродвигателя; m – количество вещественных чисел, необходимое для представления меток классов, соответствующих агрегированным состояниям электродвигателя.

В соответствии с представленной формализацией в алгоритмах диагностики технического состояния электродвигателя реализуются:

1) методы поиска и формализации закономерностей в эмпирических данных, представленных в виде набора прецедентов;
2) формальное представление найденных закономерностей и методы для их использования.

Для повышения эффективности формализации закономерностей, отражающих зависимости между контролируемыми параметрами и оценкой технического состояния электродвигателя, предлагается схема формирования и анализа данных для выделения информативных признаков, которая рассматривается в следующем разделе.

## 1 Анализ данных о техническом состоянии электродвигателя

Для решения задачи ранней диагностики электродвигателя с использованием методов ИАД предлагается следующая схема формирования и анализа данных.

Решение задачи состоит из нескольких этапов: получение эмпирических данных и их обработка, создание первичного набора векторов из эмпирических данных, оптимизация множества признаков и обучение алгоритмов ИАД на данных.

Первый этап связан с получением эмпирических данных. Эти данные были получены в ходе экспериментов с синхронным электродвигателем U5 KV400, разработанным фирмой T-motor для применения в коптерах.

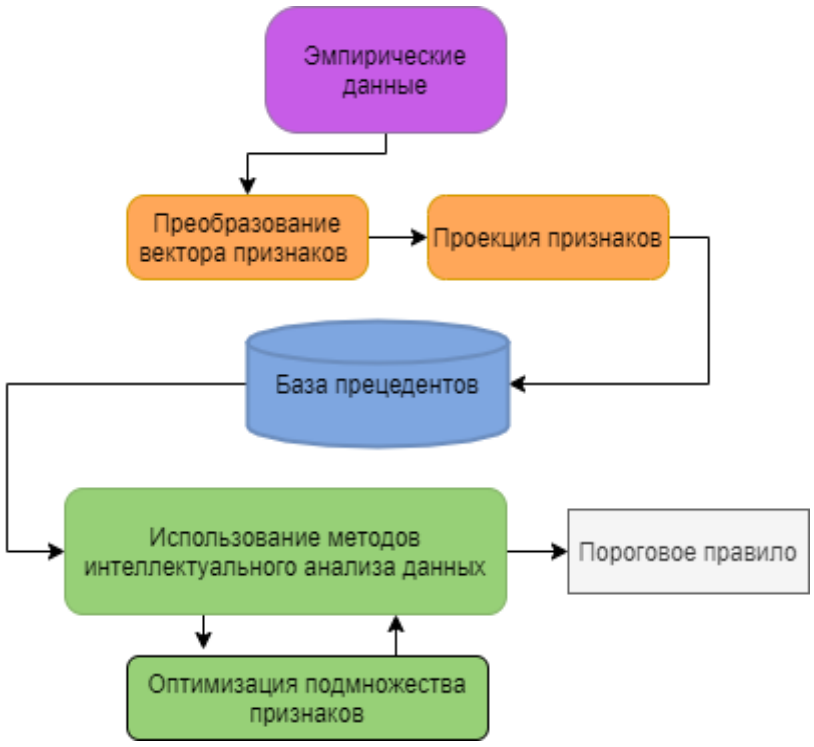


*Рис. 1. Схема анализа данных в задаче диагностики электродвигателя*

Основные характеристики электродвигателя:

1) Двигатель трёхфазный, обмотки соединены в треугольник;
2) Конфигурация 12N14P, статор 12-ти полюсной, на роторе имеется 14 постоянных магнитов;
3) Мощность – 400 Вт;
4) Частота вращения ротора в 7 раз меньше частоты питания.

Двигатель рассчитан на питание от 6-секционного Li-полимерного аккумулятора через специальный мостовой преобразователь, формирующий три биполярных ШИМ-модулированных напряжения, частота которых задается отдельным входным сигналом. Для проведения ресурсных испытаний был создан стенд с жестко закрепленным двигателем. Нагрузкой служит пропеллер, создающий поток воздуха. На стенде установлены:

1) Датчик виброускорений для измерения вибраций статора двигателя;
2) Измерительный микрофон для акустических шумов двигателя;
3) Фотодатчик, фиксирующий пересечение светового потока лопастью пропеллера;
4) Делители напряжения для записи фазных токов и напряжений.

Исходные данные эксперимента были представлены в виде бинарных файлов, по файлу на ежедневную 10 секундную запись с учётом 9 каналов. Частота дискретизации – 100 КГц. Размер каждого файла составляет $10^6$ элементов на каждом канале.

Обработка исходных данных проводилась на языке программирования Python с использованием библиотек Scikit-learn, Numpy и Pandas для организации данных и работы с ними. Данный выбор был продиктован требованиями по скорости обработки данных и их визуализации.

На этапе обработки эмпирических данных были рассмотрены данные за последние 150 дней. По каждой записи для каждого канала были построены с помощью преобразования Фурье спектры сигналов.

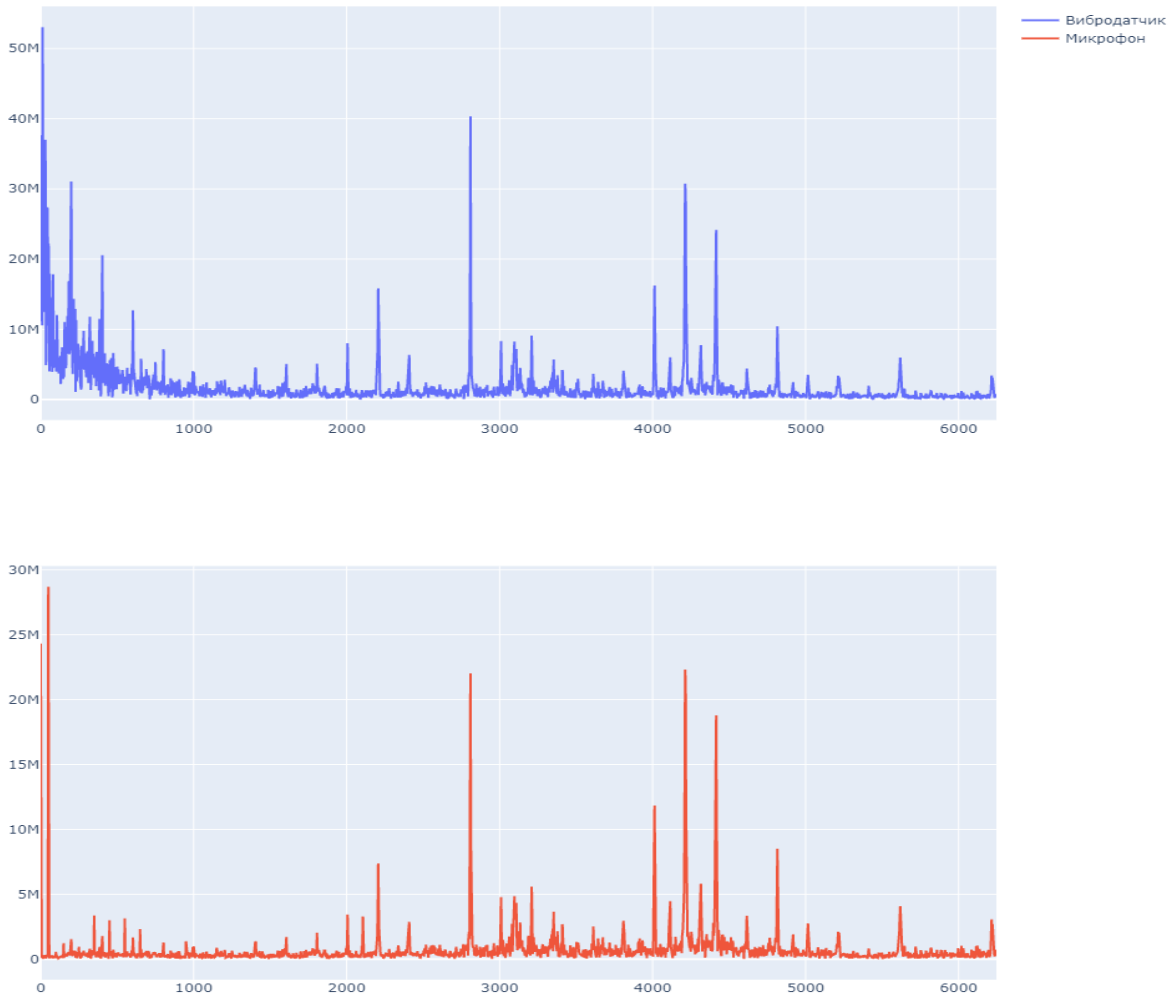


*Рис. 2. Пример преобразования Фурье вибрационного (сверху) и звукового (снизу) сигналов*

Так как в ходе эксперимента ток, напряжении, а также сигнал от фотодатчика в основном были фиксированы, то эти каналы не рассматривались в данной работе.

Таким образом для формирования набора данных для использования алгоритмов ИАД был использован метод скользящего окна. По эмпирическим данным скользящим окном выбирались временные отрезки по нескольким каналам, которые с помощью преобразования Фурье сохранялись спектрами сигналов вибродатчика и микрофона и помещались в набор данных как вектор с длинной в 4000 значений и целевым признаком. Таким образом сформировался набор данных в количестве около 7000 прецедентов. Для того, чтобы убрать избыточность в данных и снизить размерность пространства признаков были использованы метод отбора признаков с помощью теста $\chi^2$ и с использованием случайного леса.

Тест $\chi^2$ используется в статистике для проверки независимости двух событий. Учитывая данные для каждых двух переменных, мы можем получить наблюдаемое число $O$ и ожидаемое число $E$. И с помощью теста $\chi^2$ возможно измерить как ожидаемое число $E$ и наблюдаемое число $O$ отклоняются друг от друга. Таким образом мы отберём только такие признаки, которые будут наиболее значимы для метки. Используя метод $\chi^2$ удалось отобрать 24 признака с уровнем значимости $P > 0.8$.

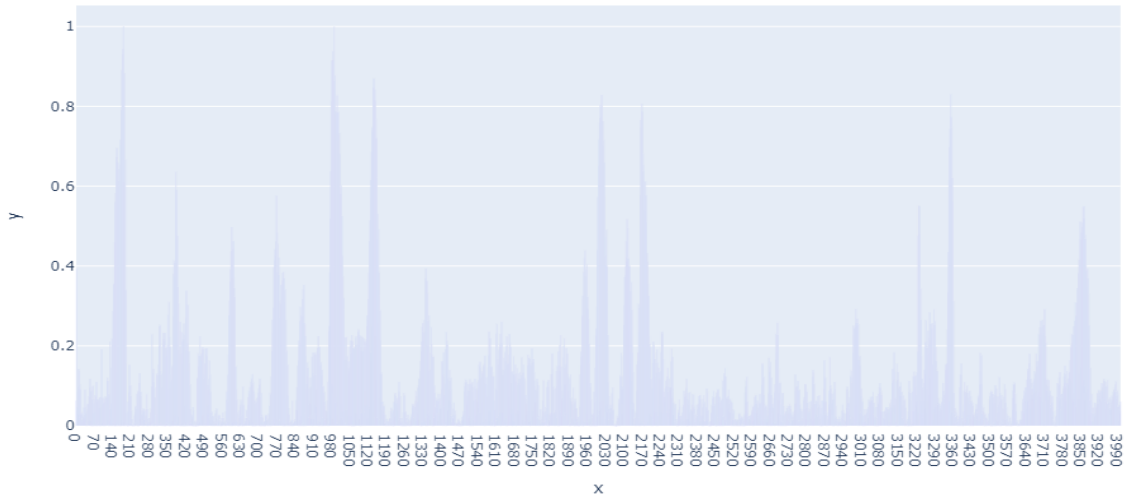


*Рис. 3. График уровня значимости признаков с использованием метода* $\chi^2$

Алгоритмы машинного обучения на основе деревьев решений или случайного леса используют набор деревьев, которые содержат узлы, полученные в результате распределения. Основная цель узлов заключается в том, что они максимально возможно, уменьшают количество «шумов». Такие деревья могут рассчитать, насколько важен признак, измерив степень уменьшения «шумов» за его счёт его отсутствия. Используя метод отбора признаков с помощью случайного леса удалось отобрать 16 признаков из набора данных.

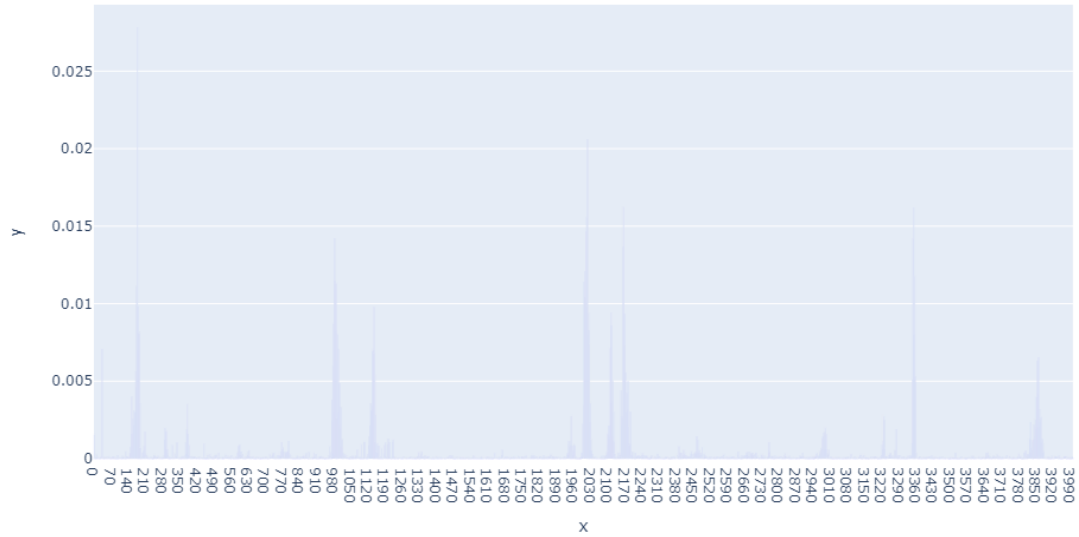


*Рис. 4. График уровня значимости признаков с использованием методов случайного леса*

После того, как были отобраны признаки из полного набора данных были сформированы 4 набора:

1) полный вектор признаков;
2) масштабированный вектор признаков;
3) масштабированные признаки, отобранные методом $\chi^2$;
4) масштабированные признаки, отобранные методом случайного леса.

Все вышеперечисленные наборы данных использовались в решении задачи классификации, где имелось два класса - А и Б. Классы были сбалансированы 55% векторов класса А к 45% векторов класса Б. Ниже приведены различные методы и их оценки на исследуемых наборах данных.

*Таблица 1. Результаты применения различных методов ИАД*

| **Модели** | **данные** | **F-мера** | **точность** | **полнота** | **валидация** | **обучение** | **тест** |
|---|---|---|---|---|---|---|---|
| Классификация случайным образом 0 | полный набор | 0.505 | 0.523 | 0.488 | 0.501 | 0.485 | 0.513 |
| Классификация случайным образом 1 | Масштабированный | 0.505 | 0.523 | 0.488 | 0.501 | 0.485 | 0.513 |
| Классификация случайным образом 2 | хи^2 | 0.505 | 0.523 | 0.488 | 0.501 | 0.485 | 0.513 |
| Классификация случайным образом 3 | Случайный лес | 0.505 | 0.523 | 0.488 | 0.501 | 0.485 | 0.513 |
| Логистическая регрессия (по умолчанию) 0 | полный набор | 0.995 | 0.993 | 0.998 | 0.995 | 1 | 0.998 |
| Логистическая регрессия (по умолчанию) 1 | Масштабированный | 0.996 | 0.993 | 1 | 0.996 | 1 | 0.996 |
| Логистическая регрессия (по умолчанию) 2 | хи^2 | 0.975 | 0.971 | 0.979 | 0.973 | 0.981 | 0.982 |
| Логистическая регрессия (по умолчанию) 3 | Случайный лес | 0.975 | 0.975 | 0.976 | 0.974 | 0.980 | 0.984 |
| Логистическая регрессия (L2, BFGS) 0 | полный набор | 0.993 | 0.991 | 0.994 | 0.992 | 1 | 0.995 |
| Логистическая регрессия (L2, BFGS) 1 | Масштабированный | 0.996 | 0.993 | 1 | 0.996 | 1 | 0.996 |
| Логистическая регрессия (L2, BFGS) 2 | хи^2 | 0.982 | 0.975 | 0.990 | 0.981 | 0.985 | 0.987 |
| Логистическая регрессия (L2, BFGS) 3 | Случайный лес | 0.974 | 0.969 | 0.979 | 0.973 | 0.982 | 0.983 |
| Метод k-ближайших соседей 0 | полный набор | 0.984 | 0.982 | 0.986 | 0.983 | 0.987 | 0.976 |
| Метод k-ближайших соседей 1 | Масштабированный | 0.995 | 0.990 | 1 | 0.994 | 0.997 | 0.994 |
| Метод k-ближайших соседей 2 | хи^2 | 0.983 | 0.967 | 1 | 0.982 | 0.987 | 0.989 |
| Метод k-ближайших соседей 3 | Случайный лес | 0.991 | 0.983 | 0.998 | 0.990 | 0.991 | 0.988 |

По результатам табл. 1 можно заметить, что при использовании полного и масштабированного наборов данных происходило переобучение на тренировочных наборах, чего нельзя сказать про результаты моделей с отобранными признаками. Стоит заметить, что оценки на тестовом наборе и F1-мера на методе k-ближайших соседей у данных с отобранными признаками выше.

## Заключение

На основе экспериментальных наборов данных о тестовой эксплуатации электродвигателя с применением преобразования Фурье получен набор прецедентов, в котором были выделены информативные признаки. С использованием отобранных информативных признаков решена задача

классификации и диагностики технического состояния. Качество классификации на тестовых данных достаточно хорошее.

Направление дальнейшей работы предполагает дополнительную обработку данных и удаление выбросов и шумов, внесённых температурой и влажностью, и оценку результатов работы моделей ИАД с информативными признаками, полученными с помощью «обёрточных методов», а также генетических алгоритмов и сверточных нейронных сетей.

**Литература**


1. *Скрябин А.В.* Системы контроля технического состояния и прогнозирования неисправностей электромеханических рулевых приводов летательного аппарата. Современный уровень развития. // Общероссийский научно-технический журнал «Полет» №2, 2018. – С. 50-64,.
2. *Мыльник В.В., Титаренко Б.П.*, Волочиенко В.А. Исследование систем управления. 2-е изд., перераб. и доп. // М.: Академический Проект, 2003. – 352 с.
3. *Уоссермен Ф.* Нейрокомпьютерная техника: теория и практика: Пер. с англ. // М.: Мир, 1992. – 157 с.
4. *Rokach L.* Data Mining With Decision Trees: Theory and Applications (2nd Edition) // World Scientific Publishing Company/ 2014, P. 328
5. *He Q. P., Wang J.* Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes // IEEE Transactions On Semiconductor Manufacturing. Vol. 20. 2007, № 4.
6. *Jayadeva, Khemchandani R., Chandra S.* Twin Support Vector Machines: Models, Extensions and Applications, // Springer. 2016, P. 211.
7. *Panthong R., Srivihok A.* Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm // The Third Information Systems International Conference // Procedia Computer Science. vol. 72. 2015, P. 162-169.
8. *Soufan O., Kleftogiannis D., Kalnis P., Bajic V.B.* DWFS: a wrapper feature selection tool based on a parallel genetic algorithm // PLoS ONE. vol. 10(2). 2015, P. e0117988.
9. *Gendreau M., Potvin J.-Y.* Handbook of Metaheuristics - Third Edition // USA, Springer International Publishing, P. 604.
10. *Xue B., Zhang M., Browne W.N.* Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach // IEEE Transactions on Cybernetics. vol. 43. issue 6, 2013. P. 1656-1671.
11. *Jeong I-S, Hong-Ki Kim H-K, Kim T-H, Lee D.H., Kim K.J., Kang S-H.* A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks // Convergence Security. vol. 1. 2018, P. 1-18.