Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка алгоритма определения аномалий в данных

Выполнила: Ханова Юлия Алексеевна, гр. 7383
Руководитель: Геппенер В.В., д.т.н., профессор
Консультант: Шевская Н.В.

Санкт-Петербург, 2021

Цель и задачи

Актуальность: выявление аномальных значений в данных может предотвратить поломку оборудования или предотвратить мошенничество.
Проблема: точность существующих методов зависит от области применения.

Цель: разработка алгоритма, определяющего аномальные значения в данных для последующего их удаления или корректировки.

Задачи:

1. проанализировать существующие методы определения аномалий в данных;
2. выбрать метод создания ансамблей моделей;
3. разработать алгоритм на основе ансамбля выбранных методов;
4. реализовать консольное приложение для демонстрации результатов работы алгоритма;
5. проанализировать качество работы разработанного инструменты.

Методы определения аномалий (задача 1)

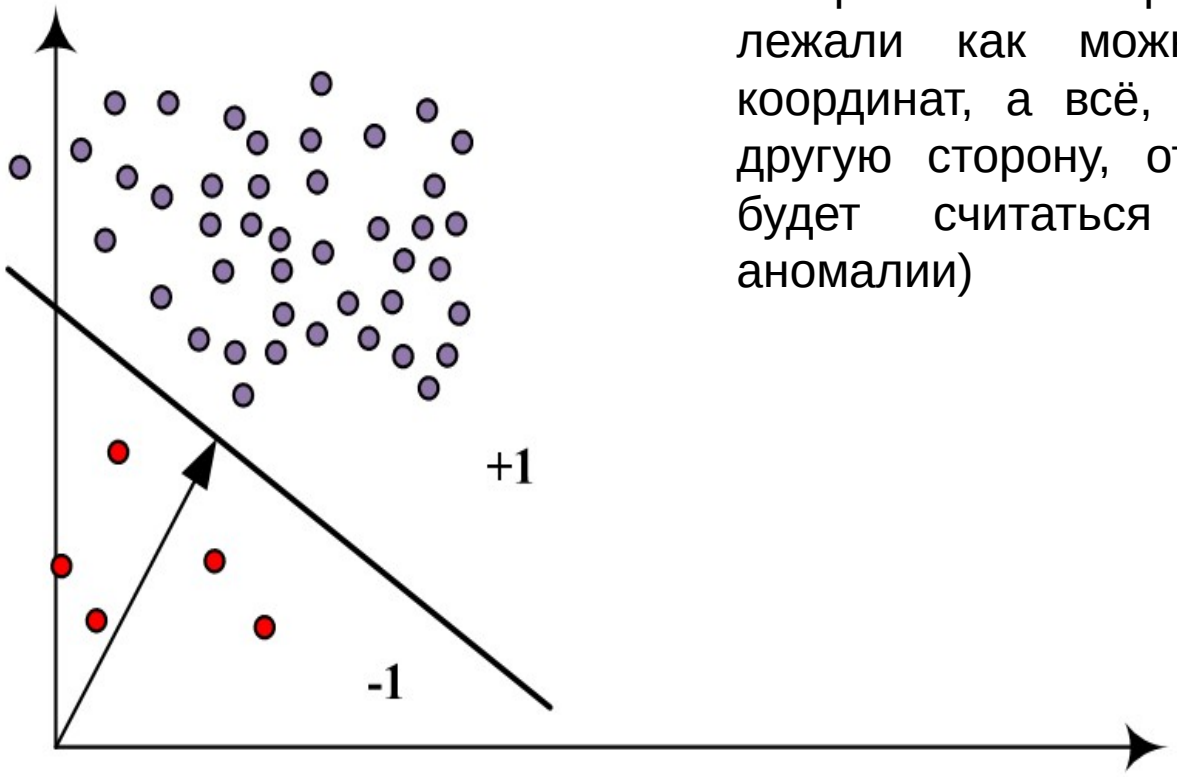
Критерии

1) Скорость работы
2) Устойчивость к неточности вводимых данных
3) Доступность формирования правил
4) Независимость от предметной области

	1	2	3	4
Статистический анализа	+	-	+	-
Машинное обучение	-	+	+	-
Индуктивный вывод	-	-	-	+
Нечеткая логика	-	+	-	-
Генетические алгоритмы	-	+	-	-
Искусственные нейронные сети	-	+	+	-
Гибридные системы	+	+	+	+

Метод опорных векторов для одного класса

Применяется разновидность классического метода – метод опорных векторов для одного класса (One-Class SVM).

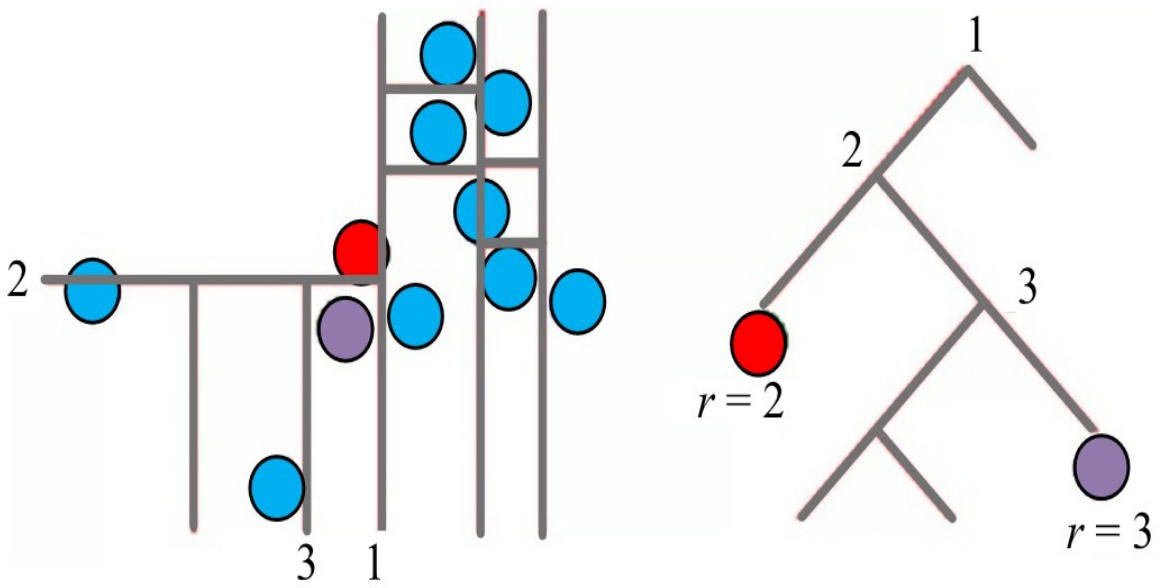


В данном методе разделяющая гиперплоскость строится так, чтобы значения лежали как можно дальше от начала координат, а всё, что будет находится по другую сторону, от разделяющей границы будет считаться аномалией. («-1» - аномалии)

Метод изолирующего леса

Изолирующий лес (Isolation Forest) – это алгоритм обучения без учителя, который принадлежит к семейству деревьев решений ансамбля.

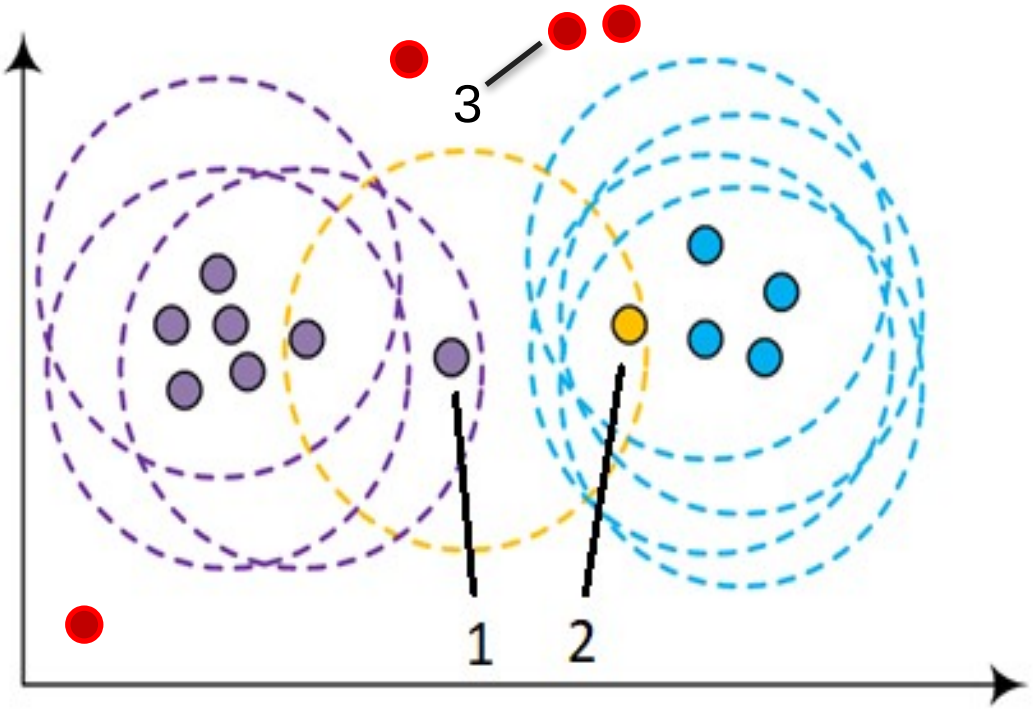
Для каждого объекта мера его нормальности - среднее арифметическое глубин листьев, в которые он попал (изолировался)



Деревья строятся до тех пор, пока каждый объект не окажется в листе. Экземпляры с меньшей мерой нормальности считаются аномалиями («r = 2» – аномалия)

Пространственная кластеризация, основанная на плотности.

DBSCAN (Density-based spatial clustering of applications with noise) – основанная на плотности пространственная кластеризация для приложений с шумами.



- Основная точка, если принадлежит плотной области (1);
- Пограничный пункт, если имеет меньше точек в своей окрестности чем min_samples, но лежит в окрестности другой точка, принадлежащей к плотной области (2);
- Аномалия, если не является основной и пограничной. (3)

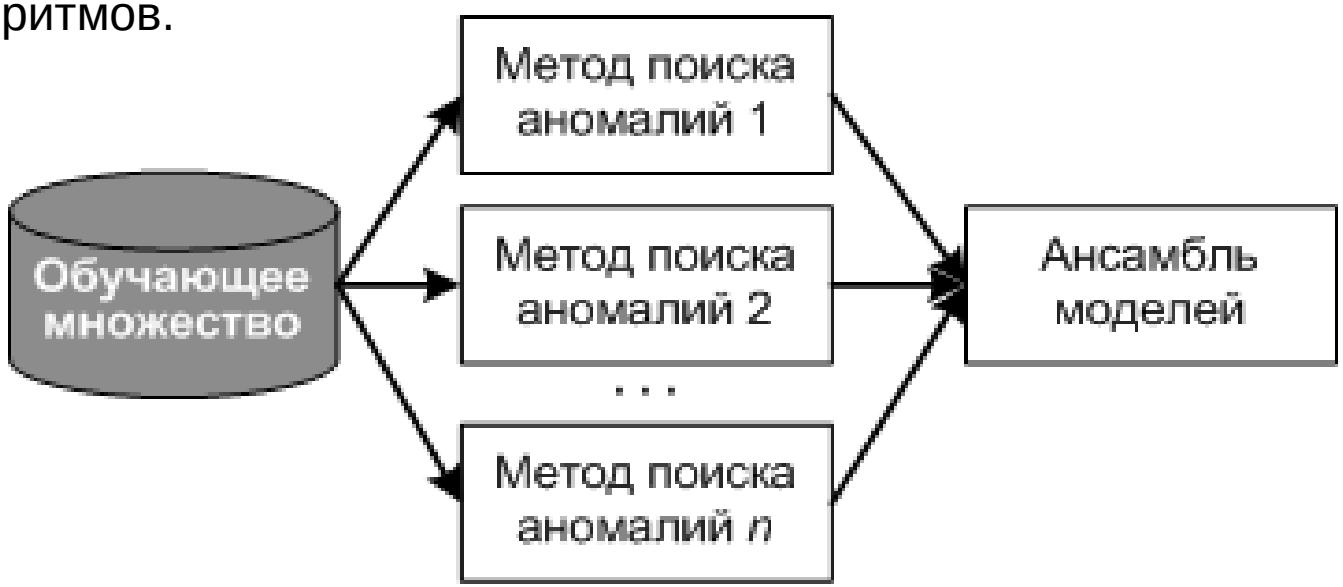
Методы создания ансамбля моделей (задача 2)

Тип ансамбля	Описание и примеры
Комитеты (голосование) / усреднение	Построение независимых алгоритмов и их усреднение / голосование.
Кодировки / перекодировки ответов	Специальные кодировки целевых значений и сведение решения задачи к решению нескольких задач.
Стекинг (stacking)	Построение мета признаков – ответов базовых алгоритмов на объектах выборки, обучение на них мета- алгоритма.
Бустинг (boosting)	Построение суммы нескольких алгоритмов. Каждое следующее слагаемое строится с учётом ошибок предыдущих.
Бэггинг (bootstrap aggregating)	Каждый базовый алгоритм обучается на случайном подмножестве обучающей выборки – бутстрэп выборки

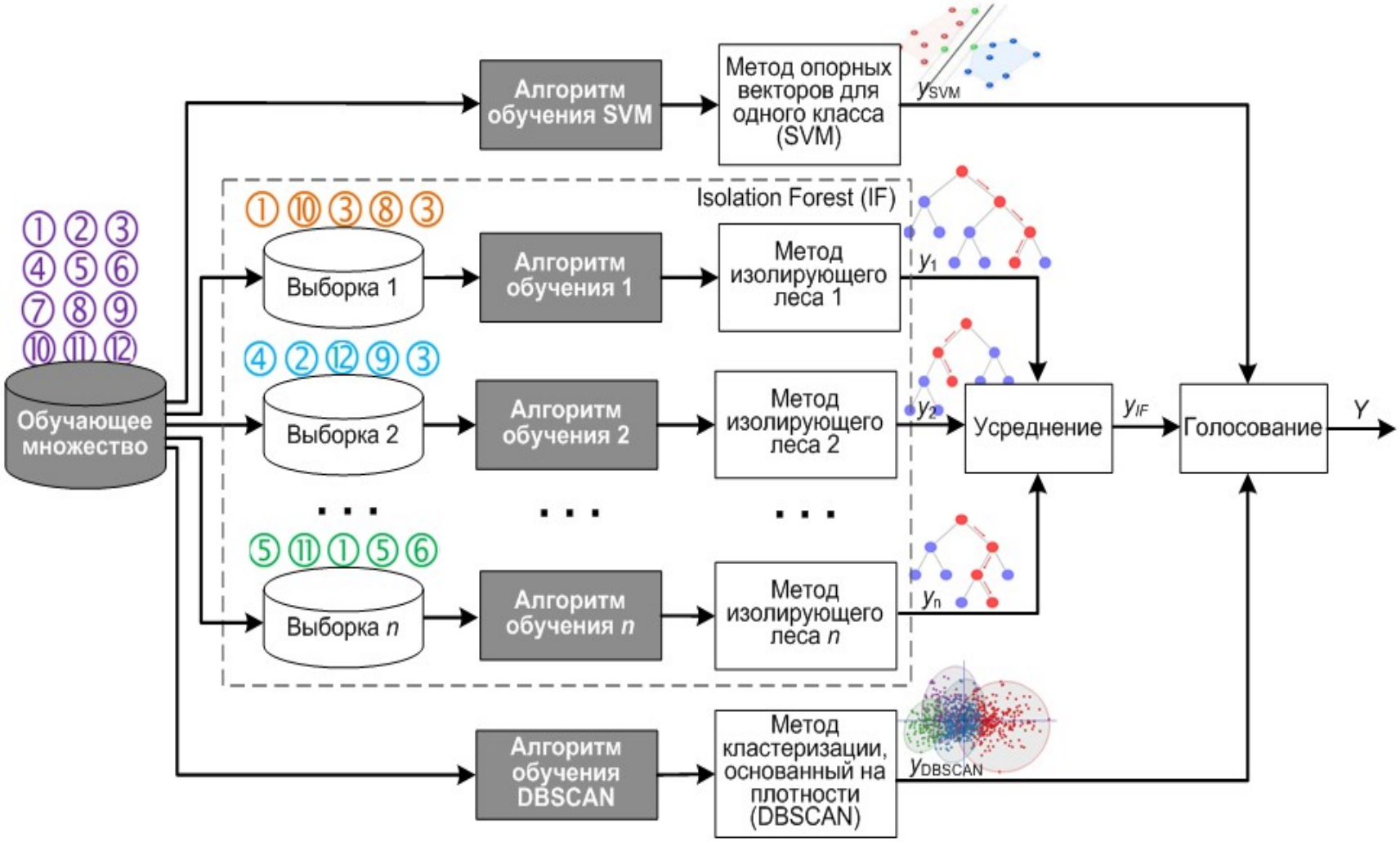
Комитеты(голосование)/усреднение

Комитеты (голосование) / усреднения имеют преимущества перед другими методами ансамблирования:

- *Статистическое* – применение ансамбля моделей усредняет ошибку каждой отдельной модели.
- *Вычислительное* – параллельно обучаются несколько базовых алгоритмов.

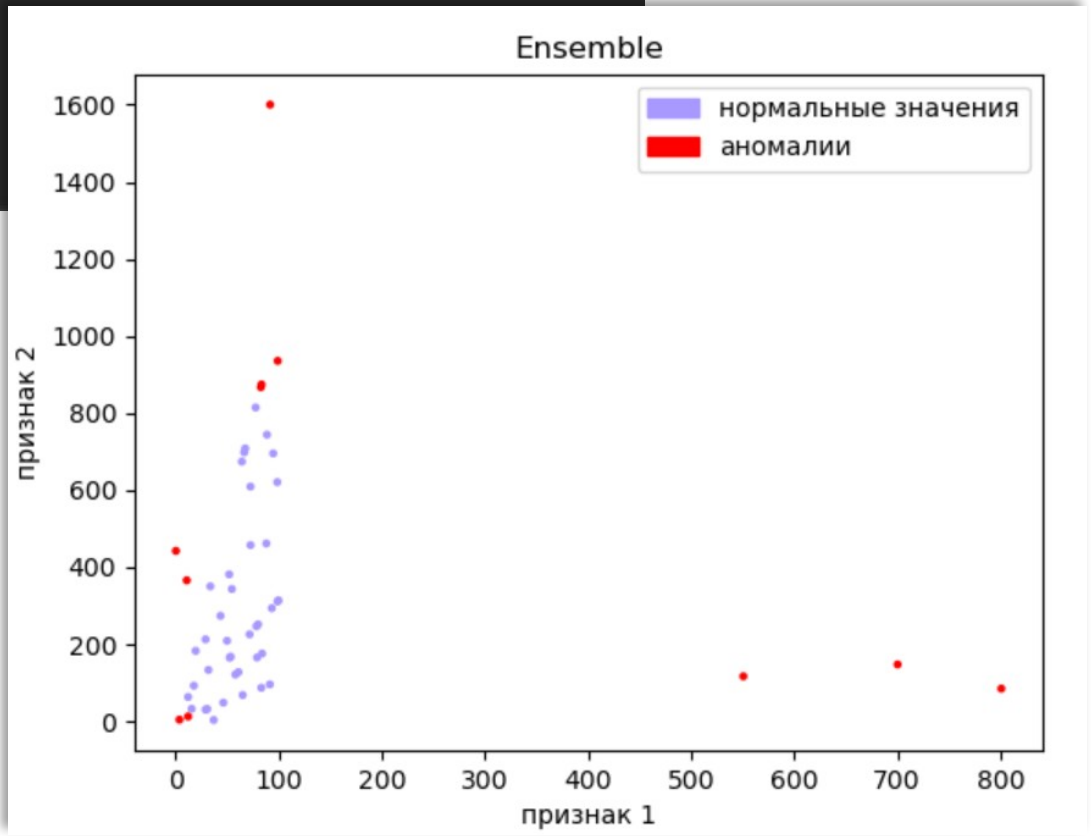


Разработка алгоритма на основе ансамбля выбранных алгоритмов (задача 3)



Консольное приложение для демонстрации результатов (задача 4)

```
Ensemble:
[ 1 -1  1  1  1  1  1  1 -1  1  1  1  1 -1  1 -1  1 -1  1  1  1  1  1 -1
  1 -1  1  1  1 -1  1 -1  1  1  1  1  1  1  1  1  1  1  1  1 -1  1  1  1
 -1  1  1]
error type 1:  2
error type 2:  4
number of detected anomalies:  11
accuracy:  88.24%
```



Оценка точности полученных результатов (задача 5)

- точность определения аномалий

Метрика / Метод	Количество аномалий (TP)	Количество нормальных значений (TN)	Количество ошибок 1-го типа (FN)	Количество ошибок 2-го типа (FP)	Точность определения аномалий (a)
OneClassSVM	8	35	1	7	84.31%
IsolationForest	6	36	3	6	82.35%
DBSCAN	4	39	5	3	84.31%
Ensemble	7	38	2	4	88.24%

Заключение

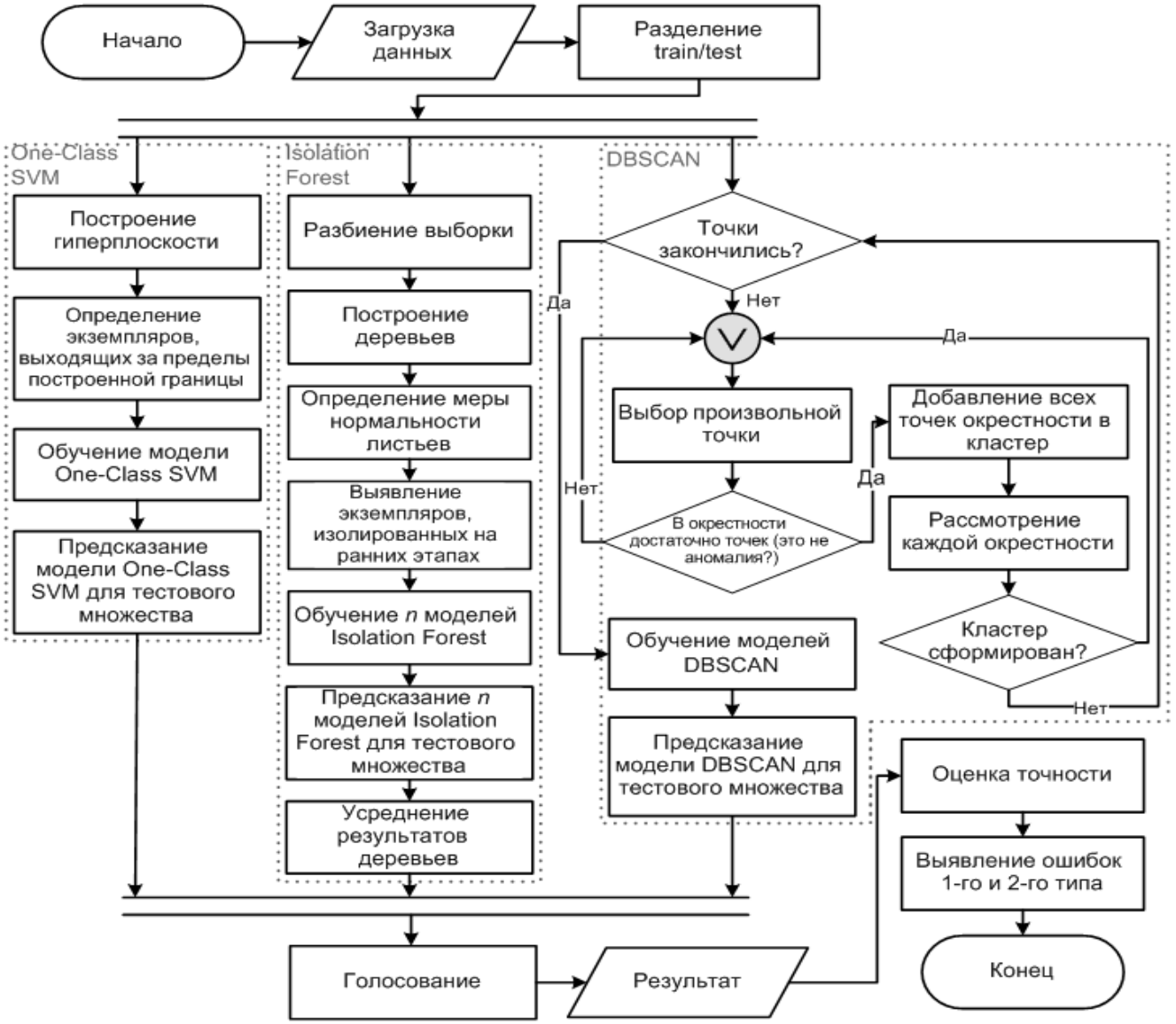
- Проделанный обзор методов показал необходимость разработки новых вариантов алгоритма определения аномалий в данных, так как для разных наборов данных применяется индивидуальный подход.
- В процессе разработки алгоритма был выбран комитет (голосование), в следствии чего три метода были объединены в ансамбль.
- Для демонстрации работы алгоритма было создано консольное приложение, содержащее меню, обеспечивающее взаимодействие пользователя и программы.
- Экспериментальное исследование точности определения аномалий показало, что точность определения аномалий ансамблем методов выше, чем точность методов по отдельности.

Дальнейшие направления исследований включают в себя повышение точности определения аномалий, использование различных видов ансамблей методов.

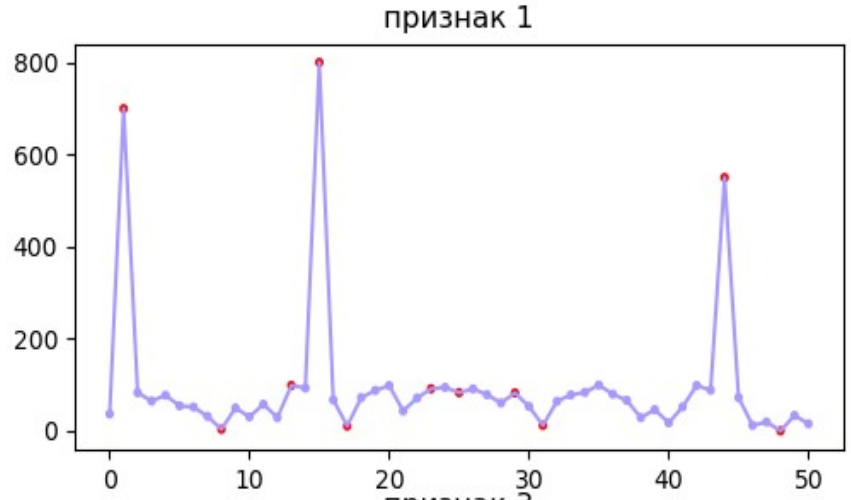
Апробация работы

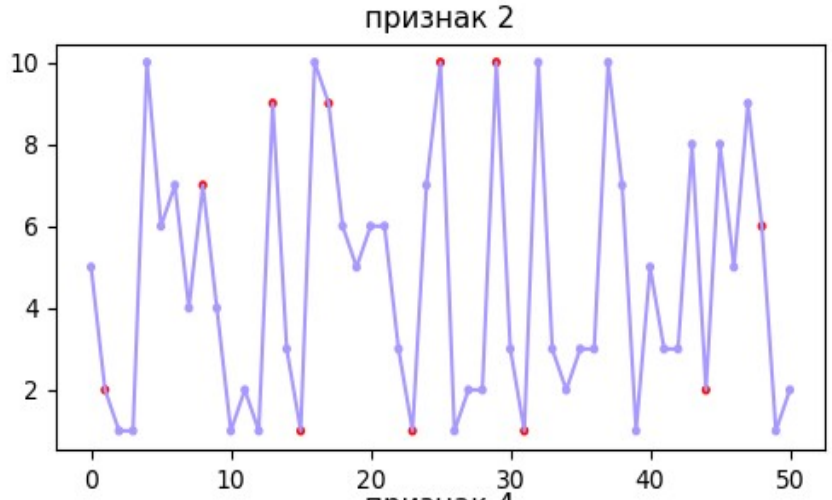
- Ханова Ю.А. Разработка процедуры бэггинга для определения аномалий в данных // IX Научно-практическая конференция с международным участием «Наука настоящего и будущего» для студентов, аспирантов и молодых ученых.
- Репозиторий проекта https://github.com/yulyakhanovagit/anomalies_in_the_data

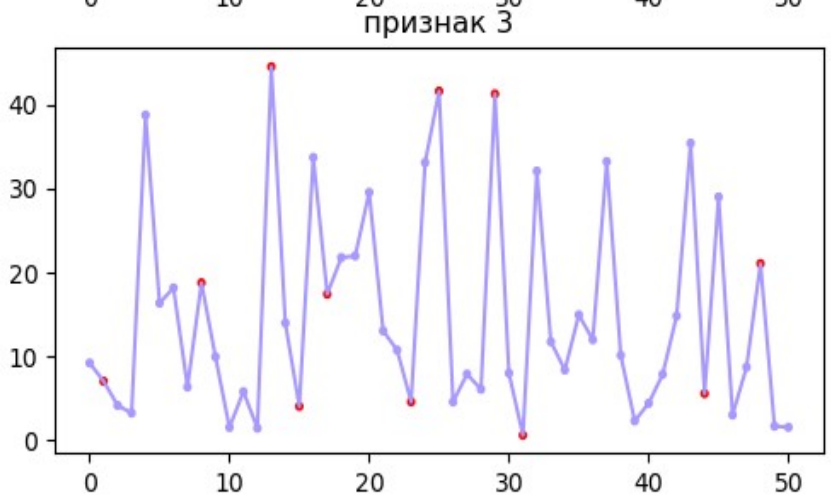
Схема работы алгоритма

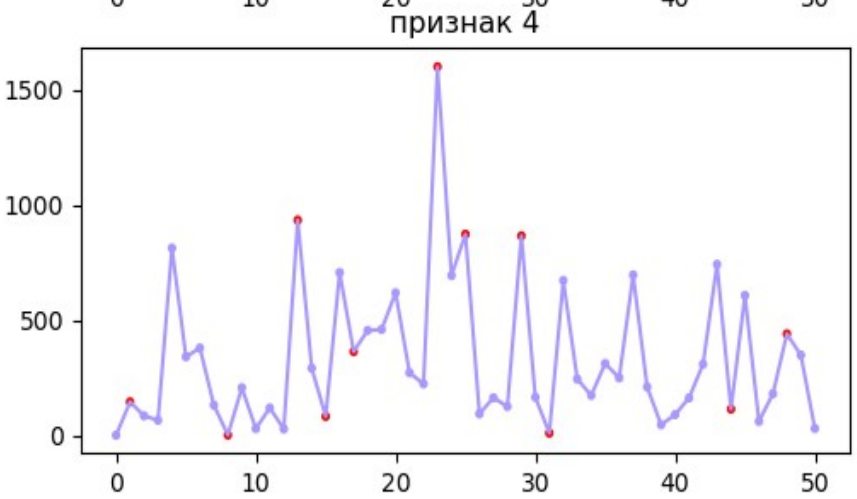


Графики отдельных признаков с отмеченными аномалиями









Меню консольного приложения

```
1 - OneClassSVM - метод опорных векторов для одного класса
2 - IsolationForest - метод изолирующего леса
3 - DBSCAN - пространственная кластеризация, основанная на плотности
4 - Ансамбль моделей
5 - Тестовые данные
0 - Выйти из программы
Выберите алгоритм выявления аномалий:
```

```
Выберите алгоритм выявления аномалий: 1
OneClassSVM:
[-1 -1  1  1  1  1  1  1 -1  1 -1  1 -1 -1  1 -1  1 -1  1  1  1  1  1 -1
  1 -1  1  1  1 -1  1 -1  1  1  1  1  1  1  1  1  1  1  1  1 -1  1  1  1
 -1  1 -1]
number of detected anomalies(OCSVM):  15
error type 1:  1
error type 2:  7
accuracy:  84.31%

1 - Вывести график полученных аномалий
2 - Вывести графики по каждому признаку
3 - Вывести матрицу путаницы
0 - Выйти к предыдущему выбору
Выберите:
```

Обозначения в матрице путаницы

- True Positive (TP): Определено как аномалия и фактически является аномалией.
- False Negative (FN): Определено как аномалия, но фактически является нормальным значением. Будем называть это ошибкой 1 типа.
- True Negative (TN): Определено как нормальное значение и фактически является нормальным значением.
- False Positive (FP): Определено как нормальное значение, но фактически является нормальным значением. Будем называть это ошибкой 2 типа.

Точность: