федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
Институт компьютерных наук и технологий

Высшая школа интеллектуальных
систем и суперкомпьютерных технологий

Выпускная квалификационная работа магистра

Применение генетического алгоритма для задачи понижения размерности пространства признаков

Выполнил студент гр. в3540203/80278
Георгиевский А. А.
Руководитель к.ф.-м.н., доцент ВШИСиСТ
Пак В. Г.

г. Санкт-Петербург
2021 год

Актуальность исследований

Актуальность исследований обусловлена необходимостью применения на практике интеллектуальных методов (подходов), позволяющих выбрать наиболее релевантные комбинации признаков с сохранением их семантики для конкретных моделей обучения и решения поставленных задач.

Цель и задачи исследования

Цель: исследование эффективности применения генетического алгоритма (ГА) и некоторых его модификаций для задачи понижения размерности пространства признаков.

Задачи:

- Реализовать исследуемый (выбранный) метод понижения размерности пространства признаков с применением ГА;
- Реализовать ГА с стандартными и модифицированными операторами;
- Выполнить оценку влияния модификации операторов ГА на эффективность понижения размерности пространства признаков;
- Выполнить сравнительный анализ исследуемого и классических методов понижения размерности пространства признаков.

Методы понижения размерности пространства признаков

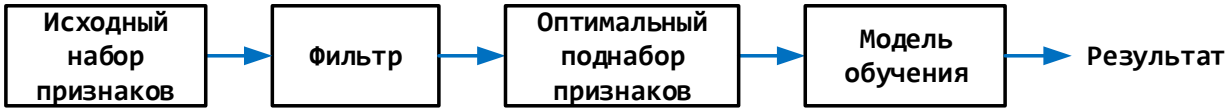
Подходы к снижению размерности

- ***Отбор признаков (feature selection)*** – это процесс выбора наиболее значимых признаков из исходного набора данных с сохранением их семантики.

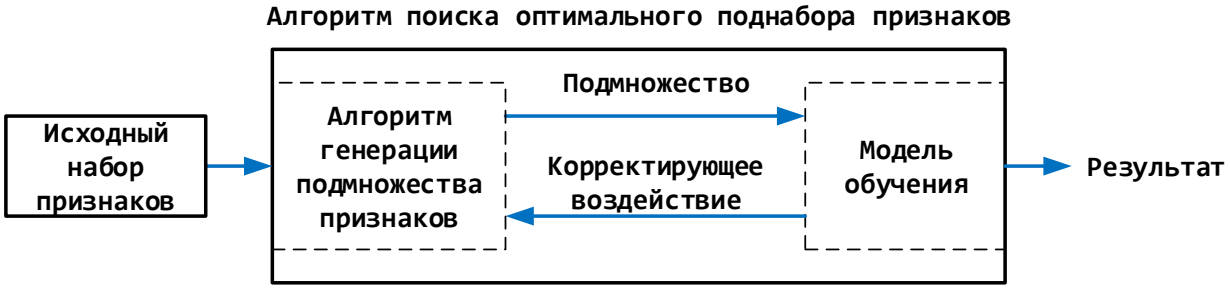
- ***Выделение признаков (feature extraction)*** – это процесс снижения размерности пространства признаков посредством специальных методов для перехода в новое пространство с иным семантическим наполнением.

Отбор признаков

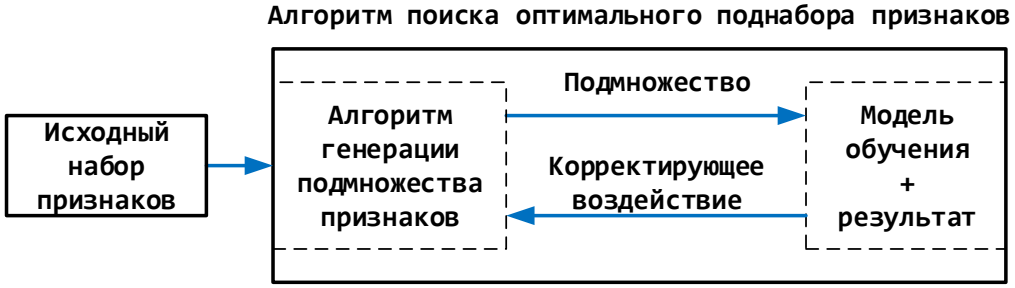
- ***Метод фильтров:***



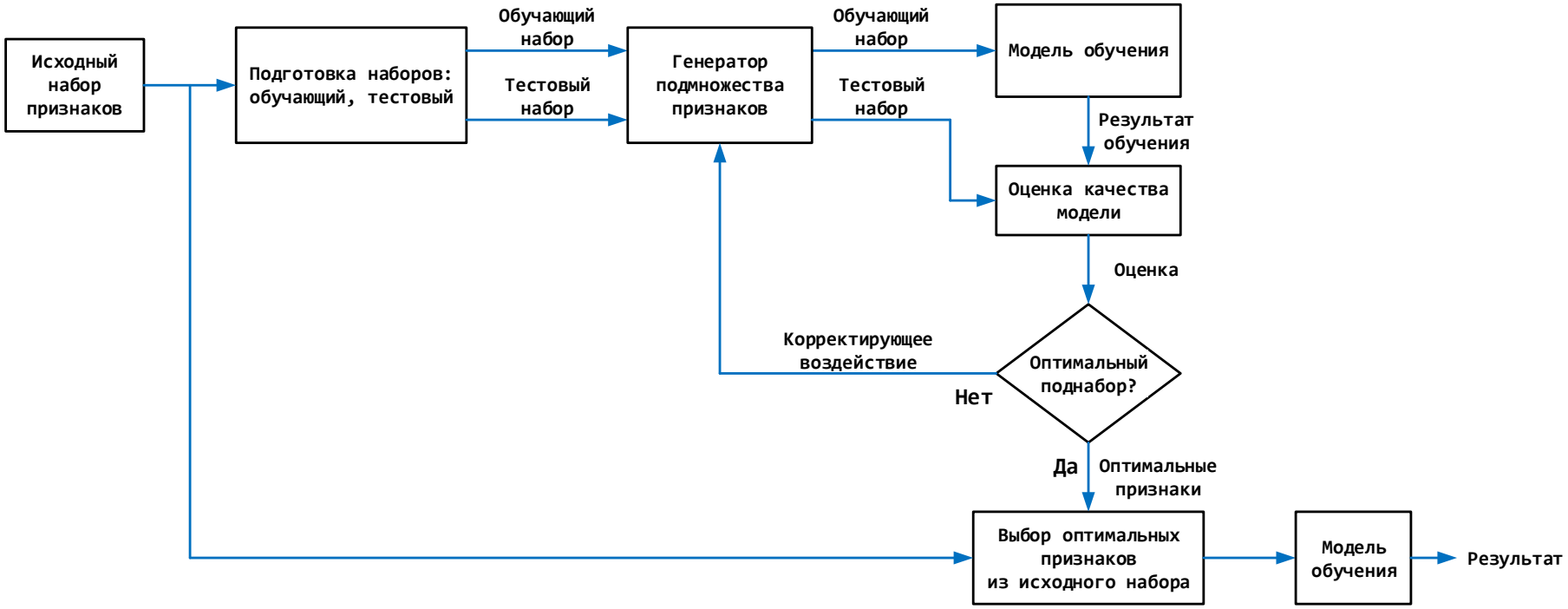
- ***Метод оболочек (оберток):***



- ***Встроенные методы:***

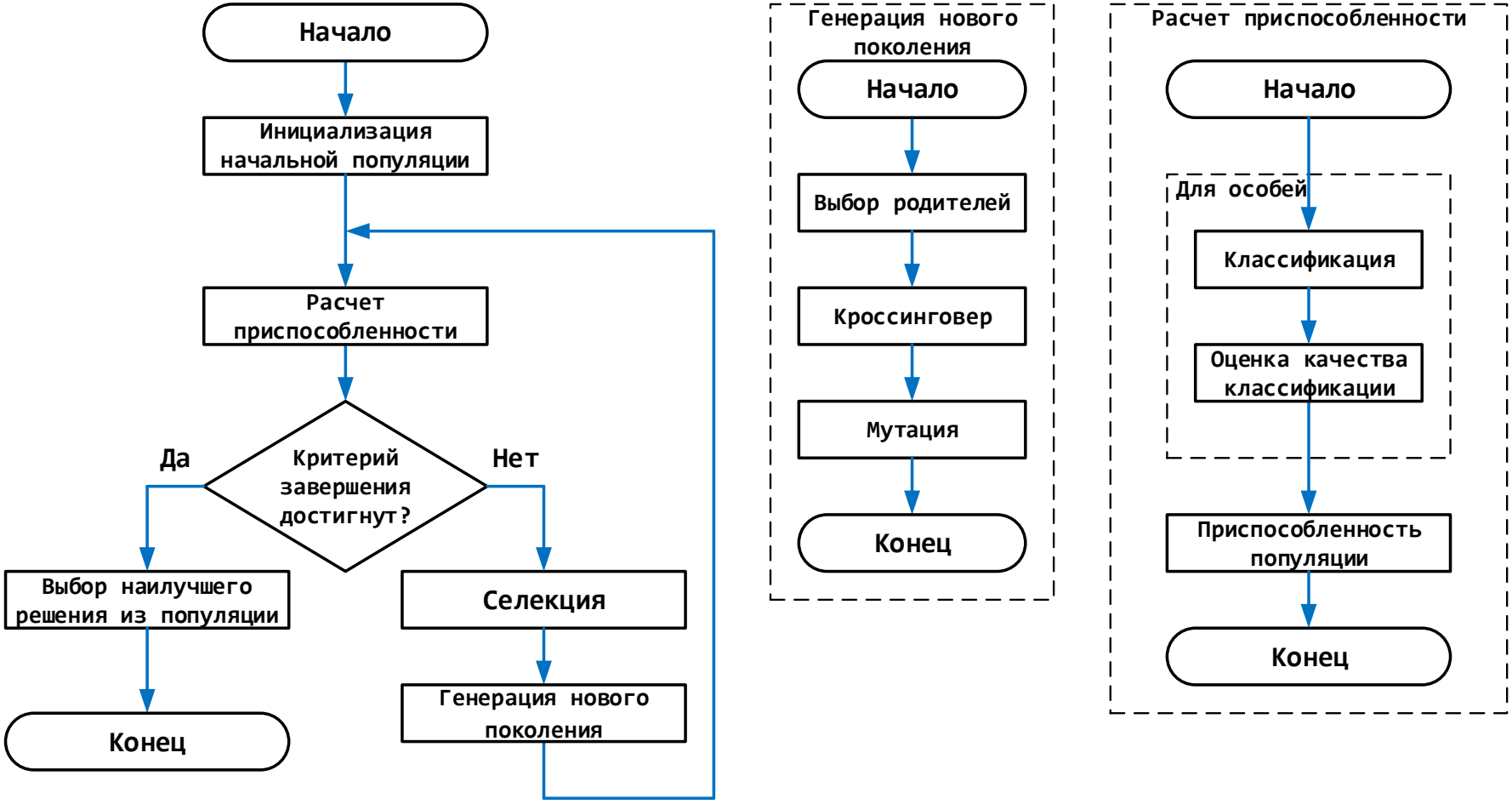


Применение генетического алгоритма в методе отбора признаков



а) Принципиальная схема отбора признаков, метод оболочек (оберток)

Применение генетического алгоритма в методе отбора признаков



b) Отбор признаков с точки зрения генетического алгоритма

Критерий оценки качества классификации в методе отбора признаков

Матрица несоответствия
(для бинарной классификации)

		Экспертная оценка	
		Класс C1	Класс C2
Оценка модели	Предсказанный класс C1	TP	FP
	Предсказанный класс C2	FN	TN

- TP (true positive) – классификатор верно отнёс объект к классу C1;
- TN (true negative) – классификатор верно отнес объект к классу C2;
- FP (falsee positive) – классификатор неверно отнёс объект к классу C1;
- FN (false negative) – классификатор неверно отнёс объект к классу C2.

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Реализация

Операторы ГА (основные):

- ***Инициализация*** – первичная популяция создается случайным образом;
- ***Приспособленность*** – классификатор kNN, оценка качества классификации F1 мера;
- ***Селекция*** – в живых остается 50 % наилучших особей из популяции;
- ***Генератор нового поколения*** – опциональный принцип элитизма;
- ***Выбор родителей*** – панмиксия (выбор родителей случайным образом);
- ***Кроссинговер*** – многоточечный, равновероятностная передача генов / передача генов с учетом значения функции приспособленности особи;
- ***Мутация*** – статическая / динамическая мутации.

Критерий завершения работы метода отбора признаков:

- MAX / Average значение приспособленности особей в популяции.

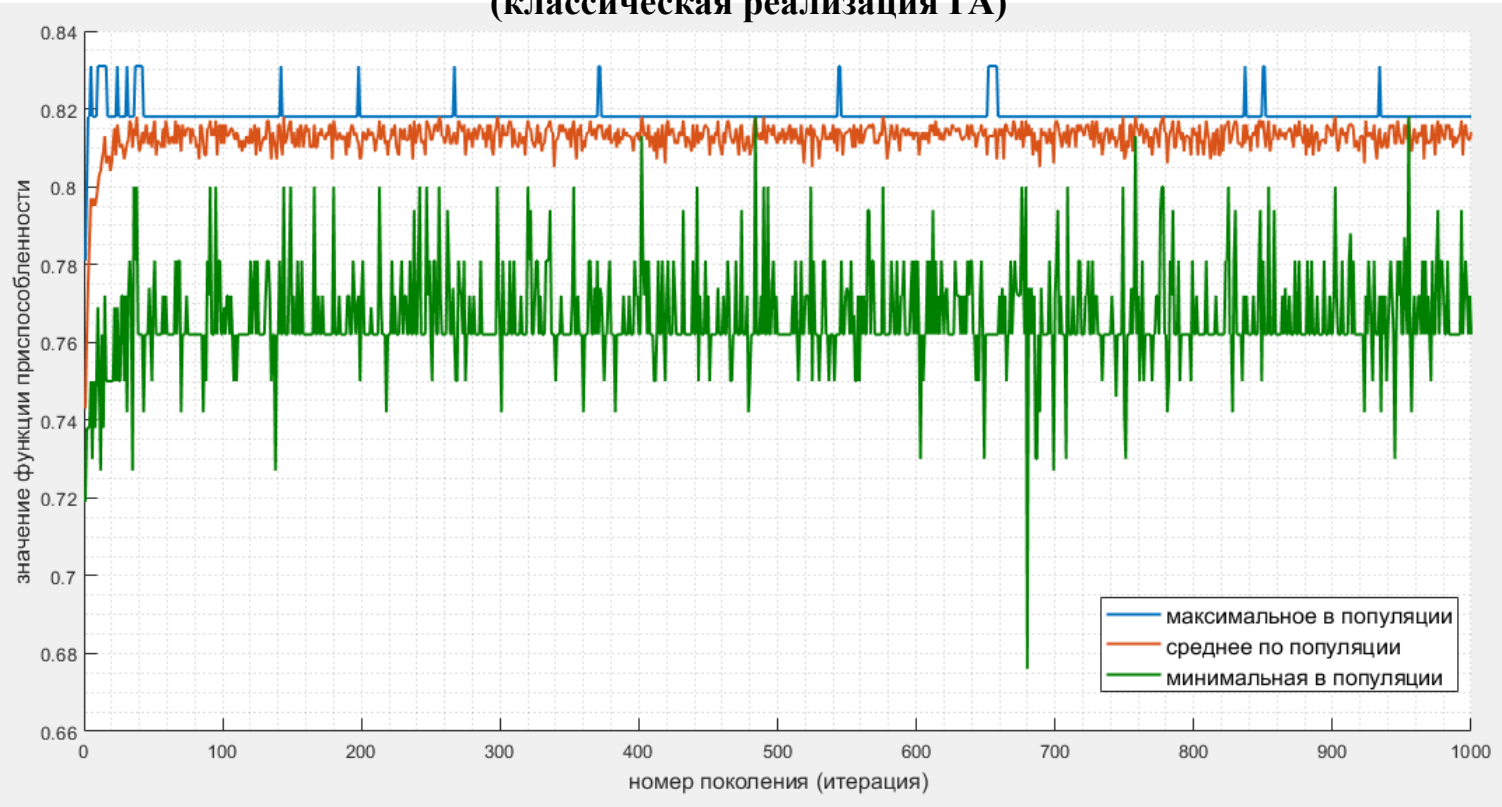
Программная реализация:

- ***Генетический алгоритм:*** .Net Core, R.Net;
- ***Модель обучения, классификация данных:*** R.

Экспериментальные исследования

Оценка влияния модификации операторов генетического алгоритма

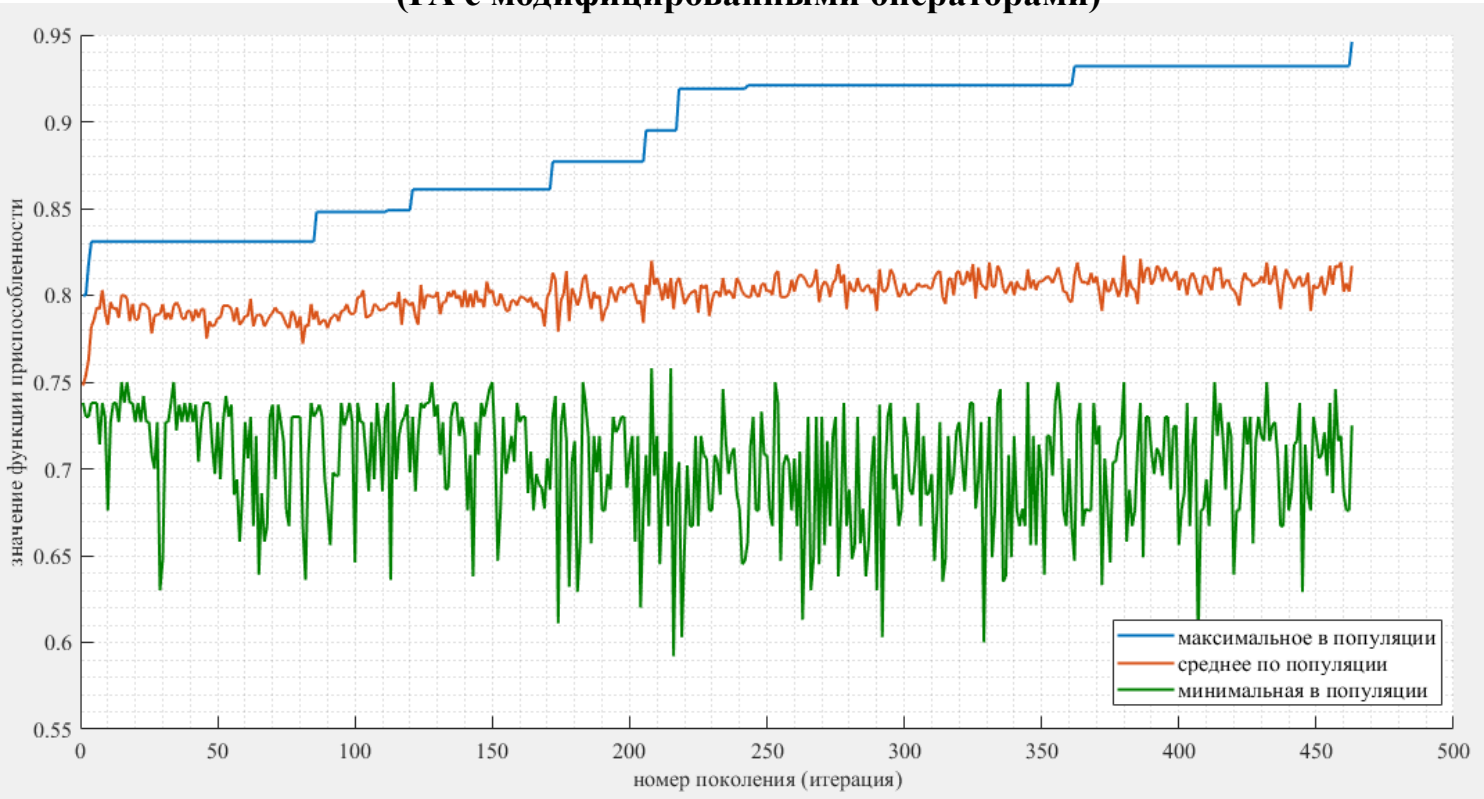
Зависимость значения функции приспособленности в популяции от номера поколения (классическая реализация ГА)



- **число параметров в исходном наборе – 37;**
- **точность классификации (F1) на исходном наборе данных, kNN – 67,70%;**
- **число параметров, отобранных алгоритмом – 9;**
- **точность классификации (F1) на отобранных параметрах, kNN – 81,80%.**

Оценка влияния модификации операторов генетического алгоритма

Зависимость значения функции приспособленности в популяции от номера поколения (ГА с модифицированными операторами)



- **число параметров в исходном наборе – 37;**
- **точность классификации (F1) на исходном наборе данных, kNN – 67,70%;**
- **число параметров, отобранных алгоритмом – 13;**
- **точность классификации (F1) на отобранных параметрах, kNN – 94,59%.**

Сравнительный анализ методов понижения размерности пространства признаков

Методы для сравнения:

- PCA; random forest; filter(information gain).

Наборы данных для бинарной классификации:

- диагностика неисправности ультразвуковых расходомеров жидкости «ultrasonic flowmeter diagnostics» (число параметров –37, число выборок –86);

- предсказание биологического ответа молекул, по их химическому свойству «bioresponse» (число параметров –1777, число выборок –3751).

Классификаторы для итоговой оценки эффективности:

- naive Bayes, kNN, tree, SVM, random forest.

Способ тестирования:

- кросс-валидация с разбиением набора на 20 частей.

Сравнительный анализ методов понижения размерности пространства признаков

Набор данных: «ultrasonic flowmeter diagnostics».

Параметр		Исходный набор данных	Методы снижения размерности			
			оболочка (ГА + kNN)	PCA	random forest	фильтр (information gain)
Число параметров в наборе		37	13	13	13	13
AUC, %	kNN	82,70	89,20	**93,30**	**75,90**	76,50
	Tree	80,40	84,60	**78,50**	**90,20**	88,60
	SVM	**98,30**	92,10	94,30	96,20	**91,90**
	random forest	**95,80**	93,80	**89,40**	94,70	93,90
	naive Bayes	82,50	**78,70**	**89,60**	83,50	84,80
Лучший результат точности классификации в строке						
Худший результат точности классификации в строке						

Для исследуемого метода отбора признаков (оболочка) :

- число параметров в исходном наборе – 37;
- точность классификации (AUC) на исходном наборе данных, kNN – 82,70%;
- число параметров, отобранных алгоритмом – 13;
- точность классификации (AUC) на отобранных параметрах, kNN – 89,20%;
- процент сжатия параметров в наборе – 64,9%.

Сравнительный анализ методов понижения размерности пространства признаков

Набор данных: «bioresponse».

* Для методов оболочка и random forest проводилась предварительная фильтрация (information gain), исходный набор сокращен до 500 параметров.

Параметр		Исходный набор данных	Методы снижения размерности			
			оболочка (ГА + kNN)*	PCA	random forest*	фильтр (information gain)
Число параметров в наборе		1777	130	130	130	130
AUC, %	kNN	**72,60**	**83,10**	78,10	82,60	80,50
	Tree	64,40	**69,70**	**60,90**	63,50	67,60
	SVM	**70,00**	**79,10**	76,80	77,30	78,00
	random forest	76,80	**81,50**	**71,00**	81,00	80,70
	naive Bayes	70,90	**77,40**	**67,40**	73,00	76,10
Лучший результат точности классификации в строке						
Худший результат точности классификации в строке						

Для исследуемого метода отбора признаков (оболочка) :

- число параметров в исходном наборе – 1777;
- точность классификации (AUC) на исходном наборе данных, kNN – 72,60%;
- число параметров, отобранных алгоритмом – 130;
- точность классификации (AUC) на отобранных параметрах, kNN – 83,10%;
- процент сжатия параметров в наборе – 92,7%.

Заключение

В ходе выполнения работы были решены поставленные задачи:

- ✓ Выполнен обзор существующих подходов к снижению размерности пространства признаков, и выбран метод для исследования.
- ✓ Реализован ГА с модифицированными операторами для исследуемого метода.
- ✓ Проведена оценка влияния модификации операторов ГА на эффективность понижения размерности пространства признаков.
- ✓ Выполнен сравнительный анализ исследуемого и часто используемых методов понижения размерности пространства признаков.
- ✓ Все исследования проводились с использованием реальных наборов данных из открытых источников.