Московский государственный университет имени М.В. Ломоносова

Факультет Вычислительной математики и кибернетики

Кафедра Математических методов прогнозирования

ДИПЛОМНАЯ РАБОТА

Сравнения моделей классификации основанных на множествах-прецедентах

Выполнил:

студент 517 группы

Антипов Алексей Алексеевич

Научный руководитель:

д.ф-м.н., профессор

Рязанов Владимир Васильевич

Москва, 2015

Оглавление

1 Введение

Двадцать первый век по праву называют веком информационных технологий. Рост компьютеризации и информатизации различных сфер общественной жизни влечет за собой взрывной рост объёмов и многообразия обрабатываемых данных, что создает необходимость развития научных областей и средств, отвечающих за анализ данных. Одной из основных таких областей является *машинное обучение*.

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения: *Обучение по прецедентам, оно* основано на выявлении общих закономерностей по частным эмпирическим данным, и *дедуктивное обучение*, которое предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Но дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.[1]

В данной работе рассматриваются задачи обучения по прецедентам. Эти задачи часто встречаются в самых разных областях человеческой деятельности.

Задача обучения по прецедентам состоит в следующем. Дано конечное множество прецедентов (объектов), по каждому из которых собраны (измерены) некоторые данные. Данные о прецеденте называют также его *описанием*. Совокупность всех имеющихся описаний прецедентов называется *обучающей выборкой*. Требуется по этим частным данным выявить общие зависимости, закономерности, взаимосвязи, присущие не только этой конкретной выборке, а всем прецедентам, в том числе тем, которые ещё не наблюдались.

Наиболее распространённым способом описания данных о выборке является признаковое описание отдельных объектов. Основная идея данной работы состоит в том, чтобы в качестве прецедентов рассматривать не отдельные объекты, а множества таких объектов. То есть группировать объекты определенным образом и описывать в качестве прецедентов уже не отдельный объект, а полученные группы объектов. Такой подход является более общим, так как в частном случае, если в качестве таких множеств взять множества, включающие в себя по одному объекту, мы вернемся к исходной задаче. Потенциально, новое описание может быть более информативным и удобным для конкретных задач, чем исходное.

В рамках данной работы в качестве множеств объектов рассматриваются множества, полученные на основе *логических закономерности классов*.

Логическая закономерность в задачах классификации — это легко интерпретируемое правило, выделяющее из обучающей выборки достаточно много объектов какого-то одного класса и практически не выделяющее объекты остальных классов.

В данной работе представлены несколько способов описания множеств-прецедентов, соответствующих логическим закономерностям. Реализованы программы (на языке *Matlab* и на языке *C++)*, которые преобразуют исходные выборки и набор логических закономерностей (найденный с помощью системы интеллектуального анализа данных, распознавания и прогноза *Recognition 2.0*) в новое представление пригодное для дальнейшего использования в данной системе. Также проведен ряд экспериментов по сравнению новых и исходных описаний.

2 Алгоритмы классификации по множествам-прецедентам

2.1 Формализация задачи распознавания по множествам-прецедентам

Рассмотрим математическую постановку задачи распознавания по прецедентам.

Пусть $\boldsymbol{X}$ — множество описаний объектов, $\boldsymbol{Y}$ — множество допустимых ответов. Существует неизвестная целевая зависимость — отображение $y^*: \boldsymbol{X} \to \boldsymbol{Y}$, значения которой известны только на объектах конечной обучающей выборки $\boldsymbol{X}^n_{train} = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Требуется построить алгоритм $a: \boldsymbol{X} \to \boldsymbol{Y}$, который приближал бы неизвестную целевую зависимость, как на элементах выборки, так и на всём множестве $\boldsymbol{X}$. Пусть $\boldsymbol{X}^m_{test} = \{x_1, \dots, x_m\}$ тестовая выборка [1]. Тогда, чтобы решить задачу распознавания тестовой выборки, достаточно применить к ней построенный по обучающей выборке алгоритм a.

Теперь введем обозначения для задачи распознавания по множествам-прецедентам.

Пусть теперь $\widetilde{\boldsymbol{X}}$ — новое множество описаний объектов, объектами которого являются подмножества $\boldsymbol{X}$. Множество допустимых ответов – $\boldsymbol{Y}$ (остается неизменным). Аналогично, известна обучающая выборка $\widetilde{\boldsymbol{X}}^n_{train} = \{(\tilde{x}_1, y_1), \dots, (\tilde{x}_n, y_n)\}$. Требуется построить алгоритм $\tilde{a}: \widetilde{\boldsymbol{X}} \to \boldsymbol{Y}$, приближающий неизвестную целевую зависимость, как на элементах выборки, так и на всём множестве $\widetilde{\boldsymbol{X}}$

Чтобы перейти от задачи распознавания по прецедентам к задаче распознавания по множествам-прецедентам необходимо задать два отображения:

1. $x_{train}: \boldsymbol{X} \times \boldsymbol{Y} \to \widetilde{\boldsymbol{X}} \times \boldsymbol{Y}$
2. $x_{test}: \boldsymbol{X} \to \widetilde{\boldsymbol{X}}$

Тогда, чтобы решить задачу распознавания по прецедентам с помощью перехода к задаче распознавания по множествам-прецедентам, наложим на отображение x_{test} условие однозначности. Алгоритм решения задачи распознавания по прецедентам с помощью

перехода к задаче распознаванию по множествам-прецедентам будет иметь следующий вид:

Вход: $\boldsymbol{X}_{train}^{n} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ – обучающая выборка

$\boldsymbol{X}_{test}^{m} = \{x_1, \dots, x_m\}$ распознаваемая выборка.

Параметры: отображения x_{train}, x_{test}

Выход: $\boldsymbol{Y}^{m} = \{y_1, \dots, y_m\}$ вектор ответов для распознаваемой выборки

1. С помощью отображения x_{train}: $\boldsymbol{X} \times \boldsymbol{Y} \rightarrow \widetilde{\boldsymbol{X}} \times \boldsymbol{Y}$**,** получаем из обучающей выборки $\boldsymbol{X}_{train}^{n} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, новую обучающую выборку $\widetilde{\boldsymbol{X}}_{train}^{l} = \{(\tilde{x}_1, y_1), \dots, (\tilde{x}_l, y_l)\}$
2. По $\widetilde{\boldsymbol{X}}_{train}^{l} = \{(\tilde{x}_1, y_1), \dots, (\tilde{x}_l, y_l)\}$ строим алгоритм $\tilde{a}$: $\widetilde{\boldsymbol{X}} \rightarrow \boldsymbol{Y}$**,** приближающий неизвестную целевую зависимость, как на элементах выборки, так и на всём множестве $\widetilde{\boldsymbol{X}}$
3. Получаем распознаваемую выборку для пространства $\widetilde{\boldsymbol{X}}$**:** $\widetilde{\boldsymbol{X}}_{test}^{m} = x_{test}(\boldsymbol{X}_{test}^{m})$ и запоминаем соответствия между объектами $\boldsymbol{X}_{test}^{m}$ и полученными $\widetilde{\boldsymbol{X}}_{test}^{m}$
4. Получаем вектор ответов для распознаваемой выборки множеств-прецедентов $\widetilde{\boldsymbol{Y}}^{m} = \tilde{a}(\widetilde{\boldsymbol{X}}_{test}^{m})$
5. Используя соответствие, полученное на *шаге 3* из вектора ответов $\widetilde{\boldsymbol{Y}}^{m}$**,** получаем вектор искомых ответов $\boldsymbol{Y}^{m}$

Предположение состоит в том, что в пространстве $\widetilde{\boldsymbol{X}}$**,** при определенном выборе отображений x_{train}, x_{test} объекты из пространства $\boldsymbol{X}$ могут классифицироваться лучше, чем в самом пространстве $\boldsymbol{X}$.

Прежде чем мы рассмотрим частный случай перехода от задачи распознавания по прецедентам к задаче распознавания по множествам-прецедентам, вспомним, что такое *логические закономерности классов* для задачи распознавания по прецедентам.

2.2 Множества-прецеденты соответствующие логическим закономерностям

Пусть $y \in \boldsymbol{Y}$ — фиксированнный класс. Объекты этого класса будем называть положительными; объекты остальных классов — отрицательными.

Под логической закономерностью класса $y \in \boldsymbol{Y}$, понимается предикат вида:

$$\varphi(x) = (a_1 \leq x_1 \leq b_1) \cap (a_2 \leq x_2 \leq b_2) \cap \dots \cap (a_k \leq x_k \leq b_k) \in \{0,1\}$$

Где $x_i -$ i-ая компонента вектора x, $k -$ количество признаков.

Такой, что:

1) Существует хотя бы один объект обучающей выборки $(\bar{x}, \bar{y}) \in \boldsymbol{X}^n_{train}$, для которого $\varphi(\bar{x}) = 1$ и $\bar{y} = y$
2) Для любого объекта обучающей выборки $(x', y') \in \boldsymbol{X}^n_{train}$ другого класса $y' \neq y$, $\varphi(x') = 0$
3) $\varphi(x)$ доставляет экстремум для некоторого критерия качества предикатов $\Phi(\varphi)$

В качестве критерия качества предикатов, обычно, используется стандартный критерий качества логических закономерностей:

$$\Phi(\varphi) = \frac{\sum_{i=1}^{n} \varphi(x_i)}{\sum_{i=1}^{n} [y_i = y]},$$

В силу многоэкстремальности задачи оптимизации $\Phi(\varphi)$, логическими закономерностями класса считаются все предикаты $\varphi(x)$, доставляющие локальный экстремум критерию $\Phi(\varphi)$. [5]

Вообще говоря, поиск логических закономерностей задача – трудная задача, которая выходит за рамки данной работы. Но так как она непосредственно связана с дальнейшими исследованиями, опишем, в чем заключается идея алгоритма поиска логических закономерностей классов.

Алгоритм состоит в решении последовательности однотипных задач. Опишем подобную задачу для фиксированного класса $y \in \boldsymbol{Y}$.

Пусть $(\bar{x}, \bar{y}) \in \boldsymbol{X}^n_{train}$ случайно выбранный обучающий объект класса $\bar{y} = y$ (будем называть его “опорным”). Поиск оптимального предиката $\Phi(\varphi)$ (т.е. значений параметров $a_1, a_2, \dots, a_k, b_1, b_2, \dots, b_k$) для опорного объекта $\bar{x}$, удовлетворяющего условию $\varphi(\bar{x}) = 1$, осуществляется сначала на некоторой неравномерной сетке пространства $\mathbb{R}^k$. После нахождения оптимального предиката $\varphi(x)$ на заданной сетке, происходит поиск оптимального предиката $\varphi(x)$ на более мелкой сетке, в окрестности ранее найденного $\varphi(x)$, и т.д. Задача поиска множества логических закономерностей, связанных с заданным опорным объектом считается решенной, если при переходе к более мелкой сетке не удается найти предикат $\varphi(x)$ с более высоким значением критерия качества $\Phi(\varphi)$. Задача поиска оптимального $\varphi(x)$ на каждой сетке состоит в поиске максимальной совместной подсистемы некоторой системы неравенств при линейных ограничениях относительно бинарных переменных и некоторого ее решения. Последняя задача сводится к решению аналогичной задачи относительно вещественных переменных. В конечном итоге задача поиска оптимального предиката $\varphi(x)$ для опорного объекта $\bar{x}$ заканчивается вычислением множества локально оптимальных предикатов $\varphi(x)$ со свойством $\varphi(x) = 1$, причем конъюнкции являются несократимыми (из них нельзя удалить какой-либо сомножитель). [6]

Все вычисления повторяются для m случайно выбранных “опорных” объектов класса y, а все найденные логические закономерности объединяются в одно множество логических закономерностей для данного класса.

Говорят, что предикат $\varphi: \boldsymbol{X} \to \{0,1\}$ покрывает объект x, если $\varphi(x) = 1$

Итак, любую логическую закономерность $\varphi(x)$ можно поставить в однозначное соответствие с подмножеством объектов, $M_\varphi \subset \boldsymbol{X}$ которых она покрывает.

Теперь рассмотрим частный случай перехода от задачи распознавания по прецедентам к задаче распознавания по множествам-прецедентам.

В качестве множеств-прецедентов, будем рассматривать логические закономерности классов.

- Логические закономерности, которые будут являться множествами-прецедентами для обучения, можно найти из первоначальной обучающей выборки описанным выше алгоритмом.
- Для объектов тестовой выборки мы можем построить их, как конъюнкции фиксированного размера задаваемого параметрами $\varepsilon_1, \dots \varepsilon_k$ с центром в тестовых объектах.

Теперь рассмотрим вопрос, как задать описание множеств-прецедентов $\widetilde{\boldsymbol{X}}$. Объектами теперь являются логические закономерности. В данной работе предложено 3 способа описания:

1. В форме бинарного вектора размерность, которого равна количеству объектов в обучающей выборке.
2. В форме многомерного параллелепипеда в первоначальном признаковом пространстве.
3. В виде центра массы логической закономерности

Распишем эти способы более подробно.

2.2.1 Первый способ описания логических закономерностей

Логическую закономерность $\varphi(x)$ можно представить в виде бинарного вектора следующего вида.

$$\varphi(x) = \left(\tilde{x}_\varphi^1, \tilde{x}_\varphi^2, \dots \tilde{x}_\varphi^n \right),$$

где n - число объектов обучающей выборки

$$\tilde{x}_\phi^i = \begin{cases} 1, & \varphi(x_i) = 1 \\ 0, & \text{иначе} \end{cases}, i = 1 \dots n$$

То есть, если объект обучающей выборки покрыт логической закономерностью, то в соответствующей координате стоит 1, в противном случае – 0.

Логические закономерности, соответствующие объектам обучающей выборки, можно найти методом, подробно описанным в [5], подробное описание которого выходит за рамки данной работы.

А каждому объекту тестовой выборки ставим в соответствие логическую закономерность (многомерные параллелепипеды) с центром в данном объекте тестовой выборки и расстоянием до границ, которое определяется при анализе логических закономерностей обучающей выборки (например, как среднее расстояние от центра до границ логических закономерностей в обучающей выборки). Хочется отметить, что полученное отображение x_{test} является взаимно-однозначным. Далее аналогично проверяем, какие объекты обучающей выборки покрывает данная логическая закономерность, и формируем бинарный вектор.

2.2.2 Второй способ описания логических закономерностей

Пусть логическая закономерность $\varphi(x)$ представлена в виде конъюнкции пороговых правил

$$\varphi(x) = \bigcap_{i=1}^{k} [a_i \leq x_i \leq b_i],$$

где k - число признаков исходной задачи.

Для однородности представлений, если какой-то признак или граница не используется, заменим ее несущественным ограничением (например, левую границу можно заменить наименьшим значением данного признака на обучающей выборке, а правую границу – наибольшим значением данного признака на обучающей выборке).

Тогда данной логической закономерности однозначно соответствует описание в виде вектора размерности $2 * k$, первые k размерностей, которого описывает центр данной логической закономерности, то есть

$$\tilde{x}_{\varphi}^{i} = \frac{a_i + b_i}{2}, i = 1 \dots k$$

А остальные k размерностей описывают размер данной логической закономерности (расстояние от центра логической закономерности до ее границ)

$$\tilde{x}_{\varphi}^{i+k} = \frac{b_i - a_i}{2}, i = 1 \dots k$$

Аналогично первому способу, логические закономерности, соответствующие объектам обучающей выборки, можно найти методом, подробно описанным в [5], подробное описание которого выходит за рамки данной работы.

А каждому объекту тестовой выборки ставим в соответствие логические закономерности (многомерные параллелепипеды) с центром в данном объекте тестовой выборки и расстоянием до границ, которое определяется при анализе логических закономерностей обучающей выборки (например, как среднее расстояние от центра до границ логических закономерностей в обучающей выборки). Хочется отметить, что полученное отображение x_{test} является взаимно-однозначным.

2.2.3 Третий способ описания логических закономерностей

Для обучающей выборки логическая закономерность описывается, как центр масс объектов, которые она покрывает.

$$\tilde{x}_{\varphi} = \frac{\sum_{i=1}^{n} \varphi(x_i) x_i}{\sum_{i=1}^{n} \varphi(x_i)},$$

Где $x_1, \dots, x_n$ – объекты обучающей выборки, $\varphi(x_i) = 1$, если объект x_i покрывается логической закономерностью, $\varphi(x_i) = 0$ в противном случае.

Объекты тестовой выборки остаются без изменений (отображение x_{test} в этом случае тоже является взаимно-однозначным).

2.3 Методы распознавания основанные на множествах-прецедентах.

Объекты в новых представлениях имеют вид числовых векторов, поэтому с ними можно работать с помощью стандартных методов машинного обучения. Но при этом часть первоначальной информации (например, о расположении объектов внутри логических закономерностей) может теряться. Поэтому в данной работе мы рассмотрим еще и пример того, как можно совместно использовать начальное описание объектов и новое описание множествами-прецедентами. Но для этого необходимо модифицировать стандартные алгоритмы классификации. В данной работе приведен пример такой модификации для алгоритма k-ближайших соседей.

3 Программные реализации алгоритмов распознавания по множествам-прецедентам

Общий алгоритм

Вход: Имя файла с обучающей выборкой; имя файла с результатами поиска логических закономерностей системой Recognition 2.0; имя файла с тестовой выборкой

Выход: Файлы обучающей и тестовой выборки в первом, втором и третьем способах представления (которые описаны выше)

- Считывание файлов с обучающей и тестовой выборками, парсинг файла с логическими закономерностями
- Предобработка:
  - Замена пропусков в данных средним значением признака (в обучающей выборке среднее значение берется по классу, в тестовой – по всем объектам).
  - Нахождение минимальных и максимальных значений признаков для каждого класса (необходимо, чтобы ограничить найденные логические закономерности для неиспользуемых признаков).
- Формирование новых способов представления обучающих выборок
- Формирование новых тестовых выборок
  - Для первого и второго способа представления размер логических закономерностей для тестовых объектов выбирается, как средний размер логической закономерности для обучающих объектов
  - Для третьего способа представления тестовая выборка остается без изменений
- Запись в файл обучающей и тестовой выборки в первом, втором и третьем способах представления

3.1 Реализация на MATLAB

При создании программы были реализованы следующие функции:

```
cellfile = freadTab(nameInputFile)
% функция считывает файл с расширением .TAB

fwriteTab(nameOutputFile, cellfile, par)
% функция записывает внутреннее представление данных в файл.

cellLogRule = freadLogRule(nameFile)
```

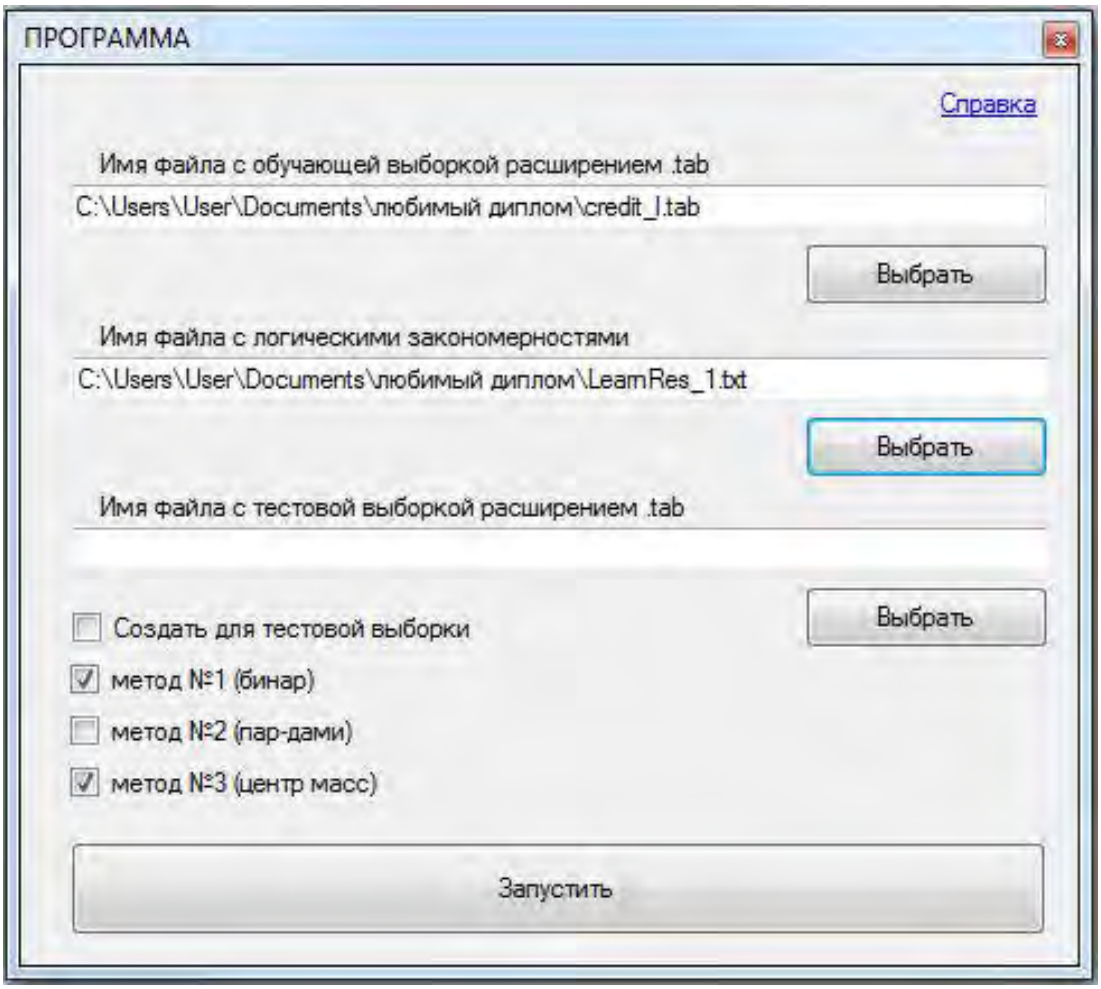
```
% функция выделяет структурированную информацию
% из текстового файла с именем nameFile полученного
% после применение Recognition 2.0 -> Логические закономерности

creatNewTable( … )
% основная функция, создающая новое представление обучающей
% и тестовой выборки описанными способами

commandFile.m
% скрипт, осуществляющий ввод параметров через управляющий файл
```

3.2 Реализация на C++

Также реализована программа на языке C++ с простым графическим интерфейсом для более удобного использования. Идейных отличий в реализации от языка MATLAB в программе нет.



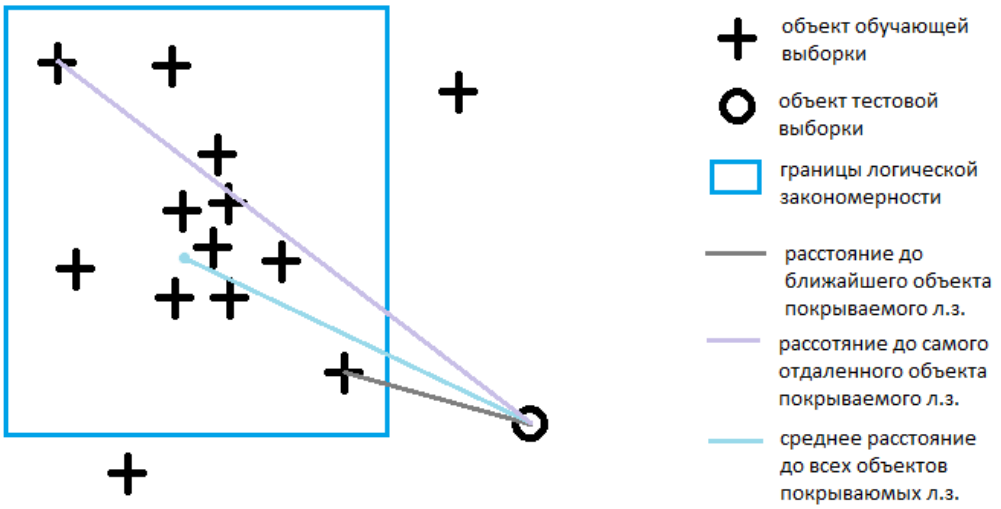
3.1 Реализация модифицированного алгоритма k-ближайших соседей

Как и в предыдущих способах в качестве объектов обучающей выборки мы рассматриваем логические закономерности.

- Находим k-ближайшие к тестовому объекту логические закономерности $\varphi_1, \ \dots, \varphi_k$ Пусть $y_1, \dots, \ y_k \ -$ классы найденных логических закономерностей
- Тестовый объект относим к классу, для которого сумма весов найденных k-ближайших логических закономерностей класса наибольшая. $a(x) = \text{argmax}_{u \in \boldsymbol{Y}} \sum_{i=1}^{k} w_i [y_i = u]$, где w_i –вес i-ой л.з.

Создается необходимость подсчета расстояние между объектом и логическими закономерностями. В данной реализации предлагаются три способа подсчета такого расстояния:

1) Расстоянием от тестового объекта до логической закономерности будем считать расстояние от тестового объекта до ближайшего к нему обучающего объекта покрываемого данной логической закономерностью
2) Расстоянием от тестового объекта до логической закономерности будем считать расстояние от тестового объекта до самого отдаленного от него обучающего объекта покрываемого данной логической закономерностью
3) Расстоянием от тестового объекта до логической закономерности будем считать среднее расстояние до всех обучающих объектов, покрываемых данной логической закономерностью

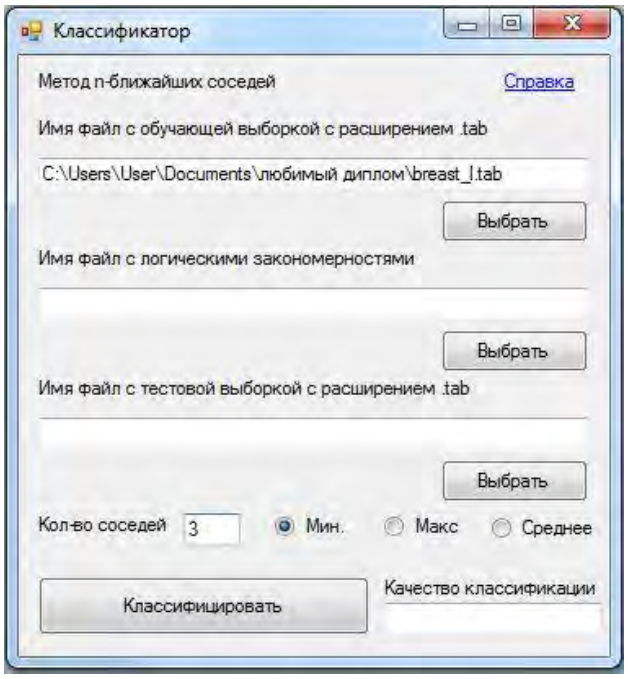


В таком подходе классификации можно учитывать веса логических закономерностей.

Пусть φ – логическая закономерность, $x_1, \dots, x_n$ – объекты обучающей выборки. Тогда вес логической закономерности

$$w_\varphi = \frac{\sum_{i=1}^{n} \varphi(x_i)}{n}$$

Для данного алгоритма также реализована программа на языке C++ с простым графическим интерфейсом для более удобного использования.



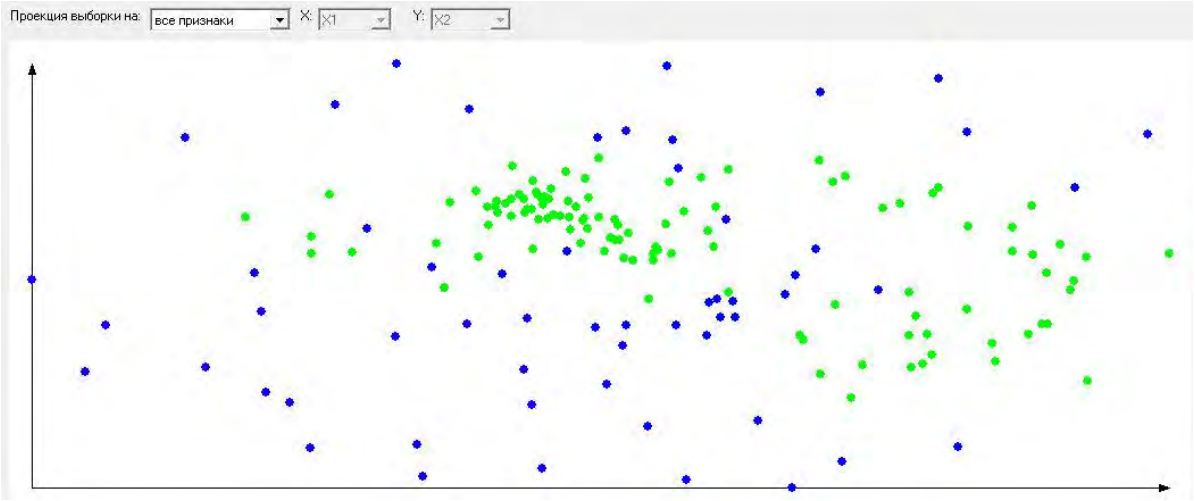
4 Экспериментальные исследования на прикладных и модельных задачах

4.1 Сравнение моделей классификации основанных на множествах-прецедентах описанных первым способом и моделей классификации основанных на первоначальном представлении данных

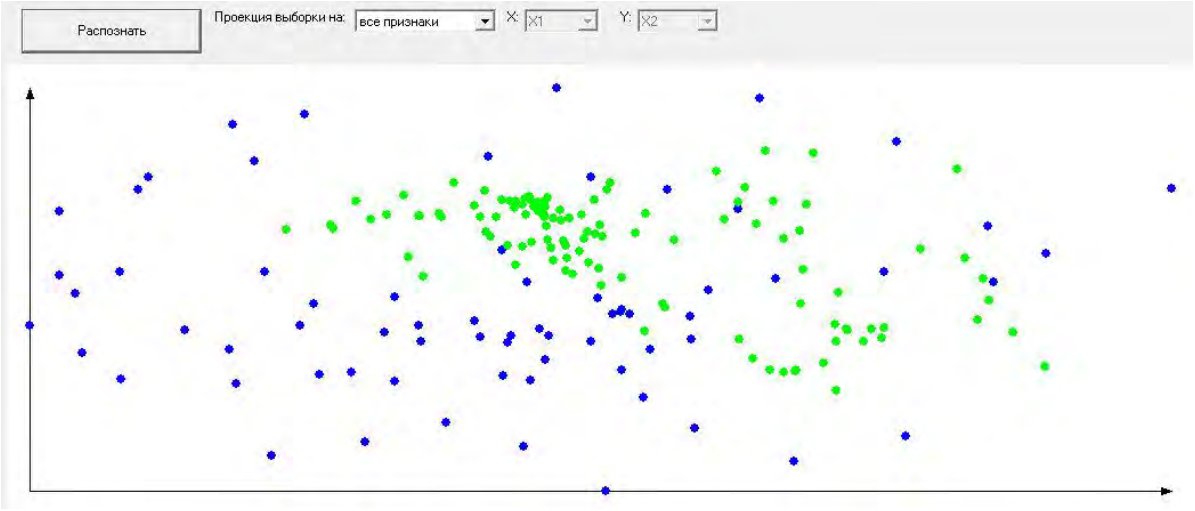
Рассмотрим данные “Ion.tab”. 2 класса. 34 признака. 169 объектов обучающей выборки: 57 объектов – 1 класса, и 112 объекта – 2 класса. 182 тестовых объекта: 69 объектов – 1 класса, 113 объектов – 2 класса.

На обучающей выборке программой Recognition 2.0 методом “Логические закономерности” было найдено 56 логических закономерностей, по 28 логических закономерностей для каждого класса.

Рассмотрим визуализацию начального представления данных. В качестве средства для визуализации выборки была использована программа Recognition 2.0.

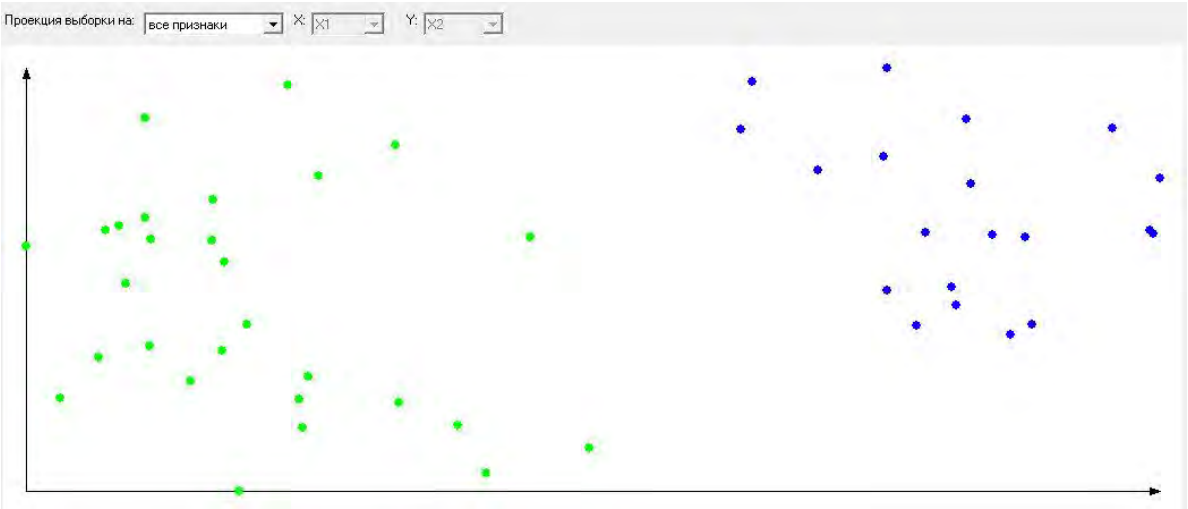


“Ion_l.tab”. Начальное представление. Обучающая выборка.

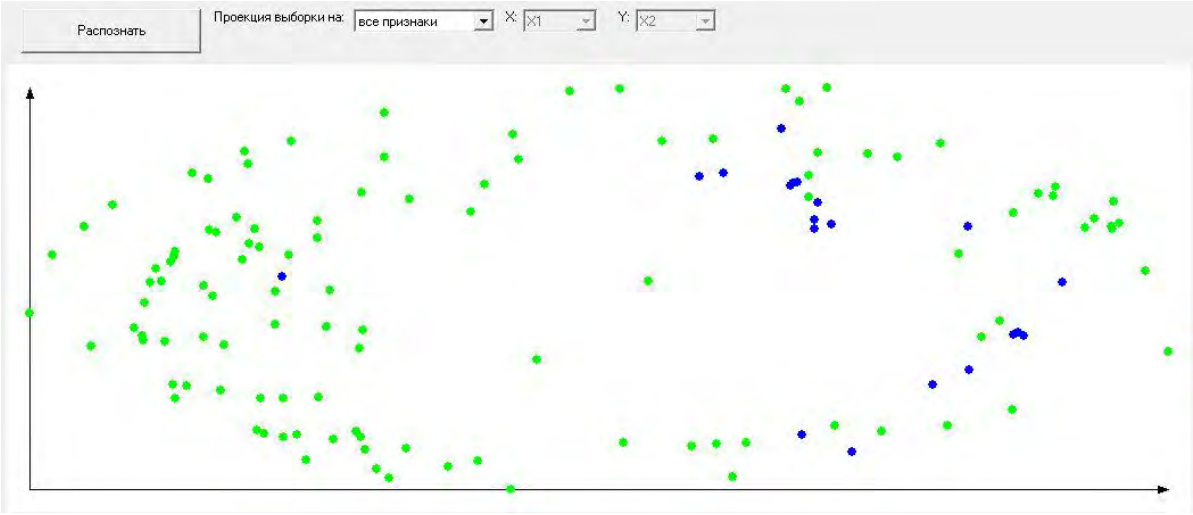


“Ion_r.tab”. Начальное представление. Тестовая выборка.

Теперь рассмотрим визуализацию данных описанную *первым способом.*



“Ion_l.tab”. *Первый способ*. Обучающая выборка.



“lon_r.tab”. *Первый способ*. Тестовая выборка.

Сравним качество этих двух представлений данных при классификации с помощью *линейного дискриминанта Фишера* (метод, разделяющий классы гиперплоскостью, построенной по статистическим свойствам выборки) и *линейной машины* (метод, строящий решающие правила в виде линейных функционалов для каждого класса).

При классификации тестовой выборки для **начального представления** данных с помощью линейного дискриминанта Фишера получаем следующие результаты:

Объектов, отнесенных в класс												
Класс	Объектов		Правильно		Ошибочно		Из класса 1		Из класса 2	
1	67	(36.8%)	**53**	**(79.1%)**	14	(20.9%)	**53**	**(76.8%)**	14	(12.4%)
2	115	(63.2%)	**99**	**(86.1%)**	16	(13.9%)	16	(23.2%)	**99**	**(87.6%)**
Отказы	0	(0.0%)					0	(0.0%)	0	(0.0%)
Итого										
	182	(100 %)	**152**	**(83.5%)**	30	(16.5%)	69	(37.9%)	113	(62.1%)

При классификации тестовой выборки для представления данных **первым способом** с помощью линейного дискриминанта Фишера получаем следующие результаты:

Объектов, отнесенных в класс												
Класс	Объектов		Правильно		Ошибочно		Из класса 1		Из класса 2	
1	82	(45.1%)	**66**	**(80.5%)**	16	(19.5%)	**66**	**(95.7%)**	16	(14.2%)
2	100	(54.9%)	**97**	**(97.0%)**	3	(3.0%)	3	(4.3%)	**97**	**(85.8%)**
Отказы	0	(0.0%)					0	(0.0%)	0	(0.0%)
Итого										
	182	(100 %)	**163**	**(89.6%)**	19	(10.4%)	69	(37.9%)	113	(62.1%)

При классификации тестовой выборки для **начального представления** данных с помощью линейной машины получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		61	(33.5%)		**53**	**(86.9%)**		8	(13.1%)		**53**	**(76.8%)**		8	(7.1%)
2		121	(66.5%)		**105**	**(86.8%)**		16	(13.2%)		16	(23.2%)		**105**	**(92.9%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		182	(100 %)		**158**	**(86.8%)**		24	(13.2%)		69	(37.9%)		113	(62.1%)

При классификации тестовой выборки для представления данных **первым способом** с помощью линейной машины получаем следующие результаты:

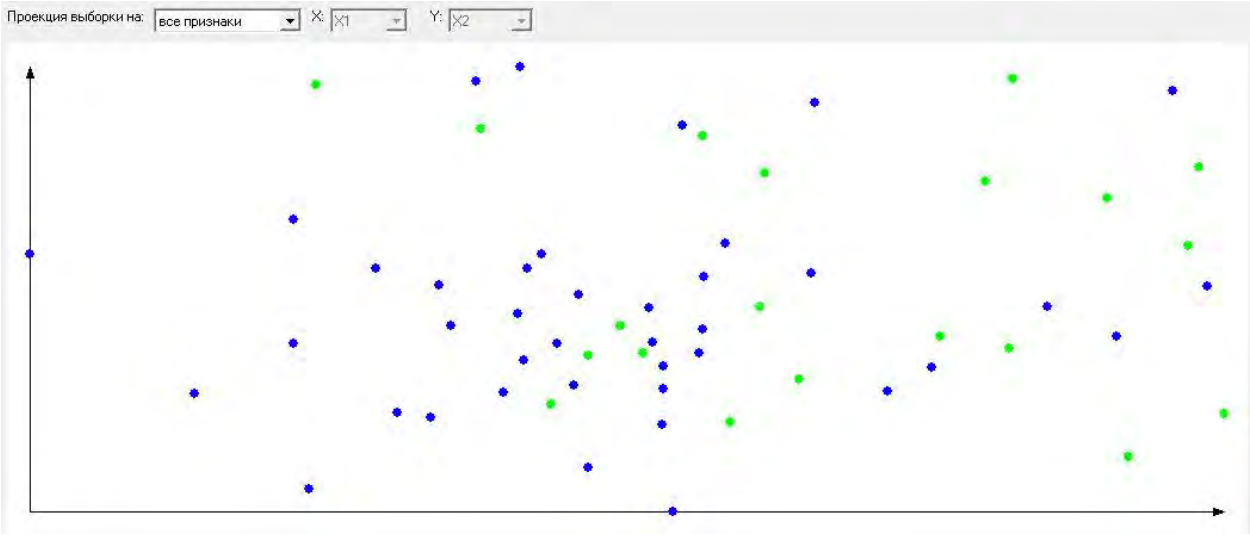
Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		96	(52.7%)		**66**	**(68.8%)**		30	(31.3%)		**66**	**(95.7%)**		30	(26.5%)
2		86	(47.3%)		**83**	**(96.5%)**		3	(3.5%)		3	(4.3%)		**83**	**(73.5%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		182	(100 %)		**149**	**(81.9%)**		33	(18.1%)		69	(37.9%)		113	(62.1%)

4.2 Сравнение моделей классификации основанных на множествах-прецедентах описанных вторым способом и моделей классификации основанных на первоначальном представлении данных

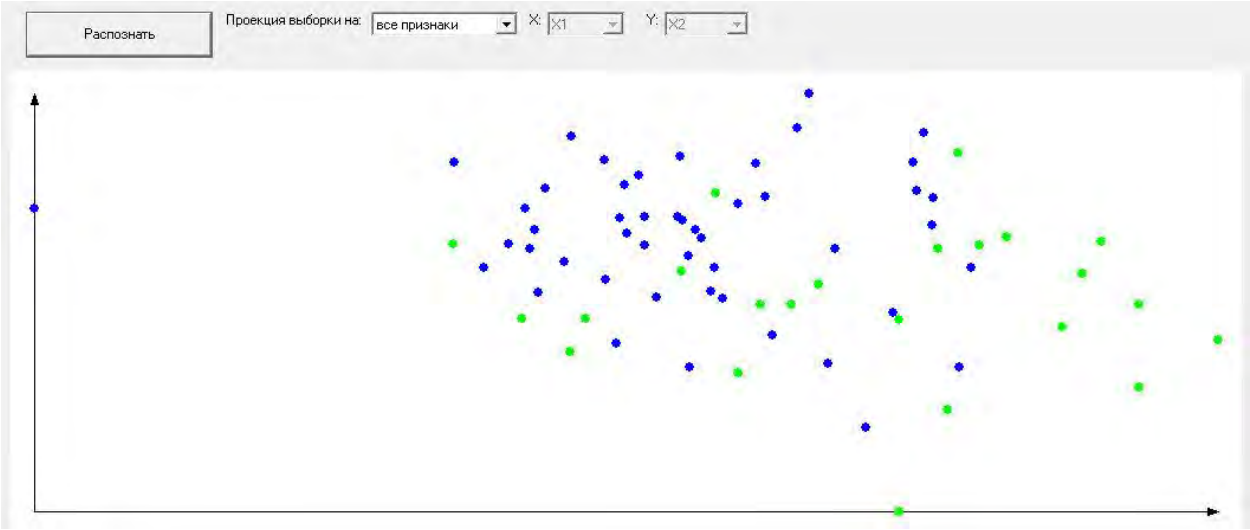
Рассмотрим данные “ech.tab”. 2 класса. 8 признаков. 60 объектов обучающей выборки: 40 объектов – 1 класса, и 20 объектов – 2 класса. 71 тестовых объекта: 48 объектов – 1 класса, 23 объектов – 2 класса.

На обучающей выборке программой Recognition 2.0 методом “Логические закономерности” было найдено 40 логических закономерностей, 24 логических закономерностей для 1 класса, и 16 логических закономерностей для 2 класса.

Рассмотрим визуализацию начального представления данных. В качестве средства для визуализации выборки была использована программа Recognition 2.0.

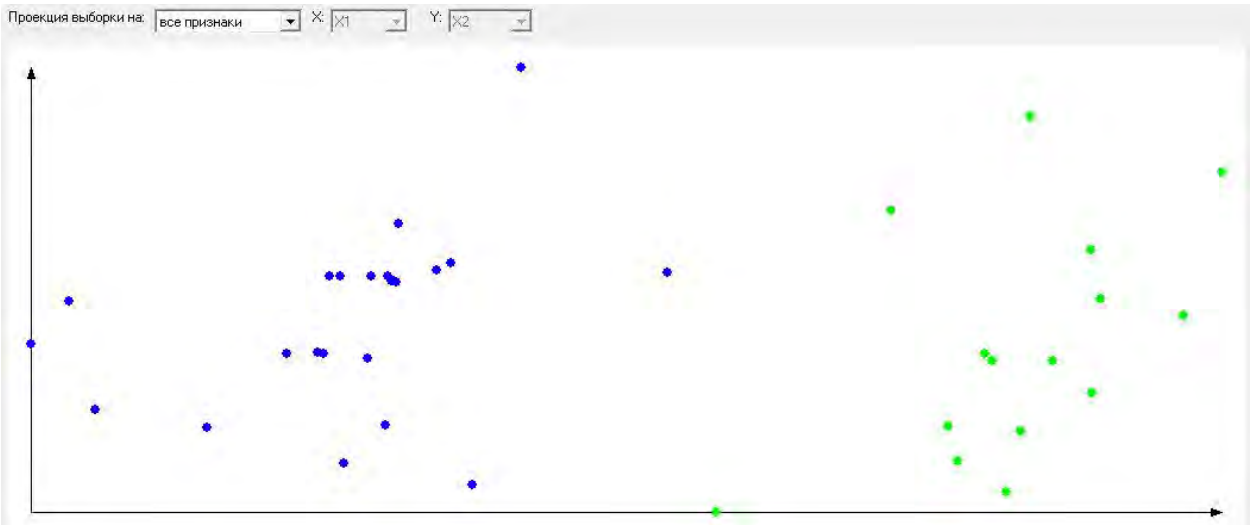


“ech_l.tab”. Начальное представление. Обучающая выборка.

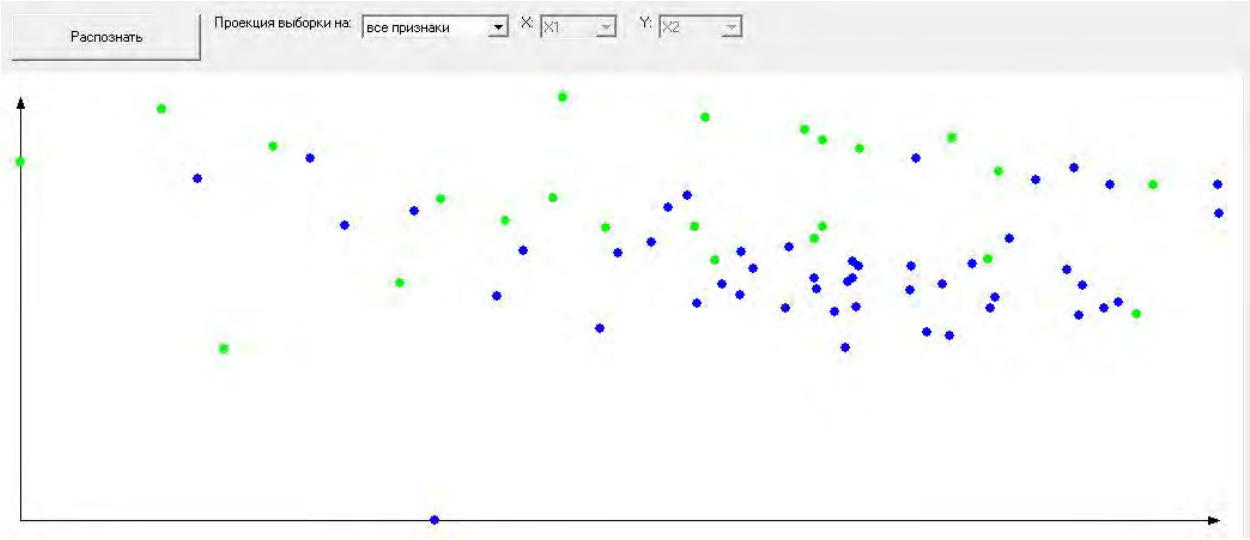


“ech_r.tab”. Начальное представление. Тестовая выборка.

Теперь рассмотрим визуализацию данных описанную *вторым способом.*



“ech_l.tab”. Второй способ. Обучающая выборка.



“ech_r.tab”. Второй способ. Тестовая выборка.

Сравним качество этих двух представлений данных при классификации с помощью *метода опорных векторов* (метод распознает образы с помощью разделяющей гиперплоскости и преобразования пространства признаков через потенциальную функцию) и *линейной машины* (метод, строящий решающие правила в виде линейных функционалов для каждого класса).

При классификации тестовой выборки для **начального представления** данных с помощью метода опорных векторов получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		55	(77.5%)		**42**	**(76.4%)**		13	(23.6%)		**42**	**(87.5%)**		13	(56.5%)
2		16	(22.5%)		**10**	**(62.5%)**		6	(37.5%)		6	(12.5%)		**10**	**(43.5%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		71	(100 %)		**52**	**(73.2%)**		19	(26.8%)		48	(67.6%)		23	(32.4%)

При классификации тестовой выборки для представления данных **вторым способом** с помощью метода опорных векторов получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		47	(66.2%)		**39**	**(83.0%)**		8	(17.0%)		**39**	**(81.3%)**		8	(34.8%)
2		24	(33.8%)		**15**	**(62.5%)**		9	(37.5%)		9	(18.8%)		**15**	**(65.2%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		71	(100 %)		**54**	**(76.1%)**		17	(23.9%)		48	(67.6%)		23	(32.4%)

При классификации тестовой выборки для **начального представления** данных с помощью линейной машины получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		57	(80.3%)		**42**	**(73.7%)**		15	(26.3%)		**42**	**(87.5%)**		15	(65.2%)
2		14	(19.7%)		**8**	**(57.1%)**		6	(42.9%)		6	(12.5%)		**8**	**(34.8%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		71	(100 %)		**50**	**(70.4%)**		21	(29.6%)		48	(67.6%)		23	(32.4%)

При классификации тестовой выборки для представления данных **вторым способом** с помощью линейной машины получаем следующие результаты:

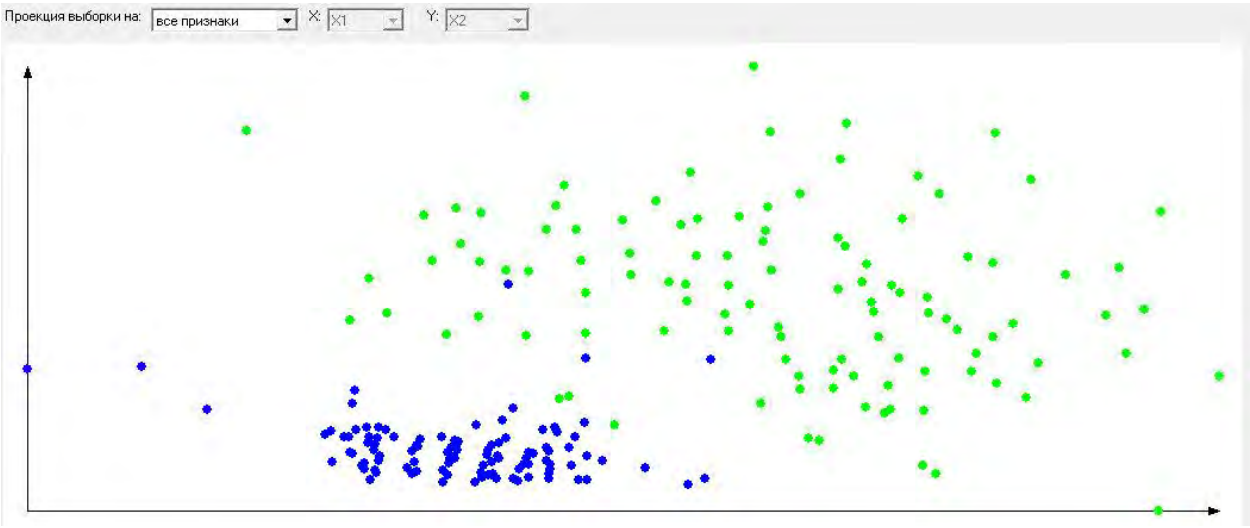
Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		44	(62.0%)		**37**	**(84.1%)**		7	(15.9%)		**37**	**(77.1%)**		7	(30.4%)
2		27	(38.0%)		**16**	**(59.3%)**		11	(40.7%)		11	(22.9%)		**16**	**(69.6%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		71	(100 %)		**53**	**(74.6%)**		18	(25.4%)		48	(67.6%)		23	(32.4%)

4.3 Сравнение моделей классификации основанных на множествах-прецедентах описанных третьим способом и моделей классификации основанных на первоначальном представлении данных

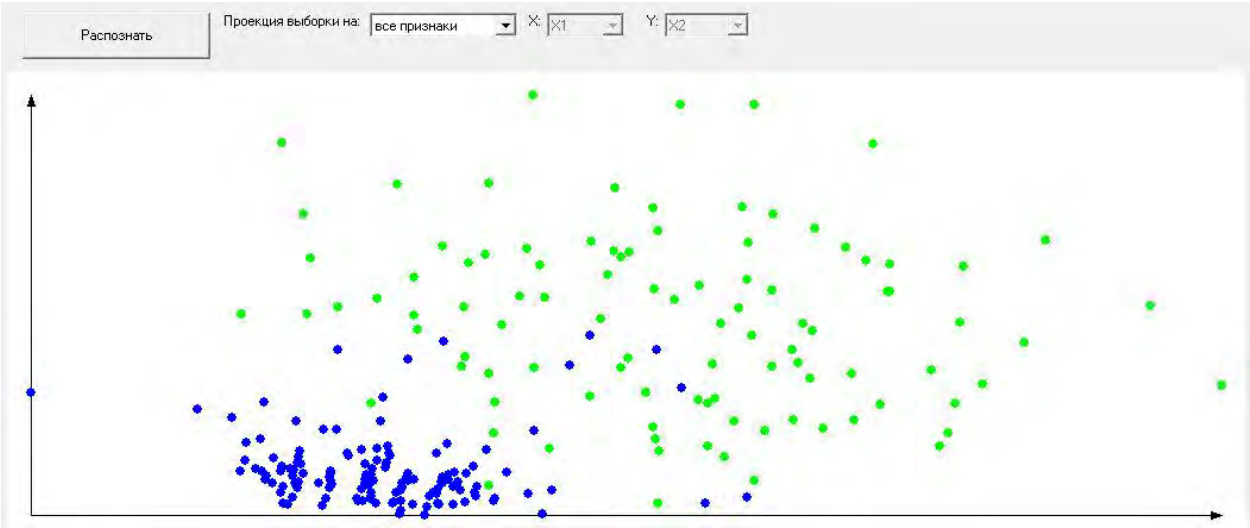
Рассмотрим данные “breast.tab”. 2 класса. 9 признаков. 344 объектов обучающей выборки: 218 объектов – 1 класса, и 126 объектов – 2 класса. 355 тестовых объекта: 240 объектов – 1 класса, 115 объектов – 2 класса.

На обучающей выборке программой Recognition 2.0 методом “Логические закономерности” было найдено 54 логических закономерностей, 28 логических закономерностей для 1 класса, и 26 логических закономерностей для 2 класса.

Рассмотрим визуализацию начального представления данных. В качестве средства для визуализации выборки была использована программа Recognition 2.0.



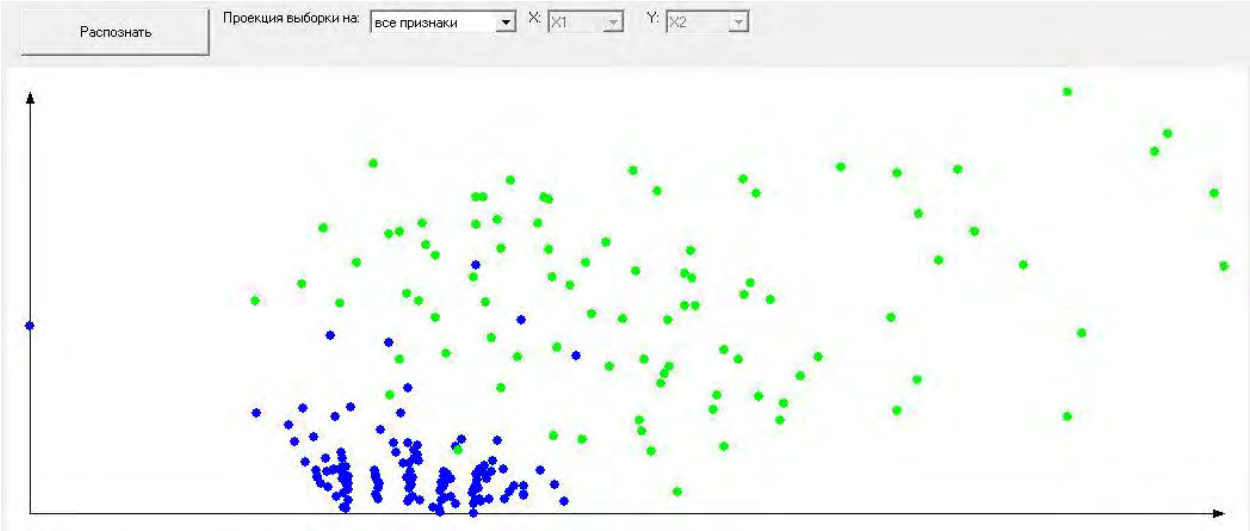
“breast_l.tab”. Начальное представление. Обучающая выборка.



“breast_r.tab”. Начальное представление. Тестовая выборка.

Теперь рассмотрим визуализацию данных описанную *третьим способом.*

“breast_l.tab”. Третий способ. Обучающая выборка.



“breast_r.tab”. Третий способ. Тестовая выборка.

Сравним качество этих двух представлений данных при классификации с помощью *метода опорных векторов* (метод распознает образы с помощью разделяющей гиперплоскости и гиперплоскости и преобразования пространства признаков через потенциальную функцию) и *двумерных линейных разделителей* (метод использует голосование по двумерным решающим правилам).

При классификации тестовой выборки для **начального представления** данных с помощью метода опорных векторов получаем следующие результаты:

Объектов, отнесенных в класс												
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		233	(65.6%)		**228**	**(97.9%)**		5	(2.1%)		**228**	**(95.0%)**		5	(4.3%)
2		122	(34.4%)		**110**	**(90.2%)**		12	(9.8%)		12	(5.0%)		**110**	**(95.7%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		355	(100 %)		**338**	**(95.2%)**		17	(4.8%)		240	(67.6%)		115	(32.4%)

При классификации тестовой выборки для представления данных **третьим способом** с помощью метода опорных векторов получаем следующие результаты:

Объектов, отнесенных в класс												
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		239	(67.3%)		**231**	**(96.7%)**		8	(3.3%)		**231**	**(96.3%)**		8	(7.0%)
2		116	(32.7%)		**107**	**(92.2%)**		9	(7.8%)		9	(3.8%)		**107**	**(93.0%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		355	(100 %)		**338**	**(95.2%)**		17	(4.8%)		240	(67.6%)		115	(32.4%)

При классификации тестовой выборки для **начального представления** данных с помощью двумерных линейных разделителей получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		224	(63.1%)		**223**	**(99.6%)**		1	(0.4%)		**223**	**(92.9%)**		1	(0.9%)
2		131	(36.9%)		**114**	**(87.0%)**		17	(13.0%)		17	(7.1%)		**114**	**(99.1%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		355	(100 %)		**337**	**(94.9%)**		18	(5.1%)		240	(67.6%)		115	(32.4%)

При классификации тестовой выборки для представления данных **третьим способом** с помощью двумерных линейных разделителей получаем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		239	(67.3%)		**231**	**(96.7%)**		8	(3.3%)		**231**	**(96.3%)**		8	(7.0%)
2		116	(32.7%)		**107**	**(92.2%)**		9	(7.8%)		9	(3.8%)		**107**	**(93.0%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		355	(100 %)		**338**	**(95.2%)**		17	(4.8%)		240	(67.6%)		115	(32.4%)

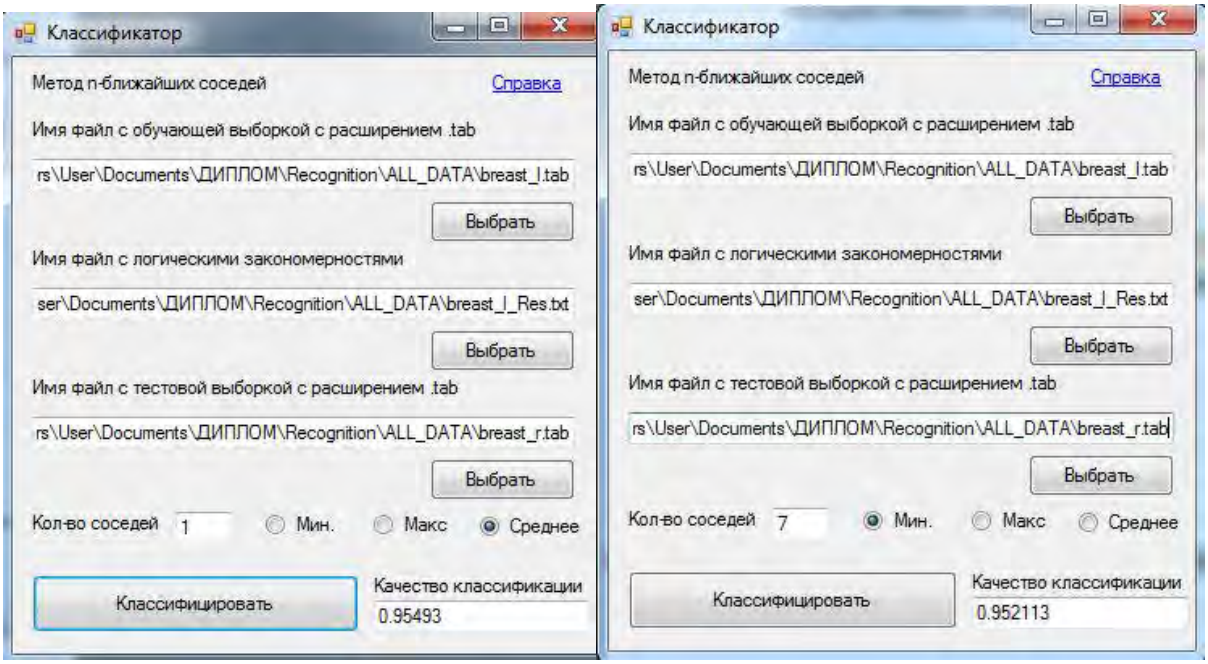
4.4 Сравнение модифицированного и стандартного алгоритма k-ближайших соседей

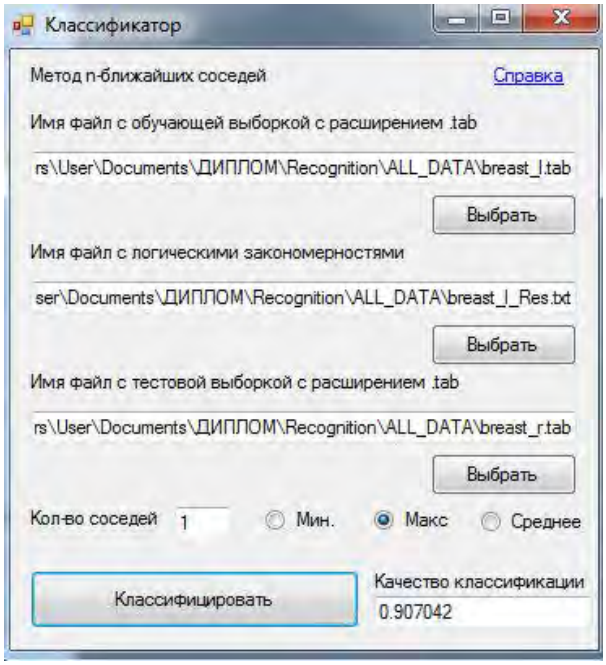
Рассмотрим те же данные, что и в прошлом пункте “breast.tab” (см. пункт 4.3).

При классификации тестовой выборки для **начального представления** данных с помощью алгоритма k-ближайших соседей с автоопределением параметра k (числа ближайших соседей) имеем следующие результаты:

Объектов, отнесенных в класс															
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		235	(66.2%)		**229**	**(97.4%)**		6	(2.6%)		**229**	**(95.4%)**		6	(5.2%)
2		120	(33.8%)		**109**	**(90.8%)**		11	(9.2%)		11	(4.6%)		**109**	**(94.8%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		355	(100 %)		**338**	**(95.2%)**		17	(4.8%)		240	(67.6%)		115	(32.4%)

При классификации тестовой выборки с помощью **модифицированного алгоритма k-ближайших соседей** для трех способов подсчета расстояния между логической закономерностью и объектом (см. пункт 3.3) получаем следующие результаты:



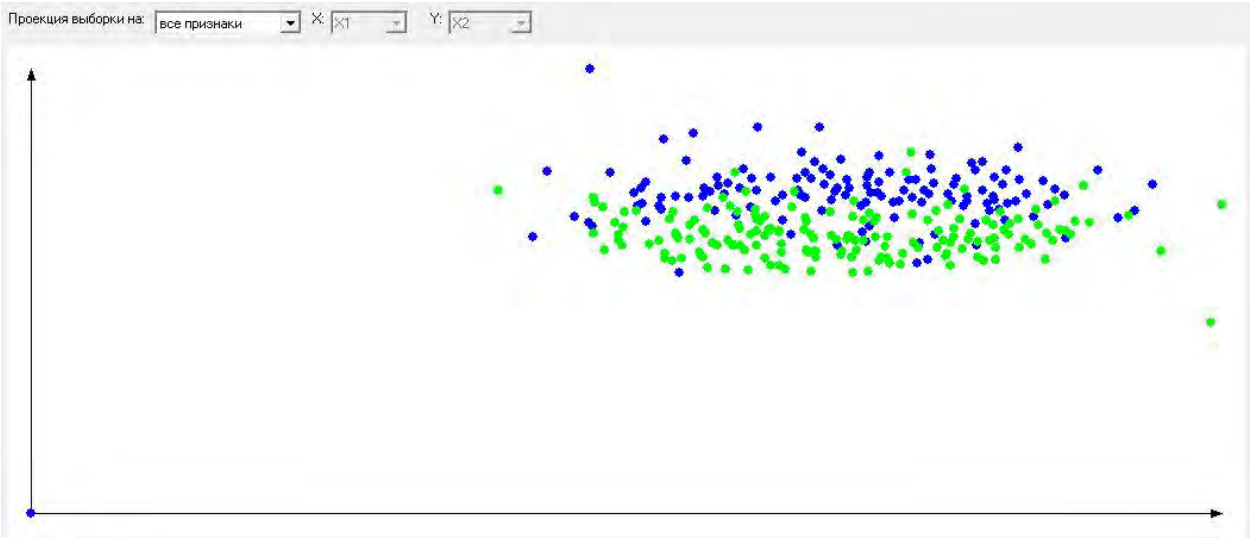


4.5 Сравнение моделей классификации основанных на множествах-прецедентах описанных первым, вторым и третьим способами

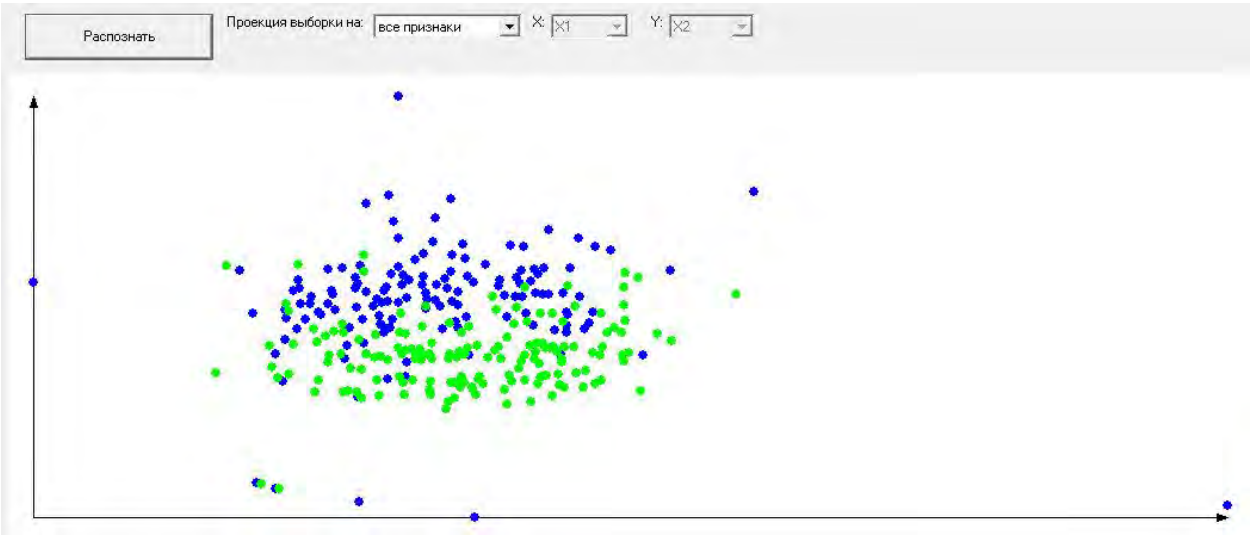
Сравним наши способы описания данных на задаче кредитного скоринга (“ *credit.tab*”).

2 класса. 15 признаков. 342 объектов обучающей выборки: 152 объектов – 1 класса, и 190 объектов – 2 класса. 348 тестовых объектов: 155 объектов – 1 класса, 193 объектов – 2 класса.

Рассмотрим визуализацию начального представления данных. В качестве средства для визуализации выборки была использована программа Recognition 2.0.



“credit_l.tab”. Начальное представление. Обучающая выборка.



“credit_r.tab”. Начальное представление. Тестовая выборка.

При классификации тестовой выборки для представления данных **первым способом** данных с помощью алгоритма k-ближайших соседей с автоопределением параметра k (числа ближайших соседей) имеем следующие результаты:

Объектов, отнесенных в класс														
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		257	(73.9%)		**144**	**(56.0%)**		113	(44.0%)		**144**	**(92.9%)**		113	(58.5%)
2		91	(26.1%)		**80**	**(87.9%)**		11	(12.1%)		11	(7.1%)		**80**	**(41.5%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		348	(100 %)		**224**	**(64.4%)**		124	(35.6%)		155	(44.5%)		193	(55.5%)

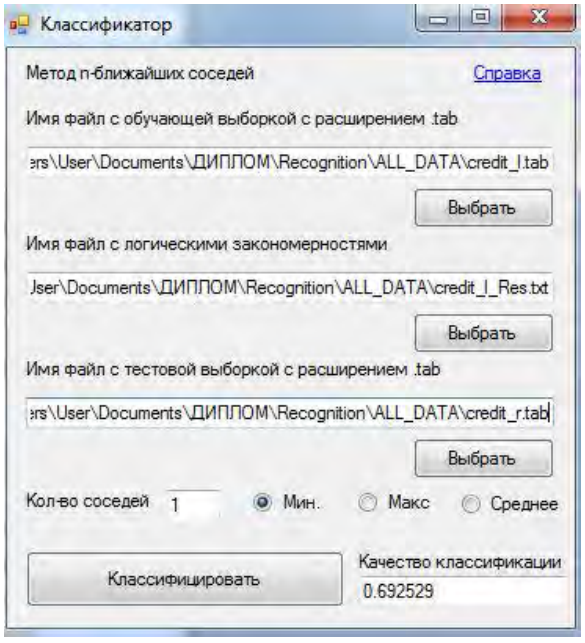
При классификации тестовой выборки для представления данных **вторым способом** данных с помощью алгоритма k-ближайших соседей с автоопределением параметра k (числа ближайших соседей) имеем следующие результаты:

Объектов, отнесенных в класс												
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		76	(21.8%)		**61**	**(80.3%)**		15	(19.7%)		**61**	**(39.4%)**		15	(7.8%)
2		272	(78.2%)		**178**	**(65.4%)**		94	(34.6%)		94	(60.6%)		**178**	**(92.2%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		348	(100 %)		**239**	**(68.7%)**		109	(31.3%)		155	(44.5%)		193	(55.5%)

При классификации тестовой выборки для представления данных **третьим способом** данных с помощью алгоритма k-ближайших соседей с автоопределением параметра k (числа ближайших соседей) имеем следующие результаты:

Объектов, отнесенных в класс												
Класс		Объектов			Правильно			Ошибочно			Из класса 1			Из класса 2	
1		154	(44.3%)		**112**	**(72.7%)**		42	(27.3%)		**112**	**(72.3%)**		42	(21.8%)
2		194	(55.7%)		**151**	**(77.8%)**		43	(22.2%)		43	(27.7%)		**151**	**(78.2%)**
Отказы		0	(0.0%)								0	(0.0%)		0	(0.0%)
Итого															
		348	(100 %)		**263**	**(75.6%)**		85	(24.4%)		155	(44.5%)		193	(55.5%)

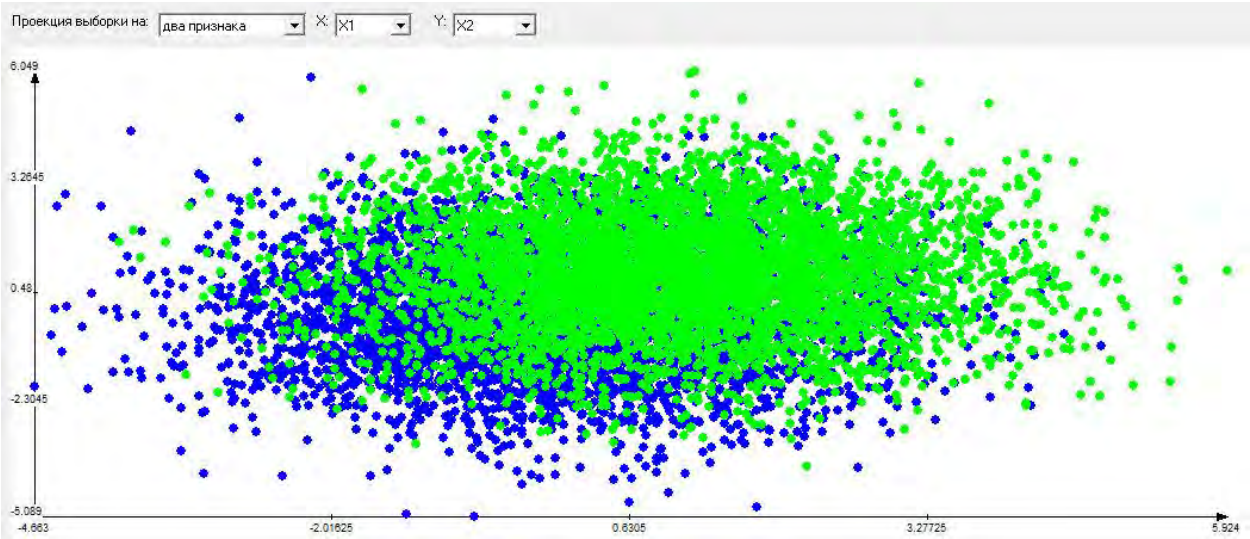
При классификации тестовой выборки с помощью **модифицированного алгоритма k-ближайших соседей** получаем следующие результаты:



4.6 Сравнение моделей классификации основанных на множествах-прецедентах на модельных данных

Сгенерируем обучающую и тестовую выборку следующим образом:

2 класса. 8 признаков. В обучающей и тестовой выборке по 10000 объектов – по 5000 объектов каждого класса. $\boldsymbol{X} \sim \mathcal{N}(x|\mu,\Sigma)$, для первого класса $\mu = (0,\dots,0), \Sigma = \begin{bmatrix} 2\ 0\ \dots\ 0 \\ 0\ 2\ \dots\ 0 \\ \dots \\ 0\ 0\ \dots\ 2 \end{bmatrix}$; для второго класса $\mu = (1,\dots,1), \Sigma = \begin{bmatrix} 2\ 0\ \dots\ 0 \\ 0\ 2\ \dots\ 0 \\ \dots \\ 0\ 0\ \dots\ 2 \end{bmatrix}$



Визуализация модельных данных

Первый способ представления для модельных данных не рассматривается из-за больших длин бинарных векторов (10000 признаков).

При классификации тестовой выборки **для начального представления** с помощью линейной машины имеем следующие результаты:

Объектов, отнесенных в класс										
Класс	Объектов		Правильно		Ошибочно		Из класса 1		Из класса 2	
1	4774	(47.7%)	**3862**	**(80.9%)**	912	(19.1%)	**3862**	**(77.2%)**	912	(18.2%)
2	5226	(52.3%)	**4088**	**(78.2%)**	1138	(21.8%)	1138	(22.8%)	**4088**	**(81.8%)**
Отказы	0	(0.0%)					0	(0.0%)	0	(0.0%)
	10000	(100 %)	**7950**	**(79.5%)**	2050	(20.5%)	5000	(50.0%)	5000	(50.0%)

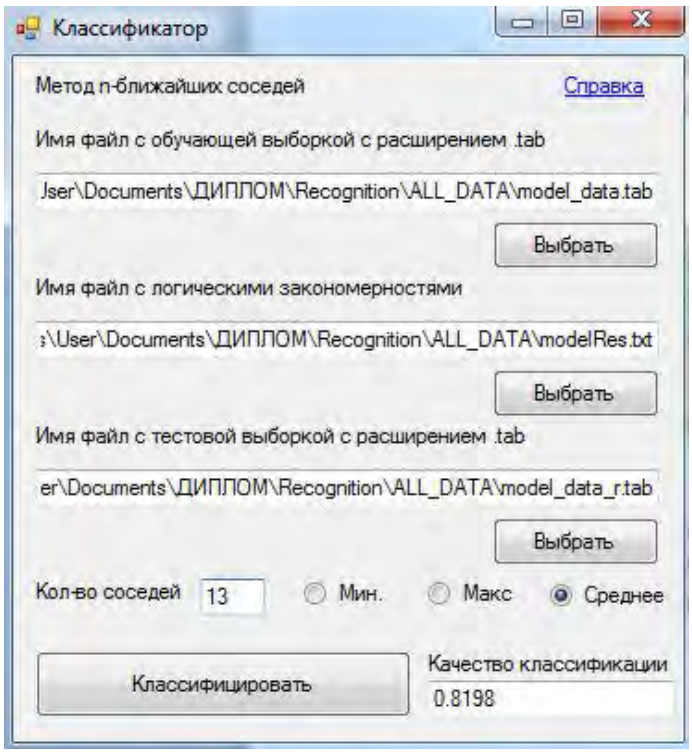
При классификации тестовой выборки **для второго представления** с помощью линейной машины имеем следующие результаты:

Объектов, отнесенных в класс										
Класс	Объектов		Правильно		Ошибочно		Из класса 1		Из класса 2	
1	3633	(36.3%)	**3299**	**(90.8%)**	334	(9.2%)	**3299**	**(66.0%)**	334	(6.7%)
2	6367	(63.7%)	**4666**	**(73.3%)**	1701	(26.7%)	1701	(34.0%)	**4666**	**(93.3%)**
Отказы	0	(0.0%)					0	(0.0%)	0	(0.0%)
	10000	(100 %)	**7965**	**(79.7%)**	2035	(20.4%)	5000	(50.0%)	5000	(50.0%)

При классификации тестовой выборки **для третьего представления** с помощью линейной машины имеем следующие результаты:

Объектов, отнесенных в класс										
Класс	Объектов		Правильно		Ошибочно		Из класса 1		Из класса 2	
1	4571	(45.7%)	**3768**	**(82.4%)**	803	(17.6%)	**3768**	**(75.4%)**	803	(16.1%)
2	5429	(54.3%)	**4197**	**(77.3%)**	1232	(22.7%)	1232	(24.6%)	**4197**	**(83.9%)**
Отказы	0	(0.0%)					0	(0.0%)	0	(0.0%)
	10000	(100 %)	**7965**	**(79.7%)**	2035	(20.4%)	5000	(50.0%)	5000	(50.0%)

При классификации тестовой выборки с помощью **модифицированного алгоритма k-ближайших соседей** получаем следующие результаты:



4.7 Анализ результатов экспериментов

- Эксперименты показали, что все рассмотренные способы описания множеств-прецедентов соответствующих логическим закономерностям на некоторых прикладных задачах могут конкурировать с первоначальным описанием данных.
- *Первый, второй* и *третий* способы описания уменьшают объемы обучающей выборки (так как количество найденных логических закономерностей, как правило, меньше количества объектов обучающей выборки, по которым их ищут)
- Самым нестабильным способом описания показал себя *первый* способ. Это можно объяснить слишком длинным описанием объектов
- Наилучшие результаты при сравнении способов описания множеств-прецедентов показывают *второй* и *третий* способы
- Наихудшим способом подсчета расстояния между объектом и логической закономерностью является расстояние от этого объекта до самого отдаленного объекта покрытого логической закономерностью
- Первый способ подсчета расстояния между объектом и логической закономерностью (*расстояние до ближайшего объекта, покрытого логической закономерностью*) лучше себя показывает на задачах с сильно пересекающимися классами. А третий способ подсчета расстояния (*среднее расстояние до всех объектов, покрываемых логической закономерностью*) лучше себя показывает на задачах с плавными границами между классами
- Проблемой *первого*, *второго* и *третьего* описания множеств-прецедентов состоит в том, что получившиеся обучающие выборки для новых представлений хорошо разделяются, но это снижает точность классификации тестовых выборок на границе классов

5 Заключение

- Предложен подход распознавания по множествам-прецедентам.
- Предложены несколько способов решения задачи распознавания по прецедентам с помощью перехода к задаче распознавания по множествам-прецедентам.
- Созданы программные реализации данных способов
- Проведены эксперименты по сравнению моделей основанных на стандартном подходе распознавания и моделей основанных на множествах-прецедентах на модельных и реальных задачах машинного обучения.

6 Список литературы

[1] www.MachineLearning.ru — профессиональный вики-ресурс, посвященный машинному обучению и интеллектуальному анализу данных

[2] Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006.

[3] Воронцов К. В. «Математические методы обучения по прецедентам».

[4] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. — 270 с.

[5] Рязанов В.В., Сенько О.В. О некоторых моделях голосования и методах их оптимизации // Распозна вание, классификация, прогноз: Матем. методы и их применение. М.: Наука, 1992. Вып. 3. С. 106-145.

[6] Ryazanov V.V. Recognition algorithms based on local optimality criteria// Pattern Recognition and Image Analys. 1994. V. 4. №2 . P. 98-109.

[7] Рязанов В. В. «Логические закономерности в задачах распознавания» Журнал вычислительной математики и математической физики, М.: Наука. Т.48, 2008, N 2, стр. 329-344

[8] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classication // John Wiley and Sons, 2nd edition, 2000

[9] Laim S.B., Ryazanov V.V. The search of precedent-based logical regularities for recognition and data analysis problems // Pattern Recognition and Image Analys. 1997. V. 7. № 3. P. 322-333. 14. Журавлёв Ю.И., Рязанов В.В. Об извлечении знаний из выборок прецедентов

[10] Н. В. Ковшов, В. Л. Моисеев, В. В. Рязанов, «Алгоритмы поиска логических закономерностей в задачах распознавания», Ж. вычисл. матем. и матем. физ., 2008, том 48, номер 2, 329–344

[11] Журавлев Ю.И., Петров И.Б., Рязанов В.В. «Дискретные методы диагностики и анализа медицинской информации» Медицина в зеркале информатики. Сб. РАН отв. ред. О.М. Белоцерковский, А.С. Холодов, Москва: Наука, 2008г., с. 113-123.

[12] Дьяконов В. П. Справочник по применению системы PC MATLAB. — М.: «Физматлит», 1993. — С. 112.