Содержание

Глава 1

Введение

Одной из важнейших задач радиолокации, наряду с обнаружением радиолокационного объекта и определением его траекторных характеристик, является идентификация типа объекта. Поскольку при решении этой задачи необходимо построить алгоритм, который ставит в соответствие некоторому набору измерений один из заранее заданных типов радиолокационных объектов, математическая составляющая алгоритма распознавания в данном случае представляет из себя типичную задачу классификации.

В случае, когда для определения типа радиолокационного объекта не используются его траекторные характеристики, распознавание осуществляется на основе измерений поляризационной матрицы рассеяния. В данной работе признаками объекта являются непосредственно измеренные элементы поляризационной матрицы рассеяния.

Цель работы заключается в том, чтобы сравнить качество работы различных алгоритмов классификации (наивный Байесовский классификатор, метод опорных векторов) в условиях данной задачи, а также составить алгоритм распознавания (включающий в себя преобразование признаков и классификацию), который будет с наибольшей вероятностью правильно классифицировать новые объекты. Для построения такого алгоритма предлагается исследовать пространственную структуру данных, составленных из измерений элементов поляризационной матрицы рассеяния.

Работа имеет следующую структуру:

В главе 2 приведена постановка задачи распознавания радиолокационного объекта в терминах математической задачи классификации.

Глава 3 посвящена обзору алгоритмов классификации, работа которых моделировалась в условиях поставленной задачи, а также описанию алгоритмов кластеризации и метода сокращения размерности данных, используемых в качестве составляющих частей предлагаемого алгоритма распознавания радиолокационного объекта.

В главе 4 приведено описание предлагаемого алгоритма.

Глава 5 содержит описание метода моделирования радиолокационных объектов и оценки работы алгоритмов классификации в условиях поставленной задачи, а также результаты моделирования.

Глава 2

Постановка задачи

В случае, когда приём сигнала осуществляется на одну линейно поляризованную антенну, принятый сигнал представляет собой линейную смесь компонент излучённого сигнала с коэффициентами, которые являются элементами поляризационной матрицы рассеяния. Если проводится предварительная обработка, включающая согласованную фильтрацию и фазовое детектирование сигнала, можно считать, что при излучении и приёме сигнала на две ортогональные линейно поляризованные антенны наблюдению подлежат непосредственно элементы матрицы рассеяния [2]. Если рассматривается другая конфигурация, например, излучение на одной линейной поляризации и приём на две ортогональные поляризации, то принимаемый сигнал можно преобразовать до линейной комбинации элементов поляризационной матрицы рассеяния. При этом мощность излучаемого импульса считается равной 1, а отношение сигнал-шум при моделировании варьируется за счёт увеличения шума, поэтому эффективность рассматриваемых алгоритмов будет зависеть только от вида алгоритма и свойств цели, а излучаемая мощность всегда одинакова.

Таким образом, наблюдаемыми признаками объекта являются элементы поляризационной матрицы рассеяния объекта:

$$\begin{pmatrix} \text{HH} & \text{HV} \\ \text{VH} & \text{VV} \end{pmatrix},$$

$$HH,\ HV,\ VH,\ VV \in \mathbb{C},\ \ HV = -VH.$$

Поскольку поляризационная матрица рассеяния антисимметрична, а каждый ее элемент является комплексным числом, вектор признаков состоит из 3 комплексных (или, соответственно, 6 действительных) чисел:

$$x = [Re\,(HH)\ \ Im\,(HH)\ \ Re\,(HH)\ \ Im\,(HH)\ \ Re\,(HH)\ \ Im\,(HH)].$$

Пусть Y – множество классов радиолокационных объектов. Задача распознавания представляет из себя типичную задачу классификации: необходимо построить функцию $f\,(x)$, которая ставит в соответствие вектору признаков один из заданных типов радиолокационного объекта (элементов множества Y).

Выбор функции $f\,(x)$ из множества возможных классификаторов подразумевает минимизацию некоторого функционала качества, зависящего от функции потерь, которая характеризует величину ошибки алгоритма на объекте. В данном случае потери при ошибочной классификации в сторону каждого из классов считаются одинаковыми, поэтому

величиной, которую необходимо минимизировать, является непосредственно доля ошибочно классифицированных объектов.

Глава 3

Обзор используемых алгоритмов и методов анализа данных

3.1. Алгоритмы классификации

В общем случае задача классификации формулируется следующим образом: имеется множество объектов X, каждый из которых относится к классу из множества классов Y:

$X^l = (x_i, y_i)_{i=1}^l$ – обучающая выборка, $f_j(x), j = 1, \ldots, n$ – признаки объекта x.

Объекты описываются некоторым количеством n признаков разных типов. Необходимо построить функцию-классификатор $a(x) \in Y$, которая каждому объекту на основе его признаков будет сопоставлять один из классов.

Как правило, при построении функции-классификатора ставится задача минимизации некоторого функционала качества:

$$Q\left(a, X^l\right) = \frac{1}{l} \sum_{i=1}^{l} \mathscr{L}(a, x_i),$$

$\mathscr{L}(a, x_i)$ – функция потерь, характеризующая величину ошибки алгоритма a на объекте x.

3.1.1. Наивный Байесовский классификатор

Одним из самых простых методов классификации является наивный Байесовский классификатор [3], основанный на применении теоремы Байеса со строгими предположениями о независимости переменных.

Пусть X – множество объектов, Y – множество классов. Тогда $X \times Y$ – вероятностное пространство с плотностью $p(x, y)$,

$$p(x, y) = p(x) \cdot P(y|x) = P(y) \cdot p(x|y),$$

где $P(y) \equiv P_y$ – априорная вероятность класса y;

$p(x|y)$ – функция правдоподобия класса y;

$P(y|x) \equiv p_y(x)$ – апостериорная вероятность класса .

Требуется найти классификатор $a : X \longrightarrow Y$ с минимальной вероятностью ошибки. Принцип максимума апостериорной вероятности:

$$a(x) = \arg\max_{y \in Y} P(y|x) = \arg\max_{y \in Y} P_y \cdot p_y(x).$$

Для использования этой формулы необходимо получить оценки априорных вероятностей классов $\hat{P}_y$ и плотностей $\hat{p}_y(x)$.

Пусть $X^l = (x_i, y_i)_{i=1}^l$ – выборка; $f_j(x), j = 1, \ldots, n$ – признаки объекта x.

Вполне естественно положить $\hat{P}_y = \frac{l_y}{l}, l_y = |X_y^l|, y \in Y$ – согласно закону больших чисел, частота появления объектов каждого из классов сходится по вероятности к P_y при $l_y \longrightarrow \infty$.

Для оценки плотности в точке можно использовать оценку Парзена-Розенблатта с ядром K и шириной окна h:

$$\hat{p}_{y,h}(x) = \frac{1}{l_y \cdot V(h)} \cdot \sum_{i=1}^{l} [y_i = y] \cdot K(\frac{\rho(x, x_i)}{h}),$$

где $\rho(x, x')$ – заданная на X функция расстояния.

Ширина окна h существенно влияет на качество восстановления плотности. Оптимальное ее значение можно найти методом скользящего контроля (Leave One Out):

$$h : LOO\left(h, X^l\right) = \sum_{i=1}^{l} \left[a\left(x_i; X^l \backslash x_i, h\right) \neq y_i\right] \rightarrow min,$$

где $a\left(x_i; X^l \backslash x_i, h\right)$ – алгоритм классификации, построенный по обучающей выборке без объекта x_i. Обычно зависимость $LOO(h)$ имеет характерный минимум, соответствующий оптимальной ширине окна.

Функция ядра K практически не влияет на качество восстановления плотности, однако определяет степень гладкости функции $\hat{p}_h(x)$ и может влиять на эффективность вычислений.

В качестве ядра $K(r)$ можно выбрать, например:

$K(r) = \frac{1}{2} \cdot [|r| \leq 1]$ – прямоугольное ядро;

$K(r) = (2\pi)^{-\frac{1}{2}} \exp(-frac12r^2) \cdot [|r| \leq 1]$ – гауссово;

$K(r) = \frac{3}{4} \cdot (1 - r^2) \cdot [|r| \leq 1]$ – Епанечникова (оптимальное).

3.1.2. Метод опорных векторов (Support Vector Machine)

Метод опорных векторов [4] [5] считается в настоящее время одним из лучших методов классификации. Обучение SVM сводится к задаче квадратичного программирования, имеющей единственное решение, которое эффективно вычисляется даже на больших выборках. Характерным свойством полученного при помощи этого метода решения является разреженность: положение оптимальной разделяющей поверхности зависит лишь от небольшой доли обучающих объектов, которые и называются опорными векторами; остальные объекты фактически не задействуются. Суть метода опорных векторов заключается в нахождении такой разделяющей поверхности между двумя классами, которая максимизирует зазор (margin) между ними: строится разделяющая полоса (которая не должна содержать точек ни одного из классов в разделимом случае и может содержать не более определенной доли точек классов в неразделимом); ширина разделяющей полосы максимизируется; объекты, которые оказываются на границе разделяющей полосы или (в неразделимом случае) внутри ее, будут опорными объектами. Итоговый алгоритм классификации представляется в виде

$$a\left(x\right) = sign\left(\sum_{i=1}^{l} \lambda_i y_i \langle x_i, x\rangle - w_0\right),$$

где суммирование идет только по опорным объектам (для них коэффициент $\lambda_i \neq 0$).

Как можно видеть, алгоритм классификации зависит только от скалярных произведений объектов, но не от самих признаковых описаний. Это дает возможность формально заменить скалярное произведение $\langle x, x'\rangle$ ядром $K\left(x, x'\right) = \langle \phi\left(x\right), \phi\left(x'\right)\rangle$ (ϕ – некоторое отображение $\phi : X \Longrightarrow H$, где H – пространство со скалярным произведением). Поскольку ядро в общем случае нелинейно, такая замена приводит к существенному расширению множества реализуемых алгоритмов и форм разделяющих поверхностей. Часто (в том числе и при решении поставленной задачи) используются такие ядра, как, например:

$K\left(u, v\right) = \langle u, v\rangle^2$ – квадратичное ядро;

$K\left(u, v\right) = \exp\left(-\frac{1}{2}\sigma||u - v||^2\right)$ – ядро с радиальными базисными функциями (rbf).

Изначально метод опорных векторов решает задачу классификации объектов между двумя классами, однако существует несколько подходов, которые позволяют обобщить метод на большее число классов (например, подход one-vs-all, в котором для каждой точки тестовой выборки строятся классификаторы каждого класса относительно остальных, и точка относится к классу, к которому она принадлежит с наибольшим значением отступа от разделяющей поверхности).

3.2. Алгоритмы кластеризации

В общем случае алгоритмы кластеризации решают задачу разбиения множества объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Эта задача также сводится к минимизации некоторого функционала, включающего в себя, в зависимости от конкретного приложения, некоторые метрики расстояния между точками внутри и вовне кластера.

3.2.1. K-means

Алгорим k-means (k средних) является одним из наиболее популярных методов кластеризации. Метод разбивает множество точек на заданное число кластеров, на каждой итерации пересчитывая центры масс кластеров и относя каждую точку к кластеру с ближайшим центром, и является, таким образом, версией EM (Estimation-Maximization) алгоритма.

Функционал, который алгоритм k-means стремится минимизировать, представляет из себя суммарное квадратичное отклонение точек кластеров от центров кластеров:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 ,$$

k – заданное число кластеров, S_i – текущие кластеры, $\mu_i, i = 1, \ldots, k$ – центры масс кластеров.

Алгоритм K-means:

1. Сформировать начальное приближение центров всех кластеров $y \in Y$;
2. Повторять, пока классы, к которым отнесены объекты, не перестанут изменяться:
   Для каждого объекта выборки:
   a. Отнести объект к ближайшему центру:

$$y := \arg\min_{y \in Y} \rho(x_i, \mu_i), i = 1, \ldots, l;$$

   b. Вычислить новое положение центров:

$$\mu_y j := \frac{\sum_{i=1}^{l} [y_i = y] f_j(x_i)}{\sum_{i=1}^{l} [y_i = y]}, y \in Y, j = 1, \ldots, n.$$

Начальное положение центров можно инициализировать случайным образом либо выбрать μ_y – наиболее удаленные друг от друга объекты выборки. Поскольку результат

сильно зависит от начального приближения центров, при случайной инициализации имеет смысл провести несколько итераций кластеризации и выбрать разбиение, наилучшим образом соответствующее выбранной метрике качества кластеризации.

3.2.2. FOREL

В отличие от алгоритма k-means, алгоритм FOREL (Формальный Элемент) разбивает множество точек на неизвестное заранее число кластеров, стремясь минимизировать по всем кластерам сумму расстояний от точек кластеров до соответствующих центров. Минимизируемый функционал качества:

$$F = \sum_{j=1} k \sum_{x \in K_j} \rho\left(x, W_j\right),$$

$K_j, j = 1, \ldots, k$ – кластеры, W_j – центры кластеров.

Использование алгоритма FOREL в связке с алгоритмом поиска кратчайшего незамкнутого пути позволяет объединять формальные элементы, найденные алгоритом, в заданное число кластеров сложной формы.

Алгоритм FOREL:

Пусть U – множество некластеризованных точек. Пока $U \neq \varnothing$, повторять:

1. Взять произвольную (случайную) некластеризованную точку $x_0 \in U$;
2. Повторять, пока центр не стабилизируется:

   a. Образовать кластер-сферу с центром в x_0 и радиусом R:

$$K_0 := x_i \in U | \rho\left(x_i, x_0\right) \leq R;$$

   b. Поместить центр сферы в центр масс кластера:

$$x_0 := \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$$

Пометить все точки K_0 как кластеризованные:
$U := U \setminus K_0$;

Применить алгоритм КНП к множеству центров всех найденных кластеров; каждый объект $x_i \in X^l$ приписать кластеру с ближайшим центром.

Алгоритм КНП (Кратчайшего Незамкнутого Пути):

1. Найти пару точек (i, j) с наименьшим $\rho_i j$ и соединить их ребром
2. Пока в выборке остаются изолированные точки, повторять:

   a. Найти изолированную точку, ближайшую к некоторой неизолированной;

b. Соединить эти две точки ребром.

3. Удалить K-1 самых длинных ребер.

3.3. Снижение размерности данных

Одним из основных способов снижения размерности данных является метод главных компонент. Суть метода заключается в поиске подпространств меньшей размерности, в ортогональной проекции на которые разброс данных максимален: на каждом шаге ищется направление, вдоль которого максимальна выборочная дисперсия, и из данных вычитается проекция на это направление.

$S_m^2[(X, a_k)] = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{n} x_{ij}a_{kj}\right)^2$ – выборочная дисперсия данных вдоль направления, заданного нормированным вектором a_k (для центрированных данных); если $A = a_1, \ldots, a_n{}^T in\mathbb{R}^{n\times n}, a_k \in \mathbb{R}^n$ – искомое преобразование, то для векторов a_k должны выполняться следующие условия:

$$a_k = \arg\max_{|a_k|=1} S_m^2[(X, a_k)];$$

$x_i := x_i - a_{k-1}\,(a_{k-1}, x_i)$ – вычитание проекции на $(k-1)$-ю главную компоненту.

Глава 4

Описание предлагаемого метода распознавания радиолокационных объектов

В качестве нового подхода к классификации было предложено исследовать кластерную структуру данных обучающей выборки, относящихся к каждому из классов. Для этого данные обучающей выборки каждого из классов, включающие измерения поляризационной матрицы рассеяния во всех ракурсах, разделяются на кластеры при помощи одного из алгоритмов кластеризации (в работе использовались алгоритмы K средних и FOREL).

Предлагаемый способ классификации заключается в следующем: для каждого из классов формируется кластерная структура, кластеры которой соответствуют наиболее типичным «состояниям» поляризационной матрицы рассеяния объекта данного типа. Эта кластерная структура в дальнейшем рассматривается как описание класса объектов данного типа. После того, как указанные описания классов сформированы, строится «общая» кластерная структура, включающая в себя кластеры, относящиеся ко всем типам объектов, которая и является алфавитом классов для распознавания; каждую новую точку можно классифицировать, отнеся ее к одному из кластеров.

Предварительная обработка признаков для такого алгоритма может также включать применение метода главных компонент для снижения размерности данных.

Таким образом, предлагаемый алгоритм обработки включает в себя следующие шаги.

На этапе обучения:

1. снижение размерности данных обучающей выборки при помощи метода главных компонент (сохраняется матрица преобразования данных из оригинального пространства в пространство главных компонент);

2. кластеризация точек обучающей выборки каждого из классов при помощи алгоритма K средних или FOREL (сохраняются координаты центров кластеров).

На этапе классификации новых данных:

1. к новым данным применяется преобразование, переводящее их в пространство главных компонент и сохраняющее заданное число компонент;

2. точка относится к ближайшему кластеру и к классу объекта, к которому относится этот кластер.

Описанный метод может быть использован в задаче классификации с произвольным числом классов, а также в случае, когда допустимой является классификация объекта как неизвестного. Поскольку алгоритм позволяет получить помимо предполагаемого класса значение меры близости классифицируемого объекта к каждому из кластеров, возможно определить степень уверенности, с которой классификатор относит точку к каждому классу, и обобщить работу алгоритма на случай, когда ни один из классов не является достаточно близким, дополнив множество значений функции-классификатора возможностью классифицировать объект как неопределенный.

Глава 5

Моделирование работы алгоритмов

5.1. Формат данных

Для оценки качества работы алгоритмов моделировалсь их работа в случае классификации между двумя классами радиолокационных объектов. Множество классов включало в себя объекты, для которых характерна зависимость значений элементов поляризационной матрицы рассеяния от угла наблюдения. Расчет поляризационной матрицы рассеяния проводился с помощью программы моделирования обратного рассеяния от объекта сложной формы, основанной на фасеточной модели [1]. Моделирование методов преобразования данных и работы алгоритмов классификации производилось в среде Matlab.

Моделировались следующие объекты.

1. Первый объект представляет собой конус радиусом 1 м и длиной 2 м с идеально проводящей поверхностью (Рис. 5.1). Объект ориентирован под углом 30 градусов по вертикальной оси. Угол наклона изменяется от 0 до 90 градусов от вертикальной оси к наблюдателю. Для расчета элементов поляризационной матрицы рассеяния использовался монохроматический сигнал, длина волны 1 м.

   Угловая диаграмма рассеяния конуса представлена на Рис. 5.2.

2. Второй объект представляет собой цилиндр радиусом 1 м и длиной 2.5 м с идеально проводящей поверхностью (Рис. 5.3). Объект ориентирован под углом 30 градусов по вертикальной оси. Угол наклона изменяется от 0 до 90 градусов от вертикальной оси к наблюдателю. Для расчета элементов поляризационной матрицы рассеяния использовался монохроматический сигнал, длина волны 1 м.

   Угловая диаграмма рассеяния цилиндра представлена на Рис. 5.4.

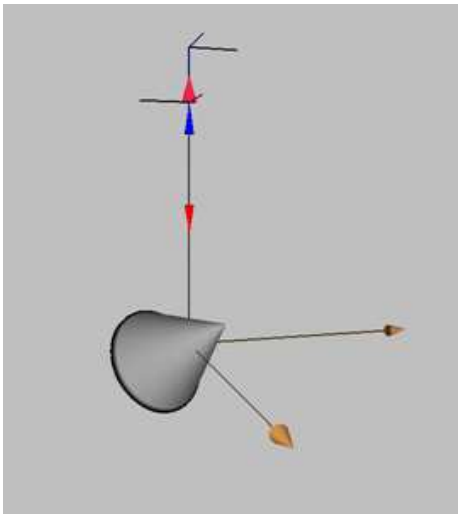

Рис. 5.1. Объект первого типа: конус.

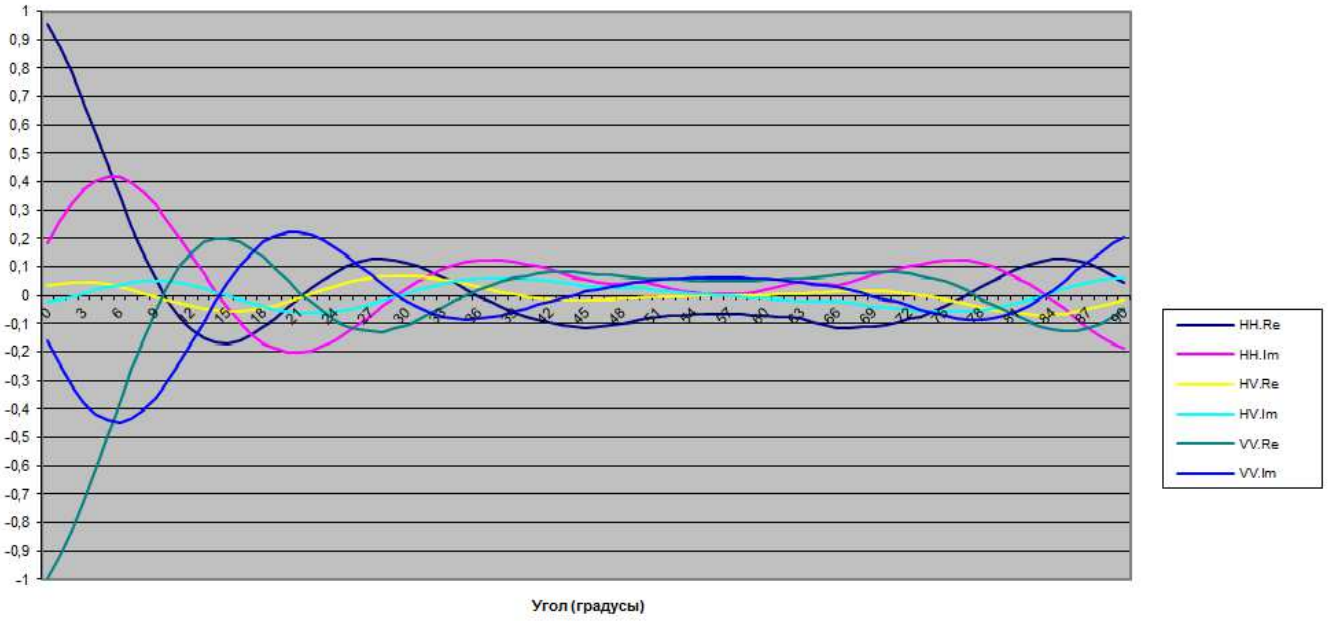

Рис. 5.2. Угловая диаграмма рассеяния конуса.

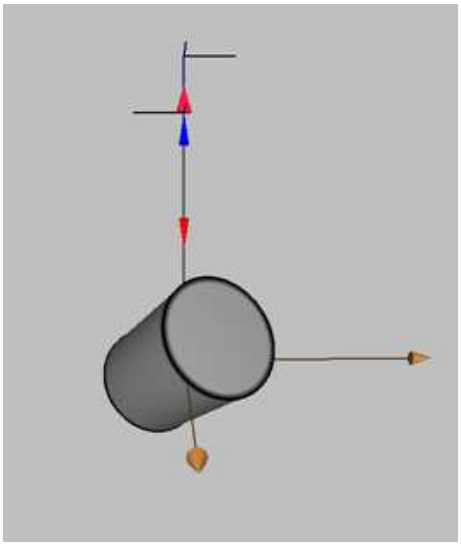

Рис. 5.3. Объект второго типа: цилиндр

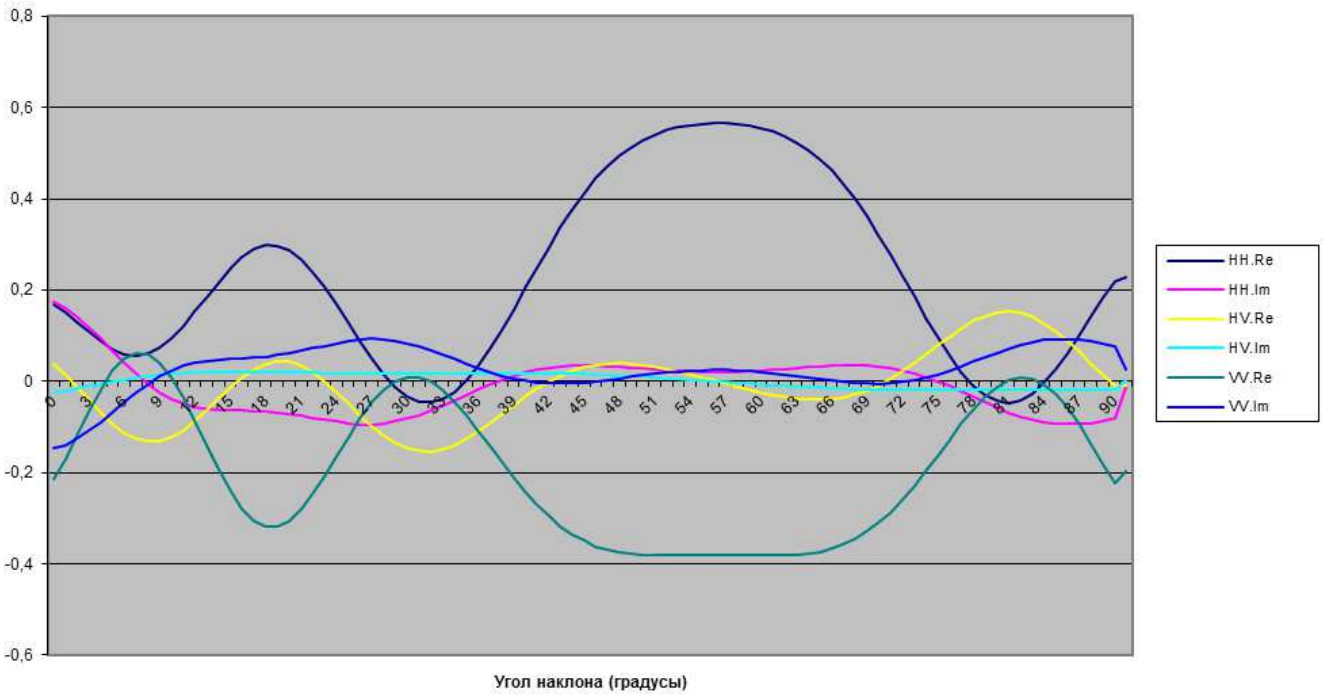

Рис. 5.4. Угловая диаграмма рассеяния цилиндра.

5.2. Снижение размерности данных

Для снижения размерности данных использовался метод главных компонент. На рис. 5.5 показана доля дисперсии, содержащаяся в каждом измерении в пространстве главных компонент. На рис. 5.6 показана доля дисперсии, сохраненная суммарно в первых n ком-

понентах после преобразования. Как можно видеть, размерность данных можно снизить до 4 компонент, сохранив при этом 99% дисперсии, или до 3 компонент, сохранив порядка 90% дисперсии.

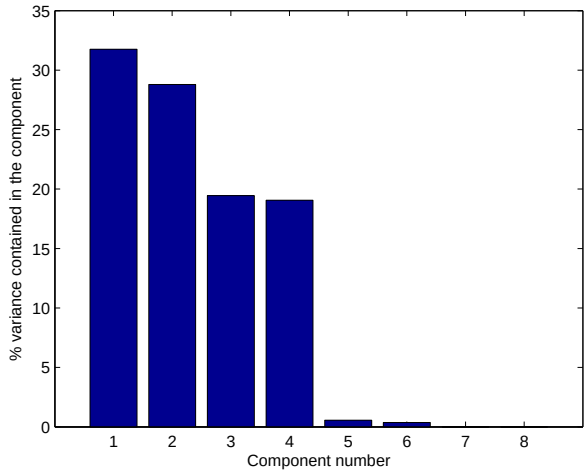

Рис. 5.5. Доля дисперсии, приходящейся на каждую из координат, найденных при помощи метода главных компонент.

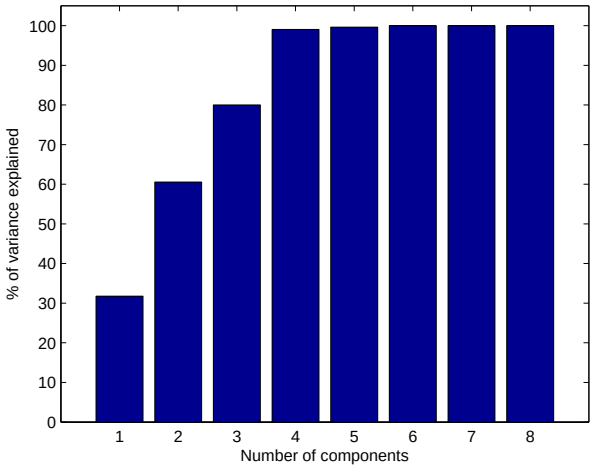

Рис. 5.6. Доля дисперсии, содержащейся суммарно в первых n координатах в пространстве главных компонент.

На рис. 5.7 приведена визуализация данных в пространстве главных компонент (для построения изображения использовались первые три компоненты).

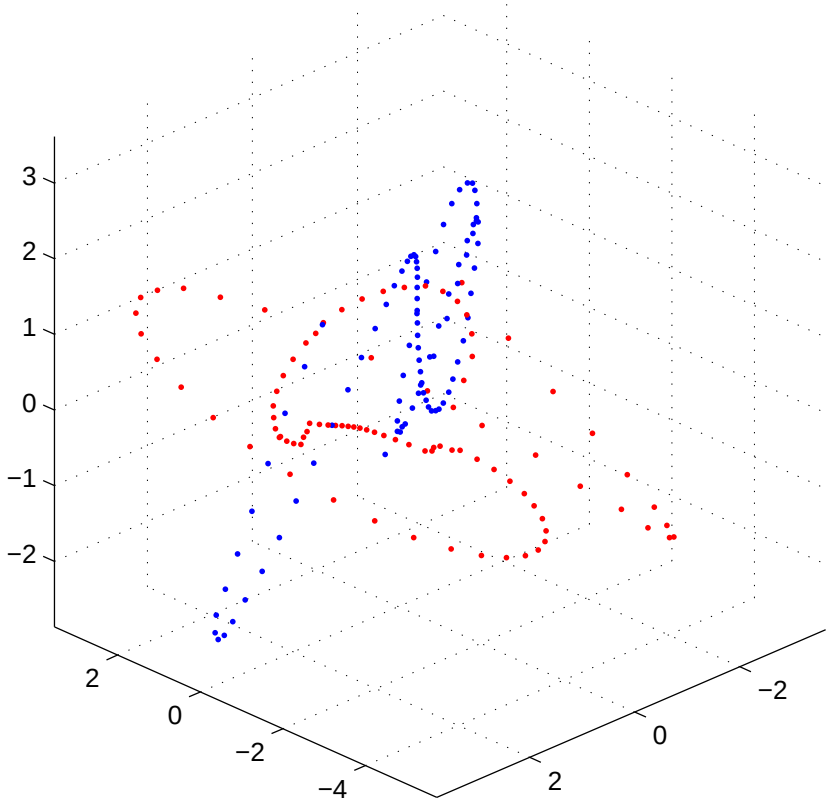

Рис. 5.7. Представление данных в пространстве трех первых главных компонент.

5.3. Кластеризация данных обучающей выборки

В качестве метрики качества кластеризации будем использовать величину, которая называется “силуэт” и определяется для каждой точки кластеризованных данных следующим образом:

$$s\left(i\right) = \frac{\left(b\left(i\right) - a\left(i\right)\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}},$$

где $a\left(i\right)$ – среднее расстояние от данной точки до других точек кластера, $b\left(i\right)$ – наименьшее (по кластерам) среднее расстояние от точки i до точек другого кластера.

Кластер можно считать удачно определенным, если для большей части его точек $s(i) > 0.6$.

Поскольку результат кластеризации значительно зависит от начальных условий, для получения наилучшего результата следует выполнять несколько итераций для каждого

набора значений параметров алгоритмов. На Рис. 5.8 изображены графики силуэта для наилучших (по среднему значению силуэта) кластеров, полученных в результате применения алгоритмов K средних (верхние два графика) и FOREL (нижние два графика) к объектам обучающей выборки двух классов. По оси y отложены точки выборки, сгруппированные по кластерам и отсортированные внутри кластера по значению силуэта; по оси x отложено значение силуэта. Графики позволяют быстро оценить качество кластеризации: видно, что для данных первого класса алгоритм K средних выделил два кластера, один из которых является уверенно определенным ($s > 0.6$), другой же значительно менее удачный. Алгоритм FOREL также определяет два кластера, однако для большинства точек обоих кластеров $s > 0.6$, что соответствует лучшему результату. Дла данных второго класса наилучший по среднему значению силуэта результат, полученный при помощи алгоритма K средних, соответствует четырем кластерам, два из которых являются не слишком удачными. Алгоритм FOREL разбивает точки на пять кластеров, для большей части точек каждого из которых $s > 0.6$, что соответствует значительно лучшему результату. Более высокое качество кластеризации, достигаемое при помощи алгоритма FOREL, можно объяснить тем, что благодаря процедуре применения к центрам кластеров алгоритма кратчайшего замкнутого пути, алгоритм FOREL создает двухуровневую систему кластеров, позволяя, таким образом, строить кластеры сложной формы: например, каждый из пяти кластеров верхнего уровня, полученных при кластеризации данных цилиндра алгоритмом FOREL, состоит из некоторого числа меньших кластеров, радиус которых ограничен параметром R алгоритма.

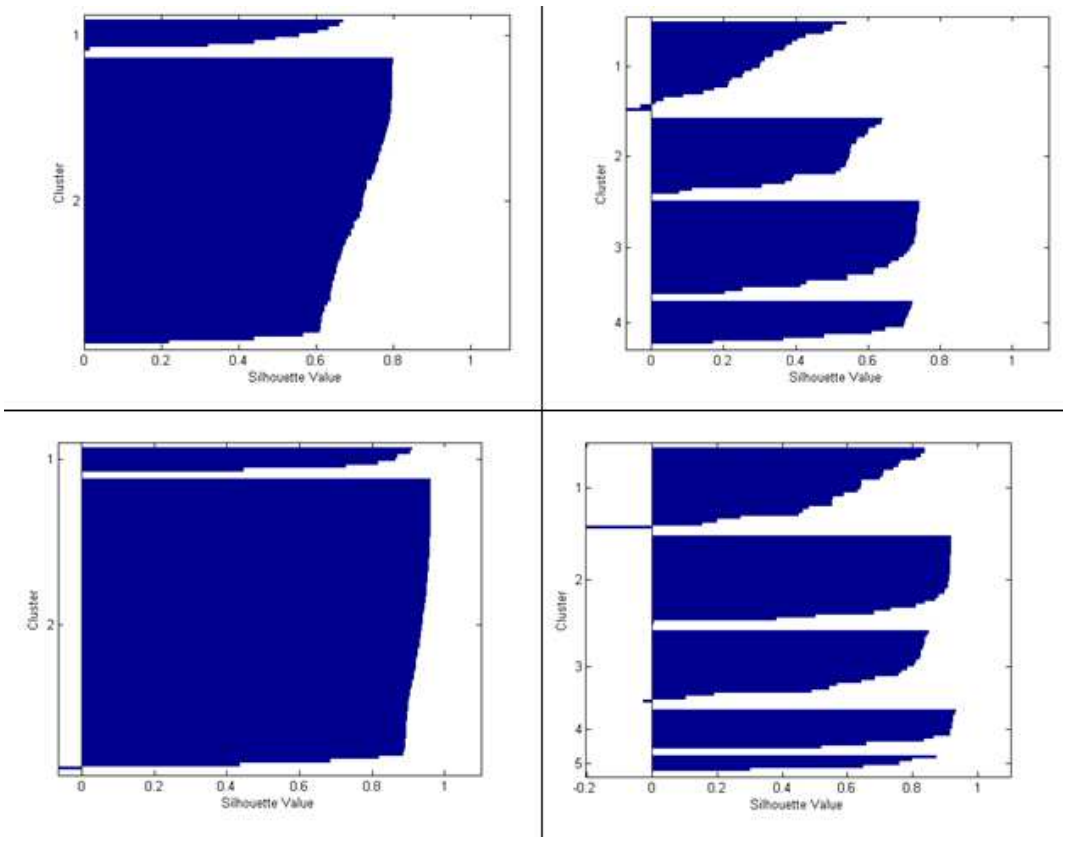

Рис. 5.8. Кластеризация данных при помощи алгоритмов k-means и FOREL.

5.4. Моделирование работы алгоритмов классификации

Для оценки качества классификации, которую позволяет достичь предлагаемый алгоритм в сравнении с другими, использовались два основных метода.

В первом методе тестовая выборка генерировалась путем наложения гауссова шума на данные обучающей выборки. Величину наложенного шума нельзя напрямую соотнести с такими показателями качества приема сигнала, как SNR (Signal to Noise Ratio – отношение сигнал/шум), поскольку точность измерения элементов поляризационной матрицы рассеяния зависит от характеристик приема косвенным образом, однако подобный способ получения тестовой выборки позволяет сравнить качество работы алгоритмов классификации в поставленных условиях и оценить их устойчивость к шуму.

Второй метод заключается в использовании кросс-валидации (метода скользящего контроля) для оценки качества работы алгоритмов. На каждой итерации выборка разделяется на обучающую, которая используется для обучения классификатора, и тестовую, которая классифицируется при помощи обученного алгоритма. Параметром кросс-валидации является величина тестовой выборки (ширина окна скользящего контроля) K, используемая на каждой итерации. Усредненная по количеству итераций ошибка алгоритма на тестовой выборке позволяет оценить out-of-sample error, ожидаемую ошибку алгоритма на новых данных, никак не использованных при обучении классификатора.

5.4.1. Сравнение работы алгоритмов классификации на модельных данных с наложенным шумом

Чтобы оценить качество работы алгоритмов классификации, используя модельные данные в качестве обучающей выборки, можно сформировать тестовую выборку из смешанных модельных данных, включающих в себя объекты обоих классов, с наложенным на них гауссовским шумом (к значениям признаков прибавляется гауссовская случайная величина со средним значением 0 и среднеквадратичным отклонением σ). На Рис. 5.9 представлена зависимость доли ошибочно классифицированных объектов в тестовой выборке от величины σ для разных алгоритмов классификации.

Поскольку при $\sigma \gg 0.1$ величина шума становится соразмерной расстоянию между точками обучающей выборки, доля ошибочно классифицированных точек становится большой для всех алгоритов, хотя Наивный Байесовский классификатор лучше других описывает форму классов в условиях, когда она определяется в равной мере исходными данными и наложенным гауссовым шумом. Видно, что при уровне шума, искажающем данные в меньшем масштабе, хорошо работает метод опорных векторов с параметром $\sigma_{rbf} = 100$, а также предлагаемый алгоритм классификации на основе кластерной структуры данных каждого класса, причем последний позволяет добиться большей точности

классификации. Эти алгоритмы характеризуются тем, что дают возможность точно описать форму классов, которые образуют точки обучающей выборки, создавая сложную разделяющую поверхность между ними.

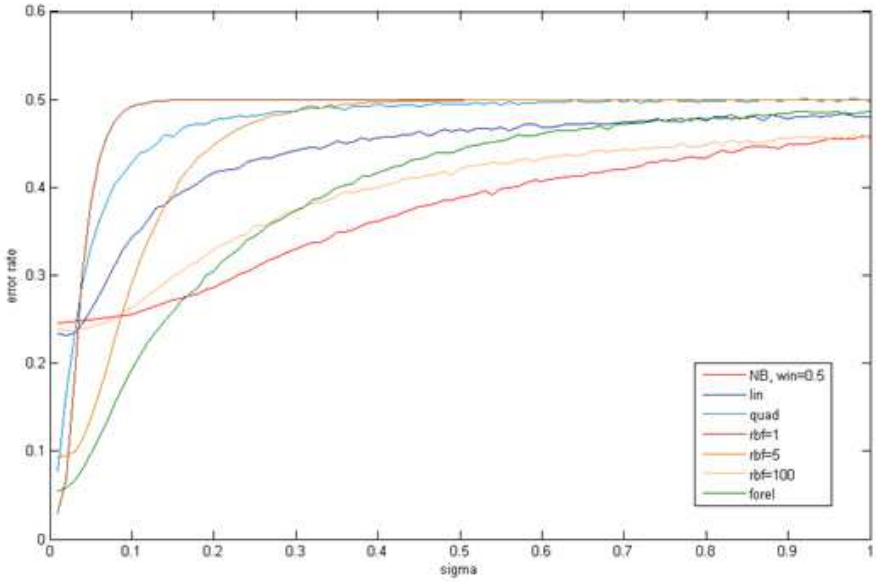

Рис. 5.9. Сравнение качества работы алгоритмов классификации на данных модели с наложенным шумом

5.4.2. Сравнение качества работы алгоритмов классификации при помощи кросс-валидации

В таблице 5.1 приведены оценки качества классификации данных, полученные при помощи кросс-валидации. Дано сравнение результатов классификации (процента неправильно классифицированных точек) для рассмотренных алгоритмов в случае использования данных в исходном пространстве признаков ($no\ PCA$, $dim = 6$), а также данных в пространстве главных компонент с сохранением 6, 4 и 3 размерностей. Параметр ширины окна кросс-валидации был выбран равным единице (каждая точка выборки классифицировалась при помощи алгоритма, обученного на остальных точках), т.е. использовался метод Leave One Out. В то время как большая величина окна позволила бы получать меньший разброс результатов между итерациями кросс-валидации, использование Leave One Out дает возможность наиболее точно оценить величину ошибки классификатора на внешних данных.

	No PCA, dim = 6	**PCA, dim = 6**	**PCA, dim = 4**	**PCA, dim = 3**
NB	18.4%	16.8%	16.6%	20.4%
SVM	12.1%	9.8%	7.4%	11.7%
clustclass	11.2%	6.9%	6.6%	11.1%

Таблица 5.1. Доля ошибочно классифицированных точек при использовании различных алгоритмов классификации и числа главных компонент. NB – Наивный Байесовский классификатор; SVM – Support Vector Machine; clustclass – предлагаемый алгоритм классификации; PCA – метод главных компонент, dim – размерность данных.

Для каждого алгоритма классификации были выбраны параметры, которые обеспечивают наилучший результат (позволяют добиться наименьшего процента ошибок):

1. для наивного Байесовского классификатора плотность распределения оценивалась при помощи окна Епанечникова (оптимального) с шириной 0.06 для исходного пространства признаков и 0.4 для пространства главных компонент;

2. для метода опорных векторов использовалось rbf-ядро с параметром $\sigma = 0.4$ для исходного пространства признаков и $\sigma = 1.3$ для пространства главных компонент.

Предложенный алгоритм обеспечивает наименьшую долю ошибок классификации, поскольку совокупность кластеров, полученная при помощи приведенной выше модификации алгоритма FOREL, позволяет точно описать структуру данных в пространстве признаков. Использование метода главных компонент дает возможность сокращения размерности данных с 6 до 3 измерений без потери качества классификации, а также повышения качества классификации при сохранении 4 и более измерений.

Глава 6

Заключение

В работе рассмотрено применение алгоритмов классификации данных к задаче определения типа радиолокационного объекта по значениям элементов поляризационной матрицы рассеяния. Предложен алгоритм, который использует кластерную структуру обучающей выборки для построения классификатора; проведено моделирование работы рассмотренных алгоритмов классификации с использованием данных поляризационной матрицы рассеяния, полученных при помощи фасеточной модели. Результаты моделирования показали, что предложенный алгоритм классификации на основе кластерной структуры данных показывает в условиях данной задачи лучшие результаты, чем наивный Байесовский классификатор и метод опорных векторов, поскольку позволяет точно описать структуру данных в пространстве признаков при помощи совокупности кластеров. Использование метода главных компонент вместе с предложенным алгоритмом дает возможность как сократить размерность данных без потери качества классификации, так и улучшить качество классификации при сохранении большей размерности данных. Предложенный метод преобразования признаков и классификации позволяет классифицировать радиолокационный объект на основе одного измерения элементов поляризационной матрицы рассеяния, без использования какой-либо априорной информации о ракурсе объекта. Предлагаемый алгоритм, так же, как и другие рассмотренные в работе алгоритмы классификации, позволяет определить как предполагаемый класс объекта, так и меру близости объекта к каждому из классов, что дает возможность расширить алгоритм на случай классификации между большим количеством классов, а также на случай, когда объект, не принадлежащий никакому из известных классов с необходимой долей достоверности, считается неизвестным. Таким образом, предлагаемый алгоритм является практически применимым в условиях поставленной задачи.

Литература

1. Олюнин Н.Н., Виноградов А.Г., Сазонов В.В. Фасеточная модель в задачах рассеяния радиолокационных сигналов. Препринт РТИ 0702. – М., 2007. – 21 с.

2. Верденская Н.В., Иванова И.А., Сазонов В.В. Сравнение радиолокаторов с различными видами поляризационного излучения и приема при использовании оптимального алгоритма обнаружения и при фиксированной излучаемой мощности. Препринт РТИ 0704. - М., 2010. - 23 с.

3. Айвазян С.А. [и др.]. Прикладная статистика. Классификация и снижение размерности. – М: Финансы и статистика, 1989. – 607 с.

4. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин).

5. Burges C.J.C. A tutorial on support vector machines for pattern recognition. – Data Mining and Knowledge Discovery, 1998. – Vol. 2, no. 2. – с. 121–167.

6. MacQueen J.B. Some Methods for classification and Analysis of Multivariate Observations. – Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967.

7. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.