5. Экспериментальное исследование алгоритмов обучения

5.1 Структура экспериментального материала

Модельные данные.

Для проверки качества решающего правила использованы следующие данные. Выборки образованы двумя нормальными распределениями с одинаковыми ковариационными матрицами в 25-мерном пространстве. Каждый объект выборки представляет собой последовательность из 25 отсчетов гладкой функции с наложенными возмущениями. Гладкость понимается в обычном смысле, т.е. функция является гладкой, если она плавно меняется при плавном изменении аргумента. В качестве таких функций выбраны синусоидальные гармоники одинаковой амплитуды, но разной частоты. Такие данные представляют собой сферические распределения, координаты центров которых определены значениям первых 25 отсчетов гладкой функции. Сами классы линейно неразделимы.

m_i^1

x_i^1

а б

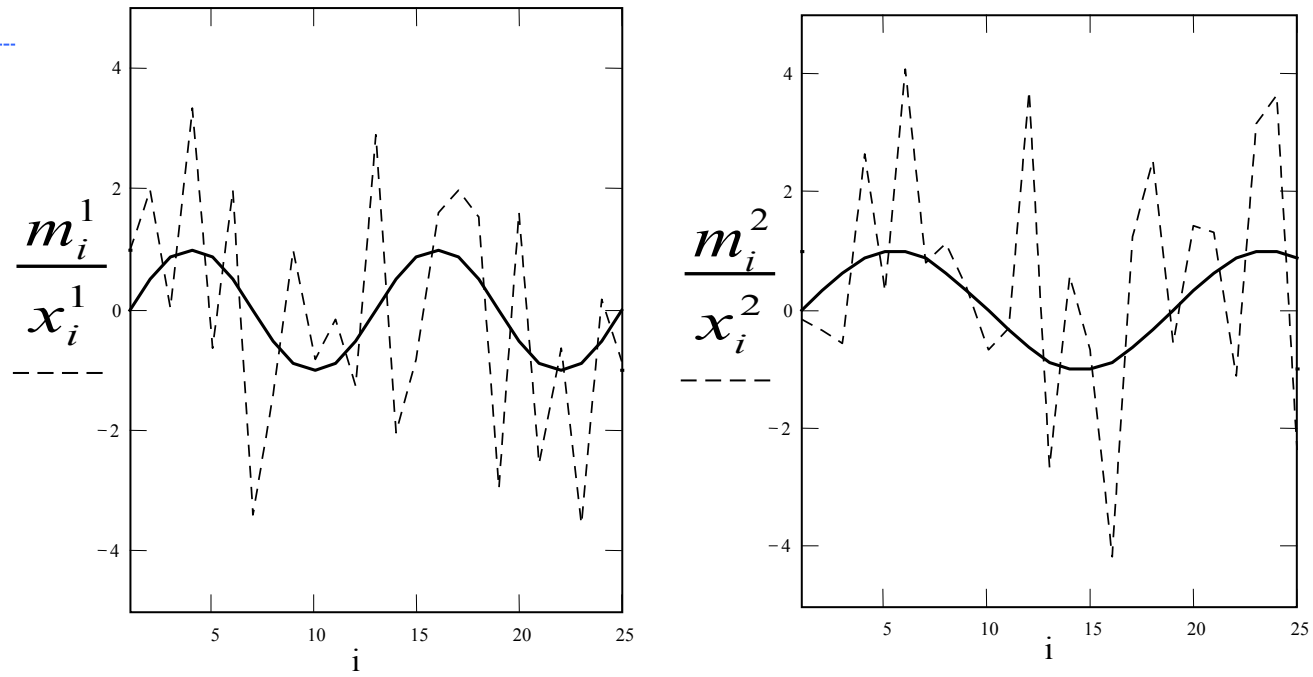

Рис.5 1.1 Отдельные признаки в составе векторов центров классов и отдельных объектов

На каждом рисунке сплошной линией показаны значения признаков центра класса, а пунктиром - значения признаков одного из объектов выборки класса.

Рукописные символы

Наглядным примером задачи распознавания образов в метрическом пространстве является задача распознавания рукописных символов, например, букв, при их вводе в ЭВМ непосредственно в процессе написания с помощью специального пера. При таком способе ввода каждый символ первоначально оказывается представленным сигналом, состоящим из двух компонент, а именно, текущих координат пера по вертикали и горизонтали, однако, может оказаться целесообразным использовать и дополнительные локальные характеристики процесса написания, например, угловой азимут мгновенного направления движения пера, его скорость, силу прижатия к бумаге, временные отрывы от нее, наклон и т.п. В качестве аргумента сигнала может выступать либо время, либо длина пути, пройденного пером от точки первого касания бумаги. На рис.5.1.2 представлен трехкомпонентный сигнал в виде функций координат и азимута движения пера от пройденного пути вдоль его траектории, полученный при написании рукописной буквы “d”.

В этой задаче трудно указать заранее фиксированное число признаков сигнала, которые могли бы сформировать пространство, удовлетворяющее гипотезе компактности. Нельзя использовать в качестве признаков и отсчеты сигнала, взятые с некоторым шагом вдоль оси аргумента, поскольку сигналы, полученные от разных написаний даже одного и того же символа неизбежно будут иметь разную длину, и, следовательно, не существует единого линейного пространства, в котором могли бы быть представлены написания распознаваемых символов [43].

Заметим, что разные варианты написания одного и того же символа естественно представить как результат некоторого нелинейного преобразования оси аргумента, приводящего к ее «короблению». Эти различия между разными написаниями, несущественные с точки зрения распознавания символов, легко компенсировать с помощью процедуры так называемого парного выравнивания (рис. 5.1.2), тогда остающееся несовпадение сигналов будет нести информацию об «истинной» непохожести сигналов, которую естественно принять в качестве рабочей метрики при построении процедуры обучения распознаванию символов.

На рис.5.1.3 приведен кадр программы, иллюстрирующий различие между введенной буквой *a* и эталонами букв *a* и *b*.

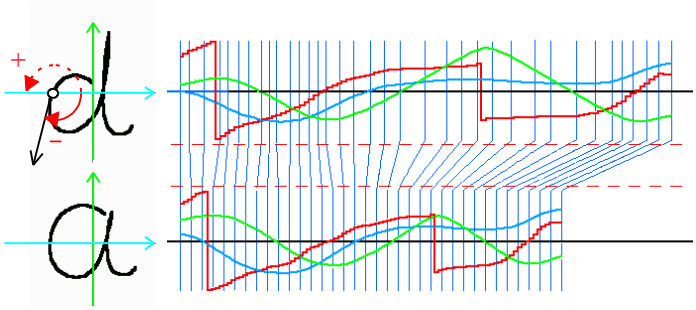

Рисунок. 5.1.2 Два сигнала в виде функций координат и азимута движения пера от пройденного пути вдоль его траектории, полученные при вводе рукописных символов в компьютер непосредственно в процессе написания. Совмещение произведено по значениям азимута.

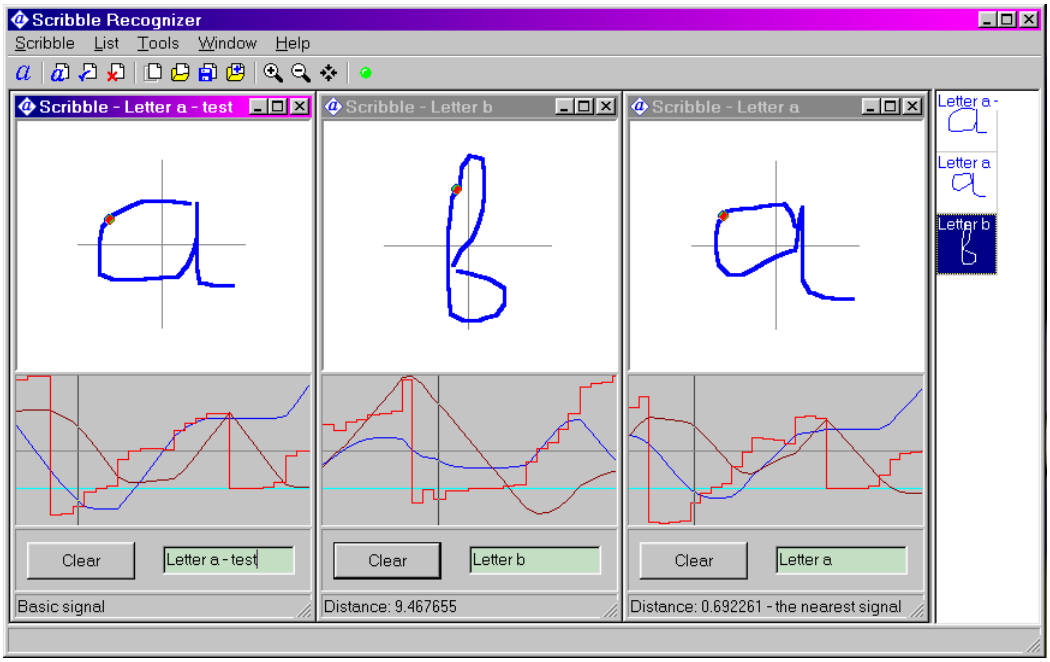

Рисунок 5.1.3. Экран работы программы “Scribble recognizer”. Нижняя часть окна показывает численное значение близости вновь введенного символа “a” к шаблонам символов “a” и “b”

Другим примером задач, требующих введения метрического пространства распознавания является задача классификации пространственной структуры белков опираясь на знание лишь первичной структуры (последовательности аминокислот). Информация о пространственной организации белка (третичная структура) является очень важной для понимания механизмов работы макромолекул и их функций. Третичную структуру макромолекул определяют экспериментальными методами (рентгеноструктурный анализ, ядерный магнитный резонанс). Эти методы являются чрезвычайно трудоемкими и требуют больших затрат времени, но позволяют получить достоверные сведения о пространственной организации молекул. Характерно, что для больших групп эволюционно родственных белков, подчас очень значительно отличающихся по первичной структуре и, значит, по распределению всех атомов в пространстве, способ укладки полипептидной цепи остается в главных чертах неизменным. С

другой стороны, при всем разнообразии пространственных структур белков удается выделить относительно небольшое число типов укладки полипептидной цепи. Налицо задача классификации – выделение групп белков, достаточно близких друг к другу по пространственной структуре. Один из фундаментальных принципов молекулярной биологии говорит о том, что последовательность аминокислотных остатков полипептидной цепи белка несет в себе всю информацию, необходимую и достаточную для формирования однозначной пространственной структуры. Учитывая это положение, в настоящее время большие усилия прилагаются для разработки методов предсказания третичной структуры молекул на основе известной первичной структуры. Разумеется, абсолютно точно произвести такой прогноз невозможно, остается надежда на то, чтобы правильно «угадать» группу к которой относится исследуемый белок.

В качестве примера, на рис. 5.1.4 представлено схематичное трехмерное изображение белка Cytochrome C4 и его первичная структура. На сегодняшний момент биологи выделяют определенные типы (семейства, фолды) известных пространственных структур [44]. Положительным результатом процедуры распознавания считается достоверное отнесение белка к одному из таких классов. Налицо задача распознавания образов.

Пространственная структура

Первичная структура:

```
1    AGDAEAGQGK
11   VAVCGACHGV
21 DGNSPAPNFP
31   KLAGQGERYL
41   LKQLQDIKAG
51 STPGAPEGVG
61   RKVLEMTGML
71 DPLSDQDLED
81   IAAYFSSQKG
91   SVGYADPALA
101  KQGEKLFRGG
111  KLDQGMPACT
121  GCHAPNGVGN
131  DLAGFPKLGG
141  QHAAYTAKQL
151  TDFREGNRTN
161  DGDTMIMRGV
171  AAKLSNKDIE
181  ALSSYIQGLH
```

Рисунок. 5.1..4. 3D представление и первичная структура белка Cytochrome C4

Имея в наличии информацию только о первичной структуре белков входящих в изучаемые данные, первым шагом представляется попытка получить некоторые количественные характеристики, которые бы отражали существо пространственной классификации. В настоящее время открыто более четырехсот признаков аминокислот (наиболее важные - гидрофобность, степень поляризации, размер и др.), однако, прямое использование этих признаков затруднено тем, что различные протеины имеют разную длину, и, следовательно, непосредственное представление первичных структур как векторных "сигналов" их свойств потребует учета специфики работы с задачами такого типа. Другой простейший подход получения количественных признаков из аминокислотной последовательности заключается в подсчете относительного числа остатков каждой из аминокислот к общей длине последовательности. В таком случае каждый протеин будет представлен точкой в двадцатимерном пространстве. Однако наиболее разумной представляется схема, использующая знания о взаимной близости аминокислотных последовательностей. Существуют априорные объективные данные о близости в химико-биологическом смысле всех пар аминокислот (210 – пар, включая близость «самой с собой»), которые обычно выражаются в виде матрицы соответствия 20×20. Для двух аминокислотных последовательностей пытаются найти такое их взаимное соответствие, чтобы величина «невязки» близостей аминокислот была по возможности минимальной. При этом, для белков различной длины более короткий приходится искусственно «вытягивать» за счет введения в определенные позиции делеций, т.е. разрывать исходную структуру. Результат такой процедуры обычно выражается величиной несходства выравниваемых последовательностей. Опираясь на какую либо процедуру выравнивания последовательностей, например Fasta3 (ftp://ftp.virginia.edu/pub/fasta), строится матрица всех взаимных расстояний между первичными структурами. Такая матрица и рассматривается как метрическое пространство.

5.2 Схема эксперимента

Оценка качества решающего правила

В настоящей работе мы не затрагивали проблему оценивания классификаторов и выбор такой оценки. Обычно, о надежности алгоритма судят по качеству распознавания, определяемому как отношение числа верно распознанных объектов к их общему числу. Достаточно часто, особенно в

зарубежной литературе, применяется термин противоположный по смыслу качеству распознавания, а именно величина (вероятность) ошибки (error rate) распознавания. Интуитивно понятно, что разработчики процедур обучения распознаванию образов стремятся улучшить качество распознавания, или, соответственно, снизить степень ошибки предлагаемого ими классификатора. Необходимо понимать, что определяемое таким образом качество распознавания зависит от выборки, на которой проводится обучение. Например, если в последовательности много раз встречается ситуация, которую машина классифицирует не так, как учитель, то процент несовпадений будет велик, в то время как при другом составе последовательности он может оказаться мал. Поэтому необходимо заранее условиться, как будет определяться качество решающего правила, т.е. по какой последовательности будет исчисляться процент несовпадений.

Наименьшая, теоретически возможная вероятность ошибки распознавания определяется т.н. ошибкой байесовского классификатора, или просто байесовской ошибкой [20].

В реальных задачах распознавания образов, часто возникает ситуация когда классы не разделимы полностью, тогда одной из проблем статистической теории распознавания является проблема определения наилучшего возможного качества распознавания. Именно байесовский классификатор и дает такую оценку. Пусть для каждого класса Ω_k существует некоторая априорная вероятность его появления $p(\Omega_k)$, $k = 1, 2$ и в n-мерном пространстве R^n задана условная плотность распределения $p(\mathbf{x} \mid \Omega_k)$ вектора $\mathbf{x}$ относительно каждого класса Ω_k. Тогда байесовское решающее правило определяется как

$$p(\Omega_k \mid \mathbf{x}) = \frac{p(\Omega_k) p(\mathbf{x} \mid \Omega_k)}{p(\mathbf{x})}, \quad p(\mathbf{x}) = \sum_{j=1}^{2} p(\Omega_k) p(\mathbf{x} \mid \Omega_j), \tag{5.2.1}$$

а классификатор, который относит вектор $\mathbf{x}$ к классу с наибольшей апостериорной вероятностью, называемый байесовским, имеет вид

$$p(\Omega_1) p(\mathbf{x} \mid \Omega_1) \underset{<}{\overset{>}{-}} p(\Omega_2) p(\mathbf{x} \mid \Omega_2) \rightarrow \mathbf{x} \in \begin{cases} \Omega_1, \\ \Omega_2 \end{cases} \tag{5.2.2}$$

и соответствующая ему ошибка называется байесовской ошибкой классификации:

$$E_b = 1 - \sum_{i=1}^{2} \int_{\Omega_i} p(\Omega_i) p(\mathbf{x} \mid \Omega_i) d\mathbf{x}. \tag{5.2.3}$$

Следует отметить, что байесовская ошибка, определяемая выражением (5.2.3) предполагает интегрирование по многомерной, не всегда точно определенной условной плотности распределения, поэтому в явном виде байесовская ошибка классификации может быть вычислена непосредственно лишь для небольшого круга задач. В частности это удается сделать, когда распределения являются нормальными с одинаковыми ковариационными матрицами. Покажем это. Решающее правило (5.2.2) можно переписать в виде

$$l(\mathbf{x}) = \frac{p(\mathbf{x}|\Omega_1)}{p(\mathbf{x}|\Omega_2)} \underset{<}{\overset{>}{-}} \frac{p(\Omega_2)}{p(\Omega_1)} \rightarrow \mathbf{x} \in \begin{cases} \Omega_1, \\ \Omega_2 \end{cases} \qquad (5.2.4)$$

Величину $l(\mathbf{x})$ называют отношением правдоподобия. Величину $p(\Omega_2)/p(\Omega_1)$ называют пороговым значением отношения правдоподобия для данного решающего правила. Нам будет удобнее вместо отношения правдоподобия $l(\mathbf{x})$ удобно использовать величину $-\ln l(\mathbf{x})$. В этом случае решающее правило (5.2.4) примет вид

$$h(\mathbf{x}) = -\ln l(\mathbf{x}) = -\ln p(\mathbf{x}|\Omega_1) + \ln p(\mathbf{x}|\Omega_2) \underset{>}{\overset{<}{-}} \ln\{p(\Omega_1)/p(\Omega_2)\} \rightarrow \mathbf{x} \in \begin{cases} \Omega_1 \\ \Omega_2 \end{cases} \qquad (5.2.5)$$

Направление неравенства изменилось, потому что использовалось отрицательное значение логарифма. Уравнения (5.2.4) и (5.2.5) называют байесовским критерием, минимизирующим ошибку решения.

Если $p(\mathbf{x}|\Omega_k)\ k = 1,2$ - нормальная случайная величина с вектором математического ожидания $\mathbf{M}_k$ и ковариационной матрицей Σ_k, то решающее правило (8) приобретает вид

$$h(\mathbf{x}) = -\ln l(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{M}_1)^T \Sigma_1^{-1}(\mathbf{x}-\mathbf{M}_1) - \frac{1}{2}(\mathbf{x}-\mathbf{M}_2)^T \Sigma_2^{-1}(\mathbf{x}-\mathbf{M}_2) +$$

$$\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \underset{>}{\overset{<}{-}} \ln\{p(\Omega_1)/p(\Omega_2)\} \rightarrow \mathbf{x} \in \begin{cases} \Omega_1 \\ \Omega_2 \end{cases} . \qquad (5.2.6)$$

Уравнение (5.2.6) показывает, что решающая граница является квадратичной формой относительно вектора **x**. В случае равных ковариационных

матриц $\Sigma_1 = \Sigma_2 = \Sigma$ граница становится линейной функцией относительно вектора **x**:

$$h(\mathbf{x}) = (\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2}\left(\mathbf{M}_1{}^T \Sigma^{-1} \mathbf{M}_1 - \mathbf{M}_2{}^T \Sigma^{-1} \mathbf{M}_2\right)$$
$$\lessgtr \ln\{p(\Omega_1)/p(\Omega_2)\} \rightarrow \mathbf{x} \in \begin{cases} \Omega_1 \\ \Omega_2 \end{cases} \qquad (5.2.7)$$

Поскольку решающее правило (5.2.7) представляет собой линейно преобразование *n*-мерного пространства в одномерное, то если **x** является нормально распределенным случайным вектором, решающее правило $h(\mathbf{x})$ также будет нормальной случайной величиной. Поскольку $E\{\mathbf{x}|\Omega_i\} = \mathbf{M}_i$, то математическое ожидание и дисперсию $h(\mathbf{x})$ можно вычислить следующим образом:

$$\eta_i = E\{h(\mathbf{x})|\Omega_i\} = (\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1} E\{\mathbf{x}|\Omega_i\} + \frac{1}{2}\left(\mathbf{M}_1^T \Sigma^{-1} \mathbf{M}_1^T - \mathbf{M}_2^T \Sigma^{-1} \mathbf{M}_2^T\right) \quad (5.2.8)$$

$$\eta_1 = -\frac{1}{2}(\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1}(\mathbf{M}_2 - \mathbf{M}_1) = -\eta \qquad (5.2.9)$$

$$\eta_2 = +\frac{1}{2}(\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1}(\mathbf{M}_2 - \mathbf{M}_1) = +\eta \qquad (5.2.10)$$

$$\sigma_i^2 = E\left[\{h(\mathbf{x}) - \eta_i\}^2 |\Omega_i\right] = E\left[\left\{(\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1}(\mathbf{x} - \mathbf{M}_i)\right\}^2 |\Omega_i\right] =$$
$$= (\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1} E\left\{(\mathbf{x} - \mathbf{M}_i)(\mathbf{x} - \mathbf{M}_i)^T |\Omega_i\right\} \Sigma^{-1}(\mathbf{x} - \mathbf{M}_i) = \qquad (5.2.11)$$
$$(\mathbf{M}_2 - \mathbf{M}_1)^T \Sigma^{-1}(\mathbf{x} - \mathbf{M}_i) = 2\eta$$

На рис. 5.2.1. изображены плотности вероятности решающего правила $h(\mathbf{x})$, причем заштрихованные площади соответствуют вероятностям ошибки, обусловленным байесовским критерием, который минимизирует ошибку решения. Эти вероятности ошибок соответственно равны

$$\varepsilon_1 = \int_t^\infty p(h|\Omega_1)dh = \int_{(\eta+t)/\sigma}^\infty (2\pi)^{\frac{1}{2}} \exp\left(-\frac{\xi^2}{2}\right) d\xi = \frac{1}{2} - \Phi\left(\frac{\eta + t}{\sigma}\right) \qquad (5.2.12)$$

$$\varepsilon_2 = \int_{-\infty}^t p(h|\Omega_2)dh = \int_{(\eta-t)/\sigma}^\infty (2\pi)^{\frac{1}{2}} \exp\left(-\frac{\xi^2}{2}\right) d\xi = \frac{1}{2} - \Phi\left(\frac{\eta - t}{\sigma}\right) \qquad (5.2.13)$$

где

$$t = \ln\{p(\Omega_2)/p(\Omega_1)\},$$
$$\sigma^2 = \sigma_1^2 = \sigma_2^2 = 2\eta.$$

Общая ошибка считается как сумма ошибок по каждому классу

$$\varepsilon = \varepsilon_1 + \varepsilon_2 \qquad (5.2.14)$$

Таким образом, если плотность вероятности отношения правдоподобия является нормальной, то вероятности ошибки можно вычислить пользуясь таблицей функции Лапласа $\Phi(\cdot)$, так как отношение правдоподобия – одномерная нормальная случайная величина. Например для используемых здесь данных наименьшая теоретически возможная вероятность ошибки составляет $P_{\min} = 2\%$

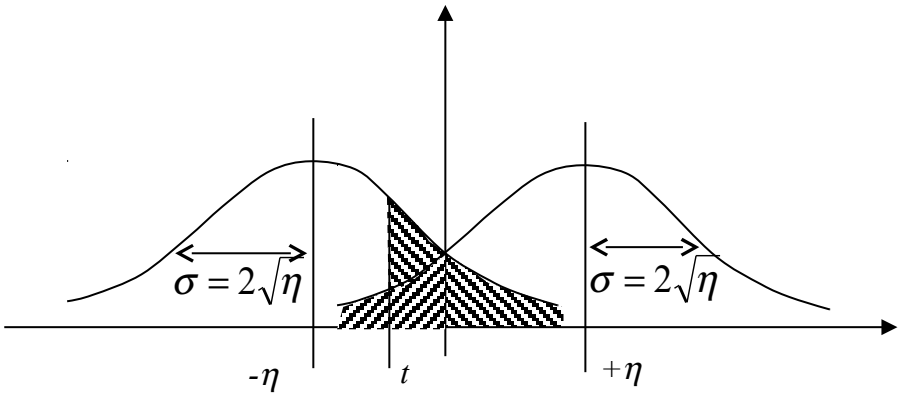

Рисунок 5.2.1. Плотность вероятности решающего правила $h(\mathbf{x})$.

Процедура проверки по генеральной совокупности

Генерируются две выборки: обучающая с малым числом объектов и контрольная с большим числом объектов каждого класса. Затем при некотором α по малонаполненной выборке проводится обучение, а полученное решающее правило проверялось на контрольной выборке с большим числом объектов. На обеих выборках по отношению числа неверно классифицированных объектов к их общему числу определялается вероятность ошибки распознавания в процентах. Далее величина штрафа α увеличивается, и эксперимент повторяется. Согласно предлагаемой гипотезе, при увеличении штрафа на негладкость вероятность ошибки распознавания на обучающей выборке должна увеличиваться из-за сужения допустимой области выбора решающего правила, а на контрольной выборке, наоборот, должна уменьшаться, из-за того, что регуляризованное решающее правило более реально отражает особенности генеральной совокупности.

Процедура «скользящий контроль»

Для эмпирической оценки качества распознавания М.Н. Вайнцвайгом была предложена процедура, впоследствии получившая название "скользящий контроль". Суть ее заключается в следующем. Из выборки удаляется один элемент и по оставшимся объектам проводится обучение. Затем на основании

полученного решающего правила удаленный элемент классифицируется. Если результат классификации совпадает с истинным классом объекта, то считается, что алгоритм дал верный результат, иначе - ошибся. Затем выбранный объект возвращается в выборку, из нее удаляется другой объект, и эксперимент повторяется. Такая процедура проводится над всеми точками выборки. Отношение числа ошибочно классифицированных векторов к размеру выборки и оценивает качество решающего правила.

В [12, 26] показано, что оценки скользящего контроля являются несмещенными, т.е. математическое ожидание результата контроля равно истинной величине качества. Полагают, хотя строгого доказательства этого утверждения не существует, что для большинства практически важных случаев дисперсия оценки "скользящий контроль" стремится к нулю с увеличением размера выборки примерно так же быстро, как дисперсия "экзамена". Отыскание дисперсии оценки метода "скользящего контроля" является одной из актуальных задач не только теории обучению распознаванию образов, но и теоретической статистики.

5.3 Результаты исследования эффективности штрафа на негладкость

Во избежание влияния на результат специфических особенностей малонаполненных выборок эксперимент проводится на пяти обучающих выборках с одинаковым числом элементов, а значение вероятности ошибки распознавания усредняется по ним. Для оценки того, как размер выборки влияет на эффективность штрафа на негладкость, опыты повторяются для различного числа элементов N в выборке, которое было как больше размерности признакового пространства, так и меньше его..

Результаты эксперимента представлены на рис. 5.3.1, где показана зависимость вероятности ошибки распознавания от величины штрафа на негладкость.

Анализируя эти зависимости, можно сделать следующие выводы. Во-первых, наибольший эффект штраф на негладкость дает для случая малонаполненных выборок, и эффективность его падает при увеличении числа объектов в выборке. Во-вторых, при увеличении α вероятность ошибки распознавания сначала уменьшается до некоторого значения, а затем начинает увеличиваться. Это связано с тем, что на область допустимых значений накладывается слишком сильное ограничение, и решающее правило становится очень “грубым”.

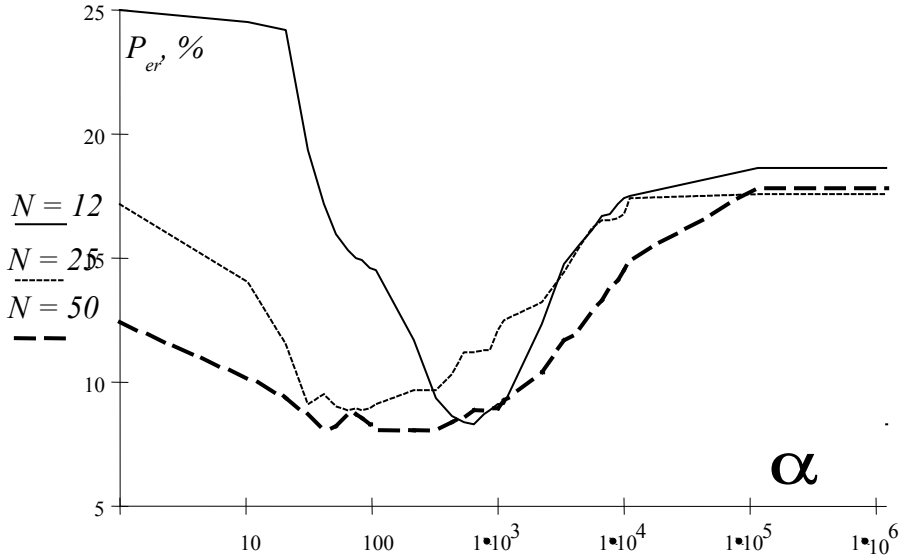

Рис. 5.3.1 Зависимость вероятности ошибки P_{er} от величины штрафа α на негладкость решающего правила.

5.4 Результаты исследования эффективности обучения распознаванию образов в метрическом пространстве.

Были сгенерированы обучающие выборки для любых двух строчных букв английского алфавита. Каждая выборка состояла из 30 символов, по пятнадцать объектов в каждом классе. Для всех выборок в метрическом пространстве было построено два решающих правила распознавания: без учета регуляризации и с учетом регуляризации. Для обоих решающих правил с помощью процедуры «скользящий контроль» была подсчитана ошибка распознавания. Результаты экспериментов представлены на рис. 5.4.1.

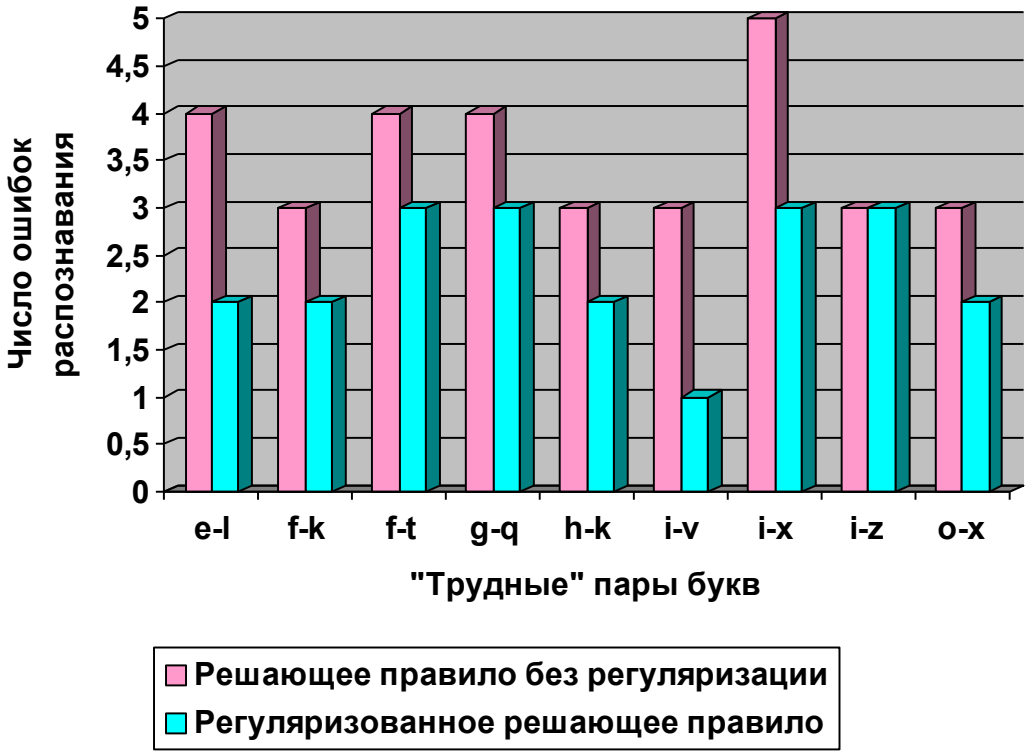

Рисунок 5.4.1 Иллюстрация необходимости регуляризации решающего правила распознавания в метрическом пространстве.

На диаграмме даны результаты процедуры «скользящих контроль» для девяти пар букв, распознавание которых в метрическом пространстве представляет собой особую сложность. Высота розовых колонок соответствует числу ошибочно классифицированных объектов в случае решающего правила без регуляризации, в то время как высота голубых колонок показывает, как ошибка распознавания может быть снижена за счет регуляризации.