**2.2 Статистические алгоритмы распознавания**

1.

Рассмотрим подход к построению алгоритмов распознавания, основанный на использовании статистического анализа.

Cтатистический анализ играет в распознавании важную роль в случаях, когда случайные факторы оказывают влияние на порождение объектов.

***Пример.***

Случайный характер природных процессов, шум аппаратуры дистанционного зондирования, некоторая неопределенность относительно правильной идентификации объектов обучающей выборке и т. д.

‒ факторы, позволяющие рассматривать набор признаков *,*  как п-мерную случайную величину, а принадлежность объекта к какому-либо классу с некоторой долей вероятности.

2.

*Статистический анализ* предоставляет возможность:

построения алгоритма оптимального в том смысле, что его применение обеспечивает в среднем наименьшую вероятность совершения ошибок распознавания.

В методах *статистического распознавания* обычно используют:

функции распределения вероятностей, связанные с классами объектов. Эти функции, как правило, не известны и должны оцениваться с помощью обучающей выборки.

Покажем, как можно использовать теорию статистических решений для построения *алгоритмов распознавания*.

3.

***Пусть:***

- имеется классов ;

- объект ‒ случайная *n*-мерная величина.

***Сделаем следующие предположения:***

) - известна или может быть оценена;

*-* известна или может быть оценена по  
обучающей выборке;

- не известна.

4.

‒ априорная вероятность класса ,вероятность наблюдения объекта из независимо от любой другой информации;

‒ функция плотности вероятностей, зависящая от объекта при условии, что он принадлежит ,

‒ вероятность принадлежности к классу (апостериорная вероятность).

5.

Если алгоритм распознавания (классификатор) принимает решение, что объект , а он принадлежит ,то классификатор несет потери .

Так как может принадлежать любому из классов, то матожидание потерь, связанных с решением можно определить выражением:

Эта величина называется *условным средним риском*.

5.

Работа алгоритма распознавания заключается в получении решения, которое бы минимизировало условный риск.

Для этого необходимо выполнить следующие действия:

**Шаг 1.** Для каждого объекта вычисляют значения условного риска

.

**Шаг 2.** Принимается решение, что ,

если для любых ).

6.

Таким образом, при построении описанного алгоритма используются решающие функции вида

Классификатор минимизирует условные потери.

Математическое ожидание полных потерь на множестве всех решений также будет минимизировано.

Такой классификатор соответствует оптимальному качеству распознавания и называется ***Байессовским классификатором.***

7.

Для реализации алгоритмов с решающими функциями (2) необходимо определить:

- способ задания функции потерь  *;*

- эффективные выражения для вычисления неизвестной величины .

Согласно известной формуле Байесса

где ‒ вероятность появления объекта *X*.

8.

Тогда

Множитель является общим для и им можно пренебречь.

Расчеты упрощаются, если функцию потерь выбрать в виде:

‒ символ Кронекера:

9.

Величина называется нуль-единичной функцией потерь:

- стоимость потерь равна нулю при правильном распознавании;

- равна единице при ошибке.

По закону вероятности:

10.

Подставляя (4) в (3), получаем

минимизируется по *j* максимизацией члена

Таким образом, при построении алгоритмов распознавания в качестве решающих можно использовать функции вида

11.

Решение, полученное с помощью таких функций, называется решением по максимуму правдоподобия.

На практике удобно использовать другое выражение:

которое приводит к тому же самому решению, так как логарифм является монотонно возрастающей функцией.

12.

*Функции распределения* не известны - могут быть оценены с помощью объектов *обучающей выборки*.

Вид функции распределения известен - *по обучающей* *выборке оцениваются параметры функций* (математическое ожидание, дисперсия и др.).

Хорошей моделью многих статистических процессов является многомерное нормальное распределение:

- просто реализуется на ЭВМ;

- удобно при аналитической обработке.

13

Пусть функция распределения вероятностей, связанных с классами , представляет многомерные нормальные плотности,

тогда решающая функция (6) примет вид:

14.

‒ *n*-мерный вектор среднего значения;

/ ‒ математическое ожидание данных *i*-гo канала сканера для класса

‒ ковариационная матрица размерности

‒ ковариация *i* и *k* каналов по классу

‒ транспонированный вектор;

‒ матрица, обратная , а ‒ ее детерминант.

На практике математическое ожидание и ковариационные матрицы классов не известны, но могут быть оценены по обучающей выборке.

15.

Пусть и ‒ несмещенные оценки соответственно для , и ,

тогда эти оценки имеют вид:

, — число объектов обучающей выборки из класса , т. е., используя полученные оценки, можно оценивать и .

16.

*Замечание.*

Для оценки матожидания и ковариационных матриц классов, необходимо иметь обучающую выборку соответствующей мощности.

Например, для данных с *n* диапазонов длин волн минимальное число обучающих объектов для класса должно быть , иначе ковариационная матрица будет вырожденной.

*На практике* для хорошей оценки параметров, необходимо иметь число объектов в диапазоне .

17.

Если тип распределения неизвестен, тогда необходимо оценить .

Одним из способов решения является *функциональная аппроксимация.*

Покажем как аппроксимировать с помощью множества функций:

(7)

‒ неизвестные коэффициенты;

‒ множество заданных базисных функций (например, полиномы Эрмита).

18.

Необходимо определить коэффициенты , для которых значение квадратичной оценки:

для всех из класса минимально.

Необходимое условие минимума для ‒

19.

Доказано, что если базисные функции ортонормированы, то

После определения коэффициентов, с помощью формулы (7) формируется оценка плотности распре деления *.*

Если объекты обучающей выборки поступают последовательно, то коэффициенты можно определить итеративно по формуле:

и - коэффициенты, полученные с помощью ; и объектов обучающей выборки соответственно.

20

***Замечание.***

Качество аппроксимации с помощью выбранной системы базисных функций зависит от *т*членов разложения.

Если для некоторой оценки качество алгоритма распознавания оказывается неудовлетворительным, то следует увеличить число базисных функций.

При отсутствии априорной информации о характере плотности распре деления базисные функции надо выбирать, исходя из простоты их реализации.