УДК 681.3.06:621.391

ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ 4-х АЛГОРИТМОВ РАСПОЗНАВАНИЯ ОБРАЗОВ

А.Н.Манохин, Н.Г.Старцева

Введение

В работе рассматривается проблема сравнения алгоритмов обучения и распознавания образов. Актуальность этой проблемы подчеркивалась на нескольких конференциях по распознаванию, в частности, на Всесоюзной конференции МОЗ в 1981 г. Обусловлена она тем, что в настоящее время имеется много различных алгоритмов обучения, а единого подхода, позволяющего оценить и сравнить эти алгоритмы, нет.

В этой работе мы попытались опробовать принципы построения модельных примеров для сравнения алгоритмов обучения, сформулированные в [1].

Постановка задачи и критерии сравнения

Пусть имеется набор алгоритмов обучения $\{D_1,\dots,D_i,\dots,D_L\}$, где D_i - алгоритм, сопоставляющий обучающей выборке $\bar{X}$ решающее правило F. При фиксированной стратегии природы P, т.е. при фиксированном распределении в пространстве $\{X_1,\dots,X_n,X_{n+1}\}$ ($X_1,\dots,X_n$ - признаки, X_{n+1} - указывает номер образа) и при заданном объеме обучающей выборки n качество алгоритма будем определять, как это широко принято [2,5], математическим ожиданием вероятности ошибочной классификации $M_0(D, P)$. Нас, как сформулировано в [1], будут интересовать c_{ij} и P_{ij}, определяемые из условия:

$$c_{ij}=\sup_{P\in\pi}(M_0(D_j,P)-M_0(D_i,P)) = M_0(D_j,P_{ij}) - M_0(D_i,P_{ij}).$$

Величина c_{ij} указывает, каков наибольший проигрыш в вероятности ошибки, если мы на всех тестовых задачах будем использовать в рас-

познавании D_j, а не D_i. Класс π - это множество возможных стратегий природы, определяемое априорными данными и предположениями о задаче, P_{ij} - стратегия, на которой достигается supremum. Аналитическое определение c_{ij} и P_{ij} - задача нерешенная. Предлагается строить их приближения на основе модельных примеров, которые формируют эксперты. Таким образом, эксперты на основе знания алгоритмов, опыта, теоретических представлений пытаются сформулировать такие гипотетические стратегии природы, которые наиболее сильно подчеркивают выигрыш алгоритма D_i у D_j. Анализ итогов эксперимента по фиксированному набору модельных примеров мы предлагаем проводить на основе игровой модели, сформулированной в [I]. Проведение достаточно широкого эксперимента, основывающегося на модели, кратко описанной выше и подробнее в [I], дает возможность зафиксировать некоторый "эталонный" набор модельных задач. Этот набор и будет служить полигоном для апробации новых алгоритмов. Далее описываются первые результаты проведенных экспериментов.

Алгоритмы обучения

Рассмотрим четыре алгоритма распознавания образов.

Первый использует дискриминантную функцию Фишера [2]. В его основе лежит байесовский подход для случая двух многомерных нормальных совокупностей с равными ковариационными матрицами $N(\mu^{(1)}, \Sigma)$ и $N(\mu^{(2)}, \Sigma)$. Дискриминантная функция алгоритма следующая:

$$Z = X' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})' S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) - \ln \frac{p_1}{p_2},$$

где $\bar{X}^{(i)}$ - оценка средних $\mu^{(i)}$ по обучающей выборке $i=\overline{1,2}$; S - оценка ковариационной матрицы Σ по обучающей выборке; X - наблюдение, которое необходимо классифицировать; p_i - априорная вероятность появления i-го класса, $i = \overline{1,2}$.

Второй алгоритм использует квадратичную дискриминантную функцию [2]. В его основе лежит байесовский подход для случая двух многомерных нормальных совокупностей с различными ковариационными матрицами $N(\mu^{(1)}, \Sigma^{(1)})$, $N(\mu^{(2)}, \Sigma^{(2)})$.

Дискриминантная функция алгоритма следующая:

$$Z = \frac{1}{2}\left[(X-\bar{X}^{(1)})' S_1^{-1}(X-\bar{X}^{(1)}) - (X-\bar{X}^{(2)}) S_2^{-1}(X-\bar{X}^{(2)})\right] + \frac{1}{2} \ln \frac{|S_1|}{|S_2|} - \ln \frac{p_1}{p_2},$$

где S_i - оценка ковариационной матрицы $\Sigma^{(i)}$ по обучающей выборке i-го класса.

Третий алгоритм распознавания основан на непараметрической оценке плотности вероятности при условии i-го класса $p_i(x)$ и подстановке этой оценки в оптимальное решающее правило $U(x) = \arg\max_i p_i\, p_i(x)$. В качестве непараметрической оценки берется оценка типа Розенблатта-Парзена [2], имеющая вид

$$\tilde{p}(x) = \tilde{p}(x_1,\ldots,x_n) = N^{-1} \sum_{l=1}^{N} \prod_{j=1}^{n} h_j\, K(h_j(x_j - x_{jl})),$$

где x_{jl} - j-я компонента l-го вектора выборки из распределения вероятностей с плотностью $p(x)$; K - ядро, равное $2^{-1}e^{-|\bar{x}|}$; $h=(h_1,\ldots,h_n)$ - вектор параметров сглаживания, определяется методом последовательных приближений [3].

Четвертый алгоритм распознавания основан на логических функциях. Решающее правило строится в виде логического дерева [4].

Модельные примеры

Количество классов в эксперименте равно 2. Объем обучающей выборки - 100 реализаций (по 50 реализаций каждого образа), объем контрольной выборки - 200 (по 100 реализаций каждого образа). Количество признаков в тестах № 1-4 - 20, а в тесте № 5 - 2.

В экспериментах вычисляются оценки вероятности ошибочной классификации на обучении P_0^{α} и на контроле P_k^{α}, усредненные по числу экспериментов (здесь 9 экспериментов, т.е. 9 раз моделируются выборки) для каждого модельного примера, где α - номер алгоритма.

Тестовые примеры строились в соответствии с принципом, сформулированном выше.

ТЕСТ № 1. Первый признак распределен следующим образом: для первого класса признак попадает в интервалы (0,1), (2,3), (4,5) с вероятностью 0,95; в интервалы (1,2), (3,4), (5,6) с вероятностью 0,05 и распределен там равномерно. Для второго класса наоборот. Остальные 19 признаков неинформативны и распределены одинаково и независимо для обоих классов N(0,20).

ТЕСТ № 2. Первые два признака распределены следующим образом: для первого класса вероятность того, что оба признака окажутся из интервала (1,2), равна 0,6, а из интервала (0,1) - 0,4. Для второго класса вероятность того, что первый признак попадет в ин-

тервал (I,2), а второй признак - в интервал (0,I), равна 0,4; а вероятность того, что первый признак попадет в (0,I), а второй в (I,2), равна 0,6. Внутри интервалов признаки распределены равномерно. Остальные I8 признаков неинформативны и распределены аналогично тесту № I.

ТЕСТ № 3. Рассматривается модель для независимых признаков с равными ковариационными матрицами. Параметры моделирования подобраны таким образом, чтобы вероятность ошибочной классификации равнялась 0,05.

ТЕСТ № 4. Для первых 2-х признаков распределение строится следующим образом: первый класс представляет собой смесь двух нормальных распределений $N(\mu_{11},\sigma^2)$ и $N(\mu_{12},\sigma^2)$, второй класс - смесь двух распределений $N(\mu_{21},\sigma^2)$ и $N(\mu_{22},\sigma^2)$, где $\mu_{11}=\begin{pmatrix}-1\\1\end{pmatrix}$; $\mu_{12}=\begin{pmatrix}1\\-1\end{pmatrix}$; $\mu_{21}=\begin{pmatrix}-1\\-1\end{pmatrix}$; $\mu_{22}=\begin{pmatrix}1\\1\end{pmatrix}$. Дисперсия σ^2 подбирается таким образом, чтобы вероятность ошибочной классификации была равна 0,05. Остальные I8 признаков - аналогично тесту № I.

ТЕСТ № 5. Аналогичен тесту № 4, но отсутствуют неинформативные признаки.

Результаты сравнения

Результаты сравнения представлены в виде таблицы, в которой указаны P_0^{α}, P_k^{α} и теоретическая вероятность ошибочной классификации P_T^{γ}. В таблице также представлены средняя величина и дисперсия вероятности ошибочной классификации на контроле для каждого алгоритма по всем примерам.

Согласно P_k^{α}, приведенным в таблице, подсчитана матрица наибольших потерь $C=\{c_{ij}\}_{i,j=\overline{1,4}}$, определенная выше,

$$C=\begin{bmatrix} - & 0,09 & 0,06 & 0,23\\ 0,49 & - & 0,I3 & 0,34\\ 0,49 & 0,03 & - & 0,2I\\ 0,42 & 0,29 & 0,29 & - \end{bmatrix}.$$

На основании таблицы можно сделать следующие выводы:

I. Если отказаться от первого алгоритма, то максимальные потери, которые можно от этого понести, будут составлять 0,06.

2. Если отказаться от второго алгоритма, то максимальные потери будут 0,I3.

Таблица

Имя алгоритма	Номер теста					Средняя величина для P_k^α	Дисперсия для P_k^α
	Тест №I	Тест №2	Тест №3	Тест №4	Тест №5		
Алгоритм №I "Фишер"	P_0^1=0,29 / P_k^1=0,46	P_0^1=0,28 / P_k^1=0,46	P_0^1=0,06 / P_k^1=0,I38	P_0^1=0,28 / P_k^1=0,49	P_0^1=0,43 / P_k^1 0,5	0,407	0,023
Алгоритм №2 "Квадратичный"	P_0^2=0,05 / P_k^2=0,47	P_0^2=0,0I / P_k^2=0,33	P_0^2=0,0I / P_k^2=0,23	P_0^2=0,003 / P_k^2=0,I7	P_0^2=0,02 / P_k^2=0,0I3	0,243	0,029
Алгоритм №3 "Парзен"	P_0^3=0,42 / P_k^3=0,47	P_0^3=0,45 / P_k^3=0,45	P_0^3=0,24 / P_k^3=0,2	P_0^3=0,24 / P_k^3=0,3	P_0^3=0,026 / P_k^3=0,0I3	0,287	0,036
Алгоритм №4 "Дерево"	P_0^4=0,07 / P_k^4=0,I8	P_0^4=0.II / P_k^4=0,I9	P_0^4=0,I2 / P_k^4=0,37	P_0^4=0,I3 / P_k^4=0,5I	P_0^4=0,003 / P_k^4=0,08	0.266	0,029
Теоретическая ошибка	P_T^{N1}=0,05	P_T^{N2}=0,0	P_T^{N3}=0,05	P_T^{N4}=0,02	P_T^{N5}=0,02		

3. Если отказаться от третьего алгоритма, то максимальные потери будут 0,0.

4. Если отказаться от четвертого алгоритма, то максимальные потери будут 0,27.

5. Если использовать только первый алгоритм, отказавшись от других, то максимальные потери будут составлять 0,49.

6. Если использовать только второй алгоритм, то максимальные потери будут составлять 0,29.

7. Если использовать только третий алгоритм, то максимальные потери будут 0,29.

8. Если использовать только четвертый алгоритм, то максимальные потери будут 0,34.

Литература

I. МАНОХИН А.Н., ПЛОТНИКОВА В.Е. Игровая имитационная модель сравнения алгоритмов обучения. -В кн.: Машинные методы обнаружения закономерностей (Вычислительные системы, вып. 88).Новосибирск, 1981, с. 85-94.

2. ФУКУНАГА К. Введение в статистическую теорию распознавания образов.-М.: Наука, 1979. - 367 с.

3. ЧЕРКАШИН Н.Г. Некоторые непараметрические алгоритмы распознавания образов большой размерности. - В кн.: Математическая статистика и ее приложения. Томск, 1979, с. 156-162.

4. МАНОХИН А.Н. Методы распознавания образов, основанные на логических решающих функциях. -В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 67). Новосибирск, 1976, с. 42-53.

5. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория распознавания образов. -М.: Наука, 1974. - 415 с.

Поступила в ред.-изд.отд.
11 марта 1984 года