

УДК 62-606.2.001:519.23

© 1995 г. А. А. БОЛОТОВ
(32 Центральный военно-морской госпиталь,
Купавна, Московская область),
О. И. ЛАРИЧЕВ, член-корреспондент РАН
(Институт системного анализа РАН, Москва)

СРАВНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ОБРАЗОВ ПО ТОЧНОСТИ АППРОКСИМАЦИИ РАЗДЕЛЯЮЩИХ ГИПЕРПЛОСКОСТЕЙ

Предлагается новый критерий сравнения методов распознавания – точность аппроксимации разделяющей гиперплоскости. С помощью статистического моделирования получены сравнительные оценки для двух методов распознавания образов (вероятностного, основанного на формуле Байеса, и линейной дискриминантной функции) при четырех типах границ между классами решений.

1. Введение

Распознавание образов как научное направление является одним из основных в проблематике искусственного интеллекта. Широкий арсенал методов распознавания образов [1 – 4] позволяет рассматривать задачи с экспертом и задачи автоматической классификации; задачи в детерминированной и статистической постановке. Для оценки эффективности методов распознавания образов выдвигались различные критерии: полнота использования имеющейся информации, соответствие сложности решающего правила размерам выборки и т.д.

В данной статье предлагается дополнительный критерий сравнения различных методов распознавания образов: точность аппроксимации разделяющей гиперплоскости (ТАРГ). Далее даны обоснования предлагаемого подхода и пример его использования для сравнения двух широко известных методов.

2. Границы между классами решений

Будем рассматривать те задачи распознавания образов, в которых объекты описываются совокупностью признаков, есть эксперт, дающий классификацию для подмножества объектов (выборки), и необходимо построить правило классификации всех объектов.

Предположим, что задача решена в том идеальном случае, когда задана вся совокупность объектов (выборка совпадает с совокупностью объектов) и эксперт безошибочно распределил их между классами решений. Тогда можно сказать, что имеются точные границы между классами решений в многомерном пространстве признаков, описывающих объекты. Мы принимаем широко распространенную в распознавании образов гипотезу о компактности, согласно которой объекты, принадлежащие к одному классу, компактно (без вкрапления объектов других классов) располагаются в многомерном пространстве.

Предположим, что признаки имеют дискретные шкалы возможных оценок. Дадим определение границы класса решений.

Граничным элементом класса решений является объект, для которого изменение оценки, хотя бы по одному признаку, переводит данный объект в какой-либо иной класс решений.

Граничные объекты по сути дела определяют правила классификации. Границы между классами отражают решающие правила эксперта. Эти правила могут иметь сложный или простой (линейная зависимость между признаками) характер.

Если задача классификации полностью решена, то границы построены. Возникает возможность использовать точность аппроксимации этих границ как тестовый пример для сравнения методов классификации.

3. Постановка задачи

Дано:

N признаков, описывающих объекты, подлежащие классификации;

W_i – число оценок на шкале i -го признака;

X_{ki} – k -я оценка по i -му признаку ($i = 1, \dots, N$; $k = 1, \dots, W_i$);

Q – число объектов, входящих в выборку;

L – число классов, к которым должны быть отнесены объекты;

$F_j X$ – граница для j -го класса решений ($j = 1, \dots, L$);

$X = (X_1, \dots, X_N)$ – векторы оценок признаков, принадлежащие границе j -го класса;

M_1, M_2 – методы классификации, строящие разделяющие гиперплоскости на основе выборки из Q объектов.

Требуется: сравнить методы M_1 и M_2 по точности аппроксимации истинной границы $F_j(X)$.

Такая постановка задачи стала возможной, в частности, в связи с появлением компьютерной системы "КЛАСС", предназначенной для построения полных и непротиворечивых баз экспертных знаний [5]. Эта система позволяет поэтапно строить классификацию, проверяя ответы эксперта на непротиворечивость. Результат работы системы может быть представлен как совокупность границ между классами решений.

Эффективность системы достаточно высока в частном случае, когда можно установить упорядочение значений на шкалах признаков по отношению к каждому из классов решений [5].

Однако в общем случае для целей сравнения методов совсем не обязательно восстанавливать истинные границы между классами, представляющие решающие правила эксперта. В тестовых примерах (см. далее) такие границы могут быть просто заданы.

4. Сравнимые методы

В качестве сравниваемых методов M_1 и M_2 нами выбраны классификаторы на базе линейной дискриминантной функции (ЛДФК) и на формуле Байеса (БК). Их выбор обусловлен рядом обстоятельств.

1. Оба метода до сих пор широко используются при построении компьютерных систем классификации в различных областях человеческой деятельности (медицина, геология, техническая диагностика и т.п.).

2. Эти методы достаточно хорошо изучены – имеют устоявшиеся критерии оптимизации, лежащие в основе их функционирования. Метод Байеса минимизирует величину общего риска [4], а ЛДФК максимизирует отношение межклассовой к внутриклассовой вариации [6].

3. Они достаточно просты для реализации и проверки в компьютерных системах.

4. Согласно многочисленным данным ряда зарубежных исследователей в задачах с плохой структурой (типа медицинской диагностики) решающие правила экспертов хорошо аппроксимируются линейными моделями [7, 8]. Такой подход, названный авторами "бутстрепингом" (bootstreping) широко используется в психофизиологии,

медицинской диагностике, банковском деле и др. и обеспечивает высокое качество решений [8].

Приведем краткое описание обоих методов. В случае использования байесовского метода задача классификации сводится к статистической задаче выбора гипотез при заданной априорной информации о каждом классе и условных вероятностях каждой оценки всех признаков по всем классам. Решение выбирается по максимуму апостериорной вероятности среди возможных классов для конкретного набора признаков. Аналитически это выражается известной формулой Байеса [4]:

$$(1) \quad P(Z_j/S) = \frac{P(Z_j)P(S/Z_j)}{\sum_j P(Z_j)P(S/Z_j)},$$

где $P(Z_j)$ – априорная вероятность j -го класса; $P(Z_j/S)$ – апостериорная вероятность j -го класса при наборе признаков S ; $P(S/Z_j)$ – условная вероятность набора признаков S для j -го класса. Будем полагать, что признаки независимы, тогда $P(S/Z_j) = \prod_i P(S_i^w/Z_j)$, где $P(S_i^w/Z_j)$ – условная вероятность w оценки (градации) i -го признака.

В качестве априорных вероятностей каждого класса использовались оценки частоты их встречаемости, при этом $P(Z_j) = n_j / \sum_j n_j$, где n_j – число объектов j -го класса среди всех возможных объектов.

В качестве условных вероятностей возможных оценок признаков использовались оценки [4]: $P(S_i^w/Z_j) = n_{ij}^w/n_j$, где n_{ij}^w – число объектов, имеющих i -й признак с w шкалой для j -го класса в заданной обучающей выборке.

Линейный дискриминантный анализ позволяет построить разделяющую гиперплоскость для задач, где заданы классы, признаки и имеется обучающая выборка. При этом граница между классами ищется в виде:

$$(2) \quad Y = \sum_i A_i X_i + C,$$

где A_i – коэффициент при i -м признаке, X_i – значение (или градация) i -го признака, C – константа.

Рассмотрим простой случай, когда имеется только два класса ($L=2$). Если для некоторого объекта $Y \geq 0$, то он считается объектом из первого класса, иначе – из второго.

Существует множество разделяющих гиперплоскостей для данной обучающей выборки [4]. В качестве оптимальной выберем гиперплоскость, которая максимизирует отношение U :

$$U = (\text{межклассовая вариация})/(\text{внутриклассовая вариация}).$$

Межклассовая вариация определяется квадратами расстояний $(m_{y_1} - m_{y_2})$ и тем больше, чем дальше отстоят средние представители (центры) m_{y_j} первого и второго классов. Внутриклассовая вариация определяется суммой квадратов расстояний от всех объектов обучающей выборки каждого класса до их центра.

В итоге коэффициенты линейной дискриминантной функции находятся из матричного уравнения [6]:

$$(3) \quad |A| = |S|^{-1} (|\bar{X}_1| - |\bar{X}_2|), \\ C = 1/2|A| (|\bar{X}_1| + |\bar{X}_2|),$$

где $|S|^{-1}$ – матрица, обратная ковариационной матрице, $|\bar{X}_1|_{(2)}$ – матрица (векторы) средних по первому (второму) классу.

5. Методика сравнения

Дадим ряд определений. Под термином “задача” будем понимать описание ее структуры: число признаков, градаций (оценок) и возможных классов. Каждая задача будет характеризоваться двумя параметрами: объемом обучающей выборки и видом границ между классами. Так как число классов фиксировано и равно двум, то задача будет определяться числом признаков и числом градаций, которые принимают значение либо 2, либо 4 ($W_k = 2; 4$). Каждую задачу будем представлять в виде 8×2 , 6×4 , что соответствует задачам: с 8 признаками и 2 градациями у каждого признака, 6 признаками и 4 градациями соответственно.

Под размерностью задачи будем понимать максимальное число возможных объектов (ситуаций) в заданном пространстве признаков.

Рассмотрим четыре вида границ, заданных в форме правил:

1) если $X_1 > W_k/2$, то это будут объекты 2-го класса, иначе 1-го класса. X_1 – номер градаций первого признака;

2) если $\sum_i X_i < N \times (1 + W_k)/2$, – первый класс, иначе – второй. X_i – значение i -го признака данного объекта;

3) если $\sum_i X_i < N \times (1 + W_k)/2$ или $X_1 \leq W_k/2$, – первый класс, иначе – второй.

4) если $\sum_i X_i < N \times (1 + W_k)/2$ или $X_1 \leq W_k/2$ или $\sum_{i=1}^{N/2} X_i - \sum_{i=N/2+1}^N X_i \geq 0$, – первый класс, иначе – второй.

Граница 1 задает границу типа “отсечка”, вторую назовем “линейной”, третью – “комбинированной из двух правил”, а четвертую – “комбинированной из трех правил”.

Как показали результаты проведенных исследований с помощью системы “КЛАСС”, из элементов подобных границ состоят реальные границы – решающие правила экспертов. Однако они, как правило, имеют более сложное строение, чем приведенные выше, и состоят из кусочно-линейных гиперплоскостей. Число таких решающих правил невелико и составляет величину не более 7 – 9 [9].

Итак, при сравнении методов будем изменять параметры задачи: объем обучающей выборки в процентах от размерности задачи и вид границы.

Для оценки точности аппроксимации разделяющей гиперплоскости ЛДФК и БК использовался метод статистических испытаний (метод Монте-Карло). Опишем его суть в виде алгоритма.

5.1. Задается структура задачи, т.е. указывается число признаков и градаций.

5.2. Для каждого из четырех рассматриваемых типов границ строится информационная база, содержащая все возможные объекты для заданного пространства признаков. При этом каждый объект характеризуется вектором признаков, истинным значением номера класса, атрибутом принадлежности его к граничному элементу. Формируются также поля для номеров классов, получаемых с помощью методов M_1 и M_2 и др. Принадлежность каждого объекта к тому или иному классу устанавливается на основании решающих правил, приведенных выше, для конкретной границы. Из них же определяются и отмечаются граничные объекты.

5.3. Задается процент обучающей выборки m_p от всех возможных объектов решаемой задачи. При этом число объектов каждого класса m_j в обучающей выборке определяется следующим образом: $m_j = m_p n_j P(Z_j)/100$, где n_j – общее число объектов j -го класса для рассматриваемой задачи и типа границы.

5.4. С помощью генератора случайных чисел определяется конкретный набор m_j объектов каждого класса. При этом вначале выбираются объекты, находящиеся около наиболее типичных представителей каждого класса. В качестве меры удален-

Таблица 1

Общие сведения по процентам расхождений линейного дискриминантного анализа и метода Байеса для 4 границ и обучающих выборок различного объема

Задача N*Wk. Число объек- тов	Обу- чаю- щая вы- борка, %	Процент расхождений для заданных границ											
		Граница 1				Граница 2				Граница 3			
		ЛДФК		БК		ЛДФК		БК		ЛДФК		БК	
		Общ	Гр	Общ	Гр	Общ	Гр	Общ	Гр	Общ	Гр	Общ	Гр
2 * 4 16	12,5	39	39	50	50	25	45	6,3	14	20	30	25	50
	25	31	31	36	36	15	28	11	24	12	18	18	31
	37,5	6,3	6	30	30	17	33	11	23	15	23	19	29
	50	1,5	2	19	19	11	21	10	21	10	19	14	25
4 * 2 16	12,5	—	—	50	50	—	—	1	1	—	—	25	27
	25	31	31	33	33	—	—	9,3	15	—	—	19	20
	37,5	4,4	4	24	24	12	15	12	19	6	6	17	18
	50	0	0	4	4	15	22	10	20	6	6	11	11
6 * 2 64	5	—	—	50	50	—	—	12	21	—	—	25	28
	12,5	10	10	23	23	18	22	16	26	18	18	18	21
	25	0	0	2	2	17	22	16	25	15	17	11	12
	37,5	0	0	0	0	11	21	8	13	9	10	7	8
4 * 4 256	5	7	13	26	28	19	37	14	33	19	31	20	30
	12,5	6	11	12	13	7	20	14	31	12	23	15	23
	25	3	6	2	2	3	8	10	24	12	24	10	17
	37,5	1	2	0	0	2	7	6	17	12	22	8	16
8 * 2 256	5	8	8	12	12	21	29	21	33	24	26	22	26
	12,5	0	0	0	0	11	21	9	16	15	17	8	9
	25	0	0	0	0	7	14	4	8	9	11	8	9
	37,5	0	0	0	0	4	8	1	2	6	7	8	9
10 * 2 1024	5	0	0	0	0	10	20	10	21	18	21	10	12
	12,5	0	0	0	0	5	10	3	6	14	19	8	12
	25	0	0	0	0	2	4	0,4	1	10	13	7	12
	37,5	0	0	0	0	1	2	0	0	8	10	5	8
6 * 4 4096	5	8	15	0	0	3	12	7	24	18	32	7	15
	12,5	7	14	0	0	1	3	3	12	17	31	6	14
	25	4	8	0	0	0,2	1	2	3	15	30	6	14
	37,5	2	0	0	0	0	0	1	2	15	28	4	10
12 * 2 4096	5	0	0	0	0	5	9	3	7	17	22	8	9
	12,5	0	0	0	0	1	3	11	14	11	14	7	9
	25	—	—	0	0	—	—	0,4	3	—	—	5	6
	37,5	—	—	0	0	—	—	0,1	1	—	—	2	2

Примечание. Общ – процент расхождений по всем возможным объектам, а Гр – только для граничных объектов.

ности была выбрана наиболее простая в вычислительном смысле метрика абсолютных значений [10].

5.5. По данной обучающей выборке определяются величины, необходимые для построения обоих классификаторов (коэффициенты линейной дискриминантной функции и матрица условных вероятностей).

5.6. Производится классификация всех возможных объектов для данной задачи с помощью ЛДФК и БК. В базе знаний отмечается номер класса (а для БК еще и апостериорная вероятность), полученного в результате работы каждого классификатора.

5.7. Находятся расхождения для обоих методов между заданной классификацией и осуществленной методами M_1 и M_2 среди всех возможных объектов и отдельно для граничных. Информация о расхождениях запоминается. Для БК подсчитываются также объекты, для которых вероятность определения каждого класса составляет величину 0,5. Ее определяем как отказ от классификации и относим к разряду расхождений.

5.8. Последовательность действий 5.4 – 5.7 повторяется 200 – 500 раз (задается количество циклов Монте-Карло), а результаты расхождений суммируются.

5.9. Вычисляются и распечатываются средние значения расхождений для данной задачи среди всех циклов Монте-Карло по всем объектам и отдельно по граничным.

Результаты такого моделирования приведены в табл. 1.

6. Анализ результатов моделирования

Анализ данных таблицы показывает, что результаты статистического моделирования можно разбить на две группы:

1) задачи малой размерности (не более 300 объектов);

2) задачи большой размерности (более 1000 объектов).

В задачах малой размерности:

1) классификатор на основе ЛДФ имеет меньше расхождений для 1-й границы (самой простой), чем байесовский;

2) ЛДФК имеет лучшие показатели качества (меньшее число расхождений) в задачах с большим числом градаций для более простых типов границ;

3) БК имеет меньшее число расхождений, чем ЛДФК в задачах одинакового размера с большим числом признаков (задачи 4×4 и 8×2);

4) в задачах с очень малым числом возможных объектов (16) при увеличении объема обучающей выборки нет монотонной зависимости качества решений в отличие от других задач, что связано с очень сильным влиянием каждого объекта выборки на параметры обоих классификаторов. Поэтому по этим задачам нельзя делать выводы общего характера.

В задачах большой размерности:

1) БК имеет лучшее качество решений для границ типа 1, 3 и 4 для всех задач;

2) в случае использования малого числа признаков (задача 6×4) ЛДФК имеет лучшее качество заключений для границы 2;

3) при увеличении числа признаков качество заключений БК растет быстрее, чем ЛДФК;

4) при увеличении объема обучающей выборки качество решений обоих классификаторов монотонно увеличивается;

5) при увеличении числа признаков для одних и тех же значений процента обучающей выборки качество заключений обоих классификаторов также увеличивается;

6) при усложнении границы между классами БК имеет лучшее качество работы, чем ЛДФК;

7) для всех задач при сложной границе (4-й тип) величина расхождений от объема ОБ изменяется медленней, чем для более простых границ;

Таблица 2

Сравнительные данные качества заключений линейного дискриминантного анализа и метода Байеса для задачи дифференциальной диагностики тромбозмболии легочной артерии и инфаркта миокарда (3 класса)

Процент обучающей выборки	ЛДФК % расхождений	БК % расхождений
12,5	— *	35
25	40,6	27
37,5	26,6	18
50	18,5	17,4
100	15,6	15,6

Примечание. Расчет для ЛДФК производился с помощью статистического пакета "STATGRAPHICS" [11] в диалоговом режиме. Для четырех верхних строк этой таблицы было отобрано только по 5 наборов обучающих выборок, подготовленных вручную. Использование такого ограниченного набора ОВ связано с высокой трудоемкостью получения результатов.

8) для сложных границ (3-й и 4-й тип) даже при полной (100%) выборке не удается добиться 100% точности аппроксимации разделяющей гиперплоскости.

Как отмечалось выше, исследования проводились для двух классов решений. Рассмотрим кратко вопрос о влиянии числа классов на качество заключений обоих методов. В задачах с двумя классами граница была одна: между 1-м и 2-м классом. Допустим, стало три класса в задачах, приведенных выше. Тогда число границ между возможными классами резко увеличивается. При трех классах их становится три: между 1-м и 2-м классом, 1-м и 3-м, 2-м и 3-м. Дополнительные границы приводят к усложнению решающих правил классификации, и следует ожидать ухудшения качества работы рассмотренных методов для задач одинаковой размерности по сравнению с задачами с двумя классами. Для проверки этой гипотезы была рассмотрена в качестве примера реальная задача 6×2 с тремя классами решений – задача дифференциальной диагностики тромбозмболии легочной артерии (ТЭЛА) и инфаркта миокарда (ИМ) (границы были получены с помощью системы "КЛАСС"). Были выделены решающие правила эксперта отнесения объектов по трем классам [9]:

I класс – ТЭЛА:

1. Первая градация (оценка) пятого признака.
2. Вторая оценка пятого признака плюс любые четыре первые оценки из пяти оставшихся признаков.

II класс – ИМ:

1. Вторая градация пятого и шестого признаков плюс не более двух новых из четырех первых признаков;

III класс – ТЭЛА и ИМ вместе:

1. Вторые оценки пятого и шестого признаков плюс одна вторая из четырех оставшихся.
2. Вторая градация по пятому признаку и первая по шестому плюс не более двух первых оценок из четырех оставшихся.

Статистическое моделирование для такой задачи было проведено по схеме, указанной в разделе 5. Результаты его приведены в табл. 2.

Сравнение данных табл. 2 и подобной задачи табл. 1 подтверждает гипотезу об ухудшении качества работы обоих методов для задач с большим числом границ.

7. Выводы

Методом статистического моделирования показано, что в задачах классификации:

1) на качество заключений сравниваемых методов влияет вид (сложность) границы между классами. При усложнении границы качество решений ЛДФК и БК ухудшается;

2) на простых границах в задачах малой размерности (число возможных объектов не более 300) ЛДФК имеет лучшие показатели качества решений, чем БК;

3) в задачах большой размерности (более 1000 объектов) БК имеет более высокую точность аппроксимации разделяющей гиперплоскости;

4) на сложных границах БК имеет лучшие показатели по критерию ТАРГ;

5) при больших обучающих выборках (более 37%) результаты расхождений ЛДФК и БК различаются меньше, чем для малых ОВ;

6) так как заранее неизвестно, какой вид границы может быть в решаемой задаче классификации, из рассмотренных двух методов предпочтительнее использовать байесовский метод, который с точки зрения критерия ТАРГ имеет лучшие показатели для более сложных границ;

7) гипотеза, выдвинутая авторами "бутстрепинга" о превосходстве линейных аппроксимирующих правил в качестве средства, заменяющего суждения эксперта [8], вызывает большие сомнения. Вероятно, ЛДФК дает действительно неплохие результаты, если граница оказывается простой (тип 1 и 2). Но это не всегда так. Поэтому при более сложных границах ЛДФК приводит к существенным ошибкам. При наличии нескольких классов оценки эксперта будут, вероятно, намного точнее, чем оценки, полученные с помощью подхода "бутстрепинг".

СПИСОК ЛИТЕРАТУРЫ

1. Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. М.: Наука, 1978. Вып. 33. С. 5-58.
2. Цыткин Я. З. Основы теории обучающихся систем. М.: Наука, 1970.
3. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.
4. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974.
5. Ларичев О. И., Мечитов А. И., Мошкович Е. М., Фуремс Е. М. Выявление экспертных знаний. М.: Наука, 1989.
6. Бонч Б., Хуань К. Дж. Многомерные статистические методы для экономики. М.: Статистика, 1979.
7. Levis R. Goldberg. Five models of clinical judgment: an empirical comparison between linear and nonlinear representations of the human inference process. // Organizational behavior and human performance. 1971. № 5. P. 458-479.
8. Robyn M. Dawes. Rational choice in an uncertain world. Carnegie Mellon University, 1988, USA, p. 346.
9. Ларичев О. И. Структура экспертных знаний в задачах классификации // Докл. РАН. 1994. Т. 336. № 6. С. 750-752.
10. Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977.
11. Григорьев С. Г., Левандовский В. В., Перфилов А. М., Юнкеров В. И. Пакет прикладных программ STATGRAPHICS на персональном компьютере. Практическое пособие по обработке результатов медико-биологических исследований. Санкт-Петербург: ВМА, 1992.

Поступила в редакцию 07.07.94