1969 г. ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ Выпуск 36

УДК 62-5:007:621.391:519.2

ПОСТРОЕНИЕ ЛИНЕЙНЫХ РЕШАЮЩИХ ФУНКЦИЙ ДЛЯ РАЗДЕЛЕНИЯ $S\,(S>2)$ КЛАССОВ

Р.М. Малкина

При числе классов $S>2$ построение решающего правила является весьма трудоемкой задачей даже тогда, когда выпуклые оболочки классов, натянутые на множества объектов LA_i обучающей последовательности $(i\leq S)$, не пересекаются.

Вместе с тем известно, что в такой наиболее простой ситуации удается построить достаточно экономные решающие правила. В работе [I] описан один алгоритм построения простых правил принятия решений, близких к оптимальным. Ниже описывается другой возможный способ построения решающего правила для разделения $S>2$ классов (классы выпуклые, матрицы ковариаций могут быть различными), с числом разделяющих гиперплоскостей, близким к минимально необходимому. Под минимально необходимым числом гиперплоскостей понимаем такое минимальное число разделяющих гиперплоскостей, при котором ещё безошибочно распознаются все точки $j\in L$. Предполагаем, что обучающая выборка L — достаточно представительна и, следовательно, решающее правило с малым числом определяет принадлежность к одному из S классов и точек контрольной последовательности, т.е. точек $j\notin L$.

Так как LA_ℓ и LA_K, $\ell\neq K$, $\ell, K\leq S$, выпуклы, всегда возможно построение разделяющей такой гиперплоскости, что

$$\Phi_j\geq 0,\quad j\in LA_\ell,$$
$$\Phi_j<0,\quad j\in LA_K.$$

При этом для остальных ($S-2$) классов возможны следующие альтернативы:

I. $\Phi_j\geq 0,\ j\in LA_i$ (все точки $j\in LA_i$ принадлежат +полупространству);
2. $\Phi_j\leq 0,\ j\in LA_i$ (все точки $j\in LA_i$ принадлежат −полупространству);
3. $\Phi_j\geq 0,\ j\in LA_i'$, $\Phi_j\leq 0,\ j\in LA_i''$, $LA_i=LA_i'\cup LA_i''$ (гиперплоскость $\Phi_j=0$ рассекает класс A_i на "осколки").

Пусть гиперплоскость $\Phi_j=0$ рассекла t классов множества LT. При этом q классов LQ попали в +полупространство, а q' множества LQ' — в −полупространство,

$$q+q'+t=S,\quad LQ\cup LQ'\cup LT=L$$

Таким образом, при разделении классов A_i и A_ℓ попутно получаем разделение $A_i\,(A_\ell)$ от множества $LQ'\,(LQ)$ соответственно.

Для отделения каждого класса A_i, $i\leq S$ от всех остальных строят последовательно гиперплоскости, разделяющие классы попарно.

Приведем возможный алгоритм.

Строим гиперплоскость I, отделяющую A_1 от A_2. Классам, попавшим в +полупространство, ставим в соответствие код +I, в −полупространство — код −I, классам, рассеченным гиперплоскостью — код 0. Проверяем необходимость построения гиперплоскости, отделяющей A_1 от A_3. Совпадение кода класса A_3 с кодом класса A_3 означает неразделенность A_1 от A_3 и необходимость построения гиперплоскости 2, разделяющей A_1 и A_3; в противном случае переходим к проверке необходимости разделения A_1 и A_4 и т.п.

Классы A_ℓ и A_K оказываются разделенными к i-ому шагу алгоритма, если хотя бы на одной из построенных к текущему шагу гиперплоскостей коды, соответствующие классам A_ℓ и A_K, не совпадают.

Таким образом, для разделения всего множества на классы потребуется построение всего R, где R меньше числа гиперплоскостей, отделяющих все классы попарно, т.е. меньше

$$C_S^2=\frac{S(S-1)}{2}.$$

Разделение обучающей последовательности L гиперплоскостями определяет решающее правило, которое в отличие от последовательного правила (I) является ветвящимся:

если для точки P_j, $j\notin L$, $\Phi_{1j}\geq 0$, то $j\in Q_1$;
если $\Phi_{1j}<0$, то $j\in Q_1'$;

· если $\Phi_{1j} \geq 0$ и $\Phi_{2j} \geq$,то $j \in Q_2$, $Q_2 \in Q_1$, и т.д. вплоть до однозначного определения принадлежности к одному из классов.

Из вышеизложенного ясно, что при прочих равных условиях величина R обусловлена числом "осколков". Поставим задачу минимизации числа "осколков", получающихся при построении каждой из R гиперплоскостей.

Требуется минимизировать число альтернатив 3, или, что то же, максимизировать число альтернатив I и 2. Сформулированная задача есть следующая задача частично целочисленного программирования:

$$\min\Big[\sum_{i=1}^{s}(\delta_i+\gamma_i)\Big|\ \Phi_j-\alpha_j+\beta_j=0\ (j\in LA_i),$$
$$\alpha_j,\ \beta_j\geq 0;\ \sum_{j\in LA_i}\alpha_j\leq\delta_i M;\ \sum_{j\in LA_i}\beta_j\leq\gamma_i M;$$
$$\delta_i\ \gamma_i-(0,1),\ i\neq e,\ k,\ i\leq s,$$
$$\Phi_j\geq 0\ (j\in LA_e);\ \ \Phi_j\leq -a\ (j\in LA_k)\Big].$$

Здесь M – верхняя граница для $|\Phi_j|$ (можно приписать M значение самой большой величины, поддающейся записи при вычислениях на конкретной ЦВМ); величина $0 < a \ll I$ вводится во избежание возможности вырожденного решения $\Lambda = 0$ (при $\Lambda = 0$ условия $\Phi_j \geq 0$, $j \in LA_e$, и $\Phi_j \leq -a$, $j \in LA_K$, несовместны).

При $\gamma_i = 0$ для $j \in LA_i$ верна альтернатива I, при $\delta_i = 0$ – альтернатива 2, при $\gamma_i = \delta_i = 1$ – альтернатива 3, выполнение условия $\gamma_i = \delta_i = 0$ практически невозможно.Таким образом, минимизируя $\sum_{i=1}^{s}(\delta_i + \gamma_i)$, действительно получаем максимум альтернатив I и 2.

При заданном расположении классов в пространстве P эффективность применения процедуры (2) больше, если больше возможности поворота гиперплоскости $\Phi_j = 0$ в пространстве. Следовательно, целесообразно перенумеровать все классы $\widetilde{A}_1$, $\widetilde{A}_2, \ldots, \widetilde{A}_S$ в соответствии с расстоянием от "самого крайнего" класса $\widetilde{A}_1$, для которого расстояние от среднего выборочного значения $\widetilde{P}_i$, до начала координат наибольшее ($\widetilde{R}_2 \leq \widetilde{R}_3 \leq \ldots \widetilde{R}_S, \widetilde{R}_i = |\widetilde{P}_i - \widetilde{P}_1|$), и построение разделяющих гиперплоскостей начинать с разделения классов $\widetilde{A}_1$ и $\widetilde{A}_S$, далее проверять необходимость разделения $\widetilde{A}_1$ и $\widetilde{A}_{S-1}$ и т.п. Эффективность процедуры (2) возрастает с уменьшением плотности "упаковки" классов и с ростом числа классов. Однако при большом числе классов, подлежащих разделению, и, следовательно, при большом числе ограничений в (2) и обращений к (2) (напоминаем, что для разделения многих классов необходимо решать последовательность задач (2)) получение решения на современных ЦВМ существующими методами частично целочисленного линейного программирования может оказаться нереальной задачей. Трудоемкость задачи можно резко снизить,предварительно выделяя и подавая на обучение лишь граничные точки каждого класса и (или) используя приближенные алгоритмы частично целочисленного программирования, например [2,3,4] или нижеописанную процедуру. При этом число разделяющих гиперплоскостей $\bar{R}$ может оказаться больше R , полученных при решении (2) конечными методами.

Процедура заключается в последовательном выделении гиперплоскостями точек класса A_i , $i \leq s$. Точки класса A_i принимаем за точки множества I, точки всех остальных невыделенных классов – за точки множества П. При выделении A_1 $\bar{II} = L \setminus LA_1$, при выделении A_2 $\bar{II} = L \setminus (LA_1 \cup LA_2)$, при выделении A_i $\bar{II} = L \setminus \sum_{k=1}^{k=i} LA_k$. На каждой итерации решаем задачу минимизации числа ошибок на II при условии правильного распознавания всех точек I (назовем её t – задачей). При этом по одну сторону от построенной гиперплоскости располагаются, в общем случае, точки, принадлежащие и I и II, а по другую – точки, принадлежащие только II, причем максимально возможное их число. Множества I и П оказываются разделенными, если II выпукло. В противном случае вновь решаем t –задачу применительно к точкам из I и тем точкам из II, которые оказались по одну сторону от гиперплоскости с точками из I. Процедуру повторяем до тех пор, пока класс A_i не окажется выделенным (II = $\emptyset$), после чего вновь решаем последовательность t -задач, полагая I = A_{i+1} , II = $L \setminus \sum_{k=1}^{i+1} LA_k$.

Для получения решения t – задачи, близкого к оптимальному, применимы полуэвристические методы [5] , существенно более простые, чем методы частично целочисленного программирования.

Число разделяющих гиперплоскостей уменьшается при упорядочивании классов в соответствии с расстоянием от наиболее удаленного класса от начала координат или от центра тяжести множества L .

Ориентировочные эксперименты по разделению двумерных (n = 2) выпуклых классов (S = 15÷30) с неравными матрицами вариаций позволяют сделать вывод о целесообразности применения изложенной процедуры. Число разделяющих прямых оказывается близким к числу оптимальных разделяющих прямых (см. [I]) и меньше последнего, если " упаковка" классов неплотна. Большая же трудоемкость оптимального разделения при неравных матрицах ковариаций классов известна (см. [I]). При случайной нумерации классов число разделяющих прямых возрастало в 2-5 раз.

Во многих практических задачах множества $L\ A_i$ равномерно распределены в объеме V, заключенном в замкнутую оболочку, натянутую на L. Пусть V представляет собой гиперсферу. Разделим радиус гиперсферы на μ равных частей. Объем гиперслоя, заключенного между гиперсферой с радиусом ri/μ, $i = 1, 2 \ldots \mu$ и гиперсферой с радиусом $r(i+1)/\mu$, тем больше, чем больше i, и, следовательно, если классы равномерно упакованы, в каждом внешнем гиперслое располагается классов больше, чем во внутреннем гиперслое. Объем $(i+1)$ -го гиперслоя тем больше объема i -го гиперслоя, чем больше размерность пространства признаков m. При достаточно большой размерности (если V и не гиперсфера) можно считать все классы граничными, расположенными вблизи оболочки, натянутой на L, тогда перенумерация классов в целях уменьшения числа разделяющих гиперплоскостей теряет смысл.

Процедура последовательного построения решающего правила пригодна и для разделения невыпуклых классов, в том числе и неодносвязных [5] . Иногда для ускорения процедуры целесообразно множества I и II поменять местами: если на какой-либо итерации число точек из II не изменилось по сравнению с предыдущей итерацией, решаем t - задачу минимизации ошибок на I при условии правильного распознавания всех точек из II.

Литература

I. Н.Г.ЗАГОРУЙКО. Линейные решающие функции, близкие к оптимальным. - Вычислительные системы, Изд-во "Наука", Сибирское отд., Новосибирск, 1965, вып. 19.

2. Dreebeek Norman J. An algorithm the solution of mixed integer programming problems. Management Science. Baltimore, 1966,12, N 7.

3. И.Н. ПЯТЕЦКИЙ-ШАПИРО, В.А. ВОЛКОНСКИЙ, Л.В. ЛЕВИНА,А.О. ПОМАНСКИЙ. Об одном итеративном методе решения задач целочисленного программирования. - ДАН, 1966, т.169, №6.

4. Ю.Ю. ФИНКЕЛЬШТЕЙН. Алгоритм для решения задач целочисленного программирования с булевыми переменными. - Экономика и математические методы, 1965, т.I, выпуск 5.

5. Р.М. МАЛКИНА, А.А.ПЕРВОЗВАНСКИЙ. Построение последовательного решающего правила в задаче распознавания образов. Изв. АН СССР, "Техническая кибернетика", 1969, № 1.

Поступила в редакцию
3.IУ.1969 г.