Литература

1. *И. М. Гельфанд, Ш. А. Губерман, М. Н. Извекова, В. И. Кейлис-Борок, Е. Я. Ранцман.* Распознавание мест возможного возникновения сильных землетрясений I. Памир и Тянь-Шань.— Сб. «Вычислительная сейсмология», вып. 6. М., «Наука», 1973.
2. *И. М. Гельфанд, Ш. А. Губерман, М. Л. Извекова, В. И. Кейлис-Борок, Е. Я. Ранцман.* О критериях высокой сейсмичности.— ДАН СССР, Геофизика, 1972, т. 202, № 6.
3. *И. М. Гельфанд, Ш. А. Губерман, М. П. Жидков, М. С. Калецкая, В. И. Кейлис-Борок, Е. Я. Ранцман.* Опыт переноса высокой сейсмичности со Средней Азии на Анатолию и смежные регионы.— ДАН СССР, 1973, т. 210, № 2.
4. *J. M. Gelfand, Sh. A. Guberman, M. L. Isvekova, V. I. Keilis-Borok, E. Ya. Ranzman.* Criteria of hich seisicity determined by pattern recognition. Tectonophysics, 1972.
5. *И. М. Гельфанд, Ш. А. Губерман, М. П. Жидков, М. С. Калецкая, В. И. Кейлис-Борок, Е. Я. Ранцман, И. М. Ротвайн.* Распознавание мест, где возможны сильные землетрясения. II. Четыре региона Малой Азии и Юго-Восточной Европы. Вычислительная сейсмология, вып. 7. М., «Наука», 1974.
6. *И. М. Гельфанд, Ш. А. Губерман, М. П. Жидков, В. И. Кейлис-Борок, Е. Я. Ранцман, И. М. Ротвайн.* Распознавание мест, где возможны сильные землетрясения. III. Случай, когда границы дизъюнктивных узлов неизвестны. Вычислительная сейсмология, вып. 7. М., «Наука», 1974.
7. *С. В. Медведев* (ред.) Сейсмическое районирование СССР. «Наука», 1964.
8. *М. М. Бонгард.* Проблема узнавания. М., «Наука», 1967.
9. *М. М. Бонгард, Н. Н. Вайнцвайг, Ш. А. Губерман, М. Л. Извекова, М. С. Смирнов.* Использование обучающейся программы для выявления нефтеносных пластов.— Геология и геофизика, 1966, № 6.
10. *Л. Н. Толстой.* Собр. соч., т. 6, стр. 81, 1951.

М. П. Полякова, М. Н. Вайнцвайг

ОБ ИСПОЛЬЗОВАНИИ МЕТОДА «ГОЛОСОВАНИЯ» ПРИЗНАКОВ В АЛГОРИТМАХ РАСПОЗНАВАНИЯ

При разработке алгоритмов обучения распознаванию, основанных на отборе признаков, соответствующих каждому из данных классов объектов и характеризующих объекты этого класса [1, 2, 3], возникает задача использования отобранных признаков для классификации (узнавания). Применение в этих целях чисто логических методов оказывается, как правило, малоэффективным. Так, например, при построении для каждого из классов дизъюнкции его признаков становится невозможной классификация объектов, характеризуемых одновременно признаками разных классов. Это приводит к тому, что вероятность ошибки при таком правиле будет относительно большой. Одним из возможных методов существенного снижения вероятности ошибки при узнавании служит так называемое «голосование» признаков.

Этот метод состоит в том, что для объекта a, подлежащего классификации, в каждом классе A_i производится подсчет числа $n_i(a)$ «голосов» за этот класс, т. е. числа отобранных признаков класса A_i, которыми обладает объект a. Объект a относится к тому из классов, для которого это число оказывается наибольшим. Использование такого метода, кроме обычного требования отбора признаков с возможно большей надежностью, накладывает некоторые дополнительные ограничения на критерии отбора совокупности признаков. Выяснение этих ограничений и является целью настоящей работы.

Будем считать, что на множестве объектов A задана вероятностная мера $\gamma(B_k)$, определенная на всех подмножествах $B_k \subset A$. Для простоты (хотя это и несущественно) рассмотрим случай, когда множество объектов разбито лишь на два класса: A_1 и A_2. Пусть в процессе обучения на множестве примеров для каждого из таких классов A_i каким-то образом уже отобран некоторый набор $\{\beta_i^1, \beta_i^2, \ldots, \beta_i^{k_i}\}$ признаков β_i^e, имеющих наилучшую оценку надежности [4]. Попытаемся оценить вероятность ошибки при классификации методом «голосования».

Обозначим через $q_i(\Delta n < 0)$ условную вероятность того, что для произвольного объекта a, принадлежащего классу A_i, разность $\Delta n = n_i(a) - n_j(a)$ числа «голосов» за этот и противоположный классы будет отрицательной. Очевидно, что $q_i(\Delta n < 0)$ есть условная вероятность ошибки для объектов класса A_i. Тогда вероятность ошибки при классификации произвольного объекта $a \in A$ методом «голосования» $q = \gamma(A_1) q_1(\Delta n < 0) + \gamma(A_2) q_2(\Delta n < 0)$ ($\gamma(A_i)$ — вероятность того, что объект a, предъявленный для распознавания, будет принадлежать классу A_i).

Следующая теорема позволяет оценить вероятности $q_i(\Delta n < 0)$.

Теорема. Пусть $M_i(\Delta n)$ — математическое ожидание разности Δn числа голосов за «свой» и «чужой» классы при условии, что объект принадлежит классу A_i, а $D_i(\Delta n)$ — дисперсия этой разности при том же условии. Тогда, если $M_i(\Delta n) > 0$, то $q_i(\Delta n < 0) \leqslant D_i(\Delta n)/(M_i(\Delta n))^2$.

Доказательство. В соответствии с неравенством Чебышева имеем

$$q_i(|M_i(\Delta n) - \Delta n| > \Sigma) \leqslant D_i(\Delta n)/\Sigma^2,$$

где Σ — любое положительное число. Полагая $\Sigma = M_i(\Delta n)$ и используя условие положительности $M_i(\Delta n)$, получаем

$$q_i(M_i(\Delta n) - \Delta n > M_i(\Delta n) > \Delta n - M_i(\Delta n)) \leqslant D_i(\Delta n)/(M_i(\Delta n))^2,$$

откуда следует, что $q_i(\Delta n < 0) \leqslant D_i(\Delta n)/(M_i(\Delta n))^2$.

Разберемся более подробно в условии положительности $M_i(\Delta n)$, которое было использовано при доказательстве теоремы. Очевидно, что

$$M_1(\Delta n) = M_1(n_1(a)) - M_1(n_2(a)),$$
$$M_2(\Delta n) = M_2(n_2(a)) - M_2(n_1(a)). \qquad (1)$$

Здесь $M_j(n_i(a))$ — математическое ожидание числа «голосов» за класс A_i при условии, что объект a принадлежит классу A_j:

$$M_j(n_i(a)) = \sum_{\beta_i^l \succ A_i} \gamma_j(\beta_i^l); \tag{2}$$

$\gamma_j(\beta_i^l)$ — мера пересечения класса A_j с областью истинности признака β_i^l, отобранного для класса A_i (обозначение $\beta_i^l \succ A_i$ указывает, что суммирование производится по всем отобранным признакам β_i^l класса A_i).

Если для каждого класса отбирать лишь достаточно надежные признаки, то можно предположить, что вероятность ошибки для каждого из них меньше 1/2. Тогда для всех признаков β_1^l класса A_1

$$\gamma_1(\beta_1^l) > \gamma_2(\beta_1^l).$$

Аналогично для всех признаков β_2^l класса A_2

$$\gamma_2(\beta_2^l) > \gamma_1(\beta_2^l).$$

Отсюда и из (2) получаем

$$M_1(n_1(a)) > M_2(n_1(a)),\ M_2(n_2(a)) > M_1(n_2(a)).$$

Таким образом, в этом случае для того, чтобы условие положительности выполнялось одновременно для $M_1(\Delta n)$ и $M_2(\Delta n)$, достаточно, как видно из (1), потребовать, например, чтобы для отобранных признаков дополнительно удовлетворялось равенство

$$M_1(n_1(a)) = M_2(n_2(a)). \tag{3}$$

Доказанная теорема позволяет дать следующие рекомендации по отбору признаков в случае применения метода «голосования».

Поскольку вероятность q ошибки при классификации равна среднему значению вероятностей q_i ($\Delta n < 0$), то для уменьшения вероятности q необходимо 1) соблюдая, например, условие (3), т. е. сохраняя одинаковой для обоих классов сумму мер γ_i (см. (2)) отбираемых признаков, стремиться отбирать признаки с наименьшей вероятностью ошибки; 2) стремиться уменьшить дисперсии $D_i(\Delta n)$, т. е. отбирать признаки таким образом, чтобы всем объектам класса A_i соответствовала примерно одна и та же разница в числе характеризующих их признаков «своего» и «чужого» классов.

Поскольку при обучении нам доступны лишь конечные множества примеров каждого из классов A_i, то в качестве приближения для мер γ_i будем пользоваться их значениями, вычисленными на множестве примеров.

Один из способов отбора признаков, позволяющих уменьшить дисперсии $D_i(\Delta n)$, состоит в следующем. Среди признаков, имеющих наименьшую оценку вероятности ошибки [3], будем отби-

рать такие, чтобы для любой пары признаков существовали как объекты, обладающие первым признаком и не обладающие вторым, так и объекты противоположного свойства — обладающие вторым признаком и не обладающие первым. Это дает возможность сделать более равномерным разброс признаков по множеству объектов. Чтобы дополнительно уменьшить дисперсии $D_i(\Delta n)$, после первоначального отбора признаков можно произвести «доучивание». Для этого в каждом классе выделим примеры, охарактеризованные числом признаков, меньшим некоторого порога, и затем проведем дополнительный отбор признаков, характеризующих эти примеры. Этот дополнительный отбор может быть повторен несколько раз.

Описанный алгоритм может быть улучшен, если после проведенного указанным выше способом отбора признаков каждому из признаков β_i^l приписать некоторый вес $r(\beta_i^l)$ и заменить подсчет числа «голосов» подсчетом суммы весов отобранных признаков. Тогда выражение для математического ожидания суммы $\sigma_i(a)$ весов признаков «за класс» A_i при условии, что объект a принадлежит классу A_j, примет вид

$$M_j(\sigma_i(a)) = \sum_{\beta_i^l \geq A_i} \gamma_j(\beta_i^l) r(\beta_i^l).$$

Очевидно, что случай подсчета числа «голосов» эквивалентен случаю приписывания признакам единичных весов. Искомые веса $r(\beta_i^l)$ можно выбирать таким образом, чтобы минимизировать величину

$$Q = \gamma(A_1) D_1(\Delta\sigma)/(M_1(\Delta\sigma))^2 + \gamma(A_2) D_2(\Delta\sigma)/(M_2(\Delta\sigma))^2,$$

являющуюся верхней оценкой вероятности ошибки при классификации (см. [1, 2]) для случая неединичных весов признаков. (Здесь $\Delta\sigma$ — разность сумм весов признаков, голосующих за «свой» и «чужой» классы). При этом, как и прежде, в качестве приближений для мер γ можно пользоваться значениями, вычисленными на множестве примеров.

Литература

1. *М. М. Бонгард*. Моделирование процесса узнавания на цифровой счетной машине.— Биофизика, 1961, № 2.
2. *М. Н. Вайнцвайг*. Об одном алгоритме распознавания двоичных кодов.— Проблемы передачи информации, 1966, № 3.
3. Алгоритмы обучения распознаванию образов. Сб. под. ред. *В. Н. Вапника*. М., «Советское радио», 1972.
4. *М. М. Бонгард, М. Н. Вайнцвайг*. Об оценках ожидаемого качества признаков.— Проблемы кибернетики, 1968, вып. 20.