Классификация образов по критерию минимума расстояния

- Случай единственности эталона
- Множественность эталонов
- Обобщение принципов классификации по минимуму расстояния

Классификация образов с помощью функций расстояния - одна из первых идей автоматического распознавания образов. Этот простой метод классификации оказывается весьма эффективным инструментом при решении таких задач, в которых классы характеризуются степенью изменчивости, ограниченной в разумных пределах. В данном параграфе подробно рассматриваются свойства и способы реализации классификаторов, работающих на основе критерия минимума расстояния. Мы начнем с рассмотрения классов, которые можно характеризовать, выбрав по одному эталонному образу из класса. Затем полученные для этого случая результаты распространяются на случай нескольких эталонов. И, наконец, рассматриваются общие свойства этого метода классификации и определяются границы его классификационных возможностей.

Случай единственности эталона

В некоторых случаях образы любого из рассматриваемых классов проявляют тенденцию к тесной группировке вокруг некоторого образа, являющегося типичным и репрезентативным для соответствующего класса. Подобные ситуации возникают, если изменчивость образов невелика, а помехи легко поддаются учету. Типичным примером этого служит задача считывания банковских чеков с помощью ЭВМ. Символы, помещаемые на чеках, сильно стилизованы и обычно наносятся магнитной печатной краской с тем, чтобы упростить процесс снятия показаний. В ситуациях, подобных этой, векторы измерений (образы) в каждом классе будут почти идентичны, поскольку одинаковые символы на всех практически используемых чеках идентичны. В таких условиях классификаторы, действующие по принципу минимального расстояния, могут оказаться чрезвычайно эффективным средством решения задачи классификации.

Рассмотрим M классов; пусть эти классы допускают представление с помощью эталонных образов $z_1, z_2, \ldots, z_m$. Евклидово расстояние между произвольным вектором образа S и i-м эталоном определяется следующим выражением:

$$D_i = \|S - z_i\| = (S - z_i)'(S - x_j). \qquad (2.8)$$

Классификатор, построенный по принципу минимума расстояния, вычисляет расстояние, отделяющее неклассифицированный образ S от эталона каждого класса, и зачисляет этот образ в класс, оказавшийся ближайшим к нему. Другими словами, образ S приписывается к классу K_i, если условие $D_i < D_j$ выполняется для всех $j \neq i$. Случаи равенства расстояний разрешаются произвольным образом.

Формуле (2.1) можно придать более удобный вид. Возведение всех членов в квадрат дает

$$D_i^2 = \|S - z_i\|^2 = (S - z_i)'(S - z_i) = S'S - 2S'z_i + z_i'z_i = ¿\, S'S - 2(S'z_i - 1/2 z_i'z_i). \qquad (2.9)$$

Выбор минимального значения D_i^2 эквивалентен выбору минимального D_i, поскольку все расстояния - величины неотрицательные. Формула (2.9), однако, показывает, что выбор минимального значения D_i^2 эквивалентен выбору максимального значения разности $(S'z_i - 1/2 z_i'z_i)$, поскольку при вычислении любых $D_i^2, i = 1,2,\ldots,M$ член $S'S$ не зависит от значения i. Следовательно, решающие функции можно определять как

$$d_i(S) = S'z_i - (1/2) z_i'z_i, i = 1,2,\ldots,M, \qquad (2.10)$$

где образ S относится к классу K_i, если условие $d_i(S) > d_j(S)$ справедливо для всех $j \neq i$.

Отметим, что $d_i(S)$ - линейная решающая функция, т.е. если $z_{ij}, j=1,2,\ldots,n$ - компоненты вектора z_i, причем

$$K_{ij}=z_{ij}, j=1,2,\ldots,n, K_{i,n+1}=-(1/2)\, z_i' z_i \qquad (2.11)$$

и $S=\left(\underset{\square}{S_1} S_2 \ldots \underset{\square}{S_n} 1\right)'$,

то (2.10) можно представить в обычной линейной форме

$d_i(S)=K_i' S, i=1,2,\ldots,M$, где $S_i=\left(S_{i_1}, S_{i_2},\ldots, S_{i,n+1}\right)$. (2.12)

На рис. 2.4 изображена разделяющая граница для примера с двумя классами, каждый из которых задавался единственным эталоном. В конце данной главы в качестве одного из упражнений предлагается показать, что линейная разделяющая поверхность, обеспечивающая разделение всех пар эталонных точек z_i и z_j, является гиперплоскостью, которая представляет собой геометрическое место точек, равноудаленных от этих двух эталонных точек.

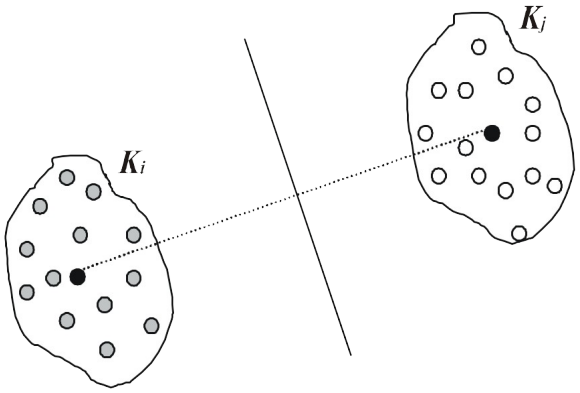

Рис. 2.4. Граница, разделяющая два класса, каждый из которых определяется одним эталоном

Мы убедились, таким образом, что классификаторы, основанные на принципе минимального расстояния, представляют собой частный случай линейного классификатора, разделяющие границы которого должны обладать указанным свойством. Поскольку классификатор, основанный на принципе минимального расстояния, классифицирует образы, исходя из наиболее полного совпадения образа с эталонами соответствующих классов, этот подход называют также корреляцией или сопоставлением с кластером.

Множественность эталонов

Допустим, что каждый класс можно охарактеризовать не единственным, а несколькими эталонными образами, т.е. любой образ, принадлежащий классу K_i, проявляет тенденцию к группировке вокруг одного из эталонов $z_i^1, z_i^2, \ldots, z_i^{Ni}$, где N_i - количество эталонных образов, определяющих i- й класс. В этом случае можно воспользоваться классификатором, подобным рассмотренному в предыдущем пункте. Запишем функцию, определяющую расстояние между произвольным образом S и классом K_i, в виде

$$D_i=\min\left\|S-z_i^l\right\|, l=1,2,\ldots,N_i. \qquad (2.13)$$

Это означает, что D_i - наименьшее из расстояний от образа S до каждого эталона класса K_i. Как и раньше, вычисляются значения расстояний D_i, $i=1,2,\ldots,M$ и классифицируемый образ зачисляется в класс K_i, если условие $D_i<D_j$ справедливо для всех $j\neq i$. В случае равенства расстояний решение принимается произвольным образом.

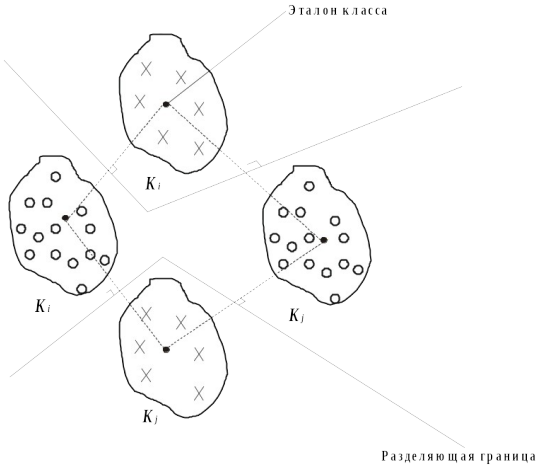

Рис. 2.5. Кусочно-линейные границы, разделяющие два класса, каждый из которых определяется двумя эталонами

Следуя процедуре, рассмотренной в п. "Случай единственности эталона", получаем решающие функции

$$d_i(S)=\max\left\{¿\,❑\left(S'z_i^l\right)-(1/2)\left(z_i^l\right)'z_i^l\right\},l=1,2,\ldots,N_i \qquad (2.14)$$

и как и раньше образ S зачисляется в класс K_i, если условие $d_i(S)>d_j(S)$ справедливо для всех $j\neq i$.

На рис. 2.5 представлены разделяющие границы для случая двух классов, когда каждый класс имеет два эталона. Обратите внимание на то обстоятельство, что границы, разделяющие классы K_i и K_j, являются кусочно-линейными. Этот случай можно было бы интерпретировать как задачу о разбиении на четыре класса, каждый из которых обладает единственным эталоном, тогда участки границ представляют собой геометрические места точек, равноудаленных от прямых, соединяющих эталоны различных классов. Это утверждение согласуется со свойствами разделяющих границ классификаторов для случая единственности эталонов, являющегося частным случаем соотношений (2.13) и (2.14).

Точно так же, как выражение (2.10) представляло частный случай линейного классификатора, выражение (2.14) является частным случаем классификаторов более общего вида - *кусочно-линейных*. Решающие функции таких классификаторов имеют следующий вид:

$$d_i(S)=\max\left\{¿\,❑d_i^l(S)\right\},i=1,2,\ldots,N_i, \qquad (2.15)$$

где функция $d_i^l(\chi)$ определяется выражением

$$d_i^l(S)=K_i^{l_1}S_1+K_i^{l_2}S_2+\ldots+K_i^{l_n}S_n+K_{i,n+1}^l=\left(K_i^l\right)'S. \qquad (2.16)$$

В отличие от решающих функций, определяемых формулой (2.14), от этих решающих функций не требуется соответствия форме, представленной на рис. 2.5.

Одной из основных проблем синтеза классификаторов образов является задача определения параметров решающей функции. Выше отмечалось, что известны универсальные итеративные алгоритмы, которые можно использовать для определения параметров линейной решающей функции. К сожалению, до сих пор не известен действительно общий алгоритм для кусочно-линейного случая (2.15), (2.16). Заметим, однако, что частные случаи (2.13) и (2.14) реализуются легко, если классы обладают относительно небольшим числом эталонов.

Обобщение принципов классификации по минимуму расстояния

Хотя идеи работы с небольшим количеством эталонов и евклидовыми расстояниями обладают геометрической привлекательностью, подход, основанный на классификации по критерию минимума расстояния, ими не исчерпывается. Для того, чтобы продолжить исследование общих свойств этой схемы классификации, рассмотрим выборку образов с известной классификацией $S_1, S_2, \ldots, S_n$, причем предполагается, что каждый образ выборки входит в один из классов $K_1, K_2, \ldots, K_M$. Можно определить правило классификации, основанное на принципе ближайшего соседа (БС-правило); это правило относит классифицируемый образ к классу, к которому принадлежит его ближайший сосед, причем образ $S_i \in \{S_1, S_2, \ldots, S_n\}$ называется ближайшим соседом образа S, если

$$D(s_i, S) = \min\{D(S_l, S)\}, l = 1,2,\ldots,N, \qquad (2.17)$$

где D- любое расстояние, определение которого допустимо на пространстве образов.

Эту процедуру классификации можно назвать 1-БС-правилом, так как при ее применении учитывается принадлежность некоторому классу только одного ближайшего соседа образа S. Нет, однако, причин, которые могли бы воспрепятствовать введению q-БС-правила, предусматривающего определение q ближайших к Sобразов и зачисление его в тот класс, к которому относится наибольшее число образов, входящих в эту группу. Сопоставление соотношений (2.17) и (2.13) показывает, что 1-БС-правило есть не что иное, как рассмотренный в предыдущем разделе случай множественности эталонов, если в качестве D выбирается евклидово расстояние.

Интересный результат, относящийся к сравнению 1-БС- и q-БС-правил, можно получить, обратившись к рис. 2.5. Допустим, что вероятность появления образов обоих представленных классов одинакова и образы K_i и K_j равномерно распределены в пределах соответствующих кругов R_i и R_j. В таком случае для выборки объема Nвероятность того, что точно α выбранных образов принадлежит классу K_i, определяется выражением

$$p_i = (1/2^N) C_n^\alpha, \qquad (2.18)$$

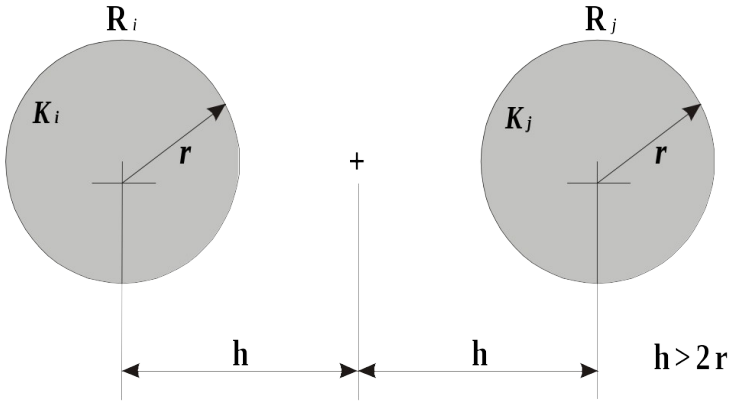

Рис. 2.6. Два класса, покрывающие идентичные области, в которых образы распределены равномерно

где $C_N^\alpha = N!/\alpha!(N-\alpha)!$ - число способов, которыми выборку объема N можно разделить на два класса, содержащих α и $N-\alpha$ элементов соответственно; 2^N определяет общее число способов разбиения N элементов на два класса. Очевидно, что вероятность p_j принадлежности α из N элементов выборки классу K_i, равна вероятности p_i.

Допустим, что классифицируемый образ S принадлежит классу K_i. При этом применение 1-БС-правила приведет к ошибке только в том случае, если ближайший сосед образа S входит в класс K_j и, следовательно, расположен в круге R_j. С другой стороны, если образ S принадлежит классу K_i, а его ближайший сосед находится в круге R_i, то в этом круге должны быть расположены все образы, что абсолютно очевидно из рис. 5. Это означает, что вероятность ошибки при применении 1-БС-правила равна в этом случае вероятности

принадлежности всех образов классу K_i, которую можно определить, положив $\alpha = N$ в выражении (2.18), т.е.

$$p_{e_1} = 1/2^N. \qquad (2.19)$$

Подобным же образом можно определить вероятность совершить ошибку при использовании q- БС-правила. Это правило зачисляет классифицируемый образ в класс, к которому принадлежит большинство его q ближайших соседей. Поскольку рассматривается случай разделения на два класса, в качестве q можно выбрать нечетное целое число, и следовательно, принцип большинства всегда будет работать.

Допустим, что образ S принадлежит классу K_i и он, следовательно, расположен в круге R_i. В таком случае применение q-БС-правила приведет к неправильной классификации только при условии, что в круге R_i находится $q - 1/2$ или меньшее количество образов. При этом нельзя располагать большинством, превышающим $q - 1/2$ ближайших соседей из круга R_i, необходимым для подтверждения правильности зачисления образа S в класс K_i. Соответствующая вероятность, являющаяся, по существу, вероятностью ошибки при использовании q-БС-правила, равна сумме вероятностей вхождения $0,1,2,\ldots,(q-1)/2$ элементов выборки в круг R_i. Следовательно, воспользовавшись уравнением (2.18), получаем выражение для вероятности ошибки при использовании q-БС-правила

$$p_{e_q} = 1/2^n \sum_{\alpha=0}^{(q-1)/2} C_N^{\alpha}. \qquad (2.20)$$

Сопоставление вероятностей ошибки классификации p_{e_1} и p_{e_q}показывает, что в данном случае 1-БС-правило характеризуется строго меньшей вероятностью ошибки, чем любое q-БС-правило $(q \neq 1)$, если все расстояния, разделяющие образы одного класса меньше всех расстояний между образами, принадлежащими различным классам.

Можно также показать, что в случае выборок большого объема $(N \to \infty)$ и при выполнении некоторых благоприятных условий вероятность ошибки 1-БС-правила заключена в следующих пределах:

$$p_B \le p_{e_1} \le p_B\left(2 - \left(M/(M-1)\right)p_B\right), \qquad (2.21)$$

где p_B - байесовская вероятность ошибки. Байесовская вероятность ошибки - наименьшая вероятность ошибки, достижимая в среднем.

Неравенство (2.14) показывает, что вероятность ошибки для 1-БС-правила превышает вероятность ошибки для правила Байеса не более чем в два раза. Это выражение устанавливает теоретические верхний и нижний пределы качества классификации с помощью 1-БС-правила. Практическим препятствием, однако, является то обстоятельство, что для достижения указанных границ необходимо сохранять в памяти большое число образов, о которых известна принадлежность их некоторому классу. Кроме того, при осуществлении классификации необходимо вычислять расстояния между каждым классифицируемым образом и всеми образами, хранящимися в памяти системы. При больших объемах обучающих выборок это обстоятельство вызывает серьезные вычислительные трудности.