Признаковое пространство.
Информативность признаков

- Множество выбора признаков
- Информативность признаков
- Методы определения информативности признаков

Выбор совокупности признаков объектов представляет собой серьезную научную проблему, предваряющую разработку самих методов распознавания образов. В этом процессе существенной проблемой являются такие вопросы; какие входные данные можно считать уместными и какая предварительная обработка приводит к получению "свойств" или "признаков"?

Априорные знания, интуиция, метод проб и ошибок, опыт, так или иначе используются при определении признаков. При наличии значительного объема априорных сведений о регулярностях, составляющих объекты, и при условии, что эти регулярности просты и имеют детерминистскую природу, отыскание признаков не составляет трудности. При распознавании приходится сталкиваться со сложными ситуациями, когда явных знаний о соответствующих регулярностях имеется немного, а сами регулярности отличаются существенными флуктуациями и изменчивостью.

В зависимости от специфики задачи используется множество типов признаков. Некоторые признаки хорошо поддаются определению и легко интерпретируются содержательно. Например, при классификации автомобилей на легковые и грузовые их длина и высота могут оказаться полезными (*информативными*) признаками. Однако более сложные признаки, основанные на форме, текстуре, статистических связях и т.д. необходимы для классификации электрокардиограмм, клеток крови, символов и т.п. Если получаемые в результате признаки можно рассматривать как *статистические величины*, то для выделения из исходного набора признаков наиболее важных можно воспользоваться статистическими методами выбора признаков. Если признаки можно рассмотреть как непроизводные элементы и их отношения, то для описания и анализа объектов можно воспользоваться лингвистическими подходами.

Пусть задана совокупность признаков $\{1, 2, \ldots, n\}$. Каждый из признаков i имеет множество допустимых значений M_i, $i=1, 2, \ldots, n$. Обычно рассматриваются признаки, имеющие следующие множества значений:

1^0. $M_i^2 = \{0, 1\}$ – признак выполнен или не выполнен на объекте;

2^0. $M_i^k = \{0, 1, 2, \ldots, k\text{-}1\}$ – признак имеет несколько градаций; $k \geq 2$;

3^0. $\widetilde{M}_i^k = \{a_1, \ldots, a_k\}$ - признак понимает конечное число значений, элементы из $\widetilde{M}_i^k$, вообще говоря, не являются числами; $k \geq 2$;

4^0. $M_i = [a, b]$, $(a, b]$, $[a, b)$, (a, b), где a, b - произвольные числа или символы $-\infty$, $+\infty$;

5^0. M_i - некоторое более сложно устроенное подмножество множества действительных чисел;

6^0. M_i^f - значениями признака i являются функции из некоторого класса функций;

7^0. M_i^μ - значениями признака являются функции распределения некоторой случайной величины.

Приведенные здесь множества 1^0 - 7^0 не исчерпывают, очевидно, многообразие признаков, встречающихся в задачах распознавания.

В дальнейшем будем считать, что множества M_i значений признака могут быть пополнены элементом Δ, означающим, что значение i-го признака неизвестно. Множество $M_i \Delta$ будем обозначать $M_i(\Delta)$.

Пусть даны наборы $(b_1,\ldots, b_n)=\tilde{b}$, $(c_1,\ldots, c_n)=\tilde{c}$, $b_i, c_i \ M_i(\ )$. Наборы $\tilde{b}$, $\tilde{c}$ называются различными, если существует хотя бы один номер i такой, что при b_i и c_i , $b_i \ c_i$, $1 \ i \ n$.

Определение 1.10. Описание

$$I(S)=(a_1(S),\ldots, a_n(S)), \ a_i(S) \ M_i(\ ) \qquad (1.17)$$

допустимого объекта S $(S\,M)$ называется стандартным описанием S. Описание I (S) называется полным, если $a_i(S)$, $i=1, 2,\ldots, n$.

По традиции существует несколько специально выделяемых классов признаков и наименований для них. Так, признаки i с множеством значений M_i^2 называются бинарными, с множеством $M_i(4^0)$ - простыми числовыми, с множеством $M_i(5^0)$ - сложными числовыми, со множеством M_i^f - функциональными, с множеством M_i^μ - вероятностными.

Особое значение в задачах распознавания имеют признаки со специальными дополнительными условиями на множество M_i.

Определение 1.11. Признак i, $1 \quad i \quad n$, такой, что M_i является метрическим пространством, называется метрическим.

Если метрика в M_i обозначена через $_i$, то признак i может в дальнейшем обозначаться как $(M_i, \ _i)$. В некоторых случаях функция $_i$ удовлетворяет всем аксиомам расстояния, кроме аксиомы треугольника. Тогда $_i$ называется полуметрикой, а признак $(M_i, \ _i)$ - полуметрическим.

В реальных задачах стандартные описания допустимых объектов обычно включают признаки разных типов.

Вопрос отбора признаков связан в первую очередь с качеством классификации. Особую важность он приобретает в системах последовательного распознавания [49], в которых на протяжении всего последовательного процесса классификации необходимо выбрать для измерения наиболее “информативный” признак с тем, чтобы процесс мог быть завершен как можно раньше. Для решения задачи так или иначе нужно привлечь понятия и меру информативности признака с точки зрения различения классов. Поэтому эту задачу будем называть I-задачей. Имеется ряд подходов к решению I-задачи [21, 24, 28, 31].

Использование функции энтропии

В работе [23] в качестве критерия отбора и упорядочения признаков предлагается использовать некоторую функцию в виде энтропии. Мерой информативности признака является число G_j, выбираемое как среднее значение этой функции.

Энтропия представляет собой статистическую меру неопределенности. Хорошей мерой внутреннего разнообразия для заданного семейства векторов образов служит энтропия совокупности, определяемая как

$$H = - E_p\{\ln p\}, \qquad (1.18)$$

где p – плотность вероятности совокупности образов, E_p – оператор математического ожидания плотности p. Понятие энтропии удобно использовать в качестве критерия при организации информативного набора признаков. Признаки, уменьшающие неопределенность заданной ситуации, считаются более информативными, чем те, которые приводят к противоположному результату.

Таким образом, если считать энтропию мерой неопределенности, то разумным правилом является выбор признаков, обеспечивающих минимизацию энтропии рассматриваемых классов. Поскольку это правило эквивалентно минимизации, дисперсии в различных совокупностях образов, то вполне можно ожидать, что соответствующая процедура будет обладать кластеризационными свойствами.

Рассмотрим l классов. Объекты характеризуются плотностями распределения соответствующей совокупности классов $p(S/K_1), p(S/K_2),\ldots, p(S/K_l)$. В силу определения энтропии, энтропия i-го класса

$$H_i = - \int_S p(S/K_i)\ln p(S/K_i)\,\mathrm{dx}, \qquad (1.19)$$

где интегрирование осуществляется по пространству образов.

Очевидно, при $p(S/K_i)=1$, т.е. при отсутствии неопределенности, имеем $H_i=0$.

В работе [45] предложен другой подход к решению задачи не требующий полного знания вероятностных описаний объектов и основанный на *использовании разложений Карунена-Лоэва*.

Основное преимущество разложения Карунено-Лоэва состоит в том, что оно позволяет обойтись без знания плотностей распределения образов, входящих в отдельные классы. Разложение вводится для случая непрерывных образов, а затем распространяется на дискретный случай, важной с точки зрения практики, машинной реализации.

Выбор признаков посредством аппроксимации функциями

Если признаки образов, составляющих некоторый класс, можно охарактеризовать с помощью функции *f(x)*, определяемой на основе результатов наблюдений, то процесс выбора признаков можно рассматривать как задачу аппроксимации некоторой функцией. В процессе обучения известны значения функции признаков *f(x)* в точках, соответствующих выборочным образам $x_1, x_2, \ldots, x_N$. Необходимо найти такую аппроксимацию *f(x)* функции *f(x)*, чтобы обеспечивалась оптимизация по некоторому критерию качества. Существуют различные методы определения аппроксимирующих функций. В [45] рассматривается метод разложения по системе функций, метод стохастической аппроксимации и метод аппроксимации с помощью ядер применительно к задаче аппроксимации функций признаков.

Выбор признаков на основе максимизации дивергенции

Соответствующий подход к выделению признаков заключается в порождении множества признаков, свойства которых позволяют максимизировать меру различия между классами. Если выделено множество признаков, которое после применения с помощью соответствующего преобразования к двум или нескольким совокупностям образов обеспечит получение множества преобразованных образов, отличающееся более заметным разделением совокупностей образов различных классов, то такие признаки можно рассматривать как характеристики, выявляющие различия совокупностей. Эта задача рассмотрена с точки зрения использования матричного преобразования для получения таких преобразованных образов, которые обеспечивают максимизацию расстояния между множествами при сохранении постоянства внутримножественного расстояния или соответственно суммы расстояний. Разделение классов, однако, можно оценивать не евклидовым расстоянием, а иными величинами. Более общим понятием расстояния является дивергенция, как мера "расстояния" или "несходства", "расхождения".

Рассмотрим две совокупности образов K_1 и K_2, характеризующиеся плотностями распределения $p_1(x)=p(x \vee K_1)$ и $p_2(x)=p(x \vee K_2)$, соответственно. Дивергенция между этими двумя классами определяется как

$$J_{12}=\int_x \left[p_1(x)-p_2(x)\right]\ln\frac{p_1(x)}{p_2(x)}\,\mathrm{dx}. \qquad (1.20)$$

Дивергенция должна быть использована в качестве функции критерия при порождении оптимального множества признаков. Более подробно данный метод рассмотрим [45,62].

Знание информативности признаков как оценку его существенности при решении классификационных задач (*I*-задачи) позволило бы успешно решить задачу отбора признаков.

В социологических исследованиях часто возникает задача выделения из исходной системы признаков наиболее эффективной подсистемы.

Пусть $(_1, _2,..., _n)$ - исходная система признаков, необходимо указать наиболее эффективную подсистему из t признаков $(t \quad n)$. Признаки исходной системы, как правило, оказывается зависимы, вследствие чего информативность признаков обуславливается тем, с какой системой он сочетается. Эту задачу можно решать полным перебором таких подсистем. Общее число таких подсистем равно числу сочетаний C_n^t. Ясно, что такой подход непригоден с вычислительной точки зрения.

Рассмотрим два алгоритма, позволяющих избежать полного перебора, близких к оптимальному (алгоритм *A*, алгоритм *B*).

В алгоритме *A* сначала перебираются все признаки исходной системы. Для каждого признака определяется значение критерия "ценности" *F*, и по этому критерию выбирается наилучший признак. Вторым выбирается признак, который в сочетании с выбранным дает наилучшее сочетание критерия. Далее, подключением по одному признаку из $(n-2)$ оставшихся к уже выбранным двум устанавливается подсистема из трех признаков и т.д. – так составляется подсистема из m признаков.

При использовании алгоритма *B* сначала делается перебор n возможных подсистем, каждая из которых состоит из $(n-1)$ признаков. Выбирается та подсистема, использование которой обеспечивает наилучшее значение критерия *F*. Далее, используя признаки выбранной подсистемы, составляют $(n-1)$ возможных подсистем, каждая из которых состоит их $(n-2)$ признаков. Из этих подсистем выбирается наилучшая. После исключения $(n-m)$ признаков исходной системы определяется подсистема из m признаков.

Когда алгоритмы *A* и *B* не дают оптимального решения, предлагается алгоритм случайного поиска с адаптацией СПА, рассмотренный в работе [28]. Адаптация связана с необходимостью увеличения на очередном шаге поиска вероятности выбора различительных признаков. Сущность метода состоит в случайном поиске эффективной подсистемы признаков с "поощрением" и "наказанием" отдельных признаков из $(_1, _2,..., _n)$.

В работе [23] рассматривается вычисление информативности признаков посредством анализа тупиковых тестов. Набор признаков $(_1, _2,..., _k)$ образует тест, если после удаления из таблицы T_{mn} всех признаков (столбцов), за исключением перечисленных, изображение объектов, относящихся к разным классам K_j, представляет собой результат локально-максимального сжатия исходной таблицы, при котором еще возможно различение объектов из разных классов.

Если некоторый признак войдет в большое число таких неизбыточных описаний, то он окажется информативным. Пусть k - общее число тупиковых тестов для таблицы T_{mn}, $k(i)$ - число тупиковых тестов, содержащих столбец, соответствующий i-му признаку, величина $p(i)=\frac{k(i)}{k}$ называется информативностью i-го признака (информационный вес признака).

Определение информационных весов признаков в модели алгоритмов, вычисление оценок (АВО) выражается через эффективные вычислительные схемы. Рассмотрим множество объектов $S_1, S_2,..., S_{m,}$ сведенных в таблицу T_{mnl}. Предлагается априори, что известно разбиение объектов на классы. Применим голосующие процедуры к самим строкам таблицы T_{mnl} и подсчитаем величины $\Gamma_1(S_q), \Gamma_2(S_q),..., \Gamma_l(S_q), \quad q=1, 2,..., m_1$ для класса K_1, и аналогичные оценки для объектов класса $K_2, K_3,..., K_l$. Вычеркнем теперь из таблицы T_{mnl} столбец с номером i и удалим его также из голосующих множеств. Строки $S_1, S_2,..., S_m$ перейдут в строки $S_1^i, S_2^i, \ldots, S_m^i$ (верхний индекс i означает, что из строки удален i-й признак). Реализуя процедуру голосования для сокращений таблицы вычисляем величины $\Gamma_u(S_q^i)$, где $q=1, 2, ..., m, u=1, 2,.., l$. Количество голосов $\Gamma_u(S_q^i)$ будут меньше, чем соответствующие $\Gamma_u(S_q)$.

Если признак i (удаленный столбец с номером i) существенный, то число голосов в среднем уменьшается сильно, и наоборот. Степень уменьшений, обработанную надлежащим образом, следует считать мерой важности изучаемого признака.

Информационный вес i-го признака вводим следующим образом.

Составляем разности

$$\Gamma_1(S_1)-\Gamma_1(S_1^i),\Gamma_1(S_2)-\Gamma_1(S_2^i),\ldots,\Gamma_1(S_{m_1})-\Gamma_1(S_{m_1}^i) \quad (1.21)$$

Вводим величину

$$\Delta_1=\frac{1}{m_1}\left\{\left[\Gamma_1(S_1)-\Gamma_1(S_1^i)\right]+\ldots+\left[\Gamma_1(S_{m_1})-\Gamma_1(S_{m_1}^i)\right]\right\} \quad (1.22)$$

Аналогично поступаем и с другими классами;
определяем $\Delta_2,\ldots,\Delta_l$. Величина

$$p_i=\Delta_1+\Delta_2+\ldots+\Delta_l=\sum_{u=1}^{l}\Delta_u \quad (1.23)$$

называется информационным весом i-го признака.
Окончательно выражение (1.23) выписывается так:

$$p_i=\sum_{u=1}^{l}\frac{1}{m_u-m_{u-1}}\sum_{q=m_{u-1}+1}^{m_u}\left[\Gamma_u(S_q)-\Gamma_u(S_q^i)\right]. \quad (1.24)$$