ОЦЕНКА АДЕКВАТНОСТИ НЕЧЕТКОГО МНОГОМЕРНОГО РАСПОЗНАВАТЕЛЯ НА ОСНОВЕ КЛАССИФИЦИРУЮЩЕГО ДЕРЕВА

Боброва Н. Л., Герман О. В.
Кафедра информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {natal123, ovgerman}@tut.by

Введение

Оценка адекватности любого распознавателя, как правило, строится на основе техники статистической проверки гипотез. Имеется некоторое обучающее множество, на основе которого построен данный распознаватель. Обучающее множество представлено множеством векторов $V_i = \langle x_{i1}, x_{i2}, \ldots, x_{in} \rangle, i = 1, N$. Для каждого вектора известна нечеткая мера принадлежности его к заданному классу $0 \le f_i \le 1$. В результате обучения распознаватель отнес V_i к классу с оценкой g_i. Таким образом, у нас есть два ряда чисел $\langle f, g \rangle$ и проблема стоит в том, чтобы оценить их статистическую адекватность. Эта задача хорошо известна. Для ее решения применяют те или иные оценки, из которых укажем следующие [1].

A) Коэффициент детерминации $r^2 = (S_t - S_r)/S_t$, где $S_t = \sum(f_i - mean(f))^2$ и $mean(f)$ представляет среднее значение по всем $f_i, i = 1, N$; $S_r = \sum(f_i - g_i)^2$. Коэффициент детерминации r^2 равен единице, если f и g полностью совпадают. Поэтому на практике считают, что значение этого коэффициента близкое к 0.8 и выше означает «очень хорошее» совпадение.

B) Средняя ошибка аппроксимации $\varepsilon = \sum(f_i^{-1}|f_i - g_i|)100\%$. На практике считают, что значение этой ошибки не должно превосходить 10-12%.

C) Критерий χ^2. Этот критерий рассчитывают по формуле $\chi^2 = \sum f_i^{-1}(f_i - g_i)^2$. Затем рассчитанное значение сравнивают с табличным с учетом величины вероятности ошибок (обычно $\alpha = 0.05$ или $\alpha = 0.01$) и числа степеней свободы k. Для принятия гипотезы необходимо, чтобы табличное значение критерия χ^2 было не меньше расчетного.

D) Используют также другие критерии, например, критерий Фишера. При использовании критерия Фишера рассчитывается отношение дисперсий $F = s^2_{\text{ад}}/s^2_{\text{общ}}$, где $s^2_{\text{ад}}$ - дисперсия адекватности, вычисляемая по формуле $s^2_{\text{ад}} = (N - p)^{-1}\sum(f_i - g_i)^2$,где p - число коэффициентов, фигурирующих в аналитическом представлении модели; $s^2_{\text{общ}}$ - общая дисперсия, значение которой находят из формулы $s^2 = (N - 1)^{-1}\sum(f_i - mean(f))^2$ и $mean(f)$ - среднее значение нечеткой меры принадлежности. Расчетное значение критерия Фишера не должно превосходить табличного значения.

Проблематичным местом в C) и D) является определение числа степеней свободы. Степень свободы – это число элементов в статистике, которые могут принимать значения безотносительно к другим элементам. В простых случаях определение числа степеней свободы не вызывает трудностей. Так, если элементами статистики являются данные о росте людей, то число степеней свободы равно числу этих людей, т.к. рост каждого отдельного человека не зависит от роста другого (если не принимать во внимание родственников). В случае (нечеткого) многомерного распознавателя на основе модели число степеней свободы равно объему обучающей выборки минус число параметров модели (назовем это правило RM). Проблема возникает тогда, когда для принятия решений используется модель, аналитическое описание которой не известно, либо оно известно, но содержит более одной формулы. В последнем случае возникает проблема выявления зависимых формул. В принципе, эта задача решена математически с помощью определителей Вронского, поэтому мы рассматриваем здесь модель распознавателя на основе дерева, описанную в работе [2], в которой общее аналитическое описание не известно. В данной модели узлам дерева приписаны линейные дискриминаторные функции вида $h = a_0 + a_1x_1 + a_1x_1 + \ldots + a_nx_n$. При реализации распознавания выполняется «проход» по узлам дерева в тупиковую вершину, ассоциированную с классом K или не $-K$. Проход по дереву начинается с корневой вершины. Вычисляется значение дискриминаторной функции, ассоциированной с этой вершиной. Если значение функции неотрицательно, то выполняется переход в левый дочерний узел данной вершины, в противном случае – в правый дочерний узел. Для выбранного узла действия повторяем, пока не доберемся до тупиковой вершины в дереве, причем для тупиковой вершины известно, какой класс K или не $-K$ она представляет, что и завершает процесс распознавания. Усовершенствованный вариант этой техники изложен в работе [3]. Цель настоящей работы представить использование кри-

терия χ^2 относительно многомерного нечеткого распознавателя [3].

I. Описание методики

Как было отмечено, [3] реализует методику распознавания на классифицирующем дереве, на базе которого построен четкий распознаватель [2]. Чтобы получить нечеткие заключения о принадлежности исходные векторы модифицируются следующим образом. Пусть, например, в обучающем множестве имеется вектор $V_i = \langle 0, 2, 3.5\rangle$ с мерой принадлежности к множеству , равной $f_i = 0.8$. Соответственно, мера принадлежности этого вектора к множеству не − равна 0.2 . Вместо вектора $V_i = \langle 0, 2, 3.5\rangle$ вводятся два четких вектора - один: $V_{i1} = \langle 0, 2, 3.5, 0.8\rangle$ и второй $V_{i2} = \langle 0, 2, 3.5, 0.2\rangle$. Вектор V_{i1} строго принадлежит , вектор V_{i2} строго не принадлежит . Таким образом, достигается переход к четкому варианту постановки задачи распознавания. По модифицированному обучающему множеству строится классифицирующее дерево [2]. Теперь на вход распознавателя последовательно подаем модифицированные векторы $V_i = \langle 0, 2, 3.5, 0\rangle$, $V_i = \langle 0, 2, 3.5, 0.05\rangle$, $V_i = \langle 0, 2, 3.5, 0.1\rangle$,. . . , $V_i = \langle 0, 2, 3.5, 1.0\rangle$. Видим, что последняя координата в этих векторах последовательно изменяется на достаточно малую величину $d > 0$ (в примере − на 0.05), начиная от наименьшего начального значения «0». Необходимо зафиксировать момент, когда четкий распознаватель первый раз «перебросит» объект V_i в противоположный класс. Значение добавленной координаты в этом векторе V_i и даст нам приближенную оценку g_i . Очевидно, что чем меньше величина шага d, тем точнее оценка g_i. Заметим, что нам не известен вид аналитической функции распознавания, реализуемой классифицирующим деревом. Поэтому применить критерий χ^2 непосредственно нельзя. Собственно, мы подошли тем самым к цели настоящего сообщения.

II. Оценка адекватности

Как уже было сказано, каждому узлу классифицирующего дерева поставлена в соответствие линейная дискриминаторная функция $h = a_0 + a_1x_1 + a_1x_1 + \ldots + a_nx_n$. Принципиально, нас даже не должны интересовать коэффициенты a_i, а только их число, которое равно разрядности вектора (вспомним, что при переходе от нечеткого случая к четкому это число возросло на единицу). Каждый узел соответствует некоторому известному заранее набору векторов. В самом деле, достаточно каждый вектор исходного обучающего множества «прогнать» через дерево, чтобы зафиксировать те узлы, по которым он перемещался. Если вектор «заходил» в узел, то он попадает в множество векторов, ассоциированных с этим узлом. Итак, мы имеем возможность применить критерий к каждому узлу дерева в отдельности, поскольку число степеней свободы для отдельно взятого узла оценивается по правилу RM. Если хотя бы для одного узла нарушается условие адекватности, то все дерево считается неадекватным и следует пересмотреть исходные данные на предмет их корректировки. Итак, с помощью описанной техники мы решаем проблему проверки адекватности нечеткого многомерного распознавателя.

Заключение

Описанная методика оценки адекватности нечеткого многомерного распознавателя обладает следующими достоинствами:

1) векторы в обучающем множестве могут иметь, вообще говоря, сколь угодно большую размерность, причем разряды векторов могут коррелировать друг с другом, что не влияет на качество распознавания;

2) закон распределения значений координат распознаваемых векторов не играет роли.

Указанные факторы существенны, например, при реализации процесса диагностики сложных объектов, в частности, в медицине, где количество психологических и психофизиологических показателей может достигать нескольких десятков, а именно: время и ошибки сенсомоторных реакций, реакции на движущийся объект и т. п.; острота зрения, контрастная чувствительность, абсолютные и дифференциальные пороги световой и цветовой чувствительности; устойчивость ясного видения; критическая частота слияния мельканий, фосфены; длительность последовательных образов; хронаксия и реобаза мышц; показатели долговременной, кратковременной и оперативной памяти на звуковые и зрительные стимулы; точность усилий и показатели двигательной памяти; устойчивость, концентрация, распределение и переключение внимания, объем внимания и другие.

Экспериментальная работа с классифицирующим деревом показала, что его размерность в среднем не превосходит корня квадратного от числа векторов в обучающем множестве. Например, если число векторов в обучающем множестве порядка 100, то число вершин в классифицирующем дерева порядка десятка. Следовательно, методика может быть предложена к практическому применению, т.к. она дает достаточно эффективный как по точности, так и по требуемым размерам памяти механизм распознавания.

III. Список литературы

1. Теория статистики / Г. Л. Громвко. – М.: ИНФА-М, 2010. – 414 с.
2. Герман, О. В. О реализации распознавателя с минимальным числом входов / О. В. Герман, Н. Л. Боброва, А. Р. Самко // Доклады БГУИР. – 2011. – июнь. – С. 86–93.
3. Герман, О. В. Многомерный нечеткий распознаватель на основе четкого распознавателя и его оценка / О. В. Герман, Н. Л. Боброва // Доклады БГУИР. – 2013. – в печати.