АНАЛИЗ ДАННЫХ И СИГНАЛОВ
(Вычислительные системы)
1998 год Выпуск 163

УДК 519

ДИСПЕРСИОННЫЙ КРИТЕРИЙ ДЛЯ ОЦЕНКИ ДОСТОВЕРНОСТИ РЕШЕНИЯ ПРИ ПРОГНОЗИРОВАНИИ И РАСПОЗНАВАНИИ С ИСПОЛЬЗОВАНИЕМ КОЛЛЕКТИВНО-ГРУППОВЫХ РЕШАЮЩИХ ПРАВИЛ[1]

Ю.И.Журавлёв, Н.Г.Загоруйко

Процесс анализа данных обычно состоит из двух этапов: обнаружение закономерностей, скрытых в имеющихся данных, и использование этих закономерностей для предсказания этих данных. Эмпирические закономерности могут иметь форму единственного предиктора, решение которого принимается в качестве окончательного значения предсказываемого факта. Но в последнее время находят все большее распространение коллективные решающие правила, имеющие форму конечного набора различных предикатов, каждый из которых выдает свой вариант предсказания. Окончательное решение принимается на основании учета этих вариантов с помощью той или иной процедуры обобщения. Такую структуру имеют алгоритмы распознавания образов с помощью "метода комитетов" [1], "коллектива решающих правил" [2] и пр. Теоретическое обобщение этого подхода в рамках алгебраической теории распознавания образов представлено в [3,4].

Множество предикторов, входящих в коллективные решающие правила могут порождаться различными способами, образуя несколько групп предикторов. При этом возникает вопрос о способах сравнения этих групп по информативности (или "компетентности") и выбора наиболее компетентных из них.

[1]Работа выполнена в рамках проекта № 97-06-80312, поддержанного Российским фондом фундаментальных исследований

Традиционно, если нужно выбирать одно из k имеющих решающих правил или стратегий $S_1, S_2, \ldots, S_i, \ldots, S_k$, используется метод ретроспективного анализа: каждая стратегия применяется для распознавания или прогнозирования верифицированной выборки и выбирается та стратегия S_i, при которой количество ошибок R_i оказалось минимальным. Делается это на основании гипотезы о сохранении "средних" закономерностей: предполагается, что ситуация при очередном событии или в следующий момент времени будет такая же, как и в большинстве прошлых случаев.

Однако, полностью согласиться с такой гипотезой мешает тот факт, что лучшие результаты на прошлом материале иногда получались при использовании не выбранной стратегии S_i, а некоторых других стратегий. Следовательно, бывают ситуации, для которых более адекватными являются некоторые из отвергнутых стратегий. Как узнать, какая стратегия является наиболее компетентной при данной конкретной ситуации, когда нам требуется сделать реальное распознавание или прогнозирование?

Вспомним гипотезу "компактности", на которой, по существу, основаны все современные методы анализа данных. В трактовке ее авторов эта гипотеза (обозначим ее через H) состоит в том, что реализации одного и того же образа обычно отображаются в признаковом пространстве в геометрически близкие точки, образуя "компактные" сгустки [5].

Конечно, она подтверждается не всегда. Если, например, среди признаков имеется много случайных, не информативных, то точки одного образа могут оказаться далекими от друга и рассеянными среди точек других образов. Это соответствует случаю, когда для распознавания обучающей выборки приходится строить "слишком вычурные" разделяющие функции, что обычно приводит к большим ошибкам распознавания контрольной выборки. При удачном же выборе признакового пространства точки одного класса действительно образуют явно выделяемые компактные сгустки.

Назовем n признаков, входящих в информативное подмножество X, "описывающими", а $(n+1)$-й признак z, указывающий имя образа или значение прогнозируемой величины, "целе-

вым". Обозначим множество объектов обучающей выборки через A, новый распознаваемый (прогнозируемый) объект через q, а тот факт, что объекты множества A "компактны" ("эквивалентны", "похожи" или "близки" друг к другу в пространстве n характеристик X, через C_A^X. Мера "компактности" может быть любой, она может характеризоваться средним расстоянием от центра тяжести до всех точек образа; средней длиной ребра полного графа или ребра кратчайшего незамкнутого пути, соединяющего точки одного образа; максимальным расстоянием между двумя точками образа и т.д. Например, "компактными" (эквивалентными) будем считать два объекта, если все признаки одного объекта равны соответствующим признакам другого. Или: объекты компактны, если Евклидово расстояние между векторами их признаков не превышает величину r.

Фактически гипотеза H равнозначна предположению о наличии закономерной связи между признаками X и z и, с учетом вышесказанного, ее тестовый алгоритм может быть представлен следующим выражением: if$(C_A^{X,z}$ & $C_{A,q}^X)$, then $C_{A,q}^z$. То есть, если объекты множества A компактны в пространстве (X, z) и добавление к множеству A объекта q не нарушает их совместной компактности в пространстве описывающих признаков X, то объекты A и q будут компактными и в пространстве целевого признака у объектов (A, q) будут одинаковыми или отличающимися на малую величину друг от друга.

Это соображение послужило основанием для предположения о том, что прогнозы компетентной группы предикатов будут отличаться от прогнозов ее менее компетентных конкурентов величиной дисперсии частных прогнозов, полученных от входящих в ее состав предикаторов.

Если соревнуются несколько групп предикаторов, то дисперсионный критерий должен позволить делать обоснованный выбор наиболее компетентной группы Если же используется одна группа (один коллектив решающих правил), то по величине дисперсии можно будет судить об ожидаемой ошибке распознавания или прогноза.

Первая проверка информативности дисперсионного критерия была сделана в рамках алгоритма заполнения пробелов в эмпири-

ческих таблицах с помощью алгоритма ZET [6]. Там используется две группы предикторов, опирающихся на "похожесть" строк и столбцов таблицы. Использование дисперсионного критерия для оценки относительной компетентности этих групп показало его явные преимущества перед тем критерием, который использовался в алгоритме ZET раньше.

Во втором эксперименте проверялась информативность дисперсионного критерия в алгоритме прогнозирования многомерных динамических рядов с помощью алгоритма GAP [7]. В этом алгоритме используется комбинаторный метод порождения групп предикторов, и количество таких групп может достигать нескольких сотен. И этот эксперимент подтвердил справедливость сделанного предположения о целесообразности использования дисперсионного критерия: ошибка прогноза полученного в результате использования частных прогнозов от предикторов данной группы, оказались прямо пропорциональной величине дисперсии этих прогнозов. Коэффициент корреляции между дисперсией и ошибкой достигает величины +0,7.

Дальнейшее развитие идеи использования дисперсионного критерия приводит к формулировке нового метода выбора компетентных предикторов: если в распоряжении имеется несколько групп предикторов (или несколько коллективов решающих правил), то используй все группы и выбирай ту из них, частные прогнозы в которой отличаются наименьшей дисперсией.

Дисперсионный критерий по отношению к группе предикторов можно использовать, как для оценки ожидаемой ошибки распознавания (прогноза), так к для оценки информативности системы признаков: тот факт, что решение сопровождается большой дисперсией говорит о малой информативности признаков для данной задачи.

На этом основании можно предложить следующую общую схему класса эффективных алгоритмов для решения задач распознавания образов и прогнозирования с помощью коллективно-групповых решающих правил (класс алгоритмов КГРП).

Алгоритмы этого класса состоят из четырех последовательных этапов:

1) генерации групп предикторов,

2) получения частных решений и оценки компетентности групп,

3) формирования обобщенного решения и

4) оценки ожидаемой ошибки.

1. На этапе генерации групп предикторов можно использовать любые способы порождения семейств решающих правил или групп предикторов. Различные схемы этого процесса предложены в [3,4,7].

Представим себе, что для распознавания используются линейные решающие правила в виде конечного набора из k гиперплоскостей. Если вместо констант в уравнениях этих плоскостей использовать непрерывно изменяемые параметры, то на базе каждой гиперплоскости можно породить группу из бесконечного числа гиперплоскостей, отличающихся друг от друга сдвигами и наклонами к координатным осям. Если использовать конечное число G дискретных значений изменяемых параметров, то каждая плоскость будет порождать группу из G различных гиперплоскостей. Средствами этого коллектива групп решающих правил можно получать $k * G$ частных результатов распознавания.

Наряду с гиперплоскостями можно использовать и другие типы решающих правил, например, наборы из квадратичных, таксономических или логических решающих правил. Каждый набор может породить свой коллектив решающих правил, и мы получим множество из W коллективов по k групп, состоящих из G решающих правил. Это множество порождает $W*k*G$ частных результатов распознавания.

Комбинаторный механизм порождения предикатов представляет собой разновидность метода параметрического расширения и состоит в следующем. Пусть таблица данных отражает значение n характеристики $x_1, x_2, \ldots\ x_j, \ldots, x_n$ состояния наблюдаемого процесса b в T последовательных моментов времени $t_1, t_2, \ldots, t_i, \ldots, t_T$. Строки упорядочены в порядке, обратном времени появления наблюдений: первая строка таблицы отражает характеристики в последний момент t_1, вторая — в предыдущий момент t_2 и так до самого первого момента t_T.

В слое таблицы с 1-й по m-ю строку содержится $S = n * m$ элементов. Выберем из них $G < S$ любых элементов $b(ij)$. Назовем этот набор элементов "базовым штаммом" мощности G. Если у каждого элемента базового штамма заменить параметр i на параметр $(i + k)$, то получится штамм той же мощности и формы (архитектуры), что и базовый, но сдвинутый в прошлое на k моментов времени. Меняя параметр k от 1 до L можно получить L штаммов, изоморфных базовому. Каждый изоморфный штамм в сочетании с базовым может служить предиктором [7]. Таким образом, мы получаем первую группу из L предикторов.

Если из указанного слоя таблицы взять другие G элементы и повторить с новым базовым штаммом описанные выше процедуры, то будет получена вторая группа предикторов. Всего таких разных групп, состоящих из L предикторов мощности G, может быть получено C_S^G штук.

Если изменить мощность базового штамма, то можно получить новый коллектив групп предикторов. Следовательно, из S элементов можно получить $\sum_{G=1}^{S} C_S^G$ разных групп предикторов.

2. Этап получения частных решение и оценки компетентности групп состоит в следующем.

Все предикторы каждой группы в отдельности принимают свои частные решения о принадлежности распознаваемого объекта к тому или иному образу или о значении прогнозируемой характеристики. В случае прогнозирования характеристики, измеренной в сильной шкале, для предикторов группы f вычисляется дисперсия D_f их частных прогнозных значений.

В случае распознавания, т.е. прогнозирования характеристики, измеренной в шкале наименований, компетентность предикторов можно оценивать через величину, близкую по смыслу к дисперсии, — через энтропию решений. Если распознается s образов, то при равномерном распределении решений в пользу всех образов энтропия решений будет максимальной и равной $H_0 = \ln s$. Если в пользу образа v высказалось pv предикторов из L, то энтропия предсказаний в группе будет равной

$$H_f = -\sum_{v=1}^{S} \frac{pv}{L} * \ln \frac{pv}{L}.$$

Компетентность Q_f предикторов группы f при распознавании образов можно принять равной $Q_f = \left(1 - \frac{H_f}{H_0}\right)$. Если энтропия H_f решений равна 0, то компетентность максимальна и равна 1. Если энтропия решений $H_f = H_0$, то компетентность такой группы предикторов естественно считать минимальной и равной 0.

Для случая прогнозирования непрерывной характеристики компетентность группы можно принять равной $Q_f = \left(1 - \frac{D_f}{D_{max}}\right)$, где D_{max} — наибольшее значение дисперсии среди всех сравниваемых групп предикторов. Неопределенность 0:0 приравнивается нулю. Как и в предыдущем случае, характеристики компетентности меняются в пределах от 0 до 1.

3. Этап получения обобщенного решения состоит в следующем. Пусть в принятии решений на предыдущем этапе принимали участие предикторы, объединенные в N групп. Вначале на базе частных решений для каждой группы f вырабатывается групповое решение B_f. При распознавании образов таким решением может быть имя образа, набравшего наибольшее количество голосов в группе. При прогнозировании в качестве группового решения может быть использовано среднеарифметическое значение частных прогнозов или их медиана.

Обобщенное решение (B) может быть получено с использованием параметрического семейства функций взвешенного усреднения: $B = \frac{\sum_{f=1}^{N} B_f * Q_f^{\alpha}}{\sum_{f=1}^{N} Q_f^{\alpha}}$.

Здесь величина показателя степени α отражает стратегию учета влияния компетентности. Если $\alpha = 0$, то решение всех групп учитываются с равными весами. С ростом α растет влияние более компетентных групп. При очень больших значениях α в усреднении будет участвовать одна или несколько самых компетентных групп.

4. На последнем этапе вырабатывается оценка ожидаемой ошибки распознавания или прогноза. С этой целью при прогнози-

ровании количественной характеристики вычисляется дисперсия (D^*) групповых решений B_f. При этом также можно учитывать компетентность групп Q_f:

$$D^* = \sqrt{\frac{\sum_{f=1}^{N} \left\{(B - B_f) * Q_f^{\alpha}\right\}^2}{\sum_{f=1}^{N} Q_f^{2\alpha}}}.$$

Для случая распознавания образов определяется количество групповых решений в пользу каждого v-го образа $(p'v)$ и вычисляется энтропия групповых решений:

$$H^* = -\sum_{v=1}^{N} \frac{p'v}{N} * \ln \frac{p'v}{N}.$$

Содержательная интерпретация величин D^* и H^* очевидна: чем больше дисперсия (энтропия) групповых решений, тем больше ожидаемая ошибка распознавания или прогноза. Вопрос же о возможности указать по дисперсии количественное значение ошибки является предметом дальнейших исследований.

Литература

1. МАЗУРОВ В.Д., ТЯГУНОВ Л.И. Метод комитетов в распознавании образов //Метод комитетов в распознавании образов. — Свердловск, 1974. — С. 10–40.

2. РАСТРИГИН Л.А., ЭРЕНШТЕЙН Р.Х. Принятие решений коллективом решающих правил в задачах распознавания образов //Известия АН СССР. Автоматика и телемеханика. — 1975, № 9. — С. 133-144.

3. ЖУРАВЛЕВ Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов, I–III //Кибернетика. — Киев. — 1977. — № 4. — С. 14–21; № 6. — С. 21–27; 1978. — № 2. — С. 35–43.

4. ZHURAVLEV Yu.I. An Algebraic Approach to Recognition or Classification Problems, Pattern Recognition and Image Analysis/ 8(10, 59–100 (1998).

5. АРКАДЬЕВ А.Г., БРАВЕРМАН Э.М. Обучение машины распознаванию образов. — М.: Наука, 1964.

6. Пакет прикладных программ ОТЭКС /Н.Г.Загоруйко, В.Н.Ёлкина, С.В.Емельянов и др. — М.: Финансы и Статистика, 1986. — 160 с.

7. ЗАГОРУЙКО Н.Г. Самообучающийся генетический алгоритм прогнозирования (GAP). — Новосибирск, 1997. — Вып. 160: Вычислительные системы. — С. 80–95.

Поступила в редакцию
30 декабря 1998 года