АНАЛИЗ ДАННЫХ И ЗНАНИЙ В ЭКСПЕРТНЫХ СИСТЕМАХ
(Вычислительные системы)
1990 Выпуск 134

УДК 519.2

СРАВНЕНИЕ АЛГОРИТМОВ РАСПОЗНАВАНИЯ С ПОМОЩЬЮ ПРОГРАММНОЙ СИСТЕМЫ "ПОЛИГОН"

Г.С.Лбов, Н.Г.Старцева

В настоящий момент известно более ста алгоритмов построения решающих правил распознавания. Поэтому возникает проблема выбора алгоритма для эффективного решения конкретной прикладной задачи. Для решения данной проблемы необходимо сравнить алгоритмы распознавания и выделить области наиболее эффективного применения того или иного алгоритма.

Трудности сравнения алгоритмов связаны со следующими вопросами: по каким принципам и критериям сравнивать алгоритмы, как формировать тестовые примеры,каков должен быть набор сравниваемых алгоритмов. Кроме того,сравнение алгоритмов предполагает большую техническую работу, связанную с решением прикладных и тестовых задач.

Для экспериментального сравнения алгоритмов необходимо создание полигона, включающего в себя алгоритмы распознавания образов, архив тестовых и прикладных задач, ряд сервисных программ, обеспечивающих сравнение.

Авторами предложен принцип сравнения алгоритмов, на основании которого разработана методика их сравнения.Создано программное обеспечение "Полигон", включающее в себя шесть алгоритмов распознавания, около пятидесяти прикладных задач и пятьдесят одну тестовую задачу, а также сервисную часть, обеспечивающую сравнение.

§1. Основные определения и постановка задачи сравнения

Пусть для описания каждого объекта $a \in \Gamma$ (Γ – генеральная совокупность) используются признаки $X_1, \ldots, X_j, \ldots, X_n$ и целевой признак $X_{n+1} \in \Omega, \Omega = \{1, \ldots, \omega, \ldots, k\}$, где k – число образов, $k \geq 2$. Если признаки являются непрерывными случайными величинами, то для них определены условные плотности распределения $p(x/\omega)$, $\omega = \overline{1,k}$, в многомерном пространстве D, где $x = (X_1(a), \ldots, X_j(a), \ldots, X_n(a))$, $x \in D$, $D = \prod_{j=1}^{n} D_j$, D_j – область значений признака X_j; $X_j(a)$ – значение признака X_j для объекта a, n – размерность пространства. Если $X_1, \ldots, X_n$ – дискретные случайные величины, то для них определены условные распределения вероятностей $P(x/\omega)$.

Под стратегией природы C будем понимать следующий набор $C = \{p(x/\omega) \cdot q_\omega, \ \omega = \overline{1,k}\}$ (для дискретных признаков $C = \{P(x/\omega) q_\omega, \ \omega = \overline{1,k}\}$), где q_ω – априорные вероятности проявления объектов из образа ω.

Введем L^0 как множество всевозможных стратегий природы в n-мерном признаковом пространстве при фиксированном k.

Под решающим правилом f будем понимать следующее отображение $f: D \to \Omega$. Решающему правилу соответствует некоторое разбиение множества D: $\alpha = \{D^1, \ldots, D^\omega, \ldots, D^k\}$, где $D^\omega = \{x: f(x) = \omega\}$. Качество решающего правила определяется вероятностью ошибочной классификации $P(f,c)$. Обозначим через $P_0(c) = P(f_0, c)$ вероятность ошибочной классификации байесовского решающего правила f_0 (байесовский уровень ошибки).

В задачах распознавания стратегия природы, как правило, неизвестна. Решающее правило строится на основе анализа обучаю-

щей выборки $V=\{x_{ij}\}$; $i=\overline{1,N}$; $j=\overline{1,n+1}$ (N - объем обучающей выборки). Алгоритмом построения решающего правила распознавания Q будем называть некоторую процедуру, которая каждой выборке V ставит в соответствие решающее правило $f \in \Phi$ (Φ - некоторый класс решающих правил), $Q(V) = f$.

Под задачей распознавания образов будем понимать построение решающего правила f с помощью алгоритма Q на основе выборки V с последующим использованием этого правила для распознавания новых объектов. Качество алгоритма Q при фиксированных c и N будем определять математическим ожиданием вероятности ошибки $EP_N(Q,c) = EP_N(Q(V),c)$.

Авторами [1] показано, что математическое ожидание вероятности ошибки отличается от байесовского уровня ошибки на некоторую величину $\varepsilon(Q,c,N) = \gamma(Q,c) + \nu(Q,c,N)$. Здесь величина $\gamma(Q,c) = P_\infty(Q,c) - P_0(c)$ - мера адекватности алгоритма Q стратегии природы c, где $P_\infty(Q,c) = \lim\limits_{N\to\infty} EP_N(Q,c)$, $\nu(Q,c,N) = EP_N(Q,c) - P_\infty(Q,c)$ - мера устойчивости алгоритма Q к объему обучающей выборки N при фиксированной c. Математическое ожидание $EP_N(Q,c) = P_0(c) + \varepsilon(Q,c,N)$.

Из множества алгоритмов распознавания $G = \{Q_1, \ldots, Q_m\}$ для фиксированной стратегии природы $c \in L^0$ при заданных размерности пространства n и объеме выборки N необходимо выбрать некоторый алгоритм $Q_{\alpha^*} \in G$ такой, что

$$EP_N(Q_{\alpha^*},c) = \min_{Q_\alpha \in G} EP_N(Q_\alpha,c)$$

или

$$\varepsilon(Q_{\alpha^*},c,N) = \min_{Q_\alpha \in G} \varepsilon(Q_\alpha,c,N).$$

При решении прикладной задачи стратегия природы c обычно неизвестна. В этих условиях проблема выбора наилучшего ал -

горитма из G при фиксированных значениях параметров n и N значительно усложняется. Возникает проблема сравнения алгоритмов на множестве стратегий природы L^0 при фиксированных n и N. Очевидно, что сравнивать алгоритмы на всем множестве L^0 невозможно. Основным вопросом для решения данной проблемы становится формирование подмножества стратегий природы, на котором можно было бы проводить сравнение алгоритмов.

Можно указать ряд существующих подходов к решению этого вопроса:

1. Предполагается, что стратегия c принадлежит достаточно узкому классу стратегий, например, когда плотности $p(x/\omega)$ описываются нормальным законом распределения или их смесями [2].

2. Предполагается, что на множестве L^0 задано равномерное распределение $p(c)$ [3].

3. Формируется конечное множество стратегий природы в рамках игровой имитационной модели [4].

Недостатки каждого из трех подходов указаны в [5].

В данной работе предлагается новый подход к решению проблемы сравнения алгоритмов распознавания. Разработана методика такого сравнения.

§2. Методика сравнения алгоритмов распознавания

Описываемая методика заключается в последовательном выполнении трех основных этапов.

На первом этапе производится деление алгоритмов на группы. В каждую группу входят алгоритмы, предназначенные для решения задач одного типа, дальнейшее сравнение алгоритмов осуществляется внутри каждой группы. Деление алгоритмов на группы проводится согласно параметрам классификации задач.

Введем параметры задачи. Любую задачу распознавания образов можно характеризовать фиксированным списком из семи пара-

метров: B - параметр, определяющий тип признаков; k - число образов; z - параметр, определяющий наличие пропусков в таблице; d - параметр, определяющий форму задания области принятия решения D^{ω} для каждого из образов; при наличии некоторой дополнительной информации о виде распределений или о виде разделяющих функций можно ввести дополнительный параметр s, характеризующий задачу (в данной работе предполагается, что такая априорная информация отсутствует); n - число признаков; N - объем обучающей выборки.

Под конкретным типом задачи распознавания образов будем понимать множество задач, имеющих одинаковое значение параметров B, k, z, d.

На втором этапе внутри каждой группы алгоритмов $G' \subseteq G$, предназначенных для решения задач одного типа, проводится отбор A-допустимых алгоритмов для дальнейшего сравнения. Алгоритм Q_r является A-допустимым, если автор алгоритма или эксперты для некоторого n могут предложить такую гипотетическую стратегию природы $c \in L^0$, заданную в виде имитационной модели, для которой хотя бы при одном N

$$\bar{E}P_N(Q_r, c) < \min_{Q_\alpha \in G'} \bar{E}P_N(Q_\alpha, c), \quad \alpha \neq r,$$

где $EP_N(Q_r, c)$ - оценка математического ожидания вероятности ошибки, полученная на основе моделирования.

Иными словами, на данном этапе проводится сравнение алгоритмов на конечном множестве гипотетических стратегий природы, сформулированных в рамках игровой имитационной модели. Очевидно, что в "Полигон" следует включать только A-допустимые алгоритмы. Впервые использование игровой модели для выбора алгоритма распознавания было предложено Т.Андерсоном [4].

На третьем этапе проводится сравнение алгоритмов на множестве "усредненных стратегий природы", заданных в виде имита-

ционных моделей, для определения значений параметров задачи (объема выборки N и числа признаков n), при которых алгоритм B-допустим.

Алгоритм Q_r является B-допустимым в n-мерном признаковом пространстве, если существуют хотя бы одна "усредненная стратегия природы" $c \in S$ (S - множество всевозможных "усредненных стратегий" в n-мерном признаковом пространстве, $S \subseteq L^0$) и такое N, что

$$\bar{E}P_N(Q_r, \bar{c}) < \min_{Q_\alpha \in G''} \bar{E}P_N(Q_\alpha, \bar{c}), \quad r \neq \alpha.$$

где G'' - множество A-допустимых алгоритмов распозна[illegible] выбранных на втором этапе сравнения.

Для формирования множества "усредненных стратегий природы" S рассмотрим класс решающих правил Φ, удовлетворяющий следующим условиям:

1) класс решающих правил Φ образует последовательность подклассов $\Phi_1 \subset \ldots \subset \Phi_i \subset \ldots \subset \Phi_g = \Phi$ такую, что каждому подклассу Φ_i можно поставить в соответствие некоторую меру сложности $\mu(\Phi_i)$, для которой:

$$\mu(\Phi_1) < \ldots < \mu(\Phi_i) < \ldots < \mu(\Phi_g),$$

g - целое больше единицы;

2) для любой $c \in L^0$ найдется такое i^* ($i^* = \overline{1,g}$), что для некоторого $f \in \Phi_{\alpha^*}$

$$P(f,c) - P_0(c) < \delta,$$

где δ - сколь угодно малое число.

Каждому Φ_i ставится в соответствие подмножество стратегий природы $L_i \subseteq L^0$ следующим образом: $L_i = \{c : \exists f \in \Phi_i, P(f,c) - P_0(c) < \delta\}$. Тогда упорядоченным подклассам решающих правил можно поставить в соответствие упорядочен-

ные подмножества стратегий природы $L_1 \subset \ldots \subset L_i \subset \ldots \subset L_g \subseteq L^0$. Меру сложности каждого подмножества L_i будем определять через меру сложности соответствующего ей подкласса решающих правил $\mu(L_i) = \mu(\Phi_i)$.

Каждому подмножеству $L_i' = L_i \setminus L_{i-1}$ $(L_0 \neq 0)$ при фиксированном байесовском уровне ошибки P_0 поставим в соответствие "усредненную стратегию природы".

"Усредненная стратегия" введена [6] пока для одномерного признакового пространства, признак количественный, $k = 2$, $\mu(L_i) < \infty$. В одномерном признаковом пространстве любое решающее правило разбивает диапазон изменения признака X на $l+1$ интервал l границами $(1 \leq l < \infty)$. Под сложностью произвольной стратегии будем понимать число границ l, соответствующих байесовскому решающему правилу. Здесь

$$P_0(c) = \sum_{i=1}^{l+1} \min \{P_i^1, P_i^2\},$$

где P_i^ω - вероятность попадания объектов образа ω, $\omega = 1,2$, в i-й интервал. Под "усредненной стратегией" при фиксированных l и P_0 будем понимать вектор $\bar{c}(l, P_0) = \{P_1^1, \ldots, P_{l+1}^1, P_1^2, \ldots, P_{l+1}^2\}$, заданный своими средними значениями компонент. Внутри каждого интервала распределение по каждому из образов равномерно. Область определения признака X - интервал $(0,1)$.

Таким образом, множество всех "усредненных стратегий природы" S, заданных в виде имитационных моделей, будет служить набором тестовых примеров для проверки алгоритмов на B-допустимость. Причем для каждого n будет свое множество S "усредненных стратегий".

На этом описание методики сравнения закончено. Необходимо лишь добавить, что для решения конкретной задачи с фиксирован-

ными значениями параметров n и N пользователю рекомендуется использовать только B-допустимые алгоритмы. Безусловно, пользователь может из B-допустимых алгоритмов исключить еще часть тех алгоритмов, которые используют слишком большие время счета или оперативную память.

§3. Результаты сравнения шести алгоритмов распознавания

Как уже отмечалось выше, в "Полигон" вошло шесть алгоритмов построения решающих правил распознавания: линейная дискриминантная функция (ЛДФ), квадратичная дискриминантная функция (КДФ), алгоритм, основанный на непараметрических оценках Розенблатта-Парзена (CANDY), и алгоритмы, основанные на логических решающих правилах (DWI3,LRP,GLRP). Описание всех этих алгоритмов дано в [5]. Трудности, связанные с передачей программного обеспечения, не позволили включить в "Полигон" ряд известных алгоритмов.

Качество решения задачи определяется оценкой вероятности ошибки на контроле, усредненной по числу экспериментов. Под экспериментом в случае решения тестовой задачи понимается последовательность следующих процедур: генерирование обучающей выборки, построение решающего правила с помощью алгоритма Q_α, генерирование контрольной выборки и вычисление оценки вероятности ошибки для заданной контрольной выборки. При решении прикладной задачи под экспериментом понимается выбор случайным образом из всей совокупности объектов обучающей выборки заданного объема, построение решающего правила с помощью алгоритма Q_α, из оставшихся объектов выбор случайным образом контрольной выборки заданного объема и вычисление оценки вероятности ошибки для заданной контрольной выборки. Качество работы алгоритма может также определяться с помощью процедуры "скользящий экзамен".

Сравнение алгоритмов на тестовых примерах проводилось согласно описанной выше методике для задач следующего типа: признаки количественные, число образов равно двум, пропуски в таблице отсутствуют, форма задания области принятия решения не учитывается.

Для проверки алгоритмов на **A**-допустимость было предложено 17 тестовых примеров. Сравнение проводилось при следующих объемах обучающей выборки N = 40, N = 100, N = 200. Объем контрольной выборки равен 200. Число объектов каждого образа одинаково. Эксперимент моделирования повторялся 9 раз для каждого варианта. Под вариантом здесь понимается некоторая стратегия природы при фиксированном объеме обучающей выборки **N**. Всего вариантов 51.

Необходимо отметить, что для алгоритма GLRP рассматривались два основных режима построения группы деревьев и для каждого из них пять процедур принятия решений.

Согласно проведенному экспериментальному сравнению было получено, что все алгоритмы, включенные в "Полигон", являются A-допустимыми.

Заметим, что для алгоритмов DW13, LRP, GLRP было проведено аналогичное сравнение для проверки их на A-допустимость в случае задач следующего типа: признаки разнотипны, **k** = 2, пропуски в таблицах отсутствуют, форма задания области принятия решения не учитывается. Было получено, что все эти алгоритмы являются A-допустимыми.

Для проверки алгоритмов на B-допустимость в одномерном признаковом пространстве без "шума" или с "шумом" (к одному информативному признаку добавлялись неинформативные признаки, для каждого из которых распределение выбиралось одинаковым, равным N(0,20)). В результате было рассмотрено 120 вариантов "усредненных стратегий природы", заданных в виде имитационных моделей. Разнообразие вариантов определялось сложностью **l**,

уровнем ошибки P_0, объемом выборки N, отсутствием или присутствием "шумов". Сложность стратегий $l = 1,2,3,4,9$; байесовский уровень $P_0 = 0; 0,05; 0,10; 0,15$. Объемы обучающей и контрольной выборок те же, что и при проверке алгоритмов на А-допустимость.

В результате сравнения было получено, что в одномерном признаковом пространстве $(n = 1)$ с "шумами" или без "шумов" при рассматриваемых объемах обучающей выборки В-допустимым оказался только алгоритм LRP (алгоритм GLRP при проверке алгоритмов на В-допустимость не рассматривался, так как в случае $n = 1$ он вырождается в алгоритм LRP).

Необходимо отметить, что так как эксперимент моделирования повторялся 9 раз для каждого варианта, то задача распознавания для проверки алгоритмов на А- и В-допустимость решалась около 13 тысяч раз. Эта цифра говорит о трудоемкости проведенного машинного эксперимента сравнения.

Программное обеспечение "Полигон" применялось также для решения около пятидесяти прикладных задач: каждая задача решалась с помощью всех алгоритмов, включенных в "Полигон". В большинстве случаев лучшие результаты (наименьшее значение оценки математического ожидания вероятности ошибки) были получены с помощью алгоритмов ЛДФ, LRP и GLRP. Алгоритм КДФ не оказался лучшим для решения ни одной из рассматриваемых задач. Очевидно, что результаты работы ЛДФ, КДФ и CANDY можно было улучшить, использовав ту или иную известную процедуру отбора информативных признаков.

В заключение отметим, что проведенное сравнение алгоритмов по вышеизложенной методике является только началом исследования сложной и трудоемкой задачи сравнения алгоритмов построения решающих правил распознавания.

Литература

1. ЛБОВ Г.С., СТАРЦЕВА Н.Г. Классификация и принципы сравнения алгоритмов построения решающих правил распознавания //Статистическая обработка информации. - Новосибирск, 1989. - С. 14-23.

2.РАУДИС Ш. Ограниченность выборки в задачах классифика - ции //Статистические проблемы управления. - Вильнюс, 1976. - Вып. 18. - С. 1-185.

3. HYGHES G.F.On the mean accuracy of statistical pattern recognizers //IEEE Trans. inform theory. - 1968.-Vol.IT-14,N1. - P. 55-63.

4. АНДЕРСОН Т. Введение в многомерный статистический анализ: Пер. с англ./Под ред. Б.В.Гнеденко. - М.: Физматлит.,1963. - 500 с.

5. СТАРЦЕВА Н.Г. Выбор алгоритма построения решающего правила в распознающей системе: Автореф. Дис... кан.техн.наук: 05.13.01 - Томск, 1988. - 18 с.

Поступила в ред.-изд.отд.

24 января 1989 года