УДК 519.95

ЛИНЕЙНЫЕ ДИСКРИМИНАНТНЫЕ ФУНКЦИИ И ВЫБОР СПРЯМЛЯЮЩЕГО ПРОСТРАНСТВА ДЛЯ ИХ РЕАЛИЗАЦИИ

Игнатьев Н. А., Саидов Д. Ю.
ignatev@rambler.ru; doniyor_2286@mail.ru
Национальный университет Узбекистана

Рассматриваются вопросы принятия и обоснования решений по результатам интеллектуального анализа данных с помощью логических закономерностей в форме полуплоскостей. Предлагается эвристический метод отбора информативных наборов признаков в спрямляющем пространстве. Востребованность метода доказывается через вычисление показателей обобщающей способности алгоритмов распознавания, основанных на принципах разделения объектов поверхностями.

Ключевые слова: линейные дискриминантные функции, информативные признаки, логические закономерности, нелинейное отображение, обобщающий способность.

Цитирование: *Игнатьев Н. А., Саидов Д. Ю.* Линейные дискриминантные функции и выбор спрямляющего пространства для их реализации // Проблемы вычислительной и прикладной математики. — 2019. — № 3(21). — С. 40–48.

1 Введение

Линейные дискриминантные функции (ЛДФ) широко используются в задачах интеллектуального анализа данных. Низкие затраты вычислительных ресурсов, возможность содержательной интерпретации результатов распознавания в качестве новых знаний являются теми свойствами, которые находят применение их при моделировании процессов и явлений в слабо формализованных предметных областях. При компьютерной реализации ЛДФ не требуется таблица прецедентов, достаточно хранить в памяти лишь веса признаков. В технических устройствах ЛДФ могут быть представлены в виде электронных схем и микросхем чипов.

Для показателей точности распознавания большое значение имеет такое свойство признакового пространства как линейная разделимость объектов классов. Одним из способов достижения этого свойства является использование нелинейных преобразований признаков.

Нелинейные преобразования признаков, как правило, приводят к описанию объектов в пространстве (обобщённом пространстве) более высокой размерности, чем исходное. В качестве примера можно привести обобщённые линейные дискриминантные функции, представляемые с помощью произведений исходных признаков степени не выше 2 и называемыми квадратичными. Обобщённое признаковое пространство можно рассматривать как линейное или спрямляющее, но значительно большей размерности, чем исходное.

Переход в спрямляющее пространство объясняется с позиций обобщающей способности алгоритмов распознавания. Теоретически такой подход приемлем, так как повышает меру статистического разнообразия (ёмкость) класса линейных алгоритмов. Доказательство этого факта можно найти в работе В.Н. Вапника [1]. Утверждается, что выборку из m объектов в пространстве из n признаков при $n \geqslant m$ всегда можно с помощью ЛДФ разделить на два класса $2m$ способами. В реальных прикладных задачах отношения между объектами выборок данных определяется скрытыми

закономерностями и нет смысла рассматривать всевозможные варианты разбиения на классы.

Обучение ЛДФ сводится к вычислению вектора весовых коэффициентов. Среди вычислительных методов безусловным лидером является линейный дискриминант Фишера. Лидерские качества демонстрируются в виде высоких относительно других методов показателях обобщающей способности к распознаванию.

Выводы о существовании признакового пространства с линейной разделимостью объектов классов скорее всего представляет теоретический интерес, но для практического использования, как правило, неприемлемы. Обобщённые функции, которые предлагаются для формирования признакового пространства, увеличивают сложность обучения и реализации ЛДФ на несколько порядков выше, чем на исходном признаковом пространстве. Исследователи искали ответ на вопрос: с помощью каких нелинейных преобразований строить спрямляющее пространство? В методе SVM [2] такой выбор был сделан на использование ядерных функций. Несмотря на наличие теоретического обоснования метода никаких рекомендаций по выбору ядерных функций не разработано. В [2] предлагалось решение проблемы линейной разделимости с помощью матриц попарного сходства объектов. Использование этих матриц рассматривалось в качестве одной из разновидностей безпризнакового распознавания. Принцип безпризнакового распознавания распространяется на такие известные методы как ближайший сосед, k ближайших соседей, базовым свойством которых является локальная компактность по мере близости.

Проблема поиска информативных признаков как в исходном, так и в расширенном (спрямляющем) пространстве оставалась открытой. Целью отбора была адаптация к той структуре признакового пространства, для которой существует линейная разделимость объектов классов.

Логические закономерности в форме полуплоскостей применяются в интеллектуальном анализе данных. Целью анализа является поиск скрытых закономерностей (новых знаний) из баз (хранилищ) данных. Результаты анализа необходимы для принятия решений в трудноформализуемых задачах.

Сложность формализации заключается в отсутствии единого критерия, для оптимизации которого можно использовать уже известные методы либо разрабатывать новые. Как правило, задачи принятия решения многокритериальные. Получить оптимальное решение по каждому критерию практически невозможно. Выбор критерия (критериев) остаётся за лицом, принимающим решение (ЛПР).

Наиболее известный и широко применяемый на практике критерий Фишера [3] не претендует на полноту исследования структуры данных с помощью логических закономерностей в форме полуплоскостей. В работе предлагаются два новых критерия для решения этой проблемы. Целью является демонстрация методологии совместного использования этих критериев для принятия решения. Предлагается эвристический метод отбора информативных наборов признаков в спрямляющем пространстве. Востребованность метода доказывается через вычисление показателей обобщающей способности алгоритмов распознавания, основанных на принципе разделения объектов поверхностями.

2 Постановка задачи

Рассматривается двухклассовая задача распознавания в стандартной постановке. Каждый из объектов выборки $E_0 = \{S_1, \ldots, S_m\}$ принадлежат одному из классов K_1 или K_2 ($E_0 = K_1 \bigcup K_2$) и описывается с помощью n количественных признаков

$X(n) = (x_1, \ldots, x_n)$. Для распознавания объектов на E_0 используются обобщённые линейные решающие функции вида $d(S) = w_1 y_1 + \cdots + w_r y_r$, где $y_c = f_c(S)$, $f_c(S) \in \left\{x_{i_1}^{a_1} x_{i_2}^{a_2} \ldots x_{i_t}^{a_t}\right\}, a_j \in \{0,1\}$, $j = 1, \ldots, t,\ t > 1$.

Считается, что для оценки выбора информативного набора признаков $Y(p) = (y_1, \ldots, y_p)$ используется функционал $F(E_0, Y(p))$. Требуется определить:

- критерии для оценки закономерностей в форме полуплоскостей;
- информативный набор признаков

$$Y(p) = \arg \max_{f_i(\mathrm{S}) \in \Omega^t} F(E_0, Y(p)),$$

где Ω^t - множество обобщённых функций степени не выше t.

3 Критерии оценки закономерностей в форме полуплоскостей

Закономерности в форме полуплоскостей представляют предикат вида $P(x) = [\sum_{i=1}^n w_i x_i \leqslant w_0] \in \{0,1\}$. Геометрическое место точек, равноудалённое от двух эталонов из разных классов, является гиперплоскость, значения весовых коэффициентов которой вычисляются через координаты эталонов [4]. В качестве таких эталонов в данной работе предлагается рассматривать векторы математических ожиданий M_1, M_2 значений признаков объектов по каждому из классов K_1 и K_2.

Пусть $m_r^1 \in M_1$, $m_r^2 \in M_2$ – математическое ожидание (среднее-арифметическое) значений признака $y_r \in \Omega^t$ соответственно в классах K_1 и K_2. Внутриклассовое сходство и межклассовое различие признака $y_r \in \Omega^t$ по объектам $S_i \in E_0$ $(S_i = (y_{i1}, ..., y_{ip}))$, $p > 1$ вычислим соответственно как $\theta_r = \sum_{j=1}^2 \sum_{S_i \in K_j} (y_{ir} - m_r^j)^2$ и $\gamma_r = \sum_{j=1}^2 \sum_{S_i \in K_j} (y_{ir} - m_r^{3-j})^2$. Для оценки веса (разделяющей способности) w_r признака $y_r \in \Omega^t$ по значениям θ_r и γ_r предлагается использовать функционал из [5]

$$J(w) = \frac{\sum_i w_i \theta_i}{\sum_i w_i \gamma_i} \to \min. \tag{1}$$

При ограничении на веса в (1) $\sum_i w_i = 1, w_i > 0$ функция Лагранжа для решения задачи условной оптимизации имела вид

$$L(w) = \frac{\sum_i w_i \theta_i}{\sum_i w_i \gamma_i} + \lambda \left(\sum_i w_i - 1 \right),$$

а значения весов вычислялись как $w_i = \frac{\gamma_i - \theta_i}{\sum_j (\gamma_j - \theta_j)}$.

Согласно доказанной в [5] теореме, необходимым и достаточным условием выбора признака $y_j \in Y(p)$ кандидатом на удаление из набора $Y(p) = (y_1, ..., y_p)$ при ограничении $\sum_i w_i = 1, w_i > 0$ является $\frac{\theta_j}{\gamma_j} = \max\limits_{y_i \in Y(p)}$. Соотношение

$$\frac{\theta_j}{\gamma_j} \tag{2}$$

даёт возможность оценивать и упорядочивать признаки по плотности их распределения вокруг математических ожиданий классов. Чем выше плотность, тем меньше значение (2).

Также как и в (2), вычисление внутриклассового сходства по отдельному признаку используется в критерии Фишера [3]

$$\frac{|m_1 - m_2|^2}{\tilde{s}_1 + \tilde{s}_2}, \tag{3}$$

в котором сумма внутриклассового разброса $\tilde{s}_1 + \tilde{s}_2 = \theta_r$, а $m_1 - m_2$ есть разность математических ожиданий классов K_1 и K_2 на числовой оси.

Третий критерий рассчитан на анализ порядка расположения объектов классов на числовой оси [6]. Пусть

$$S_{r_1}, S_{r_2}, ..., S_{r_m} \tag{4}$$

последовательность объектов, упорядоченная по невозрастанию значений признака $y_r \in Y(p)$. Упорядоченное множество значений (4) разделяется на два непересекающихся интервала $[c_1, c_2], (c_2, c_3]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границы c_2 основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов содержит значения количественного признака объектов только одного класса.

Пусть u_i^1, u_i^2 – количество значений (4) некоторого количественного признака $y \in Y(p)$ класса $K_i, i = 1, 2$ соответственно в интервалах $[c_1, c_2], (c_2, c_3]$, $|K_i| > 1$,v – порядковый номер элемента упорядоченной по возрастанию последовательности (4) из E_0, определяющий границы интервалов как $c_1 = S_{r_1}, c_2 = S_{r_v}, c_3 = S_{r_m}$. Критерий позволяет вычислять оптимальное значение границы между интервалами $[c_1, c_2], (c_2, c_3]$. Выражение в левых скобках (5) представляет внутриклассовое сходство, в правых – межклассовое различие.

$$w(y) = \left(\frac{\sum_{d=1}^{2} \sum_{i=1}^{2} u_i^d \left(u_i^d - 1 \right)}{\sum_{i=1}^{2} |K_i| \left(|K_i| - 1 \right)} \right) \left(\frac{\sum_{d=1}^{2} \sum_{i=1}^{2} u_i^d \left(|K_{3-i}| - u_{3-i}^d \right)}{2 |K_1| |K_2|} \right) \rightarrow \max_{c_1 < c_2 < c_3} \tag{5}$$

Значение $w_r = w(y_r)$ рассматривается как вес признака $y_r \in Y(p)$, а границы интервалов могут использоваться для нормирования значений признака объекта $S_i = (y_{i1}, ..., y_{ip})$ по формуле $\overline{y_{ir}} = \frac{y_{ir} - c_2}{c_3 - c_1}$.

4 Отбор информативных наборов признаков

Задача поиска информативных наборов признаков для линейной разделимости объектов классов K_1 и K_2 является NP-полной. Из этого следует вывод, что кроме полного перебора других способов решить задачу поиска глобального экстремума функционала $F(E_0, Y(p))$ не существует. Используя некоторые эвристики, можно получить локальный экстремум функционала.

Смысл использования эвристик для решения проблемы линейной разделимости сводится к следующему. Пусть Ω^t – множество обобщённых функций степени не выше t. На множестве пар $(y_i, y_j) \subset Y(p)$ рассматривается сокращённый перебор с целью поиска экстремума по критерию Фишера

$$\Phi(w) = \frac{|m_1 - m_2|^2}{\tilde{s}_1 + \tilde{s}_2} \rightarrow \max. \tag{6}$$

Критерий (6) отличается от (3) тем, что для линейной проекции описаний объектов на числовую ось необходимо вычислять значения вектора весов w. Приемлемым считается результат, при котором точность распознавания на обучении по $(y_i, y_j) \subset Y(p)$ не ниже, чем на исходном наборе $X(n)$.

Для вычисления коэффициентов дискриминантной функции $d(y) = w_1 y_i + w_2 y_j + w_0$ по паре $(y_i, y_j) \subset Y(p)$ сформируем матрицу ковариаций $Z = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{pmatrix}$ и

вектор-столбец разностей $\begin{pmatrix} m_i^1 - m_i^2 \\ m_j^1 - m_j^2 \end{pmatrix}$, где $m_i^1, m_j^1, m_i^2, m_j^2$ – математические ожидания по признакам y_i, y_j соответственно в классах K_1 и K_2. Решение системы линейных алгебраических уравнений

$$\begin{cases} w_1 z_{11} + w_2 z_{12} = m_i^1 - m_i^2 \\ w_1 z_{21} + w_2 z_{22} = m_j^1 - m_j^2 \end{cases} \tag{7}$$

даёт искомые значения весов w_1, w_2 дискриминантной функции.

Выбор коэффициентов линейного дискриминанта Фишера по (7) связан с предположением, что выборка данных распределена по нормальному закону. Исходя из этого предположения, выбор порога дискриминантный функции $d(y)$ производится как

$$w_0 = -(w_1(m_i^1 |K_1| + m_i^2 |K_2|) + w_2(m_j^1 |K_1| + m_j^2 |K_2|))/m. \tag{8}$$

Способ выбора порога без всяких предположений о природе среды впервые был предложен в [7]. Значение порога вычислялось по границе c_2 интервалов $[c_1, c_2], (c_2, c_3]$ по (5) как

$$w_0 = \frac{c_2 + u(S)}{2}, \tag{9}$$

где $u(S) = w_1 s_i + w_2 s_j$, и $S = (s_1, ..., s_p)$, $u(S) \in (c_2, c_3]$ - ближайщий к c_2 объект E_0 на числовой оси.

5 Вычислительный эксперимент

Для вычислительного эксперимента были взяты 4 выборки данных из [6, 8, 9], содержащих представителей двух непересекающихся классов. Для описания объектов использовались признаки, измеренные в интервальных шкалах. Параметры выборок представлены в табл. 1.

Таблица 1 Параметры выборок данных

№	Выборка данных	Количество	
		объектов	признаков
1	Australian	690	14
2	Chelust	42	6
3	Gipertaniya	147	29
4	Seeds	140	7

Вычислительный эксперимент проводился на объектах выборок с описанием как в исходном, так и в спрямляющем пространстве. Спрямляющее пространство было представлено обобщенными функциями степени не выше 2. Сравнительный анализ данных на выборке Chelust по критериям (2), (3), (5) представлен в табл. 2.

Нетрудно заметить, что между значениями критериев (см. табл. 2) нет линейной или квазилинейной зависимости. Многообразие отношений на множестве значений свидетельствует как о сложности структуры данных, так и о сложности принятия решения по ним.

Значения 0.8001 по критерию (5) на признаках $x_4 * x_5$, $x_5 * x_6$и x_6 указывает на то, что порядок расположения объектов одного класса относительно другого не изменился. Изменения есть у показателей плотности распределения объектов относительно математических ожиданий (центров) классов, вычисляемых по критериям (2) и (3).

Таблица 2 Сравнительный анализ данных Chelust по критериям (2), (3), (5)

Признаки (обобщенные функции)	Значения по критериям		
	(3)	(5)	(2)
x_1	0.0074	0.3781	0.7617
$x_1 * x_2$	0.0106	0.3781	0.6910
$x_1 * x_3$	0.0090	0.3620	0.7258
$x_1 * x_4$	0.0679	0.5757	0.2596
$x_1 * x_5$	0.0341	0.4196	0.4113
$x_1 * x_6$	0.0398	0.4355	0.3740
x_2	0.0134	0.3892	0.6391
$x_2 * x_3$	0.0158	0.3892	0.6014
$x_2 * x_4$	0.0811	0.6241	0.2270
$x_2 * x_5$	0.0439	0.4905	0.3517
$x_2 * x_6$	0.0499	0.4683	0.3230
x_3	0.0048	0.2884	0.8320
$x_3 * x_4$	0.0806	0.5312	0.2281
$x_3 * x_5$	0.0412	0.5180	0.3664
$x_3 * x_6$	0.0457	0.4743	0.3424
x_4	0.2669	0.8965	0.0819
$x_4 * x_5$	0.2278	0.9107	0.0946
$x_4 * x_6$	0.1797	0.8001	0.1170
x_5	0.1108	0.6254	0.1769
$x_5 * x_6$	0.1108	0.8001	0.1769
x_6	0.0787	0.8001	0.2322

Из максимального значения 0.9107 по критерию (5) следует вывод, что свойство линейной разделимости среди всех признаков из табл. 2 наиболее выражено у $x_4 * x_5$. Несмотря на менее выраженное свойство линейной разделимости, показатель плотности распределения (средне-квадратичное отклонение от математических ожиданий классов) у x_4 по (2) выше чем у $x_4 * x_5$. Относительно малое значение отклонения (0.0819 относительно 0.0946) указывает на более высокую плотность распределения.

Влияние выбора порога дискриминантной функции w_0 с предположением о нормальности распределения выборки по критерию Фишера [3] и по критерию (5) по аналогии соответственно с (8) и (9) на исходных наборах признаков приводится в табл. 3.

Значение критерия (5), равное 1.0, по определению означает, что представители классов на числовой оси не пересекаются между собой. Корректное (без ошибок) распознавание объектов на выборках Chelust и Seeds служат подтверждением этому определению.

Эффект от выбора значения порога по (8) или (9) в спрямляющем пространстве на данных Chelust демонстрируется в табл. 4 и рис.1.

На рис. 1 показана последовательность расположения объектов классов по первой и второй паре признаков из табл. 4. В границах интервалов $[c_1, c_2], (c_2, c_3]$ по (5) содержатся представители соответственно классов K_2 и K_1. При пороге, вычислен-

ном по (8) (на рис. 1 указан жирной чертой), число ошибок равно соответственно 2 и 4.

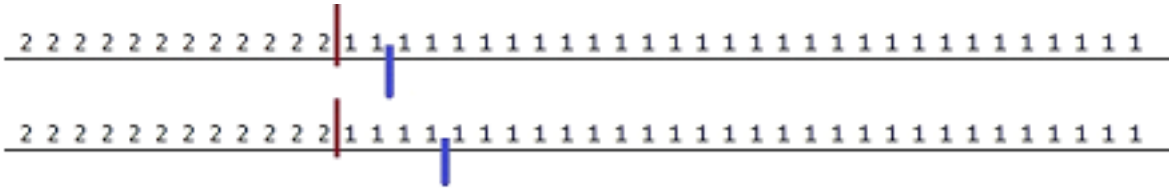

Рис. 1 Последовательность расположения объектов классов и границы порогов на числовой оси

Сравнивая результаты по данным Chelust из табл. 3 и табл. 4, отметим, что корректное распознавание в спрямляющем пространстве достигнуто за счёт использования обобщённых функций не выше 2 степени. Для вычисления этих функций будет достаточно задать значения исходных признаков x_1, x_4, x_5 или x_2, x_4, x_5, x_6.

Таблица 3 Точность распознавания в исходном пространстве признаков

№	Выборка данных	Значение критерия		Число ошибок при выборе порога по критерию	
		(3)	(5)	(6)	(5)
1	Australian	0.0088	0.6176	96	84
2	Chelust	0.5446	1.0000	3	0
3	Gipertaniya	0.2180	0.9672	8	1
4	Seeds	0.1616	1.0000	2	0

Таблица 4 Точность распознавания в спрямляющем пространстве

№	Комбинации из пар признаков	Значение критерия		Число ошибок при выборе порога по критерию	
		(3)	(5)	(8)	(9)
1	$x_1 * x_5,\ x_4 * x_5$	0.5426	1.0	2	0
2	$x_2 * x_6,\ x_4 * x_5$	0.2870	1.0	4	0

6 Заключение

В работе рассмотрена проблема выбора решений в трудноформализуемых задачах путём анализа закономерностей в форме полуплоскостей. Предложены критерии для анализа и методология их использования, которая востребована для разработки и управления военными техническими средствами на основе систем искусственного интеллекта.

Литература

[1] *Вапник В. Н.* Восстановление закономерностей по эмпирическим данным. – М.: Наука, 1979. 447 с.

[2] *Середин О. С.* Линейные методы распознавания образов на множестве объектов произвольной природы. представленные попарными сравнениями. Общий случай // Известия Тульского государственного университета. Естественные науки. 2012. Вып. 1. С. 141-152.

[3] *Дуда Р.,Харт П.* Распознавание образов и анализ сцен. – М.: Мир, 1976. 512 с.

[4] *Ту Дж., Гонсалес Р.* Принципы распознавания образов. – М.: Мир, 1978. 416 с.

[5] *Игнатьев Н. А.* Выбор минимальной конфигурации нейронных сетей // Вычислительные технологии. Т.6. №1. 2001. С. 23-28.

[6] *Игнатьев Н. А.* Вычисление обобщённых показателей и интеллектуальный анализ данных // Автоматика и телемеханика, 2011. №5. С.183-190.

[7] *Игнатьев Н. А.,Нуржонов Ш. Ю.* Выбор параметров регуляризации для повышения обобщающей способности дискриминантных функций //Послания Академии Вооруженных Сил Республики Узбекистан, 2014. Ч.1. №1(14). С. 81-87.

[8] *Ignat'ev N. A., Adilova F. T., Matlatipov G. R., Chernysh P. P.* Обнаружение знаний по клиническим данным на основе решения задач классификации // Мединфо. – Амстердам: иос пресс, 2001. P. 1354–1358.

[9] UCI Хранилище для машинного обучения. URL: `http://archive.ics.uci.edu/ml/machine-learning-databases/`

Поступила в редакцию 15.04.2019

UDC 519.95

LINEAR DISCRIMINANT FUNCTIONS AND THE CHOICE OF RECTIFYING SPACE FOR THEIR IMPLEMENTATION

Ignat'ev N. A., Saidov D. Y.
`ignatev@rambler.ru; doniyor_2286@mail.ru`
National University of Uzbekistan

It is considered the adoption and justification of decisions on the results of data mining using logical regularities in the form of half-planes. A heuristic method is proposed for selecting informative feature sets in a rectifying space. The relevance of the method is proved through the calculation of indicators of the generalizing ability of recognition algorithms based on the principles of the separation of objects by surfaces

Keywords: linear discriminant functions, informative features, logical regularities, nonlinear mapping, generalizing ability.

Citation: Ignat'ev N. A., Saidov D. Y. 2019. Linear discriminant functions and the choice of rectifying space for their implementation. *Problems of Computational and Applied Mathematics.* 3(21): 40–48.

References

[1] Vapnik V. N. *Restoration of patterns by empirical data* // M.: The science. 1979. - 447 s.

[2] Seredin O. S. *Linear methods of pattern recognition on the set of objects of arbitrary nature. presented by pairwise comparisons. General case* // News of Tula State University. Natural Sciences. 2012.Vip. 1.S. 141-152.

[3] Duda R.,Xart P. *Pattern recognition and sen analysis* // Mir. 1976.– 512 s.

[4] Tu Dj., Gonsales R. *Clarification Recognition Guidelines* //M: Mip, 1978. - 416 s.

[5] Ignat'ev N. A. *The choice of the minimum configuration of neural networks* // Computational Technologies. T.6. N 1 2001. S. 23-28.

[6] Ignat'ev N. A. *Calculation of generalized indicators and data mining* // Automation and Remote Control. N 5. 2011. S.183-190.

[7] Ignat'ev N. A.,Nurjonov Sh. Yu. *Selection of regularization parameters to increase the generalizing ability of discriminant functions* // Messages from the Academy of the Armed Forces of the Republic of Uzbekistan.N 1(14) –last. 1 section. Tashkent. 2014. B. 81-87.

[8] Ignat'ev N. A., Adilova F. T., Matlatipov G. R., Chernysh P. P. *Knowledge discovering from clinical data based on classification tasks solving* // Medinfo. - Amsterdam: ios press, 2001. P. 1354–1358.

[9] `http://archive.ics.uci.edu/ml/machine-learning-databases/`

Received April 15, 2019