УДК 519.2

01.00.00 Физико-математические науки

БАЗОВЫЕ РЕЗУЛЬТАТЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ КЛАССИФИКАЦИИ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru
Московский физико-технический институт, 141700, Моск. обл., г. Долгопрудный, Институтский пер., 9

Математическая теория классификации содержит большое число подходов, моделей, методов, алгоритмов. Эта теория весьма многообразна. Выделим в ней три базовых результата - оптимальный метод диагностики (дискриминантного анализа), адекватный показатель качества алгоритма дискриминантного анализа, утверждение об остановке после конечного числа шагов итерационных алгоритмов кластер-анализа. А именно, на основе леммы Неймана - Пирсона показано, что оптимальный метод диагностики существует и выражается через плотности распределения вероятностей, соответствующие классам. Если плотности неизвестны, следует использовать их непараметрические оценки по обучающим выборкам. Часто используют такой показатель качества алгоритма диагностики, как «вероятность (или доля) правильной классификации (диагностики)» – чем этот показатель больше, тем алгоритм лучше. Показана нецелесообразность повсеместного применения этого показателя и обоснован другой – «прогностическая сила», полученная путем пересчета на модель линейного дискриминантного анализа. Остановка после конечного числа шагов итерационных алгоритмов кластер-анализа продемонстрирована на примере метода k-средних. По нашему мнению, эти результаты являются основными в теории классификации, с ними должен быть знаком каждый специалист, развивающий эту теорию или применяющий её

Ключевые слова: МАТЕМАТИЧЕСКАЯ ТЕОРИЯ КЛАССИФИКАЦИИ, МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, ПРИКЛАДНАЯ СТАТИСТИКА, ДИАГНОСТИКА, ДИСКРИМИНАНТНЫЙ АНАЛИЗ, ЛЕММА НЕЙМАНА - ПИРСОНА, ПОКАЗАТЕЛЬ КАЧЕСТВА АЛГОРИТМА ДИАГНОСТИКИ, ВЕРОЯТНОСТЬ ПРАВИЛЬНОЙ КЛАССИФИКАЦИИ, ПРОГНОСТИЧЕСКАЯ СИЛА, КЛАСТЕР-АНАЛИЗБ ОСТАНОВКА ИТЕРАЦИОННОГО АЛГОРИТМА, МЕТОД *k*-СРЕДНИХ

UDC 519.2

Physics and Math

BASIC RESULTS OF THE MATHEMATICAL THEORY OF CLASSIFICATION

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
Bauman Moscow State Technical University, Moscow, Russia

Moscow Physics-Technical Institute; Moscow region, Dolgoprudny, Russia

The mathematical theory of classification contains a large number of approaches, models, methods, algorithms. This theory is very diverse. We distinguish three basic results in it - the best method of diagnosis (discriminant analysis), an adequate indicator of the quality of discriminant analysis algorithm, the statement about stopping after a finite number of steps iterative algorithms of cluster analysis. Namely, on the basis of Neyman - Pearson Lemma we have shown that the optimal method of diagnosis exists and can be expressed through probability densities corresponding to the classes. If the densities are unknown, one should use non-parametric estimators of training samples. Often, we use the quality indicator of diagnostic algorithm as "the probability (or share) the correct classification (diagnosis)" - the more the figure is the better algorithm is. It is shown that widespread use of this indicator is unreasonable, and we have offered the other - "predictive power", obtained by the conversion in the model of linear discriminant analysis. A stop after a finite number of steps of iterative algorithms of cluster analysis method is demonstrated by the example of k-means. In our opinion, these results are fundamental to the theory of classification and every specialist should be familiar with them for developing and applying the theory of classification

Keywords: MATHEMATICAL THEORY OF CLASSIFICATION, MATHEMATICAL STATISTICS, APPLIED STATISTICS, DIAGNOSTICS, DISCRIMINANT ANALYSIS, NEYMAN - PEARSON LEMMA, INDICATOR OF THE QUALITY OF DIAGNOSTIC ALGORITHM, PROBABILITY OF CORRECT CLASSIFICATION, PREDICTIVE POWER, CLUSTER ANALYSIS, STOPPING THE ITERATIVE ALGORITHM, k-MEANS

1. Введение

Методы классификации - неотъемлемая часть математических методов исследования, интересная теоретически и важная практически. Обзоры этой научной области даны в [1 - 3]. Многие математические методы классификации относятся к непараметрической статистике [4] и к нечисловой статистике [5], т.е. являются неотъемлемой составной частью основного потока современных научных исследований, порожденных новой парадигмой прикладной статистики [6].

В многообразии результатов математической теории классификации выделим три - оптимальный метод диагностики (дискриминантного анализа), адекватный показатель качества алгоритма дискриминантного анализа, доказательство сходимости итерационных алгоритмов кластер-анализа. По нашей оценке, эти результаты являются основными в теории классификации, с ними должен быть знаком каждый специалист, развивающий эту теорию или применяющий ее результаты.

2. Оптимальный метод диагностики основан на непараметрических оценках плотности

Рассмотрим задачу диагностики с двумя классами. Решение принимают по основе значения x - элемента некоторого пространства. Элементы первого класса имеют плотность $f(x)$, элементы второго - плотность $g(x)$. Поступает на рассмотрение новый объект со значением X. К какому классу его отнести?

Задачу диагностики можно переформулировать в терминах теории проверки статистических гипотез. Пусть согласно нулевой гипотезе H_0 результат наблюдения X имеет распределение с плотностью $f(x)$, а согласно альтернативной гипотезе H_1 результат наблюдения X имеет распределение с плотностью $g(x)$. Отнесение X к первому классу соответствует принятию гипотезы H_0 (и отклонению гипотезы H_1), а отнесение X ко второму классу

соответствует принятию гипотезы H_1 (и отклонению гипотезы H_0).

В теории проверки статистических гипотез выявлена важная роль критерия отношения правдоподобия (см., например, [7]). Статистика этого критерия имеет вид

$$Q(x) = \frac{f(x)}{g(x)}. \qquad (1)$$

Правило принятия решения основано на сравнении с порогом C значения статистики критерия $Q(X)$, рассчитанного для поступившего на рассмотрение нового объекта со значением X. Таким образом, если $Q(X) > C$, то X относят к первому классу, в противном случае - ко второму.

С точки зрения здравого смысла критерий отношения правдоподобия является естественным, как отношение шансов (вероятностей) за то, что новый объект со значением X относится к первому или ко второму классу соответственно. Важно, что согласно лемме Неймана-Пирсона этот критерий является наиболее мощным критерием среди всех статистических критериев, имеющих один и тот же заданный уровень значимости (понятия "уровень значимости" и "мощность критерия" - базовые в математической статистике). (Строго говоря, под термином "лемма" понимают верное (т.е. доказанное) утверждение, полезное не само по себе, а для доказательства других утверждений. Однако лемма Неймана-Пирсона - основной результат математической статистики, важный сам по себе. Поэтому лемму Неймана-Пирсона часто называют фундаментальной леммой математической статистики.)

Итак, *оптимальный метод диагностики существует* и задается с помощью статистики $Q(X)$ (см. формулу (1)).

Однако при решении практических задач диагностики плотности $f(x)$ и $g(x)$ обычно неизвестны. В таких случаях строят правило диагностики на основе обучающих выборок. А именно, предполагается, что имеются m объектов из первого класса (обучающая выборка для первого класса) и n

объектов из второго класса (обучающая выборка для второго класса). В вероятностно-статистической теории принимают, что обучающую выборку можно моделировать как совокупность независимых одинаково распределенных случайных объектов с соответствующей плотностью. Развиты непараметрические методы состоятельного оценивания неизвестной плотности [8, 9]. Пусть $f_m(x)$ и $g_n(x)$ - состоятельные оценки плотностей $f(x)$ и $g(x)$ соответственно по обучающим выборкам. Рассмотрим выборочный аналог статистики критерия отношения правдоподобия

$$Q_{mn}(x) = \frac{f_m(x)}{g_n(x)}. \qquad (2)$$

Из состоятельности $f_m(x)$ и $g_n(x)$ вытекает, что $Q_{mn}(x)$ для того же элемента x является состоятельной оценкой $Q(x)$ при безграничном росте объемов обучающих выборок. При справедливости обычно выполненного предположения равномерной сходимости из оптимальности критерия отношения правдоподобия для полностью известных плотностей вытекает асимптотическая оптимальность выборочного аналога этого критерия, основанного на сравнении с порогом C значения статистики (2).

В задачах диагностики со многими классами оптимальное решение также выражается через плотности, соответствующие классам. Например, при постановке задачи в терминах статистических решающих правил [10, 11]. Во всех таких случаях асимптотически оптимальное решение получаем путем замены неизвестных плотностей их состоятельными оценками [8, 9].

Наличие описанных выше оптимальных и асимптотически оптимальных правил диагностики (дискриминантного анализа, распознавания образов с учителем) не означает, что не следует разрабатывать новые алгоритмы диагностики. Исходя, например, из необходимости сокращения машинной памяти и времени на расчеты.

Однако, на наш взгляд, *необходимо сравнивать новые алгоритмы с известными оптимальными и асимптотически оптимальными алгоритмами* по тем или иным показателям качества.

3. Прогностическая сила - адекватный показатель качества алгоритма диагностики

Часто используют такой показатель качества алгоритма диагностики, как «вероятность (или доля) правильной классификации (диагностики)» [12, 13] – чем этот показатель больше, тем алгоритм лучше. Цель настоящего раздела статьи – показать нецелесообразность повсеместного применения этого показателя и обосновать другой – «прогностическую силу», найденную путем пересчета на модель линейного дискриминантного анализа.

Опишем используемую в дальнейшем вероятностно-статистическую модель диагностики. Пусть классифицируемые объекты описываются переменными x, лежащими в некотором пространстве Z, два класса – это два распределения вероятностей с плотностями $f(x)$ и $g(x)$, $x \in Z$, соответственно (если Z дискретно, то под $f(x)$ и $g(x)$ понимаем не плотности, а вероятности попадания в точку x).

Типичная схема разработки конкретного математического метода диагностики такова. С помощью специалистов соответствующей прикладной области составляют две обучающие выборки – объема m_0 из первого класса и объема n_0 из второго класса. На их основе определяют решающее правило $A: Z \to \{1,2\}$, ставящее в соответствие результату наблюдения $x \in Z$ номер класса, к которому его следует отнести. Качество работы алгоритма проверяют по контрольной выборке, состоящей из m элементов первого класса и n элементов второго. Результаты проверки удобно записать в виде табл.1.

Таблица 1. Результаты работы алгоритма диагностики

	Всего	Отнесено к первому классу	Отнесено ко второму классу
Элементы первого класса	m	a	b
Элементы второго класса	n	c	d

Ясно, что о качестве алгоритма диагностики, т.е. решающего правила A, надо судить на основе данных, приведенных в табл.1. Естественно использовать

$\kappa = a/m$ - долю правильной диагностики в первом классе;

$\lambda = d/n$ - долю правильной диагностики во втором классе.

Доля правильной диагностики μ равна

$$\mu = \frac{a+d}{m+n} = \pi_1\kappa + \pi_2\lambda,$$

где $\pi_1 = m/(m+n)$ - априорная доля первого класса, $\pi_2 = n/(m+n)$ - априорная доля второго класса. Очевидно, $\pi_1 + \pi_2 = 1$. В вероятностной модели речь идет об априорных вероятностях классов. Удобно и в статистической постановке (табл.1), и в вероятностной использовать единый термин «доля», поскольку это не приводит к недоразумениям.

Рассмотрим сначала случай, когда $\min(m_0, n_0) \to \infty$, т.е. плотности f и g можно считать известными. При применении теории статистических решений к задачам диагностики [13] считаются известными также априорные вероятности классов и потери $C(j|i)$ от ошибочной диагностики – от отнесения объекта i-го класса к классу с номером j. Доказано (при любом числе классов), что в случае, когда все потери $C(j|i)$ равны между собой, минимизация суммарных потерь эквивалентна максимизации вероятности правильной диагностики [13], состоятельной оценкой которой

служит μ. *Видимо, этот факт, вместе с вычислительной простотой и наглядностью показателя* μ, *является причиной широкого использования доли (вероятности) правильной классификации* μ *как показателя качества алгоритма диагностики.*

Оптимальное решающее правило при совпадающих потерях $C(1|2) = C(2|1)$ имеет вид: если

$$\frac{f(x)}{g(x)} > \frac{\pi_2(\infty)}{\pi_1(\infty)}, \qquad (3)$$

то отнести x к первому классу, в противном случае – ко второму; здесь $\pi_1(\infty)$ и $\pi_2(\infty)$ - априорные вероятности первого и второго классов соответственно. Как показано выше, лемма Неймана-Пирсона, исходя из другой оптимизационной постановки (из минимизации ошибки второго рода при ограничении на уровень значимости), также дает правило, основанное на отношении плотностей вероятностей. В большинстве прикладных задач плотности вероятности неизвестны. Однако их, при $\min(m_0, n_0) \to \infty$, можно заменить состоятельными оценками, например, непараметрическими ядерными оценками плотности [8, 9, 14], и получить *универсальное асимптотически оптимальное правило диагностики*, с которым необходимо сравнивать все другие правила диагностики [1, 15]. Выбор других правил диагностики для решения конкретных прикладных задач должен быть обоснован на основе соответствующих задаче критериев, например, быстродействия алгоритмов и объемов имеющейся информации [16].

Обратим внимание, что в случае, когда априорная вероятность одного из классов существенно больше априорной вероятности другого, правило (3) может *любое* наблюдение относить к классу с наибольшей априорной вероятностью.

Разберем ситуацию подробнее. Пусть имеется некоторый алгоритм диагностики на два класса с долями правильной диагностики κ - в первом

классе и λ - во втором. Сравним его с двумя тривиальными алгоритмами диагностики. Первый тривиальный алгоритм относит все классифицируемые объекты к первому классу, для него κ = 1 и λ = 0, следовательно, $\mu = \pi_1$. Второй тривиальный алгоритм относит все классифицируемые объекты ко второму классу, для него κ = 0 и λ = 1, следовательно, $\mu = \pi_2$.

В качестве показателя качества алгоритма диагностики будем использовать долю правильной диагностики μ. Когда первый тривиальный алгоритм лучше исходного? Когда $\pi_1 > \kappa\pi_1 + \lambda\pi_2$, т.е.

$$\frac{\lambda}{1-\kappa+\lambda} < \pi_1$$

(с учетом того, что $\pi_1 + \pi_2 = 1$). Когда второй тривиальный алгоритм лучше исходного? Когда $\pi_2 > \kappa\pi_1 + \lambda\pi_2$, т.е.

$$\pi_1 < \frac{1-\lambda}{1+\kappa-\lambda}.$$

Таким образом, для любого заданного алгоритма диагностики существуют границы $d(1)$ и $d(2)$ для доли первого класса π_1 в объединенной контрольной выборке такие, что при $\pi_1 < d(1)$ рассматриваемый алгоритм хуже второго тривиального алгоритма, а при $\pi_1 > d(2)$ он хуже первого тривиального алгоритма.

Разобранная ситуация встречается на практике. В конкретной прикладной медицинской задаче величина μ оказалась больше для тривиального прогноза, согласно которому у всех больных течение заболевания (инфаркта миокарда) будет благоприятно. Тривиальный прогноз сравнивался с алгоритмом выделения больных с прогнозируемым тяжелым течением заболевания. Он был разработан группой математиков и кардиологов под руководством И.М. Гельфанда. Применение этого алгоритма с медицинской точки зрения вполне оправдано [17 - 19]. Итак, по доле правильной классификации μ алгоритм группы И.М. Гельфанда

оказался хуже тривиального - объявить всех больных легкими, т.е. не требующими специального наблюдения. Этот вывод очевидно нелеп. И причина появления нелепости вполне понятна. Хотя доля тяжелых больных невелика, но смертельные исходы сосредоточены именно в этой группе больных. Поэтому целесообразна *гипердиагностика* - рациональнее часть легких больных объявить тяжелыми, чем сделать ошибку в противоположную сторону.

Поэтому мы полагаем, что использовать в качестве показателя качества алгоритма диагностики долю правильной диагностики μ нецелесообразно.

Работы группы И.М. Гельфанда [17 - 19] показывают также, что теория статистических решений не может быть основой для выбора показателя качества диагностики. Применение этой теории требует знания потерь от ошибочной диагностики, а в большинстве научно-технических и экономических задач определить потери, как уже отмечалось, сложно. В частности, из-за необходимости оценивать человеческую жизнь в денежных единицах. По этическим соображениям это, на наш взгляд, недопустимо. Сказанное не означает отрицания пользы страхования, но, очевидно, страховые выплаты следует рассматривать лишь как способ первоначального смягчения потерь от утраты близких. Следовательно, применение теории статистических решений в рассматриваемой постановке вряд ли возможно, поскольку оценить количественно потери от смерти больного нельзя по этическим соображениям.

4. Прогностическая сила

С целью поиска приемлемого показателя качества диагностики рассмотрим восходящую к Р. Фишеру [20] широко известную параметрическую вероятностную модель (модель линейного дискриминантного анализа), в которой Z – конечномерное пространство,

$f(x)$ и $g(x)$ – многомерные нормальные плотности с математическими ожиданиями m_1 и m_2 соответственно и совпадающими ковариационными матрицами Σ. Тогда при произвольных априорных вероятностях и потерях оптимальное решающее правило определяется плоскостью

$$H(x) = \left(x - \frac{m_1 + m_2}{2} \right)^T \Sigma^{-1}(m_1 - m_2) = C, \ x \in Z,$$

где константа C зависит от априорных вероятностей классов $\pi_1(\infty)$ и $\pi_2(\infty)$, а также от потерь $C(1|2)$ и $C(2|1)$ [10, с.186]. Основной параметр модели – расстояние Махаланобиса между классами

$$d = \left\{ (m_1 - m_2)^T \Sigma^{-1}(m_1 - m_2) \right\}^{1/2}.$$

(Величину d^2 нельзя называть «расстоянием», как это делается в [10, с.187], поскольку для d^2 не выполнено неравенство треугольника. Всем аксиомам, задающим метрику, удовлетворяет d.)

Через d и C/d выражаются вероятности правильной диагностики $\kappa(\infty)$ и $\lambda(\infty)$, являющиеся пределами при $\min(m,n) \to \infty$ ранее введенных долей κ и λ. Если X – случайная величина с описанной выше плотностью $f(x)$, то случайная величина $H(X)$ имеет нормальное распределение с математическим ожиданием $d^2/2$ и дисперсией d^2. Аналогично для случайной величины Y с плотностью $g(x)$ случайная величина $H(Y)$ имеет нормальное распределение с математическим ожиданием $(-d^2/2)$ и дисперсией d^2 [10, с.187]. Поэтому

$$\kappa(\infty) = P(H(X) > C) = \Phi\left(\frac{d}{2} - \frac{C}{d} \right), \ \lambda(\infty) = P(H(Y) \le C) = \Phi\left(\frac{d}{2} + \frac{C}{d} \right), \qquad (4)$$

где $\Phi(x)$ - функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из (4) вытекает, что при любых потерях $C(1|2)$ и $C(2|1)$ и любых априорных вероятностях π_1 и π_2 при использовании оптимального решающего правила величина

$$d_0 = \Phi^{-1}(\kappa(\infty)) + \Phi^{-1}(\lambda(\infty))$$

постоянна и равна d, где $\Phi^{-1}(y)$ - функция, обратная к $\Phi(x)$. Следовательно, именно расстояние Махаланобиса d целесообразно рассматривать как меру различия между классами, заданными плотностями рассматриваемого вида. Чтобы выразить меру различия в тех же единицах, что и вероятности правильной диагностики, введем «прогностическую силу»

$$\delta = \Phi\left(d/2\right),$$

при $C = 0$ равную совпадающим значениям правильной диагностики $\kappa(\infty)$ и $\lambda(\infty)$ из (4).

Мы предлагаем в качестве показателя качества произвольного алгоритма диагностики использовать величину

$$\delta^* = \Phi\left(d^*/2\right) - \qquad (5)$$

- эмпирическую прогностическую силу, где

$$d^* = \Phi^{-1}(\kappa) + \Phi^{-1}(\lambda) \qquad (6)$$

- выборочная оценка расстояния Махаланобиса, доли κ и λ правильной диагностики в классах определены по данным табл.1. Таким образом, вместо взвешенного по априорным долям классов среднего арифметического μ долей κ и λ правильной диагностики в классах предлагаем использовать их среднее по Колмогорову с весовой функцией Φ^{-1} (о средних по Колмогорову см., например, [5]).

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется классический линейный дискриминантный анализ Р.Фишера [10, 20], то величина d^* представляет собой состоятельную статистическую оценку расстояния Махаланобиса между двумя рассматриваемыми совокупностями, причем независимо от порогового значения, определяющего конкретное решающее правило. В общем случае показатель d^* вводится как эвристический.

Пример. Если κ = 0,90 и λ = 0,80, то $\Phi^{-1}(\kappa)$ = 1,28 и $\Phi^{-1}(\lambda)$ = 0,84, откуда d^* = 2,12 и эмпирическая прогностическая сила δ^* = $\Phi^{-1}(1,06)$ = 0,86. При этом доля правильной диагностики μ может принимать любые значения между 0,80 и 0,90, в зависимости от долей классов в объединенной совокупности π_i, $i=1,2$; $\pi_1+\pi_2=1$.

Изучено асимптотическое распределение δ^*, разработаны методы расчета доверительных границ для прогностический силы по данным табл.1 [21, 22].

Как проверить обоснованность применения прогностической силы, т.е. допустимость пересчета на модель линейного дискриминантного анализа? В ряде прикладных задач диагностики вычисляют значение некоторого прогностического индекса (фактора, переменной) y и решение принимают на основе его сравнении с некоторым заданным порогом c. Объект относят к первому классу, если $y \leq c$, ко второму, если $y > c$. Прогностический индекс – это обычно линейная функция от характеристик рассматриваемых объектов. Другими словами, от координат векторов, описывающих объекты. Возьмем два значения порога c_1 и c_2. Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для построения прогностического индекса применяется классический линейный дискриминантный анализ Р.Фишера, другими словами, если пересчет на модель линейного дискриминантного анализа обоснован, то, как можно показать, «прогностические силы» для обоих правил совпадают: $\delta(c_1)=\delta(c_2)$. Выполнение этого равенства можно проверить как статистическую гипотезу. Если эта гипотеза принимается, то целесообразность использования прогностической силы подтверждается, и есть основания значение $\delta^*(c_1)\approx\delta^*(c_2)$ рассматривать как объективную оценку качества алгоритма диагностики. Если же рассматриваемая гипотеза отклоняется,

т.е. значения $\delta^*(c_1)$ и $\delta^*(c_2)$ сильно различаются, то пересчет на модель линейного дискриминантного анализа и использование расстояния Махаланобиса для измерения различия классов и прогностической силы как показателя качества диагностики некорректны. Способ проверки, т.е. соответствующий критерий проверки статистической гипотезы $\delta(c_1) = \delta(c_2)$, включая алгоритмы расчетов, разработан в [21, 22].

5. Сходимость итерационных алгоритмов кластер-анализа

Сначала обсудим один из широко применяемых методов кластер-анализа - с метода *k*-средних. Он предназначен для разбиения исходного множества элементов (объектов или признаков) x_1, x_2, ... , x_n, лежащих в некотором пространстве *Z*, на *k* кластеров. Опишем его в предлагаемой нами общей формулировке.

Метод основан на использовании функции $f : Z^2 \to [0,+\infty)$ в пространстве *Z*, имеющей смысл показателя различия (меры близости, расстояния), т.е. чем элементы *x* и *y* дальше отстоят друг от друга (в смысле, принятом в предметной области), тем *f*(*x*, *y*) больше, причем *f*(*x*, *x*) = 0 для любого *x* из *Z*.

Метод *k*-средних - итерационный. Каждая итерация состоит из двух шагов - распределения элементов x_1, x_2, ... , x_n по *k* кластерам (первый шаг) и расчете центров кластеров (второй шаг).

Исходная информация перед началом первого шага - центры *k* кластеров, т.е. точки a_1, a_2, ... , a_k пространства *Z*. Для каждого из элементов x_1, x_2, ... , x_n находим ближайший центр. Для каждого из центров формируем кластер, состоящий из тех элементов x_1, x_2, ... , x_n, которые ближе к этому центру, чем к другим центрам. Для ревнителей строгости уточним: если некий элемент находится на равном расстоянии от нескольких центров, то относим его к кластеру, центр которого имеет наименьший номер.

Исходная информация перед началом второго шага итерации - разбиение на k кластеров (полученное на первом шаге). На каждом шаге рассчитываем его центр. В соответствии с методологией статистики объектов нечисловой природы [5] в качестве центра кластера будем использовать эмпирическое среднее элементов, включенных в кластер. Таким образом, для кластера A в качестве центра используем решение оптимизационной задачи

$$f(A, y) = \sum_{x \in A} f(x, y) \to \min_{y \in Z}. \qquad (7)$$

Приведем примеры [23]. Если Z - конечномерное евклидово пространство, $f(x, y)$ - квадрат евклидова расстояния между точками x и y, то решением задачи (7) является центр тяжести точек, включенных в кластер. Другими словами, для каждой координаты в этой пространстве надо взять значения этой координаты для точек, включенных в кластер, и рассчитать среднее арифметическое этих значений. Это и будет значением первой координаты искомого центра. Если же $f(x, y)$ - блочное расстояние между точками x и y, то решением задачи (7) является точка, каждая координата которой - выборочная медиана для точек, включенных в кластер. Если Z - то или иное пространство бинарных отношений, $f(x, y)$ - расстояние Кемени, то решением задачи (7) является медиана Кемени.

Условия существования решения задачи (7) найдены в статистике объектов нечисловой природы [5, 23]. Если решение задачи (7) не единственно, то правила однозначного выбора центра кластера должны быть специально указаны. Здесь нет необходимости останавливаться на этих подробностях.

Итогом второго шага итерации являются (новые) центры k кластеров.

Переходим к следующей итерации. Строим (новые) кластеры. Распределение элементов по (новым) кластерам, вообще говоря, изменится

(по сравнению с распределением в начале предыдущей итерации). Находим для (новых) кластеров центры согласно формуле (7). Затем - следующий шаг.

Для запуска алгоритма необходимо перед началом первой итерации тем или иным способом задать центры k кластеров. Можно взять первые k из элементов $x_1, x_2, \dots, x_n$, или случайно выбрать k из них, или задать k точек из пространства Z (например, стараясь равномерно охватить естественную область изменения подлежащих кластеризации элементов), и т.д. Влияние начального задания элементов уменьшается при увеличении числа итераций.

Проблема сходимости итерационного алгоритма k-средних такова: остановится ли процесс итераций? Остановка возможна, если для некоторой итерации "новые" кластеры совпадут со "старыми", тогда и "новые" центры совпадут со "старыми".

Априори есть две возможности. Либо итерационный алгоритм остановится, дав точное решение задачи кластер-анализа в рассматриваемой обстановке. Либо итерации могут продолжаться бесконечно, и тогда для получения приближенного решения надо тем или иным способом его останавливать.

Во всех проведенных расчетах итерационный алгоритм метода k-средних останавливался. Но ограниченный прошлый опыт не может дать гарантий на будущее. Покажем, что рассматриваемый алгоритм всегда остановится.

Разбиение исходного множества элементов $x_1, x_2, \dots, x_n$, на кластеры, полученное на итерации с номером t, обозначим $\Psi = \{A_1, A_2, \dots, A_k\} = \{A_1(t), A_2(t), \dots, A_k(t)\}$. Рассмотрим один из основных показателей качества кластеризации - внутрикластерный разброс

$$g(\Psi) = \sum_{i=1}^{k} f(A_i, y_i), \qquad (8)$$

где функция $f(A, y)$ определяется формулой (7), а $y_i = y_i(t)$ - центр кластера A_i , полученный на итерации с номером t путем решения оптимизационной задачи (7), $i = 1, 2, ..., k$. На втором шаге очередной итерации каждое слагаемое в правой части (8) либо уменьшается (если новый центр отличен от предыдущего), либо остается тем же самым (если новый центр совпадает с предыдущим). Следовательно, и сумма этих слагаемых - внутрикластерный разброс $g(\Psi)$ - либо уменьшается, либо остается тем же самым. При этом $g(\Psi)$ не меняется тогда и только тогда, когда все слагаемые не меняются, т.е. все центры кластеров не меняются. Такое бывает только при остановке процесса итераций. Другими словами, при продолжении итераций внутрикластерный разброс $g(\Psi)$ монотонно уменьшается. С другой стороны, число различных значений внутрикластерный разброс $g(\Psi)$ конечно, оно не превышает числа разбиений исходного множества элементов $x_1, x_2, ... , x_n$, на k кластеров. Следовательно, число возможных итераций конечно, рассматриваемый алгоритм всегда остановится, что и требовалось доказать.

Итерационные процедуры применяются в различных методах кластер-анализа. Публикация [24] посвящена проблеме остановки алгоритмов – доказательству того, что итерации эталонных алгоритмов (типа «Форель» и метода k-средних) прекращаются через конечное число шагов (оцененное сверху в этой работе). Обобщение было получено в докладе [25]. Итоги многолетних работ по различным вопросам теории классификации подведены в работе [26].

Основные наши результаты по теории классификации отражены в обширных статьях [15, 27, 28]. Подчеркнем, что все методы классификации, основанные на использовании расстояний (мер различия или близости), естественно рассматривать как часть статистики объектов

нечисловой природы [5, 29]. Недавно полученные результаты по теории классификации отражены в работах [30, 31]. Заслуживает дальнейшего развития и широкого применения метод когнитивной кластеризации, обоснованный в системно-когнитивном анализе и реализованный в его программном инструментарии – интеллектуальной системе «Эйдос» [32]. В этом методе, в частности, критерий сходства объектов кластеризации - не евклидово расстояние или его варианты, а интегральный критерий неметрической природы - «суммарное количество информации».

Математические методы классификации - важная составная часть системной нечеткой интервальной математики [33 - 35]. Они входят в число наиболее перспективных математических и инструментальных методов контроллинга [36]. Необходимо дальнейшее методологическое осмысление состояния и перспектив развития математической теории классификации, входящих в нее моделей и методов.

Литература

1. Орлов А.И. О развитии математических методов теории классификации // Заводская лаборатория. Диагностика материалов. 2009. Т.75. №7. С.51-63.
2. Новиков Д.А., Орлов А.И. Математические методы классификации // Заводская лаборатория. Диагностика материалов. 2012. Т.78. №4. С.3-3.
3. Орлов А.И. Математические методы теории классификации // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 95. С. 423 – 459.
4. Орлов А.И. Структура непараметрической статистики // Заводская лаборатория. Диагностика материалов. 2015. Т.81. №7.
5. Орлов А.И. Тридцать лет статистики объектов нечисловой природы (обзор) // Заводская лаборатория. Диагностика материалов. 2009. Т.75. №5. С.55-64.
6. Орлов А.И. Новая парадигма прикладной статистики // Заводская лаборатория. Диагностика материалов. 2012. Том 78. №1, часть I. С.87-93.
7. Леман Э.Л. Проверка статистических гипотез. 2-е изд., испр. — М.: Наука, 1979. — 408 с.
8. Орлов А.И. Оценки плотности в пространствах произвольной природы // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. / Перм. гос. нац. иссл. ун-т. – Пермь, 2013. – Вып. 25. – С.21-33.
9. Орлов А.И. Предельные теоремы для ядерных оценок плотности в пространствах произвольной природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 108. С. 316–333.

10. Андерсон Т. Введение в многомерный статистический анализ. - М.: Физматгиз, 1963. - 500 с.

11. Рао С.Р. Линейные статистические методы и их применения. - М.: Наука, 1968. - 548 с.

12. Алгоритмы и программы восстановления зависимостей / Под ред. В.Я. Вапника. – М.: Наука, 1984. – 816 с.

13. Горелик А.Л., Скрипкин В.А. Методы распознавания: учеб. для вузов. – М.: Высшая школа, 1984. – 208 с.

14. Орлов А.И. Ядерные оценки плотности в пространствах произвольной природы // Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Пермский госуниверситет, 1996. - С.68-75.

15. Орлов А.И. Математические методы исследования и диагностика материалов // Заводская лаборатория. Диагностика материалов. 2003. Т.69. № 3. С.53-64.

16. Толчеев В.О. Модифицированный и обобщенный метод ближайшего соседа для классификации библиографических текстовых документов // Заводская лаборатория. Диагностика материалов. 2009. Т.75. № 7. С.63-70.

17. Алексеевская М.А., Гельфанд И.М., Губерман Ш.А., Мартынов И.В., Ротвайн И.М., Саблин В.М. Прогнозирование исхода мелкоочагового инфаркта миокарда с помощью программы узнавания // Кардиология. 1977. Т.17. № 7. С.26-71.

18. Гельфанд И.М., Губерман Ш.А., Сыркин А.Л., Головня Л.Д., Извекова М.Л., Алексеевская М.А. Прогнозирование исхода инфаркта миокарда с помощью программы «Кора-3» // Кардиология. – 1977. – Т.17, № 6. – С.19-23.

19. Гельфанд И.М., Розенфельд Б.И., Шифрин М.А. Очерки о совместной работе математиков и врачей (2-е, дополненное издание). - М. УРСС, 2004. – 320 с.

20. Фишер Р.Э. Использование множественных измерений в задачах таксономии // Современные проблемы кибернетики. - М.: Знание, 1979. - С. 6 - 20. (*Fisher R.A.* The use of multiple measurements in taxonomic problems // Ann. Eugenics. 1936, September. V.7. P. 179 - 188.)

21. Орлов А.И. Прогностическая сила как показатель качества алгоритма диагностики // Статистические методы оценивания и проверки гипотез: межвуз. сб. науч. тр. Вып.23. – Пермь: Перм. гос. нац. иссл. ун-т, 2011. – С.104-116.

22. Орлов А.И. Прогностическая сила – наилучший показатель качества алгоритма диагностики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2014. № 99. С. 15–32.

23. Орлов А.И. Организационно-экономическое моделирование : учебник : в 3 ч. Ч. 1. Нечисловая статистика. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2009. — 541 с.

24. Орлов А.И. Сходимость эталонных алгоритмов // Прикладной многомерный статистический анализ. Ученые записки по статистике, т.33. - М.: Наука, 1978. С.361-364.

25. Орлов А.И. Остановка после конечного числа шагов для алгоритмов кластер-анализа // Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, т.36. - М.: Наука, 1980. С.374-377.

26. Орлов А.И. Некоторые вероятностные вопросы теории классификации // Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. С.166-179.

27. Орлов А.И. Классификация объектов нечисловой природы на основе непараметрических оценок плотности // Проблемы компьютерного анализа данных и

моделирования: Сборник научных статей. - Минск: Изд-во Белорусского государственного университета, 1991. С.141-148.

28. Орлов А.И. Заметки по теории классификации // Социология: методология, методы, математические модели. 1991. № 2. С.28-50.

29. Орлов А.И. О развитии статистики объектов нечисловой природы // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2013. № 93. С. 273 – 309.

30. Орлов А.И., Толчеев В.О. Об использовании непараметрических статистических критериев для оценки точности методов классификации (обобщающая статья) // Заводская лаборатория. Диагностика материалов. 2011. Т.77. № 3. С.58-66.

31. Орлов А.И. Устойчивость классификации относительно выбора метода кластер-анализа // Заводская лаборатория. Диагностика материалов. 2013. Т.79. № 1. С.68-71.

32. Луценко Е.В., Коржаков В.Е. Метод когнитивной кластеризации или кластеризация на основе знаний (кластеризация в системно-когнитивном анализе и интеллектуальной системе «Эйдос») // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2011. № 71. С. 528 – 576.

33. Орлов А.И., Луценко Е.В. О развитии системной нечеткой интервальной математики // Философия математики: актуальные проблемы. Математика и реальность. Тезисы Третьей всероссийской научной конференции; 27-28 сентября 2013 г. / Редкол.: Бажанов В.А. и др. – Москва, Центр стратегической конъюнктуры, 2013. – С.190–193.

34. Орлов А.И., Луценко Е.В. Системная нечеткая интервальная математика (СНИМ) – перспективное направление теоретической и вычислительной математики // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2013. № 91. С. 255 – 308.

35. Орлов А.И., Луценко Е.В. Системная нечеткая интервальная математика. Монография (научное издание). – Краснодар, КубГАУ. 2014. – 600 с.

36. Орлов А.И., Луценко Е.В., Лойко В.И. Перспективные математические и инструментальные методы контроллинга. Под научной ред. проф. С.Г. Фалько. Монография (научное издание). – Краснодар, КубГАУ. 2015. – 600 с.

References

1. Orlov A.I. O razvitii matematicheskih metodov teorii klassifikacii // Zavodskaja laboratorija. Diagnostika materialov. 2009. T.75. №7. S.51-63.

2. Novikov D.A., Orlov A.I. Matematicheskie metody klassifikacii // Zavodskaja laboratorija. Diagnostika materialov. 2012. T.78. №4. S.3-3.

3. Orlov A.I. Matematicheskie metody teorii klassifikacii // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 95. S. 423 – 459.

4. Orlov A.I. Struktura neparametricheskoj statistiki // Zavodskaja laboratorija. Diagnostika materialov. 2015. T.81. №7. S.

5. Orlov A.I. Tridcat' let statistiki ob#ektov nechislovoj prirody (obzor) // Zavodskaja laboratorija. Diagnostika materialov. 2009. T.75. №5. S.55-64.

6. Orlov A.I. Novaja paradigma prikladnoj statistiki // Zavodskaja laboratorija. Diagnostika materialov. 2012. Tom 78. №1, chast' I. S.87-93.

7. Leman Je.L. Proverka statisticheskih gipotez. 2-e izd., ispr. — M.: Nauka, 1979. — 408 s.

8. Orlov A.I. Ocenki plotnosti v prostranstvah proizvol'noj prirody // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. / Perm. gos. nac. issl. un-t. – Perm', 2013. – Vyp. 25. – S.21-33.

9. Orlov A.I. Predel'nye teoremy dlja jadernyh ocenok plotnosti v prostranstvah proizvol'noj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2015. № 108. S. 316–333.

10. Anderson T. Vvedenie v mnogomernyj statisticheskij analiz. - M.: Fizmatgiz, 1963. - 500 s.

11. Rao S.R. Linejnye statisticheskie metody i ih primenenija. - M.: Nauka, 1968. - 548 s.

12. Algoritmy i programmy vosstanovlenija zavisimostej / Pod red. V.Ja. Vapnika. – M.: Nauka, 1984. – 816 s.

13. Gorelik A.L., Skripkin V.A. Metody raspoznavanija: ucheb. dlja vuzov. – M.: Vysshaja shkola, 1984. – 208 s.

14. Orlov A.I. Jadernye ocenki plotnosti v prostranstvah proizvol'noj prirody // Statisticheskie metody ocenivanija i proverki gipotez. Mezhvuzovskij sbornik nauchnyh trudov. - Perm': Permskij gosuniversitet, 1996. - S.68-75.

15. Orlov A.I. Matematicheskie metody issledovanija i diagnostika materialov // Zavodskaja laboratorija. Diagnostika materialov. 2003. T.69. № 3. S.53-64.

16. Tolcheev V.O. Modificirovannyj i obobshhennyj metod blizhajshego soseda dlja klassifikacii bibliograficheskih tekstovyh dokumentov // Zavodskaja laboratorija. Diagnostika materialov. 2009. T.75. № 7. S.63-70.

17. Alekseevskaja M.A., Gel'fand I.M., Guberman Sh.A., Martynov I.V., Rotvajn I.M., Sablin V.M. Prognozirovanie ishoda melkoochagovogo infarkta miokarda s pomoshh'ju programmy uznavanija // Kardiologija. 1977. T.17. № 7. S.26-71.

18. Gel'fand I.M., Guberman Sh.A., Syrkin A.L., Golovnja L.D., Izvekova M.L., Alekseevskaja M.A. Prognozirovanie ishoda infarkta miokarda s pomoshh'ju programmy «Kora-3» // Kardiologija. – 1977. – T.17, № 6. – S.19-23.

19. Gel'fand I.M., Rozenfel'd B.I., Shifrin M.A. Ocherki o sovmestnoj rabote matematikov i vrachej (2-e, dopolnennoe izdanie). - M. URSS, 2004. – 320 s.

20. Fisher R.Je. Ispol'zovanie mnozhestvennyh izmerenij v zadachah taksonomii // Sovremennye problemy kibernetiki. - M.: Znanie, 1979. - S. 6 - 20. (Fisher R.A. The use of multiple measurements in taxonomic problems // Ann. Eugenics. 1936, September. V.7. P. 179 - 188.)

21. Orlov A.I. Prognosticheskaja sila kak pokazatel' kachestva algoritma diagnostiki // Statisticheskie metody ocenivanija i proverki gipotez: mezhvuz. sb. nauch. tr. Vyp.23. – Perm': Perm. gos. nac. issl. un-t, 2011. – S.104-116.

22. Orlov A.I. Prognosticheskaja sila – nailuchshij pokazatel' kachestva algoritma diagnostiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2014. № 99. S. 15–32.

23. Orlov A.I. Organizacionno-jekonomicheskoe modelirovanie : uchebnik : v 3 ch. Ch. 1. Nechislovaja statistika. – M.: Izd-vo MGTU im. N.Je. Baumana, 2009. — 541 s.

24. Orlov A.I. Shodimost' jetalonnyh algoritmov // Prikladnoj mnogomernyj statisticheskij analiz. Uchenye zapiski po statistike, t.33. - M.: Nauka, 1978. S.361-364.

25. Orlov A.I. Ostanovka posle konechnogo chisla shagov dlja algoritmov klaster-analiza // Algoritmicheskoe i programmnoe obespechenie prikladnogo statisticheskogo analiza. Uchenye zapiski po statistike, t.36. - M.: Nauka, 1980. S.374-377.

26. Orlov A.I. Nekotorye verojatnostnye voprosy teorii klassifikacii // Prikladnaja statistika. Uchenye zapiski po statistike, t.45. - M.: Nauka, 1983. S.166-179.

27. Orlov A.I. Klassifikacija ob#ektov nechislovoj prirody na osnove

neparametricheskih ocenok plotnosti // Problemy komp'juternogo analiza dannyh i modelirovanija: Sbornik nauchnyh statej. - Minsk: Izd-vo Belorusskogo gosudarstvennogo universiteta, 1991. S.141-148.

28. Orlov A.I. Zametki po teorii klassifikacii // Sociologija: metodologija, metody, matematicheskie modeli. 1991. № 2. S.28-50.

29. Orlov A.I. O razvitii statistiki ob#ektov nechislovoj prirody // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2013. № 93. S. 273 – 309.

30. Orlov A.I., Tolcheev V.O. Ob ispol'zovanii neparametricheskih statisticheskih kriteriev dlja ocenki tochnosti metodov klassifikacii (obobshhajushhaja stat'ja) // Zavodskaja laboratorija. Diagnostika materialov. 2011. T.77. № 3. S.58-66.

31. Orlov A.I. Ustojchivost' klassifikacii otnositel'no vybora metoda klaster-analiza // Zavodskaja laboratorija. Diagnostika materialov. 2013. T.79. № 1. S.68-71.

32. Lucenko E.V., Korzhakov V.E. Metod kognitivnoj klasterizacii ili klasterizacija na osnove znanij (klasterizacija v sistemno-kognitivnom analize i intellektual'noj sisteme «Jejdos») // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2011. № 71. S. 528 – 576.

33. Orlov A.I., Lucenko E.V. O razvitii sistemnoj nechetkoj interval'noj matematiki // Filosofija matematiki: aktual'nye problemy. Matematika i real'nost'. Tezisy Tret'ej vserossijskoj nauchnoj konferencii; 27-28 sentjabrja 2013 g. / Redkol.: Bazhanov V.A. i dr. – Moskva, Centr strategicheskoj kon#junktury, 2013. – S.190–193.

34. Orlov A.I., Lucenko E.V. Sistemnaja nechetkaja interval'naja matematika (SNIM) – perspektivnoe napravlenie teoreticheskoj i vychislitel'noj matematiki // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta. 2013. № 91. S. 255 – 308.

35. Orlov A.I., Lucenko E.V. Sistemnaja nechetkaja interval'naja matematika. Monografija (nauchnoe izdanie). – Krasnodar, KubGAU. 2014. – 600 s.

36. Orlov A.I., Lucenko E.V., Lojko V.I. Perspektivnye matematicheskie i instrumental'nye metody kontrollinga. Pod nauchnoj red. prof. S.G. Fal'ko. Monografija (nauchnoe izdanie). – Krasnodar, KubGAU. 2015. – 600 s.