Труды Карельского научного центра РАН
№ 3. 2010. С. 93–97

УДК 512.2

О НОВОМ МНОГОМЕРНОМ СТАТИСТИЧЕСКОМ КРИТЕРИИ ОДНОРОДНОСТИ ДВУХ ВЫБОРОК

С. П. Чистяков

*Институт прикладных математических исследований
Карельского научного центра РАН*

В статье предложен многомерный статистический критерий однородности (являются ли две выборки выборками из одного и того же многомерного распределения). Критерий основан на понятиях многомерной квантильной функции, множеств минимального объема и одноклассовом методе опорных векторов. Статистика критерия фактически представляет собой статистику критерия о равенстве параметра биномиального распределения определенному значению. Проведено экспериментальное сравнение критерия с некоторыми другими статистическими критериями однородности, такими как критерий Хотеллинга, многомерный ранговый критерий и ядерный критерий Крамера. Для проведения экспериментов использовался пакет **R –** свободно распространяемое программное обеспечение для статистических вычислений и графики, доступное на многих платформах.

К л ю ч е в ы е с л о в а: многомерные критерии однородности, многомерная квантильная функция, множества минимального объема, одноклассовый метод опорных векторов.

S. P. Chistiakov. ON A NEW MULTIVARIATE STATISTICAL HOMOGENEITY TEST

In this paper a multivariate homogeneity test (of whether two sets of observations arise from the same distribution) is proposed. The test is based on the concepts of multidimentional quantile function, minimum volume sets, and one-class support vector machines. The test statistic is as a matter of fact, the test statistic for binomial proportions. We conducted experimental comparison of our test with some statistical tests such as Hotelling test, multivariate rank test, and kernel Cramer test. For the experiments we used package **R** – open source environment for statistical computing and graphics which is freely available for most computing platforms.

K e y w o r d s: multivariate homogeneity tests, multidimensional quantile function, minimal volume sets, one-class support vector machines.

Введение

В статье предложен многомерный статистический критерий однородности двух выборок, т. е. критерий проверки нулевой гипотезы о том, что выборки являются выборками из одного распределения. Критерий основан на понятиях многомерной квантильной функции (обобщении понятия квантиля распределения на многомерный случай), множеств минимального объема и одноклассовом методе опорных векторов. Ключевую роль в построении критерия играет одноклассовый метод опорных векторов. Именно, предположим, что имеется случайная выборка из распределения P на некотором множестве X и мы хотим построить по данной выборке некоторое «достаточно простое» множество C, такое, что вероятность его дополнения была бы близка к некоторому заранее фиксированному числу ν между 0 и 1. Для решения этой задачи в работе [Scholkopf et al., 1999] был разработан метод, в оригинальной транскрипции носящий название novelty detection, а в отечественной литературе носящий несколько названий: одноклассовая классификация, обнаружение новизны (или нетипичности).

Метод основан на построении функции f, положительной на множестве C и отрицательной на его дополнении. Такая функция ищется в классе функций, допускающих представление в виде ядерного разложения по некоторому подмножеству векторов наблюдаемой выборки (называемых опорными векторами). Метод основан на построении гиперплоскости в так называемом (гильбертовом) «спрямляющем» пространстве высокой (возможно бесконечной) размерности, отделяющей выборку от нуля, и алгоритмически сводится к задаче квадратичного (выпуклого) программирования.

Несмотря на сравнительно недавнее появление, метод получил широкое распространение для решения задач обнаружения аномальных наблюдений, особенно для задач обнаружения аномального поведения сложных систем. Учитывая, что задача проверки на аномальность (т. е. на принадлежность данному распределению) является фактически задачей проверки гипотез, в которой вторая выборка состоит из одного наблюдения, естественно предположить, что данный метод может быть использован и для построения многомерных критериев однородности, в которых аномалией уже является целая выборка. Действительно (как показано в статье), использование одноклассового метода опорных векторов позволяет свести задачу проверки однородности двух многомерных выборок к задаче проверки гипотезы о том, равно ли значение параметра биномиального распределения (вероятность успеха) заданному числу, для которой соответствующий критерий хорошо известен.

Одноклассовый метод опорных векторов реализован в пакете **kernlab** [Karatzoglou et al., 2004], который является открытым пакетом, реализующим ядерные методы машинного обучения в среде **R** – свободно распространяемом программном обеспечении для статистических вычислений и графики, доступном на многих платформах (Linux, Windows, Macintosh). Данный пакет и был использован в проведенных нами экспериментах по сравнению предлагаемого критерия с несколькими другими критериями однородности, такими, как критерий Хотеллинга [Anderson, 2003] (многомерное обобщение критерия Стьюдента), многомерный ранговый критерий [Puri, Sen, 1971] и ядерный критерий Крамера [Baringhaus, Franz, 2004]. Отметим, что данные критерии также реализованы в среде **R** (пакеты **ICSNP** и **cramer**, которые также были нами использованы) и являются, по-видимому, единственными из многомерных критериев однородности реализованными в **R**).

Статья организована следующим образом. В первом разделе приводятся определения основных используемых понятий (многомерные квантильные функции, множества минимального объема) и дано описание предложенного критерия. Второй раздел посвящен экспериментальным результатам. В заключении кратко обсуждаются возможности применения критерия и направления дальнейших исследований.

Построение критерия

Введем понятия многомерной квантильной функции и множества минимального объема [Einmal et al., 1992]. Пусть P – некоторое распределение на множестве X и $\mathbf{C}$ – некоторый класс измеримых подмножеств X. Далее, пусть $\lambda(C)$ – некоторая функция, определенная на множествах $C \in \mathbf{C}$. Многомерная квантильная функция по отношению к $(P, \lambda, \mathbf{C})$ определяется как

$$U(\alpha) = \inf\{\lambda(C) : P(C) \geq \alpha, C \in \mathbf{C}\}.$$

Из определения следует, что многомерная квантильная функция фактически измеряет, насколько большое множество необходимо для того, чтобы оно содержало не менее заданной доли рассматриваемого распределения. Обозна-

чим $C_P(\alpha) \in \mathbf{C}$ (не обязательно единственное) множество, на котором достигается инфимум (если он достижим). Если $X \subseteq \mathbf{R}^d$ и λ – мера Лебега, то множество $C_P(\alpha)$ является множеством минимального объема, содержащим не менее доли α вероятностного распределения P. Был разработан ряд методов для построения оценок таких множеств по имеющейся выборке из распределения P. Так в работе [Nolan, 1991] в качестве $\mathbf{C}$ рассматривался класс эллипсоидов, а в [Tsybakov, 1997] для оценивания $C_P(\alpha)$ были использованы кусочно-полиномиальные приближения. Множество минимального объема впервые было предложено использовать для построения многомерных критериев, по-видимому, в [Polonik, 1999].

В работе [Scholkopf et al., 1999] был разработан метод обнаружения новизны, позволяющий строить оценки множеств минимального объема. Метод основан на обобщении метода опорных векторов на случай обучающих выборок, содержащих примеры только одного класса. В результате применения метода по случайной выборке $D_P = \{x_1, x_2, ..., x_n\}$ из распределения P могут быть построены множества $C_P^*(\alpha, n)$ такие, что $P\{C_P^*(\alpha, n)\} \to P\{C_P(\alpha)\}$ при $n \to \infty$. Множество $C_P^*(\alpha, n)$ определяются условием $C_P^*(\alpha, n) = \{x \in X \mid f(x) > 0\}$, где

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), \; x \in X$$

и $k(x, x')$ неотрицательно определенная функция, носящая название ядерной фунции или ядра. Вектора x_i, для которых $\alpha_i \neq 0$, называются опорными векторами.

Перейдем теперь непосредственно к построению критерия. Пусть Q – некоторое другое распределение на X и $D_Q = \{y_1, y_2, ..., y_m\}$ – случайная выборка из распределения Q. Мы хотим проверить нулевую гипотезу $H_0 : P = Q$. Обозначим $S_1 = \#\{x_i \in D_P \mid x_i \in C_Q^*(1/2, m)\}$ – количество элементов выборки D_P, принадлежащих множеству $C_Q^*(1/2, m)$ и, аналогично, $S_2 = \#\{y_i \in D_Q \mid y_i \in C_P^*(1/2, n)\}$ – количество элементов выборки D_Q, принадлежащих множеству $C_P^*(1/2, n)$.

Предположим теперь, что нулевая гипотеза $H_0 : P = Q$ истинна. Тогда нетрудно видеть, что в случае достаточно больших объемов выборок n и m будет справедливо $P\{C_Q^*(1/2, m) \approx Q\{C_P^*(1/2, n) \approx 1/2$.

Следовательно, случайная величина S_1 будет распределена по биномиальному закону с параметрами n и $p_1 \approx 1/2$, а случайная величина S_2 будет распределена по биномиальному закону с параметрами m и $p_2 \approx 1/2$. Поэтому случайная величина $S = S_1 + S_2$ будет распределена также по биномиальному закону с параметрами $n + m$ и $p \approx 1/2$ и для проверки нашей нулевой гипотезы $H_0 : P = Q$ можно использовать критерий для проверки нулевой гипотезы о равенстве параметра биномиального распределения p значению $1/2$. Тестовая статистика хорошо известна [Agresti, 1996]. Именно

$$z = \frac{\frac{S}{n+m} - \frac{1}{2}}{s},$$

где стандартная ошибка

$$s = \sqrt{\frac{p(1-p)}{n+m}} \approx \frac{1}{2}\sqrt{\frac{1}{n+m}}.$$

Хорошо известно, что (в условиях истинности нулевой гипотезы) для больших $n + m$ статистика z распределена приближенно по стандартному нормальному закону $N(0,1)$. Нетрудно видеть, что если нулевая гипотеза неверна, то статистика S будет распределена по биномиальному закону с параметрами $n + m$ и $p < 1/2$. Поэтому критическую область критерия целесообразно определить как $z < t_\alpha$, где t_α – квантиль уровня α стандартного нормального закона.

Экспериментальные результаты

Для сравнения предложенного нами критерия с другими многомерными критериями однородности проведен ряд имитационных экспериментов. Использовались критерий Хотеллинга [Anderson, 2003], многомерный ранговый критерий (обобщение рангового критерия Вилкоксона на многомерный случай) [Puri, Sen, 1971] и ядерный критерий Крамера [Baringhaus, Franz, 2004].

Для оценки множеств минимального объема $C_P^*(\alpha, n)$ и $C_Q^*(\alpha, m)$, что, конечно, является центральной задачей при построении предложенного критерия, использовалась функция

ksvm пакета **kernlab** [Karatzoglou et al., 2004]. Данная функция допускает использование различных ядер. Нами использовалось ядро гауссовской радиальной базисной функции (Gaussian radial basis function). Одним из важных параметров такого ядра является его ширина. Для того чтобы избежать субъективности при выборе этого параметра и сделать проведенные нами эксперименты воспроизводимыми, использовалась функция автоматической оценки ширины ядра sigest, алгоритм которой основан на работе [Caputo et al., 2002] (также входящая в пакет **kernlab**).

Во всех экспериментах рассматривался случай многомерных гауссовских распределений P и Q на $X \subseteq \mathbf{R}^d$. Эксперименты проводились на вычислительном кластере Института прикладных математических исследований КарНЦ РАН.

Методика проведения экспериментов заключалась в следующем. Задавались модели генерации распределений P и Q, зависящие от некоторого параметра, и исследовалась мощность упомянутых выше критериев в зависимости от этого параметра. Для каждого значения параметра проводилось 500 имитационных экспериментов. Оценка мощности при каждом значении параметра определялась отношением числа отклонений нулевой гипотезы к числу экспериментов. Исследовалось несколько уровней по объемам выборок (рассматривался только случай $n = m$) и размерности d. Уровень значимости критериев во всех экспериментах выбирался равным 0,05. По этой схеме было проведено три группы экспериментов (за исключением первой).

Первая группа экспериментов проводилась для исследования скорости сходимости $P\{C_P^*(1/2, n)\}$ при $n \to \infty$ к значению $1/2$ при различных значениях d. При этом нас интересовала не скорость сходимости сама по себе, а выявление условий, гарантирующих заданный уровень значимости критерия. Результаты экспериментов показали, что условие $n/d > 50$ достаточно для этого.

Вторая группа экспериментов проводилась для исследования мощности критериев против альтернатив сдвига распределений. В этом случае моделировались выборки из распределений $N(\mathbf{0}, \mathbf{I})$ и $N(\gamma\mathbf{1}, \mathbf{I})$ (где $\mathbf{1}$ – единичный вектор размерности d, а $\mathbf{I}$ – единичная диагональная матрица) с различными значениями γ, и осуществлялась оценка мощности в зависимости от евклидова расстояния между средними двух распределений. Результаты показали, что предложенный тест реагирует на данный класс альтернатив, однако, мощность, близкая к единице, достигается при больших евклидовых расстояниях между средними, чем у остальных критериев.

Третья группа экспериментов проводилась для исследования поведения мощности в зависимости от дисперсии компонент распределений. Конкретно рассматривалась следующая модель генерации данных. Моделировались выборки $N(\mathbf{0}, \mathbf{I})$ и $N(\mathbf{0}, \sigma^2\mathbf{I})$ для различных значений σ^2. Как и следовало ожидать, критерий Хотеллинга и ранговый критерий имеют в этом (и в рассматриваемом ниже) случае мощность, близкую к нулевой, поскольку они основаны на статистиках, реагирующих только на альтернативы сдвига распределений. Критерий Крамера имеет несколько большую мощность чем предложенный нами критерий, однако близкая к единице мощность предложенного критерия достигается и в этом случае (при увеличении σ^2).

В четвертой группе экспериментов модель генерации данных имела следующий вид. Пусть $d = 2r$. Моделировались выборки $N(\mathbf{0}, \Sigma_1)$ и $N(\mathbf{0}, \Sigma_2)$, где Σ_1 – диагональная матрица с элементами

$$\sigma_{11} = \sigma^2, ..., \sigma_{rr} = \sigma^2, \sigma_{r+1r+1} = 1/\sigma^2, ..., \sigma_{dd} = 1/\sigma^2,$$

а Σ_2 – диагональная матрица с элементами

$$\sigma_{11} = 1/\sigma^2, ..., \sigma_{rr} = 1/\sigma^2, \sigma_{r+1r+1} = \sigma^2, ..., \sigma_{dd} = \sigma^2.$$

В этом случае определители ковариационных матриц Σ_1 и Σ_2, определяющих степень разброса выборок относительно общего нулевого среднего, совпадают, и разница между распределениями состоит только в ориентации эллипсоидов рассеяния. Как показали эксперименты, в этом случае предложенный критерий имеет более высокую мощность, чем критерий Крамера.

Заключение

По мнению автора, результаты экспериментов позволяют считать, что предложенный критерий может найти применение в статистической практике (по крайней мере в тех случаях, когда альтернативная гипотеза состоит в неравенстве ковариационных матриц распределений), однако, для полного исследования его свойств и определения классов альтернатив, для

которых критерий «хорошо работает», требуются дополнительные исследования. Исследование свойств данного критерия предполагается продолжать в следующих направлениях:

1. Рассмотрение случаев негауссовских многомерных распределений;
2. Рассмотрение случаев наличия дискретных компонент в распределениях.

Необходимо также, конечно, расширить список критериев для сравнения.

Автор выражает признательность начальнику Центра коллективного пользования КарНЦ РАН И. А. Фалько за оперативную установку программной среды **R** и ряда пакетов на вычислительный кластер и полезные консультации.

Литература

Anderson T. W. An introduction to multivariate analysis. New Jersey: Wiley, 2003. 453 p.

Baringhaus L., Franz C. On a new multivariate two-sample test // Journal of Multivariate Analysis. 2004. Vol. 88. P. 190–206.

Caputo B., Sim K., Furesjo F., Smola A. Appearance-based object recognition using SVMs: which kernel should I use? // Proceedings of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision. Whistler, 2002. P. 111–119.

Einmal J. H. J., Mason D. M. Generalized quantile processes // Annals of Statistics. 1992. Vol. 20(2). P. 1062–1078.

Agresti A. An Introduction to Categorical Data Analysis. New York : Wiley, 1996. 290 p.

Karatzoglou A., Smola A., Hornik K., Zeileis A. kernlab – An S4 Package for Kernel Methods in R // Journal of Statistical Software. 2004. Vol. 11 (9). P. 1–20.

URL: http://www.jstatsoft.org/v11/i09/ (дата обращения 18.05.2010).

Nolan D. The excess mass ellipsoids // Journal of Multivariate Analysis. 1991. Vol. 39. P. 348–371.

Polonik W. Concentration and goodness-of-fit in higher dimensions: Asymptotically distribution-free methods // The Annals of Statistics. 1999. Vol. 27. P. 1210–1229.

Puri M. L., Sen P. K. Nonparametric Methods in Multivariate Analysis. New York: Wiley, 1971. 342 p.

R Development Core Team. R: A language and environment for statistical computing // R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0, URL: http://www.R-project.org (дата обращения 18.05.2010).

Scholkopf B., Platt J., Shawe-Taylor J. et al. Estimating the support of a high-dimensional distribution // TR MSR 99-87. 1999. Microsoft Research: Redmond.

Tsybakov A. B. On nonparametric estimation of density level sets // Annals of Statistics. 1997. Vol. 25(3). P. 948–969.

СВЕДЕНИЯ ОБ АВТОРЕ:

Чистяков Сергей Павлович
младший научный сотрудник, к. т. н.
Институт прикладных математических исследований КарНЦ РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185910
эл. почта: chistiakov@krc.karelia.ru
тел.: (8142) 763370

Chistiakov, Sergey
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Science
11 Pushkinskaya St., 185910 Petrozavodsk, Karelia, Russia
e-mail: chistiakov@krc.karelia.ru
tel.: (8142) 763370