ON A MULTIVARIATE HOMOGENEITY TEST BASED ON ONE-CLASS SUPPORT VECTOR MACHINES

S.P. Chistiakov
Institute of Applied Mathematical Research
Petrozavodsk, RUSSIA
e-mail: chistiakov@krc.karelia.ru

Abstract

In this paper a new multivariate homogeneity test (whether two sets of observations arise from the same distribution) is proposed. The test is based on concepts of minimum volume sets and one-class support vector machines. The test statistic as a matter of fact is test statistic for binomial proportions. We conducted experimental comparison our test with some statistical tests such as Hotelling test (the classical test for testing the mean difference for two multivariate populations) [1], multivariate rank test [8] and kernel Cramer test [2]. For experiments we used package **R** [9] — open source environment for statistical computing and graphics which is freely available for most computing platforms.

1 Introduction

In this paper we propose a new multivariate homogeneity test. The test is based on concepts of minimum volume sets [4] and one-class support vector machines. Let $\mathcal{D}$ be the random sample from some distribution P on the set $\mathcal{X}$. Let us we want to estimate a "simple" subset C of input space $\mathcal{X}$, such that the probability that a test point drawn from P lies outside of C equals some a priori specified value between 0 and 1. The appproach to solve this problem was proposed in [10]. The approach consists in constructing of a function f which is positive on C and negative on $\mathcal{X} \backslash C$. The functional form of f is given by a kernel expansion in terms of support vectors. The approach was realized in a package **kernlab** [6], which is an extensible package for kernel–based machine learning methods in **R**. We used above concepts to reduce a problem of the statistical homogeneity test construction to a testing problem for the binomial parameter. Experimental results show that such approach may be used in statistical practice.

2 Statistical test construction

In this section we present our approach to statistical test construction based on one-class support vector machines. We first introduce the concept of the multidimensional quantile function and minimum volume sets [4]. Let P be some distribution in a set $\mathcal{X}$ and let $\mathcal{C}$ be a class of measurable subsets $\mathcal{X}$. Let $\lambda(C)$ be a real–valued function defined on the sets $C \in \mathcal{C}$. The multidimensional quantile function with respect to $(P, \lambda, \mathcal{C})$ is defined as

$$U(\alpha) = \inf\{\lambda(C) : P(C) \geqslant \alpha, C \in \mathcal{C}\}, \quad 0 < \alpha \leqslant 1.$$

It is evident that quantile function measures how large a set one needs in order to capture a certain amount of probability mass of P. Denote by $C_P(\alpha) \in \mathcal{C}$ the (not necessarily unique) set that attain the infimum (when it is achievable). If λ is Lebesgue measure then $C_P(\alpha)$ is the minimum volume set which contains at least a fraction α of the probability mass P. One of the reasons to use the minimal volume sets for constructing two-sample tests is that they are capable of discriminating different distributions. In [7], one approach to this problem is proposed . In [10], method novelty detection to construct estimators of the minimal volume sets was proposed. The method enable possibility to constructing sets $\tilde{C}_P(\alpha, n)$ such that $P\{\tilde{C}_P(\alpha, n)\} \to P\{C_P(\alpha)\}$ as $n \to \infty$. It uses a random sample $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ from the distribution P and based on generalization of the support vector machines on the case of unlabelled training data. Now let Q be another distribution on $\mathcal{X}$ and we want testing null hypothesis $H_0 : \ P = Q$. Denote by $\mathbf{Y} = \{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}\}$ a random sample from the distribution Q. Also denote by $\tilde{C}_P(0.5, n)$ and $\tilde{C}_Q(0.5, m)$ the minimal volume estimators of sets $C_P(0.5)$ and $C_Q(0.5)$. Put in the following notations. Denote by

- $S_1 = \#\{\mathbf{x_i} \in \mathbf{X} | \mathbf{x_i} \in \tilde{C}_Q(0.5, m\}$ — the number of elements of $\mathbf{X}$ that belong to the set $\tilde{C}_Q(0.5)$;
- $S_2 = \#\{\mathbf{y_i} \in \mathbf{Y} | \mathbf{y_i} \in \tilde{C}_P(0.5, n\}$ — the number of elements of $\mathbf{Y}$ that belong to the set $\tilde{C}_P(0.5)$.

Suppose the null hypothesis H_0 is true; then $\tilde{C}_Q(0.5, m) \approx \tilde{C}_P(0.5, n)$ (in the case of large n, m). Therefore it is clear that the random variable $S = S_1 + S_2$ has a binomial distribution with the parameters $n + m$ and $p \approx 0.5$. So for testing the null hypothesis $H_0 : \ P = Q$ we can use a statistical test on equality binomial proportion to value 0.5. The test statistic is [5]

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

where standard error

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

It is well known that for large samples the null distribution of the test statistic z is the standard normal, having the mean of 0 and standard deviation of 1. So, we define the rejection region (using an α significance level) as $z < t_\alpha$, where t_α is α–quantile of standard normal distribution. For samples that are too small one can use the binomial distribution directly for calculating p–values.

3 Experimental results

We conducted distribution comparisons using our test and some other tests on artificial data sets. For comparisons we used Hotelling test (the classical test for testing the mean difference [1] for two multivariate populations.), multivariate rank test [8] and kernel Cramer test [2]. For estimating $\tilde{C}_P(0.5, n)$ and $\tilde{C}_Q(0.5, m)$ package **kernlab** [6] was used, which is an extensible package for kernel-based machine learning methods in **R** [9]. In all our experiments we used the kernel of the Gaussian radial basis function. Also we used automatic selection of the kernel width in correspondence with [3]. Experiments were realized on the computational claster of Institute of Applied Mathematical Research of the Karelian Research Centre of the RAS. We have considered the case $\mathcal{X} = R^d$ and distributions P and Q are Gaussian.

In our first experiments we investigated the rate of convergence the $P\{\tilde{C}_P(0.5, n)\}$ to $P\{C_P(0.5, n)\}$ as $n \to \infty$ for various values of d. From these experiments it is follows that values $n \approx m$ and d such that $n/d > 50$ ensure that proposed test has significance level $\alpha \leqslant 0.05$.

Secondly, samples were drawn from distributions $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(\mathbf{m}, \mathbf{I})$ with various $\mathbf{m}$ (here $\mathbf{I}$ is identical diagonal matrix). The experimental rezults show that all tests under considerations is high superior our test in that case.

Thirdly, samples were drawn from distributions $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$ with various σ. The rezults show that only Cramer test is somewhat superior our test in that case.

Also we realize group of experiments in the case dissimilarity between two distributions lies in distributions forms. For example we have investigated the following case. Let $d = 2r$ be even. Consider samples that were drawn from distributions $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_1})$ and $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_2})$, where $\mathbf{\Sigma_1}$ is diagonal matrix with elements $\sigma_{11} = \sigma, \ldots, \sigma_{rr} = \sigma, \sigma_{r+1r+1} = 1/\sigma, \ldots, \sigma_{dd} = 1/\sigma$ and $\mathbf{\Sigma_2}$ is diagonal matrix with elements $\sigma_{11} = 1/\sigma, \ldots, \sigma_{rr} = 1/\sigma, \sigma_{r+1r+1} = \sigma, \ldots, \sigma_{dd} = \sigma$. Our rezults show that our test is more powerful than kernel Cramer test in this case.

4 Conclusion

On author's opinion the proposed test may be used in statistical practice in cases when dissimilarity between two distributions lies in distributions forms.

References

[1] Anderson T.W. (2003). *An introduction to multivariate analysis.* Wiley, New Jersey.

[2] Baringhaus L., Franz C. (2004). On a new multivariate two–sample test. *Journal of Multivariate Analysis.* Vol. **88**, pp. 190-206.

[3] Caputo B., Sim K., Furesjo F., Smola A. (2002). Appearance–based object recognition using SVMs: which kernel should I use? *Proc of NIPS workshop on Statitsical*

methods for computational experiments in visual processing and computer vision. Whistler.

[4] Einmal J. H. J., Mason D. M. (1992). Generalized quantile processes. *Annals of Statistics* Vol. **20(2)**, pp. 1062–1078.

[5] Johnson N.L., Leone F.C. (1977). *Statistics and experimental design in engineering and the physical sciences*. John Wiley.

[6] Karatzoglou A., Smola A., Hornik K.,Zeileis A. (2004). kernlab–An S4 Package for Kernel Methods in R. *Journal of Statistical Software*. Vol. **11(9)**, pp 1–20. URL http://www.jstatsoft.org/v11/i09/.

[7] Polonik W. (1999). Concentration and goodness-of-fit in higher dimensions: (Asymptotically) distribution-free methods. *The Annals of Statistics*. Vol. **27**, pp. 1210-1229.

[8] Puri M.L., Sen P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

[9] R Development Core Team (2009). 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.

[10] Schölkopf B., Platt J., Shawe-Taylor J., Smola A. J., Williamson R. C. (1999). Estimating the support of a high-dimensional distribution. *TR MSR 99–87*. Microsoft Research, Redmond.