Uniformity Testing Using Minimal Spanning Tree

Anil. K. Jain, Xiaowei Xu, Tin Kam Ho*, and Fan Xiao

Dept. of Computer Science & Engineering
Michigan State University
East Lansing, MI 48823, USA
{jain, xuxiaowe, xiaofan}@cse.msu.edu

*Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974, USA
tkh@bell-labs.com

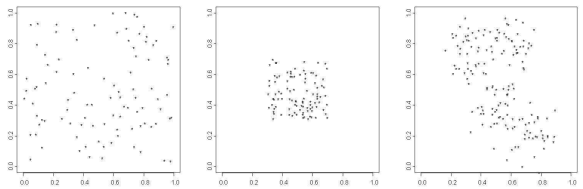
Abstract

Testing for uniformity of multivariate data is the initial step in exploratory pattern analysis. We propose a new uniformity testing method, which first computes the maximum (standardized) edge length in the MST of the given data. Large lengths indicate the existence of well-separated clusters or outliers in the data. For the data passing this edge inconsistency test, we generate two sub-samples of the data by a weighted re-sampling method, where the weights are computed based on the normalized edge lengths of MST of the entire data. The uniformity of the data is estimated by running the two-sample MST-test on these two sub-samples. Experiments with simulated and real data show the potential of the proposed test in identifying uniform or weakly clustered data. This test can also be used to rank various data sets based on their degree of uniformity.

1. Introduction

Although there exists a significant interest in exploratory pattern analysis to describe the structure of a data set, such as "how many clusters in the data?", the question of whether there exists any structure (clusters) in the data at all is still an open problem [1], [5], [6]. If the data is described by only two or three features, we can plot the patterns and use our innate perceptual organization to capture the structure of patterns. However, most pattern recognition problems deal with a large number of features simultaneously, so the approach used for the two- or three-dimensional data is generally not applicable to d-dimensional data. Presence of clusters in data can be generally identified by density variations of patterns. Hence, our meaning of lack of structure in the data corresponds to a uniform distribution of data; departure from uniformity (Fig. 1(a)) indicates the existence of possible clusters.

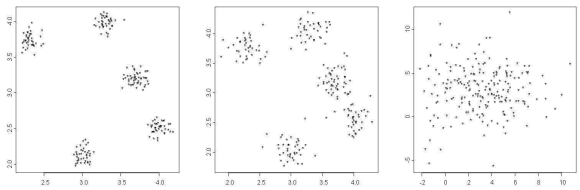
What is meant by the uniformity of data? Smith and Jain [7] argue that this notion involves specifying a "sampling window" which is defined as the "support" set of the underlying distribution of points. In the absence of prior knowledge, it is reasonable to make an assumption of connectedness of the sampling window. Consider the data shown in Fig. 1(b) that contains uniformly distributed points inside a small square enclosed by a unit square. If we assume the small square as the sampling window, the data is uniformly distributed. On the other hand, if the unit square is the sampling window, the data may not be considered as uniform. In practice, the support set is unknown and can be of arbitrary shape as in Fig. 1(c). To summarize, we say that the given set of d-dimensional data has no structure if the data is uniformly distributed on a connected set in $\Re^d$, called the sampling window. Hence, our goal is to test uniformity on the sampling window (H_0) against any structure due to density variations (H_1). This test can also be used to determine the "degree of clustering" present in the data such as in Fig. 2.



(a) (b) (c)

Figure 1. Notion of uniformity

Smith and Jain [7] proposed a method of testing uniformity in multidimensional data, which uses Friedman-Rafsky's MST test [2], [3], abbreviated as the two-sample MST-test below. The two-sample MST-test can be summarized as follows. Suppose we have two samples of size m and n, respectively, from distributions F_x and F_y, both defined on $\Re^d$. The null hypothesis H_0 is $F_x = F_y$ and the alternative hypothesis H_1 is $F_x \neq F_y$. We generate a weighted graph whose nodes represent the data points of the



(a) (b) (c)

Figure 2. Degree of clustering. Data from a Neyman-Scott process with s.d. (σ): (a) 0.08, (b) 0.16, (c) 2.56.

pooled samples, where the edge weights are the Euclidean distance between the points. We remove all the edges in the MST for which the incident nodes originate from different samples and define the test statistic, R, as the number of disjoint subtrees. We reject H_0 for a small number of subtrees. If both the samples are clustered around different cluster centers, this would produce a large number of MST edges from the same sample, resulting in a small value of the statistic R.

Let $N = m + n$, C be the number of edge pairs of MST that share a common node, and d_i be the degree of the i^{th} node. Then $C = \frac{1}{2}\sum_{i=1}^{N} d_i(d_i - 1)$. Under H_0,

$$E[R] = \frac{2mn}{N} + 1, \text{ and}$$

$$Var[R|C] = \frac{2mn}{N(N-1)} \times \left\{ \frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)} \times [N(N-1) - 4mn + 2] \right\}.$$

Under the asymptotic normality of R under H_0, we reject the null hypothesis H_0 at the significance level α if and only if

$$\frac{R - E[R]}{\sqrt{var[R|C]}} < \Phi^{-1}(\alpha),$$

where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution. For simplicity, it is assumed that $m = n$.

The above test is a two-sample test, but we only have one sample (n points in d-dimensional space) available. So, we need to generate a second sample. Smith and Jain [7] approximate the sampling window by a convex hull and generate the second sample uniformly over this convex hull. Their experiments show that the power of this test is good. The main problems with the convex hull assumption of the sampling window are: (i) large computational requirement, and (ii) uniform data inside non-convex sampling window may be identified as non-uniform as in the case of Fig. 1(c). We propose a new method for testing the uniformity where the two samples needed in the "two-sample" test are generated from the given data using a re-sampling technique.

2. Proposed Uniformity testing

Given n patterns $\{x_1, x_2, \ldots, x_n\}$ in a d-dimensional Euclidean space, construct its MST, $T = (\{x_i\}_{i=1}^{n}, \{e_j\}_{j=1}^{n-1}, \{l_j\}_{j=1}^{n-1})$, where l_j is the weight of edge e_j. We normalize l_j as $\tilde{l_j} = \frac{l_j - \mu}{\sigma}$, where μ is the mean of $\{l_j\}$ and σ is the standard deviation of $\{l_j\}$. The proposed uniformity test is composed of following two phases: (i) test if the data is well separated into two or more subsets; (ii) if the data is not well separated (is distributed on a connected set), we test its uniformity by a re-sampling technique-based MST-test.

Suppose the data contains two clusters that are well separated. The MST of this data will contain an "inconsistent" edge [4], [8] whose length is significantly larger than the average length of the nearby edges. We use the maximum of the normalized edge lengths of MST to define an inconsistent edge. Let $\rho = \max_{1 \le i \le n} \tilde{l_j}$ be the normalized length of the longest edge. We have observed that the distribution of ρ does not change much with respect to the underlying uniform distribution. Based on this observation, we choose a threshold $\rho_0 = 6$ and reject the claim that the given data is uniformly distributed on a connected set if $\rho > \rho_0$. Once the data has passed the edge inconsistency test, we subject it to a re-sampling-based two-sample test as described below.

Regions of points with low average edge length in MST represent regions of high density and vice versa. Based on this observation, we consider points with small average edge length as defining one sample and the points with large average edge length as the second sample. Now, if the data are not uniform and we run the two-sample MST-test on the pooled sample, it should produce a large number of joins from the same sample, thus reducing the value of the statistic R defined earlier, which leads to the rejection of the null hypothesis. The re-sampling method discussed below provides a possible solution to identify points with low (high) average edge length.

Let $E_i = \{j | e_j \in T \text{ is incident to } x_i, 1 \le j \le n-1\}$. Compute two weights for x_i as $\bar{\omega}_i = \frac{1}{d_i}\sum_{j \in E_i} e^{\tilde{l_j}}$ and $\underline{\omega}_i = \frac{1}{d_i}\sum_{j \in E_i} e^{-\tilde{l_j}}$, where d_i is the degree of x_i in T and $e^{(.)}$ is the exponential function. Notice that $\bar{\omega}_i(\underline{\omega}_i)$ is positively (negatively) correlated to the edge length for edges incident to point x_i. Let $\bar{S} = \sum_{i=1}^{n} \bar{\omega}_i$ and $\underline{S} = \sum_{i=1}^{n} \underline{\omega}_i$. We obtain two probability distributions on $\{x_i\}_{i=1}^{n}$ as $\bar{p}_i = \bar{p}(x_i) = \bar{\omega}_i / \bar{S}$ and $\underline{p}_i = \underline{p}(x_i) = \underline{\omega}_i / \underline{S}$. Now using the probabilities $\{\underline{p}_i\}_{i=1}^{n}$ and $\{\bar{p}_i\}_{i=1}^{n}$, respectively, we resample the given patterns $\{x_i\}_{i=1}^{n}$ without replacement and obtain two sub-samples $\{\bar{x}_i\}_{i=1}^{n_r}$ and $\{\underline{x}_i\}_{i=1}^{n_r}$, where n_r is the re-sampling size. Hence, $\{\bar{x}_i\}_{i=1}^{n_r}(\{\underline{x}_i\}_{i=1}^{n_r})$ represents the points in the original data set with high (low) average edge length in the low (high) density regions. Finally, we run MST-test on the two samples $\{\bar{x}_i\}_{i=1}^{n_r}$ and $\{\underline{x}_i\}_{i=1}^{n_r}$ and

compute the statistic R and reject the null hypothesis (the uniformity of patterns) for small values of R.

3. Experimental Results

Let the number of simulation trials $n_s = 1,000$, the number of patterns per dimension $n_d = 100$ and the sample size $n = n_d \times d = 100d$ where d is the dimensionality. We fix the resampling size as $n_r = 50$. Let r_0 be the percentage of rejections of the null hypothesis H_0 using the edge inconsistency test. Let $r(\alpha)$ be the percentage of rejections of H_0 at the α level using the two-sample MST test. Then the total percentage of rejections of H_0 at the α level is $r = r_0 + r(\alpha)$. The entries in Tables 1 and 2 are $(r_0, r(0.01))$. First we consider the performance of the test on synthetically generated uniform and normal data. Table

Table 1. Uniformity test on uniform and normal data

	2-dim	5-dim	10-dim
$U([0,1]^d)$	(0.1%, 1.5%)	(0%, 5.2%)	(0%, 4.1%)
N(**0**,**I**)	(45.5%, 12.6%)	(9.7%, 36.4%)	(2.5%, 14.4%)
N(**0**,**3I**)	(43.7%, 13.6%)	(10.6%, 38.9%)	(1.4%, 15.5%)

1 shows that our uniformity test works well. The rejection percentage of normal data due to edge inconsistency reduces dramatically with dimensionality, which is also observed in Table 2. One possible reason could be the sparsity of normal data in a high dimensional space.

Next, we consider the following three mixtures of normal data.

$$MN_1 = 0.5N(\mathbf{0}, \mathbf{I}) + 0.5N(\frac{1}{\sqrt{d}}\mathbf{1}, \mathbf{I}),$$

$$MN_2 = 0.5N(\mathbf{0}, \mathbf{I}) + 0.5N(\frac{4}{\sqrt{d}}\mathbf{1}, \mathbf{I}),$$

$$MN_3 = 0.5N(\mathbf{0}, \mathbf{I}) + 0.5N(\frac{8}{\sqrt{d}}\mathbf{1}, \mathbf{I}),$$

where $\mathbf{1}$ is the unit vector and $\mathbf{I}$ is the unit covariance matrix. The cluster centers in these three mixtures are separated by a distance of 1, 4 and 8, respectively. Table 2 shows that rejection by edge inconsistency decreases with dimensionality significantly. Once the data pass the edge inconsistency test, rejection by the resampling method with $d = 10$ is significantly lower than rejection with $d = 5$. This is an indication that sparsity of data in high dimensional spaces could reduce the power of our test significantly for nonuniform data, and it could also change the rejection error of uniform data slightly.

Table 2. Uniformity test for mixture of normal data

	2-dim	5-dim	10-dim
MN_1	(43.3%, 14.7%)	(10.2%, 37.8%)	(2.6%, 14.5%)
MN_2	(28.1%, 10.4%)	(6.8%, 28.1%)	(2.1%, 10.8%)
MN_3	(99%, 0.2%)	(59.4%, 10.8%)	(1.5%, 8.3%)

We apply our test on two real datasets: BUPA liver disorder data set[1] and Fisher's iris data. BUPA data contains 345 patterns and six features. Since the features in BUPA have a large difference of scales, we normalize each feature so that it has mean zero and standard deviation one. A projection of the data on the first two principal components shows that the data are not uniformly distributed and there is a dense core with a sparse and long tail. The maximum edge length $\rho = 5.33$, so the data set passes the edge inconsistency test. Two subsamples with size $n_r = 50$ are drawn representing the dense region and the sparse region. The test statistic $R = -3.636$ and the P-value is 0.000138. Hence, the null hypothesis is rejected and we claim that the BUPA data is not uniformly distributed.

The Fisher's iris data consists of 150 patterns in 4 dimensions. Among the three classes, Setosa, Virsicolor and Virginica, the latter two are well separated from the first one. The maximum edge length is 7.8. Hence, the null hypothesis is rejected and we claim that the Iris data set is not uniform. Next, we consider the Iris23 data, which consists of patterns belonging to Versicolor and Virginica. Now, the maximum edge length is 3.7. Hence, Iris23 passes the edge inconsistency and we now subject it to the two-sample MST test. Since the test statistic depends on the specific subsamples randomly drawn (this seldom happens when n is large and the P-value of the MST-test is significantly large or small), we compute 20 values of the test statistic, R. The average P-value of 20 test statistics is 0.29. Hence, there is no significant evidence to reject the null hypothesis although the P-value is quite low.

To investigate how well the R statistic can be used as a descriptive measure of the degree of uniformity (or clustering tendency) of a data set, we designed three experiments each using a series of data sets featuring a controlled, continuous degradation of a uniform distribution. The continuous degradation of uniformity is introduced by mixing samples from a uniform distribution (D1) with samples from a nonuniform distribution (D2) with varying proportions. In Experiment 1, D1 is taken to be the uniform distribution in the unit square $U([0,1] \times [0,1])$ and D2 is a normal distribution $N(0.5\mathbf{1}, 0.1\mathbf{I})$. In Experiment 2, D1 is again $U([0,1] \times [0,1])$ and D2 is a second uniform distribution clipped to a smaller region $U([0,0.5] \times [0,0.5])$. Both these

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html.

two experiments use two-dimensional data. Experiment 3 is done in several dimensionalities $d = 2, 3, 4, 5$ and in each case D1 and D2 are clipped to a concave sampling window that is the union of hyper-rectangles $R_i (i = 1, 2, \ldots, d)$, where R_i has the i^{th} coordinate restricted to $[0.45, 0.55]$.

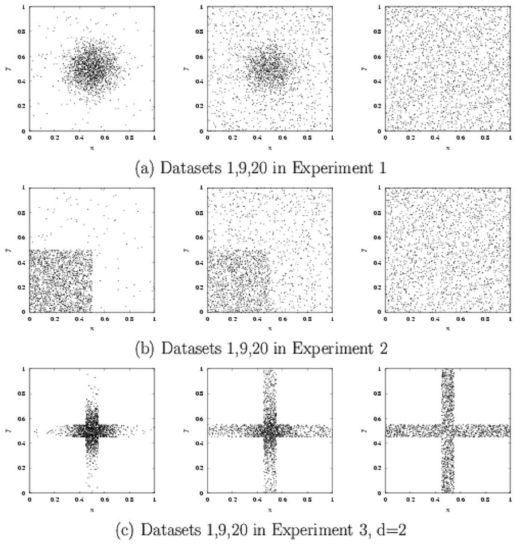
In each experiment, 20 data sets are created with varying numbers of samples from D1 and D2 as shown in Figure 3. Thus, in each case, first we start with a non-uniform data in set 1 and evolve into a uniform distribution in set 20. For each data set, the R statistic is calculated using both the re-sampling method and the convex hull method. The re-sampling method uses two subsamples that are $1/16^{th}$ of the size of the data set ($125/2000$ points). Because of random variations involved in the re-sampling procedure and the realization of random samples in the convex hulls, we apply both methods for 10 independent passes on each data set and take the mean value of the statistic to be the descriptive measure.

In each of the two-dimensional experiments we observe a nearly monotonic decrease in the maximum edge length as the data approach a globally uniform distribution, indicating that this is a good measure for assessing the degree of uniformity. However, with higher dimensional data, monotonicity of the trend is maintained only for highly nonuniform data. The R statistic given by the re-sampling method also increases nearly monotonically. Using a significance level $\alpha = 0.01$ (critical $z = -2.33$), the null hypothesis is rejected for data sets $1 - 15$. For data sets 16 and 17, the 10-pass mean of the statistic is very close to the critical value. For data sets $18 - 20$ the null hypothesis is accepted in most passes. This reconfirms that the test is generally useful. Interestingly, stronger discrimination between uniform and non-uniform data sets is achieved with the higher dimensional data in Experiment 3.

The convex hull statistic also increases monotonically as the data sets become uniform. However, with concave sampling windows (Experiment 3), the statistic is not effective for testing the uniformity (null) hypothesis. On the other hand, the resampling method performs consistently well in these cases, regardless of the sampling window geometry and data dimensionality. This demonstrates an advantage of the resampling statistic as a descriptor of uniformity over both the maximum edge length and the convex hull statistics.

4. Conclusions

We have developed a test based on the minimal spanning tree to determine whether a give set of multidimensional patterns is distributed uniformly on a connected sampling window. Simulation studies on the power and rejection error of the test are encouraging but some sensitivity to data dimensionality is observed. Finally, we found that the proposed statistic correlates well with continuous degradations of uniformity. As a measure of degree of uniformity, it compares favourably with a previously proposed statistic based on convex hull sampling in terms of robustness to variations in the sampling window geometry and data dimensionality.



(a) Datasets 1,9,20 in Experiment 1

(b) Datasets 1,9,20 in Experiment 2

(c) Datasets 1,9,20 in Experiment 3, d=2

Figure 3. Data used for degree of uniformity experiments.

References

[1] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, NY, 2001.

[2] J. Friedman and L. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist*, 7:697–717, 1979.

[3] J. Friedman and L. Rafsky. Graphics for the multivariate two-sample problem. *JASA*, 76:277–295, 1981.

[4] R. Hoffman and A. K. Jain. A test of randomness based on the minimal spanning tree. *Pattern Recognition Letters*, 1:175–180, 1983.

[5] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, NJ, 1988.

[6] B. D. Ripley. *Spatial Statistics*. Wiley-Interscience, New York, 1981.

[7] S. Smith and A. Jain. Testing for uniformity in mutidimensional data. *IEEE Trans. Pattern Anal. Machine Intell.*, 6:73–80, 1984.

[8] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comp*, C-20:68–86, 1971.