بازشناسی الگو

درس ۱۹

اعتبارسنجی خوشه‌ها

Cluster Validity

کاظم فولادی
دانشکده مهندسی برق و کامپیوتر
دانشگاه تهران

http://courses.fouladi.ir/pr

اعتبارسنجی خوشه‌ها

۱

مقدمه

CLUSTER VALIDITY

❖Clustering tendency

<u>Facts</u>

➢ Most clustering algorithms impose a clustering structure to the data set X at hand.

➢ However, X may not possess a clustering structure.

➢ Before we apply any clustering algorithm on X, it must first be verified that X possesses a clustering structure. This is known as clustering tendency.

➢ Clustering tendency is heavily based on hypothesis testing.

Specifically, it is based on testing the randomness (null) hypothesis (H_0) against the regularity (H_1) hypothesis and the clustering (H_2) hypothesis.

• Randomness hypothesis (H_0): "The vectors of X are randomly distributed, according to the uniform distribution in the sampling window of X (the compact convex support set for the underlying distribution of the vectors of the data set X)".

• Regularity hypothesis (H_1): "The vectors of X are regularly spaced (that is they are not too close to each other) in the sampling window".

• Clustering hypothesis (H_2): "The vectors of X form clusters".

اعتبارسنجی خوشه‌ها

۲

آزمون فرض (بازبینی‌شده)

- $p(q|H_0)$, $p(q|H_1)$ and $p(q|H_2)$ are estimated via Monte Carlo simulations

Some tests for spatial randomness,
when the input space dimensionality greater than or equal to 2 are:

- Tests based on structural graphs
    - *Test that utilizes the idea of the minimum spanning tree (MST)*
- Tests based on nearest neighbor distances
    - *The Hopkins test*
    - *The Cox-Lewis test*
- A method based on sparse decomposition.

➢ Important notes:
- Clustering algorithms should be applied on *X*, only if the **randomness** and the **regularity** hypotheses are <u>rejected</u>. Otherwise, methods different than clustering must be used to describe the structure of *X*.
- Most studies in clustering tendency focus on the detection of **compact clusters.**

➢ The basic steps of the clustering tendency philosophy are:
- Definition of a test statistic *q* suitable for the detection of clustering tendency.
- Estimation of the pdf of *q* under the null (H_0) hypothesis, $p(q|H_0)$.
- Estimation of $p(q|H_1)$ and $p(q|H_2)$ (they are necessary for measuring the power of *q* (the probability of making a correct decision when H_0 is rejected) against the regularity and the clustering tendency hypotheses).
- Evaluation of *q* for the data set at hand, *X,* and examination whether it lies in the critical interval of $p(q|H_0)$, which corresponds to a predetermined significance level ρ.
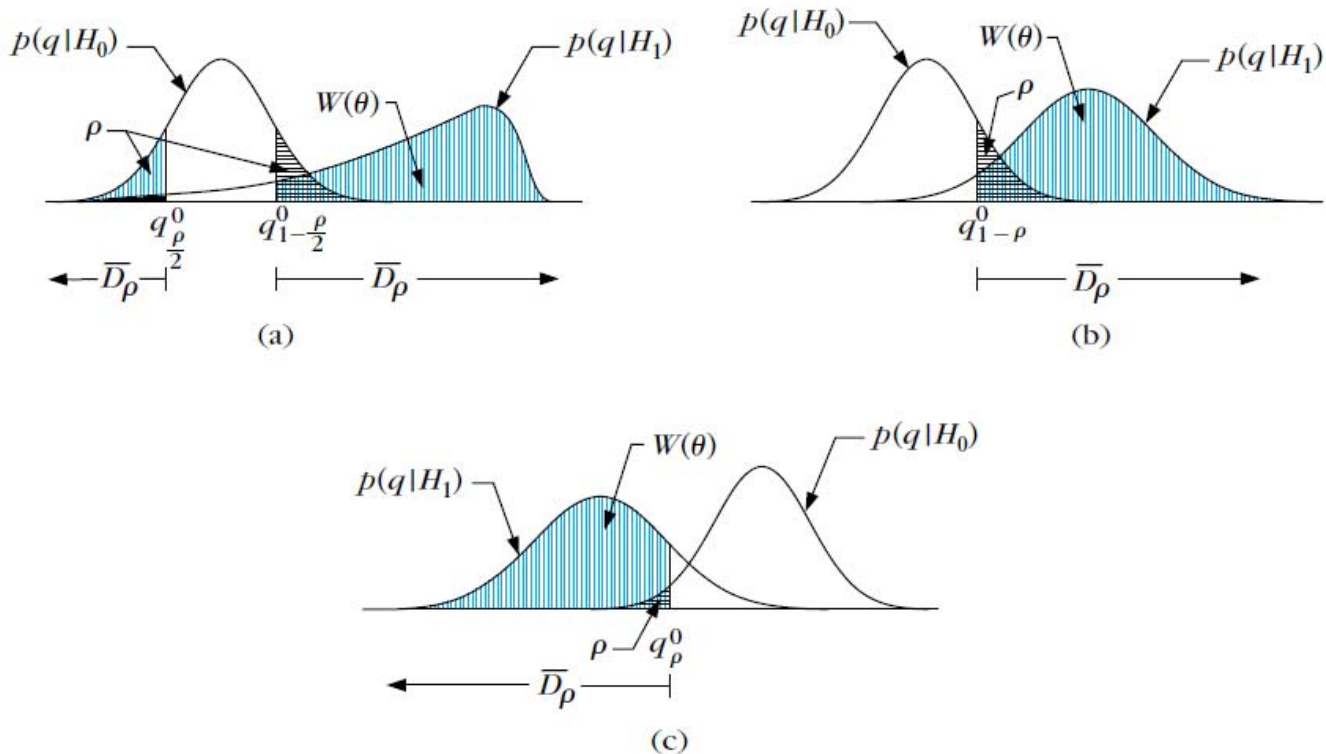
FIGURE 16.1

Critical regions of (a) A two-tailed test, (b) A right-tailed test, and (c) A left-tailed test. q_a^0 is the a percentile of q under H_0.[1]

اعتبارسنجی خوشه‌ها

۳

اعتبارسنجی خوشه‌ها

❖ Cluster validity

➢ In the sequel it is assumed that the clustering tendency procedure indicated the existence of a clustering structure in X.

➢ Applying a clustering algorithm on X, with inappropriate values of the involved parameters, poor results may be obtained. Hence the need for further evaluation of clustering results is apparent.

➢ Cluster validity:
  a task that evaluates quantitatively the results of a clustering algorithm.

➢ A clustering structure C, resulting from an algorithm may be either
  • A hierarchy of clusterings or
  • A single clustering.

➢ Cluster validity may be approached in three possible directions:

- C is evaluated in terms of an independently drawn structure, imposed on X *a priori*.
  * The criteria used in this case are called external criteria.

- C is evaluated in terms of quantities that involve the vectors of X themselves (e.g., proximity matrix).
  * The criteria used in this case are called internal criteria.

- C is evaluated by comparing it with other clustering structures, resulting from the application of the same clustering algorithm but with different parameter values, or other clustering algorithms, on X.
  * Criteria of this kind are called relative criteria.

➢ _Cluster validity for the cases of **external** and **internal** criteria_

- **Hypothesis testing is employed**.
- The null hypothesis H_0, which is a statement of randomness concerning the structure of X, is defined.
- The generation of a reference data population under the random hypothesis takes place.
- An appropriate statistic, q, whose values are indicative of the structure of a data set, is defined. The value of q that results from our data set X is compared against the values obtained for q when the elements of the reference (random) population are considered.

Ways for generating reference populations under the null hypothesis (each one used in different situations):
- Random position hypothesis.
- Random graph hypothesis.
- Random label hypothesis.

➢ Statistics suitable for <u>external criteria</u>
- For the comparison of *C* with an independently drawn partition *P* of *X*
  - Rand statistic
  - Jaccard statistic
  - Fowlkes-Mallows index
  - Hubert's Γ statistic
  - Normalized Γ statistic
- For assessing the agreement between *P* and the proximity matrix *P*.
  - Γ statistic.

➢ Statistics suitable for <u>internal criteria</u>
- Validation of hierarchy of clusterings
  - Cophenetic correlation coefficient (*CPCC*)
  - γ statistic
  - Kudall's τ statistic.
- Validation of individual clusterings
  - Γ statistic
  - Normalized Γ statistic

اعتبارسنجی خوشه‌ها

۴

ضابطه‌های نسبی

➤ _Cluster validity for the cases of **relative criteria**_

Let **A** denote the set of parameters of a clustering algorithm.

Statement of the problem

- _"Among the clusterings produced by a specific clustering algorithm, for different values of the parameters in **A**, choose the one that best fits the data set X"._

We consider two cases

(a) **A** does not contain the number of clusters m.

The estimation of the best set of parameter values is carried out as follows:

- Run the algorithm for a wide range of values of its parameters.
- Plot the number of clusters, m, versus the parameters of **A.**
- Choose the widest range for which m remains constant.
- Adopt the clustering that corresponds to the values of the parameters in **A** that lie in the **middle** of this range.

(b) **A** *does contain the number of clusters m.*

The estimation of the best set of parameter values is carried out as follows:

- Select a suitable performance index q
  (the best clustering is identified in terms of q).

- For $m = m_{min}$ to m_{max}
    - Run the algorithm r times using different sets of values for the other parameters of **A** and each time compute q.
    - Choose the clustering that corresponds to the best q.
- End for

- Plot the best values of q for each m versus m.
- The presence of a significant knee indicates the number of clusters underlying X. Adopt the clustering that corresponds to that knee.
- The absence of such a knee indicates that X possesses no clustering structure.

➤ *Statistics suitable for **relative criteria***

- **Hard** clustering
  - Modified Hubert Γ statistic
  - Dunn and Dunn-like indices
  - Davies-Bouldin (DB) and DB-like indices

- **Fuzzy** clustering
  - Indices for clusters with point representatives
    - o Partition coefficient (PC)
    - o Partition entropy coefficient (PE)
    - o Xie-Beni (XB) index
    - o Fukuyama-Sugeno index
    - o Total fuzzy hypervolume
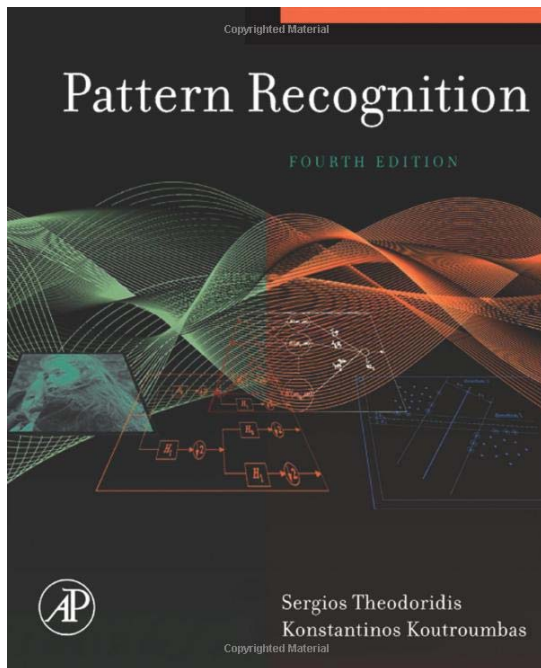   - o Average partition density
    - o Partition density

➤ *Statistics suitable for **relative criteria** (cont.)*

- Fuzzy clustering (cont.)
    - Indices for shell-shaped clusters
        - o Fuzzy shell density
        - o Average partition shell density
        - o Shell partition density
        - o Total fuzzy average shell thickness

اعتبارسنجی خوشه‌ها

۵

منابع

منبع اصلی

S. Theodoridis, K. Koutroumbas,
Pattern Recognition,
Fourth Edition, Academic Press, 2009.

Chapter 16

CHAPTER

Cluster Validity

16

16.1 INTRODUCTION

A common characteristic of the majority of the clustering algorithms, discussed in the previous chapters, is that they *impose* a clustering structure on the data set X, even though X may not possess such a structure. In the latter case, the results produced after the application of a clustering algorithm on X are not indicative of the structure of X. In other words, *cluster analysis is not a panacea*. That is, we must have an indication that the vectors of X form clusters before we apply a clustering algorithm. The problem of verifying whether X possesses a clustering structure, without identifying it explicitly, is known as *clustering tendency* and is discussed at the end of the chapter.

Let us now assume that X possesses a clustering structure and we want to unravel it. A different kind of problem is encountered now. Recall that all the clustering algorithms require knowledge of the values of specific parameters and, in addition, some of them impose restrictions on the shape of the clusters (e.g., compact, hyperellipsoidal). As already shown in the previous chapters, poor estimation of these parameters and inappropriate restrictions on the shape of the clusters (wherever such restrictions are required) may lead to incorrect conclusions about the clustering structure of X. Thus, the need for further evaluation of the results of a clustering algorithm is apparent.

In this chapter, we discuss methods suitable for quantitative evaluation of the results of a clustering algorithm. This task is known under the general term *cluster validity*. However, it must be emphasized that the results obtained by these methods are *only* tools at the disposal of the expert in order to evaluate the resulting clustering.

Let C denote the clustering structure resulting from the application of a clustering algorithm on X. This may be a hierarchy of clusterings, as is the case with the hierarchical algorithms, or a single clustering, as happens with all the other algorithms discussed in the previous chapters. Cluster validity can be approached in three possible directions. First, we may evaluate C in terms of an independently

863