Uncertainty in clustering

Miroslav Sabo, Oľga Nánásiová

Slovak University of Technology in Bratislava

FSTA 2012, Liptovský Ján

Outline of presentation

- 1. Similarity coefficients from probabilistic point of view
- 2. Uncertainties in clustering
- 3. Simultaneous clustering using similarity coefficients

What are similarity coefficients ?

- Measures designed to assess degree of similarity between some objects
- Objects can be humans, countries, clusters, clustering methods – anything where one can compare entities w.r.t. similarity

Example: Jaccard coefficient (1912)

- Assume 2 objects x and y described by vectors x=(1,0,0,1,1), y=(1,0,0,0,1), where ones may denote that object has some property and zeros mean opposite
- Then, Jaccard similarity between x and y is defined as

$$S_{Jaccard}(x,y) = \frac{a}{a+b+c}$$

where

$$a = \left| j \in \{1,2,\ldots,5\} : x_j = y_j = 1 \right|$$

$$b = \left| j \in \{1,2,\ldots,5\} : x_j = 1,\ y_j = 0 \right|$$

$$c = \left| j \in \{1,2,\ldots,5\} : x_j = 0,\ y_j = 1 \right|$$

Problems with similarity coefficients

- Since 19th century, there were proposed too many similarity coefficients
- Therefore, there is some need to categorize them
- Moreover, do they measure the same or not?
- In addition, can we extend them to compare clusters instead of objects?

Similarity coefficients from probabilistic point of view (Nánásiová et al., 2010)

- At first, assume rational function of the form

$$f(k,q,t,x)=\frac{k(1-x)}{t+qx}$$

where k>0, q≥0 are constants and 0≤x≤1

- For k=t, this function has good properties to be a similarity function, because

1) $f(k,q,t,1)=0 \quad f(k,q,t,0)=1$

2) $\forall x_1 < x_2 \quad f(k,q,t,x_1) > f(k,q,t,x_2)$

Two general forms of similarity coefficients

- Consider probability space (Ω, S, P) and functions $s, g : S^2 -> R$ defined as

$$s(A,B) = \frac{P(A \cap B)}{P(A \cap B) + \alpha P(A \cap B^C) + \beta P(B \cap A^C)}$$

$$g(A,B) = \frac{P(A \cap B) + \gamma P(A^C \cap B^C)}{P(A \cap B) + \alpha P(A \cap B^C) + \beta P(B \cap A^C) + \delta P(A^C \cap B^C)}$$

- For $\alpha = \beta$ (equal weights) and using trivial identity

$$P(A\Delta B) = P(A \cap B^C) + P(A^C \cap B) = P(A \cup B) - P(A \cap B)$$

we can rewrite first function as

$$s(A,B) = \frac{P(A \cup B) - P(A\Delta B)}{P(A \cup B) + (\alpha - 1)P(A\Delta B)}$$

In addition, for $P(A \cup B) > 0$ and using another trivial identity

$$\frac{P(A\Delta B)}{P(A \cup B)} = P(A\Delta B \mid A \cup B)$$

we obtain $s(A,B) = \dfrac{1 - P(A\Delta B \mid A \cup B)}{1 + (\alpha - 1)P(A\Delta B \mid A \cup B)}$

- After replacing $P(A\Delta B \mid A \cup B)$ by x, one obtain new form of function s(A,B)

$$s(A,B) = \frac{1-x}{1+(\alpha-1)x} = f(1,\alpha-1,1,x)$$

- As can be seen, first general form of similarity coefficients is rewritten using rational function as its special form

Summary

- If we replace x by $P(A\Delta B \mid A \cup B)$ in the function

$$f(k,q,t,x) = \frac{k(1-x)}{t+qx}$$

then for some values of constants k, q, t we obtain G1 family of similarity coefficients. Examples of two members of this family are Jaccard and Dice coefficient

$$S_{Jaccard} = f(1,0,1,x) \qquad S_{Dice} = f(2,-1,2,x)$$

- If we replace x by $P(A\Delta B)$ in the function

$$f(k,q,t,x) = \frac{k(1-x)}{t+qx}$$

then for some values of constants k, q, t we obtain G2 family of similarity coefficients. Examples of two members of this family are Rogers-Tanimoto and Sokal-Sneath coefficient

$$S_{Rogers-Tanimoto} = f(1,1,1,x) \qquad S_{Sokal-Sneath-1} = f(2,-1,2,x)$$

Equivalence between coefficients from the same group

- Theorem: Coefficients from the same group preserve ordering

- What is preserved ordering? Assume 2 pairs of objects x, y and u, v. If x, y are more similar than u, v according to first coefficient, then the same finding must result from any other coefficient from the same group

What is clustering?

- Cluster analysis belongs to multivariate statistical methods
- Since 1950’s there have been proposed many different algorithms
- Aim is to assign n p-dimensional objects into k groups (called clusters) where k<<n

Uncertainties in clustering

- 1. In most situation, number of clusters is not known in advance
- 2. Clusters may have nonlinear structure (not only globular shape)
- 3. Clusters may overlap and may be noised
- 4. Which is the best algorithm for clustering our data?
- 5. How to assess accuracy of results obtained?

- Result: Clustering deals with the most difficult task in statistics (compare it with classification, where number of clusters and also true assignments of some objects are known)

Fuzziness in clustering

- In addition to previous problems, some objects may belong to more than one cluster (imagine overlapping clusters)

- Bezdek (1981) proposed fuzzy clustering approach to deal with this situation

Simultaneous clustering using similarity coefficients (Sabo and Nánásiová, 2012)

- In 2012, we proposed clustering by two methods simultanously
- Why? To produce more accurate results than results obtained by clustering with only one algorithm
- Our algorithm uses results from two clustering algorithms and similarity coefficients to produce third partition of objects (and more accurate)
- Pseudocode of our algorithm can be described as follows

Simultaneous clustering

- Input: Original dataset X

  k=1

  Repeat steps 1-5 until some stopping criterion is satisfied
- 1. Use two clustering methods to obtain two partitions of X
- 2. Find two most similar clusters according to any similarity coefficient
- 3. Find their intersection Y(k) (k-th cluster)
- 4. Create new dataset X=X\Y(k)
- 5. k=k+1

- 6. Assign all unclassified objects to the nearest cluster

- Output: Partition of X

Uncertainty and simultaneous clustering

- Clustering by only one algorithm may produce very inaccurate results. When we use another algorithm, partition may be very different. Applying simultaneous clustering on these two partitions may decrease uncertainty and fuzziness and produce more stable solution.

References

- BEZDEK, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms.
- JACCARD, P. (1912). The distribution of the flora in the Alpine zone. The New Phytologist, 11, 37-50.
- NÁNÁSIOVÁ, O., PASTUCHOVÁ, E., VÁCLAVÍKOVÁ, Š. One Expression of Associative Coefficients. Forum Statisticum Slovacum Roč, 6, č. 5. 176—179.
- SABO, M., NÁNÁSIOVÁ, O. (2012): Clustering by two methods simultaneously. Forum Statisticum Slovacum Roč, 7, č. 1. zv. 7, s. 86-92.

Thanks for attention