Análise de agrupamentos consensuais

XX JOCLAD 2013

Lúcia Sousa - ESTGV e Fernanda Sousa - FEUP e CITTA

13 de Março de 2013

- Consenso de agrupamentos
- Técnicas de obtenção de consenso de agrupamentos
- Metodologia de simulação
- Discussão dos resultados e comentários finais

Na análise de dados, os métodos de agrupamento constituem uma ferramenta poderosa que visam, muitas vezes sem nenhuma informação à priori da estrutura dos dados, identificar grupos naturais.

Na análise de agrupamento, para um dado conjunto de elementos, pretende-se identificar grupos tais que:

- Elementos de um mesmo grupo apresentem elevadas semelhanças (grupos homogeneos);
- Elementos de diferentes grupos sejam bem distintos (grupos bem separados).

Existe um conjunto vasto de métodos de agrupamento que incluem diferentes escolhas de parâmetros ou medidas, sendo que a obtenção de um agrupamento depende do método usado e das escolhas feitas.

A utilização de diferentes métodos de agrupamento ou diferentes escolhas, para um mesmo conjunto de dados, produz frequentemente diferentes agrupamentos, colocando-se o problema da escolha de um deles ou da determinação de um consenso dos mesmos.

Consenso de agrupamentos

Um agrupamento que combina os diferentes agrupamentos individuais, refletindo as principais estruturas neles inerentes, numa perspetiva de obter um agrupamento:

- Mais estável;
- Mais robusto;
- Mais consistente.

Dados vários agrupamentos sobre o mesmo conjunto de elementos, obtidos por diferentes métodos, diferentes parâmetros ou inicializações, a teoria de consenso define nova(s) classificação(ões) que capta(m) os aspetos estruturais comuns aos vários agrupamentos.

Para atingir estes objetivos são estabelecidas funções consenso, que definem o modo de combinar os agrupamentos individuais.

Fred e Jain [1],[2],[3]

- O objetivo é encontrar consensos de agrupamentos consistentes e robustos.
- Os agrupamentos individuais são obtidos pelo algoritmo K-médias.
- A função consenso é baseada num mecanismo de voto e matriz de co-associação:

Sejam, E - conjunto a classificar e r agrupamentos sobre E,

Para cada par (i, j) de elementos de E,

1) Calcula $co - assoc(i, j) \rightarrow$ n° de vezes, nos r agrupamentos, que o par (i, j) pertence ao mesmo grupo.

2) Se $co - assoc(i, j)$ fôr elevado então o par (i, j) pertencerá ao mesmo grupo no consenso de agrupamentos.

Posteriormente,

Em [2] surgiu o conceito de acumulação por evidência, *Evidence Accumulation Clustering* - EAC, cuja ideia é, aplicar o algoritmo *single linkage* (ligação única) à matriz de co-associação.

Em [3] generalizam a aproximação EAC a outros critérios de agregação de classificação hierárquica, AL - *Average Linkage* (ligação média) e WL - *Ward criteria* (critério de Ward).

Strehl e Ghosh [4],[5]

- O objetivo é obter consensos de agrupamentos robustos.
- A função consenso é baseada no conceito de Informação Mútua e hipergráficos, associando-lhes um problema de otimização.
- A Informação Mútua mede a informação partilhada entre pares de agrupamentos.

$$IM(C_1, C_2) = H(C_1, C_2) - H(C_1) - H(C_2)$$

- Pretende-se encontrar um (novo) agrupamento, seja λ, que maximize a Informação Mútua média entre cada agrupamento e λ, isto é,

$$\max_{\lambda} \frac{\sum_{i=1}^{r} IM(C_i, \lambda)}{r},$$

λ percorre todos os agrupamentos possíveis sobre E.

- Problema de otimização que levanta enormes problemas computacionais.

Representação hipergráfica de agrupamentos

Exemplo: $E = \{e_1, ..., e_7\}$; $C_1, C_2, C_3, C_4 \rightarrow$ agrupamentos sobre E, $v_i \rightarrow$ vértices e $h_i \rightarrow$ arestas do hipergráfico.

	C_1	C_2	C_3	C_4
e_1	1	2	1	1
e_2	1	2	1	2
e_3	1	2	2	1
e_4	2	3	2	1
e_5	2	3	3	2
e_6	3	1	3	2
e_7	3	1	3	1

	H_1			H_2			H_3			H_4	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}
v_1	1	0	0	0	1	0	1	0	0	1	0
v_2	1	0	0	0	1	0	1	0	0	0	1
v_3	1	0	0	0	1	0	0	1	0	1	0
v_4	0	1	0	0	0	1	0	1	0	1	0
v_5	0	1	0	0	0	1	0	0	1	0	1
v_6	0	0	1	1	0	0	0	0	1	0	1
v_7	0	0	1	1	0	0	0	0	1	1	0

1) Obtenção do hipergrafo;
2) Redução do hipergrafo por recurso a algoritmos de partição de gráficos (METIS, HMETIS);
3) Aplicação de algoritmos de agrupamento ao hipergrafo reduzido (CSPA - *Cluster-based Similarity Partitioning Algorithm*, HGPA - *HyperGraph Partitioning Algorithm*, MCLA - *Meta-Clustering Algorithm*);
4) A maximização da Informação Mútua média é agora procurada nestes 3 agrupamentos reduzidos;
5) Obtido o consenso de agrupamento reduzido, são aplicados os algoritmos METIS ou HMETIS para recuperar o hipergrafo completo (com todos os vértices e arestas).

Desenvolvimento/adaptação de códigos em Matlab e R que incluem:
A obtenção de hierarquias, abordagem tradicional de classificação hierárquica, considerando: Medida de proximidade - Distância Euclidiana 3 métodos de agregação: sl, cl,al. A obtenção de partições, para os diferentes métodos de agregação usando, i) a abordagem SEP/COP,ii) a abordagem tradicional com o nº de classes da partição de referência. Comparação das partições

Validação externa e validação relativa, usando o índice ARI.

[1] A. Fred, 2001, "Finding consistent clusters in data partitions", in J. Kittler and F. Roli, editors, Multiple Classifier Systems, volume LNCS, Springer, pp. 309-318.
[2] A. Fred e A. K. Jain, 2002, "Data Clustering Using Evidence Accumulation", in proc. of 16th Int'l Conference on Pattern Recognition, pp. 276-280.
[3] A. Fred e A. K. Jain, 2005, "Combining Multiple Clusterings Using Evidence Accumulation", IEEE Trans Pattern Analysis and Machine Intelligence 27(6), pp. 835-850.
[4] A. Strehl e J. Ghosh, 2002, "Cluster Ensembles - A Knowledge Reuse Framework For Combining Multiple Partitions", Journal of Machine Learning Research 3, pp. 583-617.
[5] A. Strehl e J. Ghosh, 2002, "Cluster Ensembles - A Knowledge Reuse Framework For Combining Partitionings", in proc. Conference on Artificial Intelligence, Edmonton, pp. 93-98.
[6] L. Hubert e P. Arabie, 1985, "Comparing Partitions", in Journal of Classification 2, pp. 193-218.