χ -Sim: A new similarity measure for the Co-clustering task

Gilles Bisson[1,2], Fawad Hussain[2]

[1] Centre National de la Recherche Scientifique, France

[2] Laboratoire TIMC-IMAG, CNRS/UJF 5525

Université de Grenoble, France

Co-clustering problem

- Simultaneously cluster rows (instances) and columns (features)
  - Allows to deal with sparse data
  - Improves clustering result
- Old Idea (Hartigan [1], etc) but has drawn recent attention.
  - Text Mining: Document Classification
  - Bio-Informatics : DNA chip analysis
- Widely used techniques :
  - Matrix Decomposition (BVD) [2]
  - Information theoretic (ITCC) [3]
  - Graph based (RSN) [4]

Proposed method: χ-Sim

M	m^1	m^2	…	m^c
m_1	1	1		0
m_2	0	0		1
…				
m_r	0	3		0

- Classical similarity as a sum of the similarities between features

$$Sim(\mathbf{m_i},\mathbf{m_j}) = F_s(m_{i1},m_{j1}) + F_s(m_{i2},m_{j2}) + \ldots + F_s(m_{ic},m_{jc})$$

- Extension to all pairs

$$Sim(\mathbf{m_i},\mathbf{m_j}) = \boxed{F_s(m_{i1},m_{i1}).sc_{11}} + F_s(m_{i1},m_{i2}).sc_{12} + \ldots + F_s(m_{i1},m_{ic}).sc_{1c} + F_s(m_{i2},m_{j1}).sc_{21} + \boxed{F_s(m_{i2},m_{i2}).sc_{22}} + \ldots + F_s(m_{i2},m_{jc}).sc_{2c} + \ldots F_s(m_{ic},m_{j1}).sc_{c1} + F_s(m_{ic},m_{j2}).sc_{c2} + \ldots + \boxed{F_s(m_{ic},m_{jc}).sc_{cc}}$$

Normalization

- Normalize by the maximum possible value which is the product of length of document vector or term vector
  - Row similarity is given by

$$\forall i,j \in 1..r,\ sr_{i,j} = Sim(\mathbf{m_i},\mathbf{m_j}) / (|\mathbf{m_i}|\,|\mathbf{m_j}|)$$

  - Column similarity is given by

$$\forall i,j \in 1..c,\ sc_{i,j} = Sim(\mathbf{m^i},\mathbf{m^j}) / (|\mathbf{m^i}|\,|\mathbf{m^j}|)$$

where $|m_j|$ = length of document i
and $|m^i|$ =number of documents featuring word I

- Normalization insures that the simimarity value is in [0,1]

How to implement efficiently

- When functions F_s is defined as a product

$$F_s(m_{ij},m_{kl}) = m_{ij} \cdot m_{kl}$$

- At any given iteration t,

$$\mathbf{SR_t} = \mathbf{M^T}.\mathbf{SC_{(t-1)}}.\mathbf{M} \text{ x } \mathbf{NR} \quad \forall i,j \in [1,r],\ NR_{i,j} = 1 / (|\mathbf{m_i}|\,|\mathbf{m_j}|) \qquad \text{Eq (1)}$$

$$\mathbf{SC_t} = \mathbf{M}.\mathbf{SR_{(t-1)}}.\mathbf{M^T} \text{ x } \mathbf{NR} \quad \forall i,j \in [1,r],\ NC_{i,j} = 1 / (|\mathbf{m^i}|\,|\mathbf{m^j}|) \qquad \text{Eq (2)}$$

"x" here represents the Hadamard matrix product

- Complexity $O(N^3)$ like any other classical similarity measure (cosine, Euclidean, etc)

χ-Sim Algorithm

```
Function χ-Sim
Input  : data matrix M , no. of iterations N_iter
Output : two similarity matrices SR and SC expressing the
         co-similarity between rows and columns of M
Initialize SR and SC with the identity matrix
  for iter = 1 to N_iter
    Update-SR_iter(M, SC_iter-1)    // Equation 1
    Update-SC_iter(M, SR_iter)      // Equation 2
  end for
```

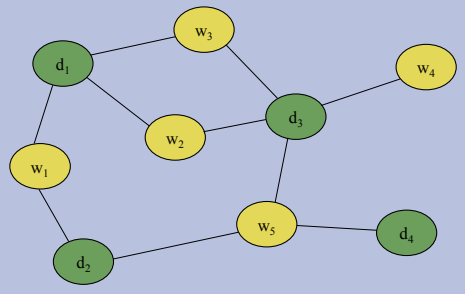
Experimental Results

- Corpus
  - Newsgroup (M2, M5, M10, NG1, NG2, NG3) and SMART (Classic3)
- Methodology
  - Generate matrices **SR** and **SC**
  - Generate clusters using Hierarchical clustering (ward's linkage)
  - Evaluate result using Precision value (Pr) and Normalized Mutual Information (NMI)

	Runtime		χ-Sim	Cosine	LSA	ITCC	BVD	RSN
M2	*12.01s*	Best Pr	**0.96**	0.72	0.94	0.92	0.95	-
		$Pr_{(avg)}$	**0.93 ± 0.01**	0.62 ± 0.04	0.78 ± 0.15	0.71 ± 0.15		
		$NMI_{(avg)}$	**0.65 ± 0.05**	0.15 ± 0.05	0.37 ± 0.23	0.21 ± 0.21		
M5	*10.93s*	Best Pr	**0.98**	0.76	0.95	0.56	0.93	-
		$Pr_{(avg)}$	**0.94 ± 0.04**	0.6 ± 0.08	0.82 ± 0.09	0.48 ± 0.06		
		$NMI_{(avg)}$	**0.85 ± 0.07**	0.55 ± 0.11	0.65 ± 0.14	0.27 ± 0.09		
M10	*13.53s*	Best Pr	**0.80**	0.55	0.67	0.33	0.67	
		$Pr_{(avg)}$	**0.75 ± 0.04**	0.45 ± 0.07	0.54 ± 0.09	0.28 ± 0.04		-
		$NMI_{(avg)}$	**0.65 ± 0.05**	0.43 ± 0.07	0.45 ± 0.08	0.16 ± 0.04		
Classic3	*164.38s*	Pr	**0.99**	0.89	0.98	0.97	-	-
		NMI	**0.96**	0.63	0.93	0.90		
NG1	*6.60s*	Best Pr	**1**	0.97	0.98	0.83		
		$Pr_{(avg)}$	**1 ± 0**	0.90 ± 0.11	0.95 ± 0.02	0.67 ± 0.12		
		$NMI_{(avg)}$	**1 ± 0**	0.62 ± 0.20	0.75 ± 0.08	0.13 ± 0.13	-	0.64 ± 0.16
NG2	*14.56s*	Best Pr	**0.94**	0.72	0.87	0.72		
		$Pr_{(avg)}$	**0.93 ± 0.01**	0.60 ± 0.08	0.82 ± 0.03	0.57 ± 0.08		
		$NMI_{(avg)}$	**0.81 ± 0.04**	0.51 ± 0.07	0.64 ± 0.03	0.35 ± 0.09	-	0.75 ± 0.07
NG3	*20.20s*	Best Pr	**0.91**	0.62	0.80	0.65		
		$Pr_{(avg)}$	**0.85 ± 0.06**	0.60 ± 0.04	0.72 ± 0.05	0.55 ± 0.06		
		$NMI_{(avg)}$	**0.77 ± 0.02**	0.55 ± 0.03	0.59 ± 0.03	0.48 ± 0.05	-	0.70 ± 0.04

Graphical Meaning of an iteration

- Matrix M can be seen as a bipartite graph
  - First Iteration: direct occurrences
  - Second Iteration: Second order occurrences
  - …
- 3-4 iterations are enough [5]



Conclusion

- New approach to use co-similarity for co-clustering
- Allows to use various similarity based clustering methods
- No need to define number of word clusters
- Good results even on smaller number of examples

References

[1] Hartigan J.A. Direct clustering of a data matrix. Journal of American Statistical Association, 67(337):123-129, 1972

[2] Long B., Zhongfei Z. and Yu P. S. Co-clustering by Block Value Decomposition. In SIGKDD, august 21-24, 2005.

[3] Dhillon I. S., Mallela S. and Modha D. S. Information Theoretic Co-clustering. Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 24-27, 2003, pp 89-98

[4] Long Bo, Wu Xiaoyun, Zhang Zhongfei and Yu Phillip: Unsupervised Learning on K- partite Graphs, Proceedings of SIGKDD August 20-23, 2006.

[5] Kontostathis A., Pottenger W.M. A framework for understanding LSI performance. Information Processing and Management. Volume 42, number 1, 2006, pp 56-73.