Multi-regularization for Fuzzy Co-clustering

Vikas K. Garg[1], Sneha Chaudhari[2], and Ankur Narang[2]

[1] Toyota Technological Institute, Chicago
vkg@ttic.edu
[2] IBM Research Lab, India
{snechaud,annarang}@in.ibm.com

Abstract. Co-clustering is a powerful technique with varied applications in text clustering and recommender systems. For large scale high dimensional and sparse real world data, there is a strong need to provide an overlapped co-clustering algorithm that mitigates the effect of noise and non-discriminative information, generalizes well to the unseen data, and performs well with respect to several quality measures. In this paper, we introduce a novel fuzzy co-clustering algorithm that incorporates multiple regularizers to address these important issues. Specifically, we propose *MRegFC* that considers terms corresponding to Entropy, Gini Index, and Joint Entropy simultaneously. We demonstrate that MRegFC generates significantly higher quality results compared to many existing approaches on several real world benchmark datasets.

1 Introduction

Co-clustering or bi-clustering is a powerful tool that alleviates notable limitations of clustering techniques such as poor scalability, lack of cluster intrepretability and sensitivity to noise [1]. Co-clustering allows simultaneous clustering of the rows and columns of a matrix, and has been used successfully in text mining [2], [3] and collaborative filtering [4]. In collaborative filtering, for example, co-clustering can be used for identifying groups of customers with similar interests or preferences toward a set of products. The co-clusters thus obtained can be leveraged for target marketing in recommender systems.

Many co-clustering methods partition the data into non-overlapping regions where each point belongs to only one cluster such as ITCC [3], Bregman co-clustering [5]. However, in real world applications, *fuzzy* co-clustering, that allows the data points to be members of two or more clusters, is more suitable. For example, when clustering documents into topics, documents may contain multiple relevant topics and hence an overlapped co-clustering is more appropriate [6]. Overlapped co-clustering algorithms also capture the vague boundaries between clusters and improve the representation and interpretability of the clusters.

Further, certain issues need to be addressed for obtaining superior performance using fuzzy co-clustering. The points or features occurring across a large number of clusters should not be allowed to dominate since they contain very little discriminative information. Also, noise in the underlying data needs to be

M. Lee et al. (Eds.): ICONIP 2013, Part II, LNCS 8227, pp. 67–75, 2013.

effectively handled. One common way to deal with such issues is to devise fuzzy techniques that focus on optimizing an objective based on some regularizer as shown in FCR [7] and FCM with Maximum Entropy regularization [8]. A major limitation of these techniques lies in the insufficiency of a single regularizer to perform well with respect to several quality measures. For example, FCR uses entropy in the objective function which helps to obtain better degree of aggregation on real datasets, but shows lower accuracy.

FCCM [9] is a fuzzy clustering algorithm that maximizes the co-occurrence of categorical attributes (keywords) and the individual patterns (documents) in clusters. However, this algorithm poses difficulties while handling large data sets and also works for only categorical data. *Fuzzy-CoDoK* [10], a scalable modification of FCCM, involves heavy parameter tuning that makes the approach data-dependent, is susceptible to variations in data and may often fail to converge. Technique such as *SCAD* [11] only works with data lying in some Euclidean space. *SKWIC* [12] overcomes this limitation but lacks in parameter tuning and scalability. Similarly, technique such as *MOCC* [13] performs poorly with respect to degree of aggregation.

In this paper, we formulate a framework, Multi-Regularization for Fuzzy Co-clustering (MRegFC), based on maximizing an objective function that incorporates penalty terms based on the Entropy, the Gini Index, and the Joint Entropy simultaneously under certain constraints. Each one of the regularizers used in MRegFC contribute to address the issues related to co-clustering, as explained later in Section 2. MRegFC can also handle high dimensional and sparse data without over-fitting. However, incorporating multiple regularizers becomes challenging as different regularizers might have contrasting behaviors and learning a good set of weights for several regularizers simultaneously is important. Our technique MRegFC alleviates both these issues. Further, MRegFC provides valid range of values for different parameters used, to obtain high quality results. In experimental evaluation, we demonstrate superior performance in terms of precision, recall, and F-measure as compared to prior approaches: MOCC [13], ITCC [3], FCR [7] and algorithms employing only one of these regularizers. Our algorithm also demonstrates better RMSE compared to FCR [7] and individual regularizers on all the datasets in consideration. To the best of our knowledge, MRegFC is the first multiple regularizer based approach for fuzzy co-clustering.

2 The Proposed Approach

In this work we propose an approach called MRegFC for fuzzy co-clustering which formulates the objective function employing the Entropy, the Gini Index, and the Joint Entropy regularizers simultaneously. A regularization term is added to the objective function in order to prevent it from being an ill-posed problem and to avoid overfitting. The regularization term based on Entropy [8] elegantly captures the notion of *purity* of a co-cluster while emphasizing the marginal coherence along the rows (points) and the columns (features). Hence, the homogeneity along the points and the features are appropriately taken into

Table 1. Notation

Symbol	**Definition**
C	Number of co-clusters
N	Number of data points (rows)
K	Number of features (columns)
u_{ci}	Membership of row i in co-cluster c
v_{cj}	Membership of column j in co-cluster c
d_{ij}	Measure of extent of correlation between row i and column j

account using the Entropy regularizer. Gini Index, despite being similar to Entropy, ensures that the points and features that occur across a large number of clusters are not provided with any unfair advantage, besides imparting numerical stability to the algorithm [10]. Joint Entropy [14] characterizes, in a natural way, the statistical dependence of the points (rows) and the features (columns) on each other. Moreover, the Joint Entropy term, in conjunction with Entropy, creates a Mutual Information term thereby lending a better generalization ability to MRegFC by making it robust against noise. It is easy to see that the Joint Entropy term is maximized when the product $u_{ci}.v_{cj}$ is evenly distributed across the different co-clusters. Thus incorporating a joint entropy fuzzifier also reduces the susceptibility of the algorithm to overfitting[1].

A typical fuzzy co-clustering algorithm strives to maximize an objective function, generally the degree of aggregation. Using the notations given in Table 1,

the degree of aggregation for cluster c can be quantified as

$$\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ci}v_{cj}d_{ij}, \quad \text{for } \; c \in \{1, 2, \ldots, C\} \tag{1}$$

The intuition is that we want to bring together rows and columns with high d_{ij} values in the same co-cluster. To maximize the value of the objective function, for such i and j, we need to set high values for both u_{ci} and v_{cj} for the same cluster c. Additionally, we impose the following constraints:

$$\sum_{c=1}^{C} u_{ci} = 1, u_{ci} \in [0, 1], i \in \{1, 2, \ldots, N\} \tag{2}$$

$$\sum_{j=1}^{K} v_{cj} = 1, v_{cj} \in [0, 1], c \in \{1, 2, \ldots, C\} \tag{3}$$

The first constraint requires that the addition of membership values of each row across all the co-clusters is equal to 1. Such a constraint is said to satisfy the *Ruspini's condition* [15]. The second constraint, on the other hand, requires that the summation of all column memberships must be one for each co-cluster

[1] The under-fitting issues are implicitly taken care of by the term corresponding to the degree of aggregation.

thereby implying a weighting scheme for the columns, instead of the partitions[2]. We now add regularization terms corresponding to Entropy, Gini Index and Joint Entropy in the objective function. Consequently, using the weight parameters T_{u_1} and T_{u_2} for Entropy, T_{v_1} and T_{v_2} for Gini Index, T_{uv} for Joint Entropy to specify the extent of fuzziness, we strive to maximize our regularized objective function, OBJ

$$\begin{aligned}
&= \sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ci}v_{cj}d_{ij} - T_{uv}\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ci}v_{cj}\log(u_{ci}v_{cj}) - T_{u_1}\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}\log(u_{ci}) \\
&- T_{u_2}\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^2 - T_{v_1}\sum_{c=1}^{C}\sum_{j=1}^{K} v_{cj}\log(v_{cj}) - T_{v_2}\sum_{c=1}^{C}\sum_{j=1}^{K} v_{cj}^2 \\
&+ \sum_{i=1}^{N}\lambda_i\left(\sum_{c=1}^{C} u_{ci} - 1\right) + \sum_{c=1}^{C}\gamma_c\left(\sum_{j=1}^{K} v_{cj} - 1\right) \qquad (4)
\end{aligned}$$

Differentiating with respect to u_{ci} we get

$$\Rightarrow \frac{\partial OBJ}{\partial u_{ci}} = \sum_{j=1}^{K} v_{cj}d_{ij} - T_{u_1}\left(1+\log(u_{ci})\right) - 2T_{u_2}u_{ci} - T_{uv}\sum_{j=1}^{K} v_{cj}\left(1+\log(u_{ci}v_{cj})\right) + \lambda_i \qquad (5)$$

For optimality of OBJ, we must have $\dfrac{\partial OBJ}{\partial u_{ci}} = 0$. Further, since $u_{ci}, v_{cj} \in [0,1]$, approximating $(1+\log u_{ci})$ and $(1+\log u_{ci}v_{cj})$ by u_{ci} and v_{cj} respectively, we have

$$\Rightarrow u_{ci} = \frac{\lambda_i + \sum_{j=1}^{K} v_{cj}d_{ij}}{T_{u_1} + 2T_{u_2} + T_{uv}\sum_{j=1}^{K} v_{cj}^2} \qquad (6)$$

Now using $\sum_{c=1}^{C} u_{ci} = 1$, and simplifying, we obtain

$$u_{ci} = \frac{1}{C} + \frac{1}{T_{u_1} + 2T_{u_2} + T_{uv}\sum_{j=1}^{K} v_{cj}^2} * \left(\sum_{j=1}^{K} v_{cj}d_{ij} - \frac{1}{C}\sum_{t=1}^{C}\sum_{j=1}^{K} v_{tj}d_{ij}\right) \qquad (7)$$

Following a similar procedure of obtaining u_{ci}, we can compute

$$v_{cj} = \frac{1}{K} + \frac{1}{T_{v_1} + 2T_{v_2} + T_{uv}\sum_{i=1}^{N} u_{ci}^2} * \left(\sum_{i=1}^{N} u_{ci}d_{ij} - \frac{1}{K}\sum_{t=1}^{K}\sum_{i=1}^{N} u_{ci}d_{it}\right) \qquad (8)$$

[2] We do not impose Ruspini's condition on the columns since then a single co-cluster containing all the rows and columns would be formed.

A good selection of the parameters T_{u_1}, T_{u_2}, T_{v_1}, and T_{v_2} can be mathematically derived in a straightforward way (omitted due to space constraints):

$$0 < T_{u_1} < \frac{\sum_{t=1}^{C}\sum_{j=1}^{K} v_{tj}P_j}{N} - T_{uv}\max_c \sum_{j=1}^{K} v_{cj}^2 \tag{9}$$

$$T_{u_2} = \frac{\sum_{t=1}^{C}\sum_{j=1}^{K} v_{tj}P_j - NT_{uv}\max_c \sum_{j=1}^{K} v_{cj}^2 - NT_{u_1}}{2N} \tag{10}$$

$$0 < T_{v_1} < \frac{\sum_{j=1}^{K} P_j}{C} - T_{uv}\max_c \sum_{i=1}^{N} u_{ci}^2 \tag{11}$$

$$T_{v_2} = \frac{\sum_{j=1}^{K} P_j - CT_{v_1} - CT_{uv}\max_c \sum_{i=1}^{N} u_{ci}^2}{2C} \tag{12}$$

Please note that this is a lateral benefit of our approach since in general, tuning the input parameters appropriately is a difficult problem, and the algorithm may not perform satisfactorily in the absence of any tuning guidelines.

Algorithm 1 describes our approach for fuzzy co-clustering. The algorithm takes as input the number of co-clusters C, the row-column correlation matrix D, and a threshold ϵ to specify the stopping criterion. It can be observed that the parameters λ and γ do not play a role in the resulting algorithm and hence show no effect on the overall performance. The different row memberships are randomly initialized subject to the constraint that their summation is equal to 1. Based on selection of T_{uv}, the values of the parameters T_{v1} and T_{v2} is chosen from the respective acceptable range. The algorithm then alternately updates the row and column memberships repeatedly, until the change in all the row memberships across two successive iterations is bounded by ϵ. At termination, the algorithm outputs appropriate row and column memberships across the different co-clusters.

3 Experimental Evaluation

In this section, we present experimental evaluation on several benchmark datasets that demonstrates a superior performance of *MRegFC* over the *FCR*, *ITCC* and *MOCC* algorithms. We also demonstrate the benefits of using multiple regularizers in *MRegFC* by presenting a comprehensive evaluation against the individual regularizers.

Algorithm 1. Multi-Regularized Fuzzy Co-clustering (MRegFC)

```
Input  : No. of co-clusters $C$, row-col matrix $D$, and threshold parameter $\epsilon$
Output : Membership values $u_{ci}$ and $v_{cj}$
1  Compute $P_j = \sum_{i=1}^{N} d_{ij}$. Initialize randomly memberships $u_{ci} \geq 0$, $c \in [C]$ and
   $i \in [N]$ such that $\sum_{c=1}^{C} u_{ci} = 1$.
2  Choose $T_{uv} \in \left( 0, \frac{\sum_{j=1}^{K} P_j}{C \max_c \sum_{i=1}^{N} u_{ci}^2} \right)$.
3  Choose $T_{v_1}$ using Eqn. (11).
4  Compute $T_{v_2}$ using Eqn. (12).
5  Compute memberships $v_{cj}$ using Eqn. (8).
6  Choose $T_{u_1}$ using Eqn. (9).
7  Compute $T_{u_2}$ using Eqn. (10).
8  $u_{ci}^{old} \leftarrow u_{ci}$
9  Update memberships $u_{ci}$ using Eqn. (7).
10 if $\left( \max_c |u_{ci} - u_{ci}^{old}| > \epsilon \right)$ then
11     Update memberships $v_{cj}$ using Eqn. (8).
12     Go to step 9
13 end
```

We conducted experimentation on the following datasets [16], [13]: (a) *Movielens* for movie recommendations, (b) *Classic3* for document collections, (c) *Jester* for joke ratings (d) *Reuters (21578)* for text categorization, and (e) *20 Newsgroups* for text classification and clustering. We used two subsets of the Movielens dataset: (a) (*Mv1*: 679 movies from 3 genres - Animation, Children and Comedy, and (b) *Mv2*: 232 movies from 3 genres - Thriller, Action and Adventure. These are similar to the ones used in [13], and therefore provide for a consistent comparison with the *MOCC* and *ITCC* algorithms. Each reported result is based on an average over 10 trials. The number of clusters chosen for experiments $E1$ and $E2$ were 8 and 16, respectively for *MRegFC*; other algorithms were represented by $(5, 5)$ and $(10, 10)$ row and column clusters. The threshold parameter ϵ was set to 0.00001. In order to compare the quality of clustering results, we use the following standard measures: RMSE, precision, recall, and F-measure [13].

3.1 Comparison with Existing Approaches

Table 2 presents the comparative results for precision, recall and f-measure on the Movielens dataset. *MRegFC* has a high average precision value of around 0.73, and consistently outperforms the other algorithms. *MRegFC* achieves a high recall value of about 0.67 on an average. Further, it can be seen that *MRegFC*

Table 2. Precision, Recall, F-measure Comparison with Existing Approaches

Dataset	Precision				Recall				F-measure			
	MRegFC	MOCC	ITCC	FCR	MRegFC	MOCC	ITCC	FCR	MRegFC	MOCC	ITCC	FCR
Mv1-E1	**0.75**	0.60	0.63	**0.75**	0.65	**0.67**	0.19	0.61	**0.69**	0.63	0.29	0.67
Mv1-E2	**0.76**	0.62	0.65	0.75	**0.71**	0.65	0.13	0.61	**0.73**	0.63	0.22	0.67
Mv2-E1	**0.70**	0.46	0.54	0.69	**0.64**	0.62	0.23	0.57	**0.66**	0.53	0.32	0.63
Mv2-E2	**0.70**	0.48	0.57	0.69	**0.69**	0.58	0.16	0.56	**0.69**	0.52	0.25	0.63

Table 3. Precision, Recall and F-measure Comparison with Individual Regularizers. (E: Entropy, GI: Gini Index, JE: Joint Entropy)

Dataset	Precision				Recall				F-measure			
	MRegFC	E	GI	JE	MRegFC	E	GI	JE	MRegFC	E	GI	JE
Reuters (21578)	**0.548**	0.409	0.31	0.43	**0.63**	0.546	0.555	0.477	**0.6**	0.46	0.49	0.38
20News Groups	**0.516**	0.3	0.304	0.3	**0.825**	0.767	0.546	0.609	**0.66**	0.43	0.39	0.4
Mv1	**0.756**	0.701	0.689	0.711	**0.653**	0.562	0.554	0.216	**0.7**	0.62	0.61	0.35
Mv2	**0.718**	0.684	0.69	0.702	**0.64**	0.515	0.558	0.268	**0.67**	0.59	0.61	0.39

has an average F-measure of 0.69, while the closest competitor *FCR* achieves a value of 0.65. This clearly demonstrates that *MRegFC* yields consistently better quality clusters compared to the existing algorithms.

Fig. 1 presents the variation of F-measure (using the *Mv2* dataset), as the number of clusters and row-clusters increases from 2 to 10, for *MRegFC* and other algorithms (*MOCC*, *FCR* and *ITCC*). This result was used to choose the number of co-clusters in the algorithm. It can be seen that as the number of clusters increases beyond 5, *MRegFC* consistently outperforms the other algorithms by a convincing margin.

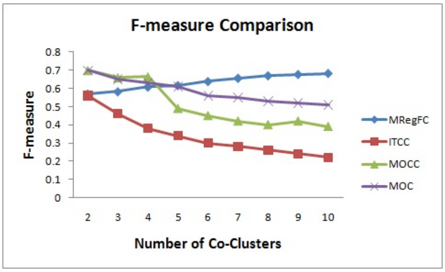

Fig. 1. F-measure vs no. of co-clusters (Mv2)

3.2 Comparison with Individual Regularizers

To quantify the benefit of incorporating multiple regularizers, we also compared *MRegFC* with similar algorithms that include only one of the *Entropy*, *Gini Index*, and *Joint Entropy* regularizers. Table 3 presents the comparison results on the different datasets in terms of F-measure. Clearly, *MRegFC* outperforms the techniques using individual regularizers. Since *MRegFC* also achieves the lowest RMSE of all techniques across all data sets (Table 4, we conclude that the need for incorporating multiple regularizers, as in *MRegFC*, cannot be overemphasized.

We also varied the parameter T_{uv} over a large range on all the datasets and observed that training time and RMSE do not vary much with change in T_{uv}. This demonstrates the robustness of the proposed approach with respect to the input parameter T_{uv}. We omit the details due to space constraints.

Table 4. RMSE comparison

Dataset	MRegFC	Entropy	Gini Index	Joint Entropy	FCR
Reuters (21578)	**1.37**	1.47	1.51	1.58	1.4
20News Groups	**1.45**	1.56	1.56	1.57	1.56
Mv1	**1.36**	1.39	1.45	1.56	1.48
Mv2	**1.23**	1.28	1.3	1.51	1.4
Jester-1	**17.68**	20.47	20.68	22.12	20.64
Jester-2	**17.55**	22.55	20.56	25.7	20.02
Classic3 (CRAN)	**1.04**	1.09	1.13	1.16	1.46
Classic3 (MED)	**1.27**	1.57	1.59	1.59	1.59

4 Conclusion

We present a novel fuzzy co-clustering framework that simultaneously incorporates multiple regularizers namely Entropy, Gini Index, and Joint Entropy while trying to maximize the degree of aggregation. The approach can handle categorical and numerical data in addition to the highly sparse high dimensional data, without over-fitting. Furthermore, unlike existing algorithms, we provide an appropriate range of values for tuning the various parameters to obtain high quality results. We demonstrate superior performance, in terms of several quality measures such as precision, recall, F-measure and RMSE compared to the prior approaches as well as algorithms using individual regularizers.

References

[1] Madiera, S., Oliveira, A.: Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Computational Biology and Bioinformatics 1, 24–45 (2004)

[2] Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2001, pp. 269–274. ACM, New York (2001)

[3] Dhillon, I.S., Mallela, S., Modha, D.: Information theoretic co-clustering. In: Proceedings of the Ninth ACM SigKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 89–98. ACM Press, New York (2003)

[4] George, T., Merugu, S.: A scalable collaborative filtering framework based on co-clustering. In: Proceedings of the Fifth IEEE International Conference on Data Mining, ICDM 2005, pp. 625–628. IEEE Computer Society, Washington, DC (2005)

[5] Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., Modha, D.S.: A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 509–514. ACM, New York (2004)

[6] Sahami, M., Hearst, M., Saund, E.: Applying the multiple cause mixture model to text categorization. In: International Conference on Machine Learning (1996)

[7] Tjhi, W.C., Chen, L.: A partitioning based algorithm to fuzzy co-cluster documents and words. Pattern Recognition Letters 27, 151–159 (2006)

[8] Miyamoto, S., Mukaidono, M.: Fuzzy c-means as a regularization and maximum entropy approach. In: Proceedings of IFSA, vol. 2, pp. 86–92 (1997)

[9] Oh, C.H., Honda, K., Ichihashi, H.: Fuzzy clustering of categorical multi-variate data. In: Proceedings of IFSA/NAFIPS, Vancouver, USA, pp. 2154–2159 (2001)

[10] Kummamuru, K., Dhawale, A., Krishnapuram, R.: Fuzzy co-clustering of documents and keywords. In: IEEE International Conference on Fuzzy Systems (2003)

[11] Frigui, H., Nasraoui, O.: Simultaneous clustering and attribute discrimination. In: Proceedings of FUZZIEEE, pp. 158–163 (2000)

[12] Frigui, H., Nasraoui, O.: Simultaneous categorization of text documents and identification of cluster-dependent keywords. In: Proceedings of FUZZIEEE, pp. 158–163 (2001)

[13] Shafiei, M.M., Milios, E.E.: Model based overlapping co-clustering. In: SDM, Maryland, USA (2006)

[14] MacKay, D.: Information theory, inference, and learning algorithms. Cambridge University Press (2003)

[15] Dumitrescu, D., Lazzerini, B., Jain, L.: Fuzzy sets and their applications to clustering and training. CRC Press, Boca Raton (2000)

[16] Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval 4, 133–151 (2001)