International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 (2018) pp. 12289-12295

Multi View Cluster Approach to Explore Multi Objective Attributes based on Similarity Measure for High Dimensional Data

Deena Babu Mandru[1] and Y.K. Sundara Krishna[2]

[1]Department of Computer Science,
Krishna University, Machilipatnam, Andhra Pradesh -521002, India.

[2]Krishna University, Machilipatnam, Andhra Pradesh -521002, India.

Abstract

Data retrieval is an aggressive concept in data mining with different attributes based on relations present in overall data from different web data sources. In data mining, clustering is an approach to explore data representation based on expressive attributes from large data sources. Clustering with attribute selection from overall data source, traditionally proposed Enhanced Feature Selection based Clustering (EFSC) to evaluate efficiency to form sub set of features with respect to quality assurance for sub set of features. All these clustering methods mainly focused to assume cluster relation based on different features among different objects. Similarity between pair of objects is either outside data relations i.e explicitly or inside data relations i.e implicitly. So that improves multi objective data relations with different attributes for data retrieval from data sources is aggressive and important concept to view data relations in different dimensions. Traditional approach is only support to single data view based on different attributes. In this paper, we propose Enhanced Multi View Voronoi based Clustering (EMVVC) approach which is extension to EFSC for multi object attributes relations. In multi view Voronoi clustering, most informative multi-object attribute similarity could be achieved, theoretical and empirical study is conducted to support this problem for different attribute relations. This approach mainly proposes on documents to retrieve multi objective attribute relations on multi view of data representation. Our experimental outcome shows coherent multi-view cluster results for multi-objects with respect to different attribute relations from different data sources.

Keywords: Data mining, similarity measure, features selection, multi-objectives, multi-view points and voronoi clustering

INTRODUCTION

For effective data collection from data resources regarding to relevant data single class learning is required to perform marked centered category with individual training series on features. For some real world data sourcing, for real-time data set portioning with irregular behavior category brand instances with expensive impossible data demonstration. To learn these types of combined series in real-time data set techniques to categorize target data into unique classifier data techniques. For variety of different programs abnormality recognition, papers category image annotation and content requirements for different data creation. Clustering is the most effective concept in data mining to group relevant elements based on similarity, main aim of clustering is to define desirable structures in representation of data and control them into meaningful sub cluster to main cluster for further study and analysis of data. There have been different types of clustering approaches were introduced traditionally; they can propose and define different types developed approaches in research fields. According to the recent discussion about clustering and their properties to explore data in different ways, more than 10-50 years k-means is the mostly used algorithm and is best data mining algorithm now a days. Another recent algorithm i.e Enhanced Feature Selection based Clustering (EFSC) algorithm states that cluster data based on different features present in database, and define features in different way to explore or combined data at different formations. Procedure of this approach is shown in fig 1 with different feature selections from original data.

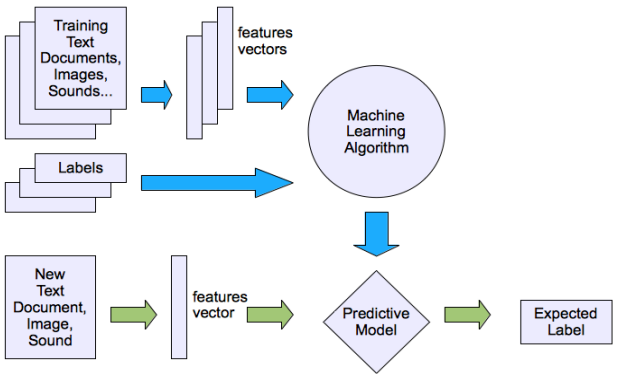

Figure 1: Feature based data extraction from different text related documents.

The main common approach to grouping is an optimization procedure from data with different attributes, optimal process to partitioned different attributes is to found by optimizing particular function like similarity or distance between different attributes from data. Basic important assumption regarding structure of data should define similarity between different attributes assumed and embedded clustering

function. However, effectiveness of clustering methods in these criteria primarily depends on similarity measure to the data at a time with different presentations of data. The main concept behind k-means and other clustering algorithms is sum of the squared error with objective function that used Euclidean distance; it is only support to simple type of data sets. For very high and sparse dimensional data like text documents, k-means and Enhanced Feature Selection based Clustering algorithms which use cosine similarity instead of Euclidean distance as a measure function is most suitable to all the data sets. In this paper we propose and implement Enhanced Multi View Voronoi based Clustering (EMVVC) approach which is extension to EFSC for multi object attributes relations. In multi view Voronoi clustering, most informative multi-object attribute similarity could be achieved, theoretical and empirical study is conducted to support this problem for different attribute relations. General procedure to retrieve data from different document and represent them into different views as shown in Fig 2

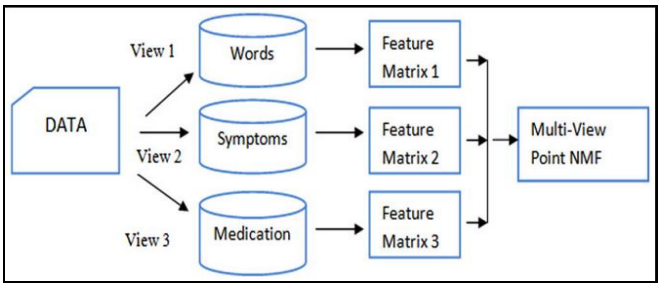

Figure 2: Overview of multi view point with factorization of data based on different attributes

The work behind this approach is motivated and investigated from different similar search finding attributes at different positions. This approach defines that nature of similarity measure and important role in success and failure rate of proposed clustering approach. Main contributations of this paper as follows

a) Derive and define novel technique for calculating similarity among different data objects in high and sparse dimensional data, particularly in text related documents.

b) We formulate new clustering approach function related to desire similarity measure with respective attributes.

c) Our approach is scalable and flexible like k-means and other clustering approaches and also provides high and sparse quality and increase the performance with respect to different parameters.

Remaining of this paper organized as follows: related work relates to different approaches and algorithms with their respective author opinions on clustering and similarity measure discussed in section 2. Background work relates to similarity clustering is shown in section 3, propose and implement novel similarity measure clustering approach discussed in section 4. Experimental evaluation with comparison of existing approaches discussed in section 5, section 6 concludes overall conclusion about discussed in this paper.

BACKGROUND WORK

Irrelevant feature removal along with repeated attributes, definably accuracy of the different machine learning approaches, Thus, feature subset selection should be able to recognize and remove as much of the unrelated and repetitive information as possible. Moreover, "good function subsets contain features extremely associated with (predictive of) the course, yet uncorrelated with (not predictive of) each other."

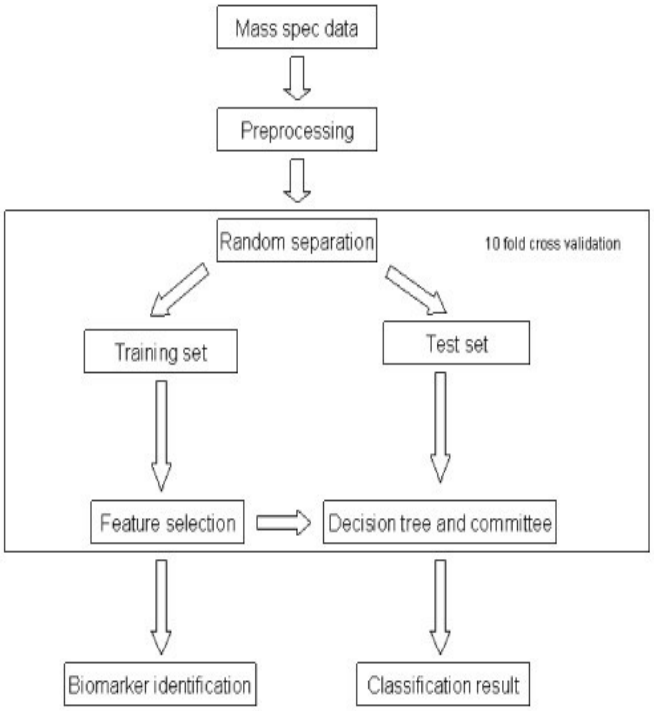

Figure 3. Procedure relates to feature selection in EFSC.

Traditional feature selection novel approaches which can proficiently and viably manage both unessential and excess highlights, and get a decent element subset. We accomplish this through another component choice system (appeared in Fig. 3) which made out of the two associated parts of insignificant element evacuation and repetitive component disposal. The previous acquires highlights pertinent to the objective idea by taking out insignificant ones, and the last mentioned expels excess highlights from pertinent ones by means of picking agents from various component bunches, and in this way produces the last subset. The unessential element expulsion is direct once the correct importance measure is characterized or chosen, while the repetitive element end is a touch of complex. In EFSC calculation, it includes 1) the development of the base spreading over tree from a weighted finish diagram; 2) the dividing of the MST into a backwoods with each tree speaking to a group; and 3) the choice of delegate features from the clusters (groups).

REVIEW OF LITERATURE

This section discussed about different authors opinion regarding clustering with respect to similarity measure and other clustering algorithms. Multi-view information is extremely regular in genuine applications in the huge information period. For example, a website page can be portrayed by the words showing up on the page itself and the words fundamental all connections indicating the site page from different pages in nature. In interactive media content comprehension, media fragments can be at the same time portrayed by their video signals from visual camera and sound signs from voice recorder gadgets. The presence of such multi-view information raised the enthusiasm of multi-view learning [2], [3], [4], which has been broadly considered in the semi-regulated picking up setting. For unsupervised adapting, especially, multi-view bunching single view based grouping strategies can't make a powerful utilization of the multi-view data in different issues. For example, a multi-view grouping issue may require to distinguish bunches of subjects that contrast in every one of the information views. For this situation, linking highlights from the diverse perspectives into a solitary association took after by a solitary view bunching technique may not fill the need. It has no instrument to ensure that the resultant bunches contrast from the majority of the perspectives on the grounds that a particular perspective of highlights may probably be weighted significantly higher than different perspectives in the element association which renders the gathering is construct just with respect to one of the views. Multi-view grouping has along these lines pulled in an ever increasing number of considerations in the previous two decades, which makes it essential furthermore, gainful to condense the best in class and outline open issues to control future headway. Like the classification of grouping calculations in [1], we partition the current MVC techniques into two classifications: generative (or show based) approaches a d discriminative (or likeness based) approaches. Generative approaches attempt to take in the principal appropriation of the information and utilize generative models to speak to the information with each model speaking to one group. Discriminative methodologies straightforwardly upgrade a target work that includes pairwise likenesses to limit the normal similitude inside bunches what's more, to amplify the normal closeness between bunches. Because of countless methodologies, in view of how they consolidate the multi-view data, we additionally isolate them into five classes: (1) basic Eigen-vector framework (chiefly multi-view ghostly grouping), (2) basic coefficient grid (primarily multi-view subspace bunching), (3) regular pointer framework (mostly multi-view non-negative network factorization grouping), (4) coordinate view mix (basically multi-piece bunching), (5) view blend after projection (for the most part canonical correlation analysis (CCA)). The initial three classes have a shared trait that they share a comparable structure to consolidate numerous perspectives.

Most provably effective bunching calculations first undertaking the information down to some low dimensional space and afterward group the information in this lower dimensional space (a calculation, for example, single linkage generally does the trick here). Regularly, these calculations likewise work under a partition necessity, which is estimated by the base separation between the methods for any two blend segments. One of the main provably effective calculations for learning blend models is because of [Das99], who takes in a blend of circular Gaussians by arbitrarily anticipating the blend onto a low-dimensional subspace. [VW02] give a calculation an enhanced detachment necessity that takes in a blend of k circular Gaussians, by anticipating the blend down to the k-dimensional subspace of most astounding fluctuation. [KSV05, AM05] stretch out this outcome to blends of general Gaussians; notwithstanding, they require a partition relative to the greatest directional standard deviation of any blend segment. [CR08] utilize a standard relationships based calculation to learn blends of pivot adjusted Gaussians to a division corresponding to $\sigma*$, the most extreme directional standard deviation in the subspace containing the methods for the circulations. Their calculation requires an organize freedom property, and an extra "spreading" condition. [BL08] propose a comparative calculation for multi-view grouping, in which information is anticipated onto the best bearings acquired by part CCA over the perspectives. They indicate exactly that for bunching pictures utilizing the related content as a second view (where the objective grouping is a human-characterized class), CCA-based grouping techniques out-perform PCA-based calculations.

PROPOSED SYSTEM DESIGN AND IMPLMENETATION

In this section, we propose a novel multi view cluster based voronoi approach is to evaluate cosine similarity between relevant documents and consecutively formulae related to document clustering. Basic parameters used in multi view cluster analysis shown in Table 1.

Table 1. Different parameter sequences used in multi view cluster for different attributes.

Parameter	Description
n,m,c,k,d	Number of documents, terms, classes, clusters, and document factor ‖d‖=1
S = {d1, . . . , dn} ,Sr	Set of documents in cluster Sr
$D=\sum_{d_i \in S} d_i$	Composite vector of documents
$D_r=\sum_{d_i \in S_r} d_i$	Composite documents for cluster r
$C = D / n$	Centroid vector documents
$C_r = D_r / n_r$	Centroid vector documents for cluster r

Basic summarizations of different aspects with different attributes based on calculation with Euclidean-distance as follows:

$$\text{Dist (di, dj)} = \|\text{di - dj}\|$$

Minimum distance for cluster formation based on different attributes

$$\min \sum_{r=1}^{k} \sum_{d_i \in S_r} \| d_i - C_r \|^2$$

Vector representation of different attributes with similar attributes as follows:

$$\text{Sim(di, dj)} = \cos(\text{di, dj}) = \text{d}^{\text{t}}_{\text{i}}\ \text{d}_{\text{j}}$$

Cosine similarity for different attributes shown in above equation presentation for k-means with Euclidian distance, similarity magnitudes are main difference between Euclidian distance and k-means distance from overall data sets. Some of the researchers define more sequential clustering data presentation to access different attributes in cosine similarity attribute presentation

Similarity Measure with Voronoi clustering

In this section, we define and present voronoi clustering procedure with factorization matrix formation to solve optimization problem and similar measure. Voronoi clustering describe data in different views would be assigned to same cluster with sparse and high probability for different data sets. Therefore matrix formation at different co-efficient matrices from different views of data to be formalized with single cluster to words common similar consensus using different situations. We present the development process of our proposed approach to define efficient data presentation in different dimensions with effective similarity measures between data objects. Multi view point similarity measure for structure documents as follows:

$$\text{MVS}(\text{d}_\text{i}, \text{d}_\text{j} \mid \text{d}_\text{i}, \text{d}_\text{j} \in \text{S}_\text{r}) = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S_r} (d_i^t d_j - d_i^t d_h - d_j^t d_h + d_h^t d_h)$$

$$= d_i^t d_j - \frac{1}{n - n_r} d_i^t \sum_{d_h} d_h - \frac{1}{n - n_r} d_i^t \sum_{d_h} d_h + 1, \| d_h \| = 1$$

Compare two similar documents with attributes relations for all documents, MVS (d_i, d_j) and MVS (d_i, d_l), documents dj is more similar to documents di than the other documents dl is, if and only if. Implementation procedure of the MVS with similar attributes is show in the following clustering algorithm.

Input: Nonnegative Matrix {X(1), X(2), . . . , X(nv)},
parameters {λ1, λ2, . . . , λnv }, variety of groups K
Output: Foundation Matrices {U (1), U(2), ... , U(nv)}, Coefficient
Matrices {V(1), V (2), . . . , V (nv)} and Consensus
Matrix V∗1: Stabilize each perspective X(v)
such that ||X(v)||1 = 1
2: Initialize U(v),V(v) and U∗(1 ≤ v ≤ nv)
3: repeat
4: for v = 1 to nv do
5: repeat
6: Solving V∗ and V(v), upgrade U(v) by above equations
7: Stabilize U(v) and V (v) as in above equation
8: Solving V ∗ and U(v), upgrade V(v)
9: until calc Sim() .
10: end for
11: Solving U(v) and V(v)(1 ≤ v ≤ nv), upgrade V∗
12: do it again and get {U (1), U(2), . . . , U(nv)} with Sim()

Algorithm 1: Procedure for clustering with different attributes to generate matrices.

1. Optimization Procedure: We will probably perform archive grouping by improving rule capacities IR and IV [clustering with MVSC]. To accomplish this, we use the consecutive and incremental variant of k-implies [4] [5], which are ensured to merge to a neighborhood ideal. This calculation comprises of various emphases: at first, k seeds are chosen haphazardly and each report is doled out to bunch of nearest seed in light of cosine likeness; in every one of the ensuing cycles, the records are picked in arbitrary request and, for each archive, a move to another group happens if such move prompts an expansion in the goal work. Especially, considering that the declaration of IV [clustering with MVSC] depends just on nr and Dr, r=1… k,

$$I_v = \sum_{r=0}^{k} I_r(n_r, D_r)$$

Think that, at starting of some version a documents di is associated with a group SP that has purpose value IP (nP, DP).di will be transferred to another group Sq that has purpose value Iq (nq,Dq) if the following situation is satisfied:

$$\Delta I_v = I_p(n_p - 1, D_p - d_i) + (I_p(n_p + 1, D_p + d_i)) - I_p(n_p, D_p) - I_q(n_p, D_p)$$

Subsequently, report di is moved to another group that gives the biggest increment in the goal work, if such an increment exists. The composite vectors of relating old and new bunches are refreshed in a split second after each move. In the event that a greatest number of cycles is come to or no more move is recognized, the technique is halted. A noteworthy favorable position of our grouping capacities under this streamlining plan is that they are exceptionally proficient computationally. Amid the improvement process, the primary computational request is from hunting down ideal bunches to move singular records to, and refreshing composite vectors because of such moves. On the off chance that T signifies the quantity of emphases the calculation takes, nz the aggregate number of non-zero sections in all archive vectors, the computational

multifaceted nature required for grouping with IR and IV is roughly.

2. Cluster Representation: Two authentic evaluation datasets are used as situations in this authenticity analyze. The first is reuters7, a part of the well known collecting, Reuters-21578 Submission 1.0, of Reuter's newswire articles1. Reuters-21578 is one of the most generally used analyze collecting for material purchase. In our authenticity analyze, we select 2,500 records from the greatest 7 classifications: "acq", "polytics, "tech", "health", "cash fx", "ship" and "exchange" to form reuters7. A part of records may display up in more than one category.

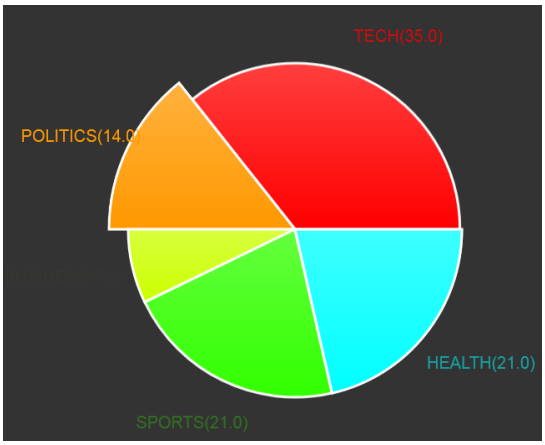

Figure 3: Multi-view data representation for different attributes.

At last, the reviews were calculated by TF-IDF and consistent to device vectors. The complete features of reuters7 and k1b are shown in Fig. 3. The validity test has shown the prospective benefits of the new multi-viewpoint centered likeness evaluate in comparison to the cosine evaluate.

EXPERIMENTAL EVALUATION

Experiments were introduced to explore the procedure of proposed approach i.e multi view voronoi clustering with comparison to existing approach i.e Feature Selection based Clustering (EFSC) to define architecture of cluster shared with multi view points.

Data sets: Basic synthetic data sets used in experiments among real world application oriented text, video and other data representations shown in Table 2.

Table 2: Statistics of different data sets used for multi view similarity index

dataset	size	# view	# cluster
Synthetic	10000	2	4
3-Sources	169	3	6
Reuters	600	3	6
Digit	2000	2	10

Fig 4 shows the accuracy of our proposed approach with different data sets evaluation procedure on text oriented documents with feasible parameters with values shown in Table 3

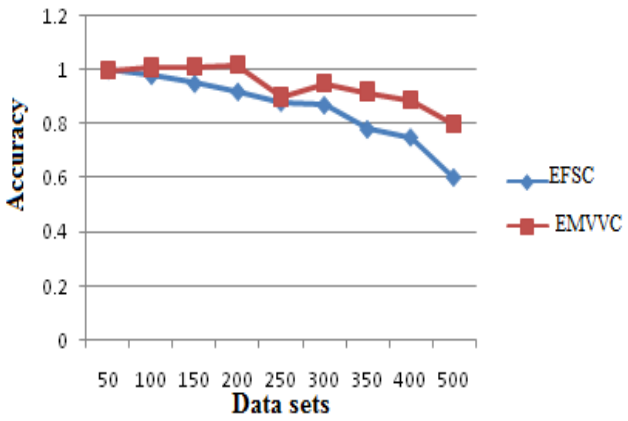

Figure 4: Accuracy of proposed approach with respect to different data views.

Table 3: Accuracy values relates to different documents

Documents	5Ws Model	SMCMV
50	1	1
100	0.98	1.01
150	0.95	1.015
200	0.92	1.02
250	0.88	0.9
300	0.87	0.95
350	0.78	0.92
400	0.75	0.89
500	0.6	0.8

Time efficiency results are plotted with following values show in Table 4. The presented of performance evaluation of our proposed approach with traditional approach shown in Fig 5 with respect to time efficiency in real time data set processing.

Table 4: Time efficiency values

Documents	SMCMV	5Ws Model
15	0.015	0.04
30	0.014	0.03
45	0.012	0.035
60	0.011	0.02
75	0.009	0.025
90	0.008	0.015

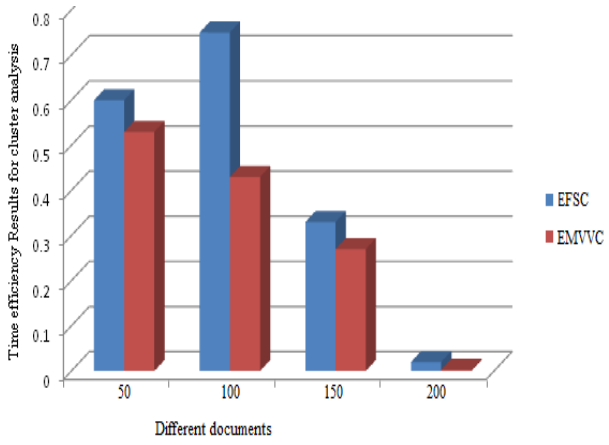

Figure 5: Performance evaluation of proposed approach with respect to different attributes.

Based on above results, finally, we describe and conclude proposed (EMVVC) approach gives better and efficiency results than EFSC for different types of documents related to dissimilar kind of documents with respect to multi view representation of different attributes.

CONCLUSION

In this paper, we propose and implement novel multi-view voronoi clustering for multi view representation of different attributes based on matrix formation. Increase efficiency learns from implementation of clustering approach in multiple views. We require different matrices learn from voronoi matrices formation of different views to regulate and combine different attributes in similar cluster. To achieve this procedure, we implement voronoi clustering approach to incorporate not only for individual data elements. We also present proposed system implementation methodology in meaningful way. Our experimental results show effective performance results worked on synthetic data sets with better accuracy when compare to existing approaches

REFERENCES

[1] Duc Thang Nguyen, Lihui Chen, "Clustering with Multi-Viewpoint based Similarity Measure" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. XX, NO. YY, 2011.

[2] Cai, X., Nie, F., Huang, H., and Kamangar, F. Heterogeneous image feature integration via multi-modal spectral

clustering. In *CVPR*, 2011.

[3] Kumar, A., Rai, P., and Daum´e III, H. Co-regularized multi-view spectral clustering. In *NIPS*, 2011.

[4] Obozinski, Guillaume, Taskar, Ben, and Jordan, Michael I. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20:231–252, 2010.

[5] Prettenhofer, P. and Stein, B. Cross-language text classification using structural correspondence learning. In *ACL*, 2010.

[6] Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon, Saykin, Andrew J, and Shen, Li. Identifying adsensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *MICCAI 2011*, pp. 115–123. Springer, 2011.

[7] Wang, Hua, Nie, Feiping, Huang, Heng, Kim, Sungeun, Nho, Kwangsik, Risacher, Shannon L, Saykin, Andrew

J, Shen, Li, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012a.

[8] Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon L, Saykin, Andrew J, Shen, Li, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012b.

[9] Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Nho, Kwangsik, Risacher, Shannon L., Saykin, Andrew J., Shen, Li, and for the Alzheimer's Disease Neuroimaging Initiative. From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics*, 28(18):i619–i625, 2012.

[10] Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Risacher, Shannon, Saykin, Andrew, and Shen, Li. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, 2012.

[11] Wang, Hua, Nie, Feiping, Huang, Heng, and Ding, Chris. Heterogeneous Visual Features Fusion via Sparse Multimodal Machine. In *CVPR 2013*, 2013.

[12] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, An Introduction toInformation Retrieval. Press, Cambridge U., 2009.

[13] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering," in IEEE ICDM, 2001, pp. 107–114.

[14] H. Zha, X. He, C. H. Q. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in NIPS, 2001, pp. 1057–1064.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 888–905, 2000.

[16] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in KDD, 2001, pp. 269–274.

[17] Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag New York, Inc., 2007.

[18] V. Sindhwani and P. Niyogi. A co-regularized approach to semi-supervised learning with multiple views. In Proceedings of the ICML Workshop on Learning with Multiple Views, 2005.

[19] A. Singh and G. Gordon. Relational learning via collective matrix factorization. In KDD, pages 650{658, 2008.

[20] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. SIGIR, pages 267{273, 2003.

[21] J. Yang, S. Yang, Y. Fu, X. Li, and T. Huang. Non-negative graph embedding. In CVPR, pages 1 {8, june 2008.

[22] D. Zhou and C. Burges. Spectral clustering and transductive learning with multiple views. ICML, pages 1159{1166, 2007.

[23] P. Lakkaraju, S. Gauch, and M. Speretta, "Document similarity based on concept tree distance," in Proc. of the 19th ACM conf. on Hypertext and hypermedia, 2008, pp. 127–132.

[24] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," IEEE Trans. on Knowl. and Data Eng., vol. 20, no. 9, pp. 1217–1229, 2008.

[25] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "Fast detection of xml structural similarity," IEEE Trans. on Knowl. And Data Eng., vol. 17, no. 2, pp. 160–175, 2005.

[26] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, "Webace: a web agent for document categorization and exploration," in AGENTS '98: Proc. of the 2nd ICAA, 1998, pp. 408–415.

[27] J. Friedman and J. Meulman, "Clustering objects on subsets of attributes," J. R. Stat. Soc. Series B Stat. Methodol., vol. 66, no. 4, pp. 815–839, 2004