Clustering Patient Medical Records via Sparse Subspace Representation

Budhaditya Saha[1], Duc-Son Pham[2], Dinh Phung[1], and Svetha Venkatesh[1,2]

[1] Center for Pattern Recognition and Data Analytics
School of Information Technology, Deakin University, Geelong, Australia
[2] Institute for Multi-sensor Processing and Content Analysis
Department of Computing, Curtin University, Western Australia
budhaditya.saha@deakin.edu.au

Abstract. The health industry is facing increasing challenge with "big data" as traditional methods fail to manage the scale and complexity. This paper examines clustering of patient records for chronic diseases to facilitate a better construction of care plans. We solve this problem under the framework of subspace clustering. Our novel contribution lies in the exploitation of sparse representation to discover subspaces automatically and a domain-specific construction of weighting matrices for patient records. We show the new formulation is readily solved by extending existing ℓ_1-regularized optimization algorithms. Using a cohort of both diabetes and stroke data we show that we outperform existing benchmark clustering techniques in the literature.

Keywords: subspace clustering, medical data, sparse representation.

1 Introduction

Traditional methods fail to manage the scale and complexity of "big data". The health sector is at the epicenter of this "big data" - data on admissions, diagnosis, outcomes, spanning a bewildering and disconnected web of images, computerized records and registries. There are no systems to manage this big data. The result is "write only data", mostly unused. Critically it has potential to identify critical safety issues, as well as service and clinical efficiency. This paper explores the pressing need, to construct data analytic to inform such clinical decisions. The outcomes are critically important from economic, patient safety and systems perspectives.

Historically, classical statistical methods have been used to verify stated hypotheses. This requires a priori assumption, for example, on data distributions. As the scale, distribution and diversity of data increase, this approach leads to sub-optimal use of this information. This paper examines new ways to analyze cohorts of patients with chronic diseases, such as Diabetes mellitus (diabetes) and stroke. Chronic care is expensive to administer. One crucial problem in the management of chronic patients is to deliver care plans, such that in majority of cases patients can be manged in the community without hospitalization.

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 123–134, 2013.

This requires us to find sub-groups of patients with same disease characteristics, without any prior assumptions on grouping.

Considering the complexity and nature of the datasets, we propose to model the data by a union of subspaces [1] where each subspace corresponds to patients with similar diagnostic conditions. This model has been used in many applications, such as lossy compression of images [2][3], motion segmentation in video sequences [4,5,6,7,8] etc. Early subspace clustering methods include mixture of Gaussian, factorization, algebraic, compressed sensing/low-rank [9] methods, and examples range from K subspaces [10], mixture of probabilistic PCA [11], multi-stage learning [12] etc. These algorithms typically require prior knowledge about the subspaces - the number of subspaces or their dimensions [13]. The computation is also exponential with the number and/or dimensions. Recently, Elhamifar and Vidal [13] propose sparse subspace clustering (SSC), in which the clustering is solved by seeking a sparse representation of data points. By computing an affinity graph on the sparse representations for all data points, SSC automatically discovers the subspaces and their dimensions. However, the previous results by SSC show that there are many instances in which the sparse coefficients corresponding to points outside a cluster of interest are significantly non-zero. This suggests that enforcing constraints that discourage points further apart will prevent them from entering the same cluster [14]. This was also exploited in [5], who propose a weighted version (WL-SSC).

Inspired by the related success of sparse subspace clustering in computer vision, this paper proposes a novel application of this powerful approach in the context of health care data. Here, it requires a careful modeling and interpretation of subspaces in health care data as well as novel construction of weighting matrices. The weighting matrix acts as the prior knowledge on the similarity between patient records and is computed directly from the data. We explore the decomposition into union of linear subspaces (WL-SSC) and extend the model to consider decomposition of a union of affine subspaces (WA-SSC). To decide on the weighting constraints, we consider three different ways of specifying proximity of points in a k-neighborhood - RBF, cosine and 0-1 matrix. We apply the models across a cohort of 1580 diabetes patients with 551 disease codes, and 1159 stroke patients with 805 codes. The data is collected over a period of 5 years, and each time the patient comes to hospital, a diagnosis code is assigned. Evaluation of such algorithms, with real-world data is notoriously hard. We propose the use of the recently introduced ρ-measure- this method allows ground-truth to be allocated based on degree of similarity between two points. Using this measure, we can compute the Rand-Index and F-measure for a given ρ. We show that our methods outperform the unweighted version and many competitive clustering methods such as affinity propagation (AP) [15], locality-preserving projection (LPP)[16] and k-means [17]. We show that further improvement can be achieved with a weighted union of affine subspace model. We also show tag clouds for clusters in the diabetes cohort and demonstrate how the sub-groups discovered are qualitatively meaningful.

The novelty in our paper is threefold: (a) it applies weighted sparse subspace clustering to a unique medical dataset problem to improve service efficiency, (b) it proposes a new affine, weighted subspace clustering method, and (c) uses a novel principled way to evaluate real world clustering results for which no ground-truth can be obtained.

The significance of the problem lies in the ability to save costs with efficient sub-group identification, leading to targeted care plans. Both chronic diseases chosen have reached epidemic status. For example, Diabetes mellitus (diabetes) is spreading so rapidly that recent studies show that the total number of diabetic people across the world was 171 million in 2001 and it is estimated to 230 million by the end of 2030 [18,19].

2 Related Background

2.1 Sparse Subspace Clustering (SSC)

Consider a set of N data points collected in a $D \times N$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbb{R}^D$ and D is the number of features. SSC [4] clusters the datapoints via the subspace principle. Intuitively, a linear representation of a datapoint with respect to the whole set gives more preferences to those points that belong to the same subspace. Denote as $\mathcal{S}_i$ the subspace (cluster) that $\mathbf{x}_i$ belongs to. Then, the linear representation of a datapoint can be written as follows:

$$\mathbf{x}_i = \sum_{j \neq i} c_{ij}\mathbf{x}_j = \sum_{i \in \mathcal{S}_i, j \neq i} c_{ij}\mathbf{x}_j + \sum_{j \notin \mathcal{S}_i} c_{ij}\mathbf{x}_j = \mathbf{X}\mathbf{c}_i \tag{1}$$

Here, $\mathbf{c}_i \doteq [c_{i1}, c_{i2}, \ldots, c_{iN}]^T$ are the coefficients of the representation. In the ideal case, the coefficients in the second summation of the right term are zero, giving rise to sparse coefficient vector $\mathbf{c}_i$. However, the solution of (1) is generally not unique when the number of features D is usually much less than the number of observations N. Recent advances in sparse learning [20,21] show that it is possible to regularize the solution and at the same time achieve sparse solution, which is consistent to the ideal case, by enforcing the ℓ_1-norm of the coefficient vector, $\|\mathbf{c}_i\|_1 = \sum |c_{ik}|$, to be small. Using this principle, SSC [4] advocates to find the solution with two variations as follows.

Linear Sparse Subspace formulation (L-SSC). Under this formulation, we assume that data points in $\mathbf{X}$ are sampled from a union of linear subspaces. Then the sparse coefficients are obtained by solving following optimization problem without employing any others constraints on coefficient vector $\mathbf{c}_i$.

$$\arg\min_{\mathbf{c}_i} ||\mathbf{c}_i||_1 \quad \text{s.t. } \mathbf{x}_i = \mathbf{X}\mathbf{c}_i, \;\; c_{ii} = 0 \tag{2}$$

Affine Sparse Subspace formulation (A-SSC). L-SSC can be extended to union of affine subspaces by enforcing an additional equality constraint over the sparse coefficient vector $\mathbf{c}_i$ as follows:

$$\arg\min_{\mathbf{c}_i} ||\mathbf{c}_i||_1 \quad \text{s.t. } \mathbf{x}_i = \mathbf{X}\mathbf{c}_i, \;\; \mathbf{c}_i^T\mathbf{1} = \mathbf{1} \quad c_{ii} = 0 \tag{3}$$

The coefficients are then used to compute a balanced affinity matrix for final spectral clustering: $\bar{\mathbf{C}} = (\mathbf{C} + \mathbf{C}^T)/2$. Then, the Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\bar{\mathbf{C}}\mathbf{D}^{-1/2}$ is computed, with $\mathbf{I}$ being the identity matrix and $\mathbf{D}$ being a diagonal matrix where $\mathbf{D}_{ii} = \sum_{j=1}^{N} \bar{c}_{ij}$. The smallest eigenvalues of $\mathbf{L}$ is used to estimate number of subspaces and the corresponding data points are obtained using the k-means algorithm.

3 Proposed Method

3.1 Weighted Sparse Subspace Clustering (W-SSC)

In the ideal case, the coefficients c_{ij} are zero if data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are sampled from two different subspaces. However, there are cases where they significantly deviate from zero due to numerical properties of the data matrix $\mathbf{X}$ [5]. To avoid undesirable sparse solutions, it has been suggested to introduce a weighting scheme in the sparse formulation [5]. Under this scheme, a weight matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is used to enforce sparse coefficients to better fall into the same subspace they deem to belong to. Such a desired solution is encouraged by minimizing the weighted ℓ_1-norm $\|\mathbf{w}_i \odot \mathbf{c}_i\|_1$ instead of $\|\mathbf{c}_i\|_1$. Here, $\odot$ denotes element-wise product of two vectors. Inspired by this principle, we also propose to employ the weighting scheme in our method. The remaining challenge is to construct a suitable weighting matrix for the data, which we detail next.

3.2 Construction of Weighting Matrix W

An optimal weighting matrix can be constructed if we have ground-truth knowledge of the clusters to suppress cross-cluster coefficients (by setting w_{ij} large or small for inter- or intra-cluster coefficients respectively). However, as this knowledge is not available, we propose to use the information within the data to approximate the optimal weighting matrix. We rely on the principle that the weights for inter-cluster coefficients are large whilst those for intra-cluster coefficients are small. Denote as $\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j)$ as $\mathbf{x}_i$ is k-nearest neighbor of $\mathbf{x}_j$, and $\mathbb{I}$ the indicator (0/1) function. We propose the following choices:

- **Inverse RBF** : $w_{ij} = \mathbb{I}_{\mathbf{x}_i \notin \mathcal{N}(\mathbf{x}_j)} \times \exp^{\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2\sigma^2}}$
- **0-1** : $w_{ij} = \mathbb{I}_{\mathbf{x}_i \notin \mathcal{N}(\mathbf{x}_j)}$
- **Cosine**: $w_{ij} = \mathbb{I}_{\mathbf{x}_i \notin \mathcal{N}(\mathbf{x}_j)} \times \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{||\mathbf{x}_i||_2 ||\mathbf{x}_j||_2}$

3.3 Weighted Formulation

Extending the basic SSC algorithms, we propose to adapt to the idea in [5] and solve the following *basic* weighted formulation with linear subspace assumption

$$\arg\min_{\mathbf{c}_i} ||\mathbf{w}_i \odot \mathbf{c}_i||_1, \quad \text{s.t.} \mathbf{x}_i = \mathbf{X}\mathbf{c}_i, c_{ii} = 0 \tag{4}$$

The above basic formulation assumes noiseless data generation. Considering noise while modeling the data points sampled from the union of subspaces, we assume that each data points $\mathbf{x}_i$ is contaminated with noise $\mathbf{e}_i$. i.e. $\mathbf{x}_i = \mathbf{x}_i^{true} + \mathbf{e}_i$ where $\mathbf{x}_i^{true}$ is the true value of the i-th variable and $\mathbf{e}_i$ is bounded: $||\mathbf{e}_i||_2 \leq \epsilon$. Thus, it is more realistic to extend the basic model to account for noise by considering the noise-aware version of the formulation

$$\arg\min_{c_i} ||\mathbf{w}_i \odot \mathbf{c}_i||_1 \quad \text{s.t. } ||\mathbf{x}_i - \mathbf{X}\mathbf{c}_i||_2^2 \leq \epsilon, \quad c_{ii} = 0 \tag{5}$$

This can be more conveniently written in a Lagrangian form

$$\arg\min_{\mathbf{c}_i} \lambda||\mathbf{w}_i \odot \mathbf{c}_i||_1 + \frac{1}{2}||\mathbf{x}_i - \mathbf{X}^{-i}\mathbf{c}_i||_2^2 \tag{6}$$

Here, λ is regularization parameter, $\mathbf{X}^{-i}$ is $\mathbf{X}$ with the ith column removed, and we implicitly ignore the i^{th} entry of $\mathbf{c}_i$. When considering the affine subspace modeling, the above Lagrangian formulation can be extended to account for the additional affine constraints as follows

$$\arg\min_{\mathbf{c}_i} \lambda||\mathbf{w}_i \odot \mathbf{c}_i||_1 + \frac{1}{2}||\mathbf{x}_i - \mathbf{X}^{-i}\mathbf{c}_i||_2^2, \ \ \mathbf{c}_i^T\mathbf{1} = 1, \tag{7}$$

Next, we discuss optimization algorithms to solve (7) (note that (6) can be readily solved by a slight modification of many efficient compressed sensing solver, such as reweighting the column of $\mathbf{X}^{-i}$ by the inverse of the corresponding weights and working on the reweighted variables [5]). As they are convex problems, off-the-shelf solvers, such as CVX, can be used, but we do not seek to use them because they are rather inefficient. We show that it is possible to solve (7) more efficiently with the alternative direction method of multipliers (ADMM) [22]. For notational simplicity, we drop the subscript/superscript of $\mathbf{c}_i, \mathbf{x}_i$ and $\mathbf{X}^{-i}$. Under the ADMM framework, we decouple the ℓ_1 regularization term from the quadratic terms by introducing a new variable $\mathbf{z}$ such that $\mathbf{z} - \mathbf{c} = 0$ and consider the augmented Lagrangian

$$\begin{aligned}\mathcal{L}(\mathbf{c}, \mathbf{z}, \mathbf{y}, v) = \frac{1}{2}\|\mathbf{x} - \mathbf{X}\mathbf{c}\|_2^2 + \lambda\|\mathbf{z}\|_1 + \mathbf{y}^T(\mathbf{c} - \mathbf{z}) + \frac{\rho_1}{2}\|\mathbf{c} - \mathbf{z}\|_2^2 \\ + v(\mathbf{1}^T\mathbf{c} - 1) + \frac{\rho_2}{2}(\mathbf{1}^T c - 1)^2.\end{aligned} \tag{8}$$

Here, $\mathbf{y}$ and v are the dual parameters corresponding to the inequality constraints $\mathbf{c} - \mathbf{z} = 0$ and $\mathbf{1}^T\mathbf{c} - 1 = 0$ respectively; ρ_1 and ρ_2 are small parameters to improve numerical stability (see [22] for ADMM background). By using the normalized dual variables $\mathbf{u}_1 = (\mathbf{y}/\rho_1)$ and $u_2 = (v/\rho_2)$ we derive the following ADMM updates that solve (7)

$$\mathbf{c}^{k+1} = (\mathbf{X}^T\mathbf{X} + \rho_1\mathbf{I} + \rho_2\mathbf{1}\mathbf{1}^T)^{-1}(\mathbf{X}^T\mathbf{x} + \rho_1(\mathbf{z}^k - \mathbf{u}_1^k) + \rho_2\mathbf{1}(1 - u_2)) \quad (9)$$

$$\mathbf{z}^{k+1} = \mathsf{S}_{\lambda/\rho_1}(\mathbf{c}^{k+1} + \mathbf{u}_1) \quad (10)$$

$$\mathbf{u}_1^{k+1} = \mathbf{u}_1^k + (\mathbf{c}^{k+1} - \mathbf{z}^{k+1}) \quad (11)$$

$$u_2^{k+1} = u_2^k + (\mathbf{1}^T\mathbf{c}^{k+1} - 1). \quad (12)$$

Here $\mathsf{S}_{\boldsymbol{\tau}}(\mathbf{c})$ is the soft-thresholding shrinkage operator, defined as a vector r such that $r_i = \mathsf{sign}(c_i)\max(|c_i| - \tau_i, 0)$ (see [22]).

Once the coefficient vectors $\mathbf{c}_i$'s are found, the spectral clustering part proceeds in the same way as the original SSC algorithm [4].

4 Experiments

4.1 Datasets

We validate our approach on two real-world datasets collected from patients having diabetes and heart (stroke) diseases collected over a period of five years from 2007 to 2011 and has diagnosis records from 9878 patients. Each patient has been diagonised several times over a period of five years and assigned unique diagnosis code(s). An example of a record for a patient over time might be (`E1172, I10, E1172, Z9222`). Table 1 and 2 shows the description of some codes. Patients may be assigned similar code more than once over time.

We remove records without codes, patients diagonised less than twice and also duplicated codes. This results in 1580 diabetes patients with 551 unique codes. We construct a code-patient matrix, where codes are used as features and each patient is an observation, analogous to term-document matrix for text data analysis. In our second data set (stroke patients), there are 1159 patients with 805 diagnostic codes.

4.2 Evaluation Method

As no ground-truth is available for latent groups, it is impossible to measure the clustering performance by standard evaluation metrics. Thus, we evaluate the performance using a novel ρ-measure method as follows:

1. Each data point $\mathbf{x}_i \in \mathbb{R}^N$ is mapped to a binary vector $\bar{\mathbf{x}}_i$ where $\bar{x}_{ij} = \mathbb{I}_{x_{ij} \neq 0}$.
2. Compute relative similarity metric $s_\rho(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j)$

$$s_\rho(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \frac{\sum(\bar{\mathbf{x}}_i \odot \bar{\mathbf{x}}_j)}{\sum_{k=1}^{N} \bar{\mathbf{x}}_{ik} + \sum_{k=1}^{N} \bar{\mathbf{x}}_{jk} - \sum(\bar{\mathbf{x}}_i \odot \bar{\mathbf{x}}_j)} \quad (13)$$

3. Construct a ground-truth matrix $\mathbf{G}_\rho \in \mathbb{R}^{N \times N}$ with element $g_{ij} = \mathbb{I}_{s_\rho(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) \geq \rho}$
4. Construct a cluster membership matrix $\mathbf{V}$ with element $v_{ij} = \mathbb{I}_{ID_K(i) = ID_K(j)}$

Table 1. Examples of code description

Codes	Description of Codes
E1172	Type 2 diabetes mellitus with features of insulin resistance
I10	Essential (primary) hypertension
Z9222	Personal history of long-term (current) use of other medicament, insulin
R63Z	Chemotherapy

Table 2. Diabetes dataset

Patient Id	Diagnosis Codes
P1	E1172,I10,E1172,Z9222
P2	M81403,Z511,R63Z,R63Z
P3	E1023,E1023,E1012

Next, we compute the standard*Precision* (P), *Recall* (R) and *F-measure* (F):

$$P = \frac{TP}{TP + FP}, \; R = \frac{TP}{TP + FN}, \; F = \frac{2 \times P \times R}{P + R} \tag{14}$$

Here, true positive (TP) is scored when two similar data points in the ground-truth are grouped together in the obtained results, a true negative (TN) is scored when two dissimilar data points are grouped separately, a false positive (FP) is scored when two dissimilar data points are grouped together and a false negative (FN) is scored when two similar data points are grouped separately. Similarly, the rand index (RI) is defined as

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where high RI and F indicates the better accuracy.

Algorithm 1 show the overall method of computing F-measure. Note that, we compute F measure over a matrix of $N \times N$ variables, instead of N number of data points.

4.3 Results and Comparisons

Performance against Other Methods. We compare our proposed clustering method against competitive sparse subspace clustering and baseline alternatives, including affinity propagation (AP) [15], locality preserving projection (LPP) [16], and k-means [17]. In all experiments, we set ρ to 0.9, regularization parameter λ to 0.001.

Table 3 presents the clustering results obtained from SSC methods for diabetes and stroke data. Clearly, our proposed method outperforms both L-SSC and A-SSC variants by obtaining larger RI and F scores. The F measure scores of WL-SSC and WA-SSC have improved over L-SSC and A-SSC by large margins of **47**% and **45**% for the diabetes data and **236**% and **257**% for the stroke data respectively.

Algorithm 1. Computing F measure

```
Input: Groundtruthed Matrix $\mathbf{G}_\rho$ and Cluster Index matrix $\mathbf{V}$.
Output: $F -$ measure
Intialize: Set TP=TN=FP=FN=0.

– for $i = 1$ to $N$
    • for $k =$ 1to $N$
        ∗ if ($\mathbf{g}_{ik} = 1$) and ($\mathbf{v}_{ik} = 1$) $TP = TP + 1$; // Two similar data points grouped together.
        ∗ else if ($\mathbf{g}_{ik} = 0$) and ($\mathbf{v}_{ik} = 0$) $TN = TN + 1$; // Two dissimilar data points grouped separately.
        ∗ else if ($\mathbf{g}_{ik} = 0$) and ($\mathbf{v}_{ik} = 1$) $FP = FP+1$;//Two dissimilar data points grouped similar.
        ∗ else ($\mathbf{g}_{ik} = 1$) and ($\mathbf{v}_{ik} = 0$) $FN = FN + 1$;//Two similar data points grouped separately.
    • end
– end
– Calculate $F$-measure following the equation 14.
```

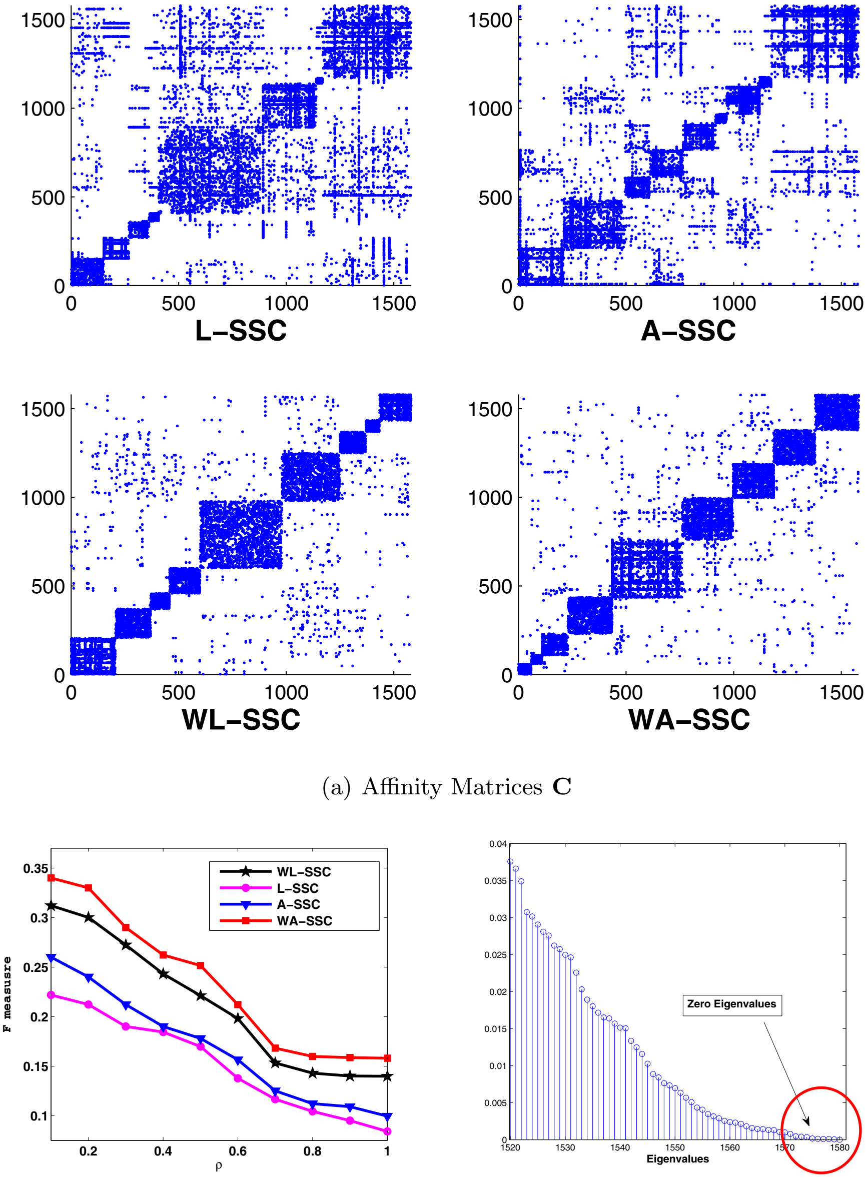
Likewise, the F measure is improved by **275**% (AP), **85**% (LPP), **388**% (k-means) for diabetics datasets, whereas the betterment in RI is **87**% (AP), **14**% (LPP), **10**% (k-means) respectively. For the strokes data, F measure is improved by **173**% (AP), **54**% (LPP), **465**% (k-means) and *Rand Index* is **71**% (AP), **13**% (LPP), **139**% (k-means) respectively.

Table 3. Performance comparison

Datasets	Diabetics Data		Strokes Data	
Methods	F measure	Rand Index	F measure	Rand Index
AP	0.0423	0.4639	0.062	0.522
LPP	0.0854	0.7654	0.11	0.8045
k-means	0.0325	0.4312	0.0294	0.3845
L-SSC	0.0951	0.7817	0.0475	0.5210
A-SSC	0.1092	0.7862	0.0619	0.7324
WL-SSC	**0.1401**	**0.8652**	**0.1597**	**0.90**
WA-SSC	**0.1587**	**0.8982**	**0.1697**	**0.91**

Table 4. Performance analysis using different weighting schemes

Datasets	Diabetes Data		Strokes Data	
Weighting Schemes	WL-SSC	WA-SSC	WL-SSC	WA-SSC
RBF	**0.1401**	**0.1587**	**0.1597**	**0.1697**
0-1	0.1199	0.1221	0.1191	0.1201
cosine	0.1352	0.1444	0.1390	0.1382

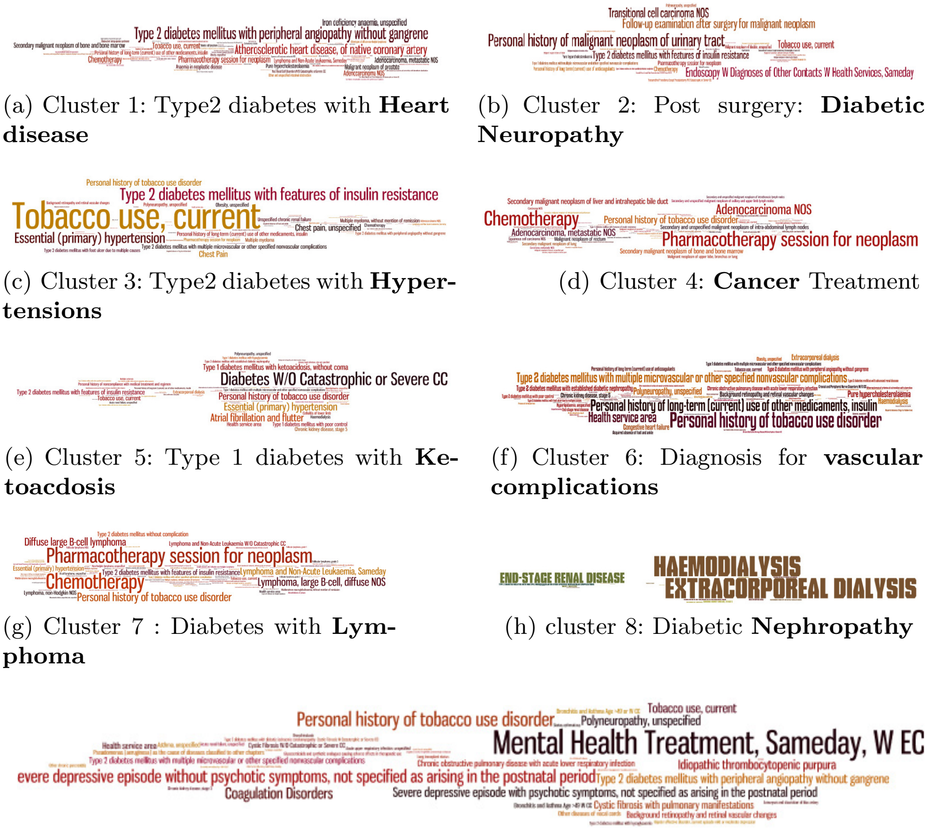


(a) Affinity Matrices **C**

(b) *F*-measure

(c) Eigenvalues of **L**

Fig. 1. Plots for Qualitative Evaluations



(a) Cluster 1: Type2 diabetes with **Heart disease**

(b) Cluster 2: Post surgery: **Diabetic Neuropathy**

(c) Cluster 3: Type2 diabetes with **Hypertensions**

(d) Cluster 4: **Cancer** Treatment

(e) Cluster 5: Type 1 diabetes with **Ketoacdosis**

(f) Cluster 6: Diagnosis for **vascular complications**

(g) Cluster 7 : Diabetes with **Lymphoma**

(h) cluster 8: Diabetic **Nephropathy**

(i) Cluster 9: Diabetes with **Psychiatric Disorders**

Fig. 2. Diagnostic Clouds

Influence of Weighting Schemes. Table 4 include the performance for different weighting schemes and it is found that the RBF choice provides better performance than the other choices.

Discovered Clusters. The number of clusters K equals to the number of zero eigenvalues of of Laplacian matrix $\mathbf{L}$. Fig. 1(c) shows the eigenvalue plot of $\mathbf{L}$ for the diabetes data where the number of zero eigenvalue equals to 9. Similarly, we found 12 sub-groups for stroke data.

Since ρ is the relative similarity between the two data points, which means high value of ρ denotes two observations are highly similar, we vary ρ varies from 0.1 to 1 in a separate experiment on diabetes data and plots are shown in . Figure 1(b). As expected, F-measure is high for small values of ρ and F-measure is low when ρ is increasing.

Figures 1 and 2 show the qualitative evaluation of clusters for the diabetes data. Figure 1(a) shows the affinity matrices, whilst Figure 2 shows the tag clouds of the diagnosis codes in each cluster. As anticipated the clusters are qualitatively different in terms of disease differentiation within diabetes: diabetes with heart disease, with cancer, with dialysis. Type 1 and 2 are clearly differentiated.

5 Conclusion

We have demonstrated a novel application of the sparse subspace clustering theory in solving the clustering problem of health care data. Our novel contributions includes special construction of the weighting matrices to obtain better sparse solution and the efficient algorithm to solve the formulation with affine constraints. To evaluate realistic health care data where no ground-truth is available, we have also suggested a novel evaluation method of clustering results. Compared with competitive alternatives in the literature, our proposed method achieve much better F and RI scores, and discovers meaningful patients subgroups.

References

1. Eldar, Y., Mishali, M.: Robust recovery of signals from a structured union of subspaces. IEEE Transactions on Information Theory 55(11), 5302–5316 (2009)
2. Hong, W., Wright, J., Huang, K., Ma, Y.: A multiscale hybrid linear model for lossy image representation. In: Proc. ICCV, pp. 764–771 (2005)
3. Yang, A., Wright, J., Ma, Y., Sastry, S.: Unsupervised segmentation of natural images via lossy data compression. Computer Vision and Image Understanding 110(2), 212–225 (2008)
4. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: Proc. CVPR, pp. 2790–2797. IEEE (2009)
5. Pham, D.-S., Saha, B., Phung, D., Venkatesh, S.: Improved subspace clustering via exploitation of spatial constraints. In: Proc. CVPR. IEEE (2012)
6. Wang, S., Yuan, X., Yao, T., Yan, S., Shen, J.: Efficient subspace segmentation via quadratic programming. In: Proc. AAAI (2011)
7. Yu, Y., Schuurmans, D.: Rank/norm regularization with closed-form solutions: Application to subspace clustering. Arxiv preprint arXiv:1202.3772 (2012)
8. Vidal, R., Tron, R., Hartley, R.: Multiframe motion segmentation with missing data using power factorization and GPCA. IJCV 79(1), 85–105 (2008)
9. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: Proc. ICML (2010)
10. Ho, J., Yang, M., Lim, J., Lee, K., Kriegman, D.: Clustering appearances of objects under varying illumination conditions. In: Proc. CVPR, vol. 1, pp. I–11. IEEE (2003)
11. Tipping, M., Bishop, C.: Mixtures of probabilistic principal component analyzers. Neural Computation 11(2), 443–482 (1999)
12. Gruber, A., Weiss, Y.: Multibody factorization with uncertainty and missing data using the EM algorithm. In: Proc. CVPR (2004)
13. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: Proc. CVPR, pp. 2790–2797 (2009)

14. Candes, E., Wakin, M., Boyd, S.: Enhancing sparsity by reweighted l1 minimization. Journal of Fourier Analysis and Applications 14(5), 877–905 (2008)
15. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science 315(5814), 972–976 (2007)
16. He, X., Cai, D., Liu, H., Ma, W.: Locality preserving indexing for document representation. In: Proc. ACM SIGIR, pp. 96–103 (2004)
17. Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 881–892 (2002)
18. Fabris, P., Floreani, A., Tositti, G., Vergani, D., De Lalla, F., Betterle, C.: Type 1 diabetes mellitus in patients with chronic hepatitis c before and after interferon therapy. Alimentary Pharmacology & Therapeutics 18(6), 549–558 (2003)
19. Young, J., McAdam-Marx, C.: Treatment of type 1 and type 2 diabetes mellitus with insulin detemir, a long-acting insulin analog. Clinical Medicine Insights. Endocrinology and Diabetes 3, 65 (2010)
20. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52(2), 489–509 (2006)
21. Donoho, D.: Compressed sensing. IEEE Transactions on Information Theory 52(4), 1289–1306 (2006)
22. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. In: Jordan, M. (ed.) Foundations and Trends in Machine Learning, vol. 3(1), pp. 1–122. Now Publisher (2011)