2013 IEEE 13th International Conference on Data Mining

Exploring Patient Risk Groups with Incomplete Knowledge

Xiang Wang, Fei Wang, Jun Wang, Buyue Qian, Jianying Hu
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{*wangxi, fwang, wangjun, bqian, jyhu*}*@us.ibm.com*

***Abstract*—Patient risk stratification, which aims to stratify a patient cohort into a set of homogeneous groups according to some risk evaluation criteria, is an important task in modern medical informatics. Good risk stratification is the key to good personalized care plan design and delivery. The typical procedure for risk stratification is to first identify a set of risk-relevant medical features (also called risk factors), and then construct a predictive model to estimate the risk scores for individual patients. However, due to the heterogeneity of patients' clinical conditions, the risk factors and their importance vary across different patient groups. Therefore a better approach is to first segment the patient cohort into a set of homogeneous groups with consistent clinical conditions, namely risk groups, and then develop group-specific risk prediction models. In this paper, we propose RISGAL (RISk Group AnaLysis), a novel semi-supervised learning framework for patient risk group exploration. Our method segments a patient similarity graph into a set of risk groups such that some risk groups are in alignment with (incomplete) prior knowledge from the domain experts while the remaining groups reveal new knowledge from the data. Our method is validated on public benchmark datasets as well as a real electronic medical record database to identify risk groups from a set of potential Congestive Heart Failure (CHF) patients.**

***Keywords*-Patient Risk Stratification; Risk Group Analysis; Electronic Medical Records; Semi-Supervised Learning**

I. INTRODUCTION

Personalized care is one of the major trends in modern medical informatics, where a key step is to segment the patient cohort into homogeneous groups so that a customized treatment plan can be constructed for each group. Patient risk stratification [1] can be viewed as a specific way of patient cohort segmentation such that patients in each group share similar risks of having an adverse outcome, e.g. the onset of Congestive Heart Failure (CHF).

A major challenge for risk stratification is the heterogeneity of patients' clinical conditions. For instance, CHF patients have different comorbidities, such as diabetes, kidney diseases, lung diseases, etc.. In different comorbidity groups, the medical features that contribute to the risk, or *risk factors*, are different. Even for the common risk factors across different patient groups, their contributions to the risk score could vary significantly. For example, asthma is a known risk factor for heart disease, but it will contribute much more to the heart disease risk for patients with (other) existing lung diseases than patients with diabetes. Therefore, constructing a universal risk prediction model using a shared set of risk factors may not be the best approach for risk stratification. It makes more sense to first segment the patient cohort into risk groups with consistent clinical conditions, and then construct the prediction model using customized risk factors from each group.

In order to accurately segment the patient cohort, we want to incorporate prior knowledge from domain experts (physicians). On the one hand, it is very important to incorporate these domain knowledge (often in the form of known risk factors) because they reflect crucial medical insights that are validated by extensive clinical studies. On the other hand, these knowledge are mostly incomplete because the domain experts can only provide guidance within their areas of expertise, which are unlikely to cover all the relevant medical aspects of any given patient cohort.

Based on the above considerations, we propose RISGAL (RISk Group anALysis), a novel semi-supervised learning framework for data- and knowledge-driven patient risk group exploration. The input of RISGAL is a graph with nodes as patients and edges as patient similarities, as well as a set of knowledge-driven risk factors (labels) provided by domain experts. The output will be a set of patient risk groups that align with those provided risk factors. The key challenge is that the label set is *incomplete*, i.e. there are unseen classes. It is worthwhile to highlight the following aspects of our proposed approach:

- Thanks to the semi-supervised learning scheme, RISGAL can discover risk groups that align with the given risk factors (labels) derived from domain knowledge.
- In the meanwhile, RISGAL can also discover data-driven risk groups that are not covered by the knowledge-driven risk factors.
- We propose an efficient algorithm based on *Block Coordinate Descent* (BCD) to solve the optimization problem of RISGAL. Our algorithm guarantees convergence to a local optimum.

We first justify the effectiveness of RISGAL on several public benchmark datasets. The empirical results validate the advantage of RISGAL as compared to existing methods. Then we apply RISGAL to a real-world electronic medical record database to stratify a set of patients with respect to their risk of CHF onset. We demonstrate that our algorithm is

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.129

able to identify both data- and knowledge-driven risk groups with rich clinical insights.

II. Related Works

Graph-based semi-supervised learning *with unseen classes* has not been well studied in the literature. Nie et al. [2] proposed a variation of the *Learning with Local and Global Consistency* (LLGC) algorithm [3] in which they initialize the algorithm by assigning all unlabeled nodes to a new class. After LLGC converges, the nodes that remain in the new class are considered as a novel class. The limitation of their approach is that it can only handle one novel class.

In a broader context, the problem of unseen classes has been addressed in both semi-supervised and supervised learning settings. For example, PU learning [4] considers the binary classification problem and uses only positive and unlabeled samples to train the classifier. Zero-shot learning [5], [6] was proposed in the computer vision society and it uses the semantic-relatedness between instance features to discover novel object classes. Other criteria used to identify novel classes from the unlabeled data include maximum margin [7] and maximum entropy [8]. The key difference between our work (and [2]) and the aforementioned techniques is that the latter are not graph-based, i.e. they require a unified vector space representation or some other auxiliary information (in the case of [8], auxiliary classes), whereas RISGAL takes a similarity graph as input.

III. The Proposed Framework

A. Objective Function

Assume we have a set of n patients with their similarity matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$, whose (i,j)-th entry encodes the clinical similarity between patient i and patient j. $\mathbf{W}$ is symmetric. Let Δ be the corresponding normalized graph Laplacian. Suppose we have c knowledge-driven risk factors, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_c] \in \{0,1\}^{n\times c}$ encodes their association to the patients, i.e., $\mathbf{y}_{ij} = 1$ means patient i has risk factor j (so that patient i belongs to risk group j, note that such group assignment can be overlapping, i.e., one patient can belong to multiple groups based on the risk factors he/she has), $\mathbf{y}_{ij} = 0$ otherwise. Let $\mathcal{L} \subset \{1, \ldots, n\}$ denote the index set of labeled patients and c' be the total number of risk groups. We assume $c' > c$, i.e. some risk groups are unseen with unknown risk factors. Let $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_c] \in \{0,1\}^{n\times c}$ be the patient assignment matrix to the knowledge-driven risk groups, and $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_{c'}] \in \{0,1\}^{n\times c'}$ be the patient assignment matrix to all potential risk groups.

We design the following objective for RISGAL:

$$\begin{aligned}\mathcal{J} = \alpha \sum_{k=1}^{c} \|\mathbf{f}_k - \mathbf{y}_k\|_{\mathcal{L}}^2 + \beta \sum_{k=1}^{c} \mathbf{f}_k^T \Delta \mathbf{f}_k \\ + \gamma \sum_{l=1}^{c'} \mathbf{g}_l^T \Delta \mathbf{g}_l - \mu \sum_{k=1}^{c}\sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{f}_k \mathbf{f}_k^T)\mathbf{g}_l\end{aligned} \tag{1}$$

$\alpha, \beta, \gamma, \mu > 0$ are all weighting parameters. Our goal is to minimize $\mathcal{J}$. The following section will introduce the meaning of each term in $\mathcal{J}$.

B. Interpretation and Discussions

Fitting Term: $\alpha \sum_{k=1}^{c} \|\mathbf{f}_k - \mathbf{y}_k\|_{\mathcal{L}}^2$. Note that $\mathbf{F}$ is the assignment of patients to the c knowledge-driven risk groups. This term governs how well $\mathbf{F}$ must fit the input knowledge $\mathbf{Y}$. The subscript $\mathcal{L}$ means the fitting only applies to labeled patients. α decides how much $\mathbf{F}$ can deviate from $\mathbf{Y}$. When $\alpha \to \infty$, the known labels are not allowed to be altered.

Smoothing Term: $\beta \sum_{k=1}^{c} \mathbf{f}_k^T \Delta \mathbf{f}_k$. This term enforces the neighborhood assumption of semi-supervised learning, i.e. if two patients are highly similar in the graph then they are likely to belong to the same risk group. Larger β will bias $\mathbf{F}$ more towards the graph structure as encoded by Δ.

Grouping Term: $\gamma \sum_{l=1}^{c'} \mathbf{g}_l^T \Delta \mathbf{g}_l$. Note that $\mathbf{G}$ is the assignment of patients to all c' potential risk groups. This term represents the data-driven exploration of the graph structure Δ. γ decides how much $\mathbf{G}$ will be biased towards the normalized min-cut of the graph.

Matching Term: $-\mu \sum_{k=1}^{c}\sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{f}_k \mathbf{f}_k^T)\mathbf{g}_l$. This term *maximizes* (note the negative sign before μ) the agreement between assignment $\mathbf{F}$ and assignment $\mathbf{G}$ in terms of pairwise relations. The value of $\sum_{k=1}^{c}\sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{f}_k \mathbf{f}_k^T)\mathbf{g}_l$ is the total number of patient pairs whose relation $\mathbf{F}$ and $\mathbf{G}$ agree on. μ decides how close $\mathbf{G}$ and $\mathbf{F}$ must be to each other.

If we treat $\mathbf{F}$ and $\mathbf{G}$ as two groups of variables, we can adopt a *Block Coordinate Descent* (BCD) type of approach to solve it. This approach is an iterative method such that at each iteration, we fix either $\mathbf{F}$ or $\mathbf{G}$ and minimizing $\mathcal{J}$ with respect to the other. In our case, fixing $\mathbf{G}$ solving $\mathbf{F}$ leads to graph transduction, and fixing $\mathbf{F}$ solving $\mathbf{G}$ leads to normalized min-cut. Unfortunately, solving either step of the alternating minimization process is NP hard in their original form. In the following section we show how to relax the objective to allow an efficient solution.

C. Efficient Solution

In this section, we show how to solve a relaxed version of Eq.(1) efficiently. Our algorithm is summarized in Algorithm 1.

First we relax $\mathbf{F}$ and $\mathbf{G}$ from binary assignment to soft assignment. The relaxed objective becomes:

$$\begin{aligned}\underset{\mathbf{f}_k, \mathbf{g}_l \in \mathbb{R}^n}{\operatorname{argmin}} \; & \alpha \sum_{k=1}^{c} \|\mathbf{f}_k - \mathbf{y}_k\|_{\mathcal{L}}^2 + \beta \sum_{k=1}^{c} \mathbf{f}_k^T \Delta \mathbf{f}_k \\ & + \gamma \sum_{l=1}^{c'} \mathbf{g}_l^T \Delta \mathbf{g}_l - \mu \sum_{k=1}^{c}\sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{f}_k \mathbf{f}_k^T)\mathbf{g}_l \\ \text{s.t.} \; & \mathbf{G}^T\mathbf{G} = \mathbf{I}_{c'}, \; \mathbf{G} \geq 0.\end{aligned} \tag{2}$$

$\mathbf{I}_{c'}$ is a $c' \times c'$ identity matrix. The orthogonality constraint on $\mathbf{G}$ stops trivial solutions. Note that it is unnecessary to pose

the same constraint on $\mathbf{F}$ because $\mathbf{F}$ is already constrained by the fitting term to approximate $\mathbf{Y}$.

After relaxation, given a fixed $\mathbf{G}$, we solve for $\mathbf{F}$:

$$\underset{\mathbf{f}_k \in \mathbb{R}^n}{\operatorname{argmin}} \alpha \sum_{k=1}^{c} \|\mathbf{f}_k - \mathbf{y}_k\|_{\mathcal{L}}^2 + \beta \sum_{k=1}^{c} \mathbf{f}_k^T (\mathbf{I}_n - (\mathbf{S} + \frac{\mu}{\beta}\mathbf{G}\mathbf{G}^T))\mathbf{f}_k. \tag{3}$$

Zhou et al. [3] showed that this objective can be solved in closed form:

$$\mathbf{F} = (1-\rho)(\mathbf{I}_n - \rho(\mathbf{S} + \frac{\mu}{\beta}\mathbf{G}\mathbf{G}^T))^{-1}\mathbf{Y}, \tag{4}$$

where $\rho = \alpha/(\alpha+\beta)$ and $\mathbf{S} = \mathbf{I}_n - \Delta$.

Given a fixed $\mathbf{F}$, we solve for $\mathbf{G}$:

$$\begin{aligned} \underset{\mathbf{g}_l \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{l=1}^{c'} \mathbf{g}_l^T \Delta \mathbf{g}_l - \frac{\mu}{\gamma} \sum_{k=1}^{c} \sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{f}_k \mathbf{f}_k^T) \mathbf{g}_l, \\ \text{s.t. } \mathbf{G}^T\mathbf{G} = \mathbf{I}_{c'},\ \mathbf{G} \geq 0. \end{aligned} \tag{5}$$

Eq.(5) is equivalent to:

$$\begin{aligned} \underset{\mathbf{g}_l \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{l=1}^{c'} \mathbf{g}_l^T (\mathbf{S} + \frac{\mu}{\gamma}\mathbf{F}\mathbf{F}^T) \mathbf{g}_l, \\ \text{s.t. } \mathbf{G}^T\mathbf{G} = \mathbf{I}_{c'},\ \mathbf{G} \geq 0. \end{aligned} \tag{6}$$

Since $\mathbf{F}\mathbf{F}^T$ is a kernel, $\mathbf{S} + \frac{\mu}{\gamma}\mathbf{F}\mathbf{F}^T$ remains a positive semidefinite kernel. Eq.(6) is a standard graph min-cut objective with nonnegativity constraint and it can be solved by the multiplicative update rule [9]:

$$\mathbf{G} \leftarrow \mathbf{G} \odot \sqrt{\frac{(\mathbf{S} + \frac{\mu}{\gamma}\mathbf{F}\mathbf{F}^T)\mathbf{G}}{\mathbf{G}(\mathbf{G}^T(\mathbf{S} + \frac{\mu}{\gamma}\mathbf{F}\mathbf{F}^T)\mathbf{G})}}. \tag{7}$$

$\odot$ is Hadamard product. $\mathbf{G}$ can be initialized by the cluster assignment from performing spectral clustering on $\mathbf{S}$.

The alternating minimization process is guaranteed to converge because the objective in Eq.(2) is lower-bounded. The proof is omitted here due to page limit.

D. Implementation Issues

Setting β, γ, μ. Since we only care about the ratio μ/β and μ/γ, without loss of generality we can fix μ to 1. $1/\gamma > 0$ decides the influence of $\mathbf{F}\mathbf{F}^T$ on $\mathbf{S}$ in Eq.(6). Smaller γ will makes $\mathbf{G}$ biased more towards $\mathbf{F}$ rather than $\mathbf{S}$. To balance the influence of the two kernels ($\mathbf{S}$ and $\mathbf{F}\mathbf{F}^T$), notice that the most significant cut of $\mathbf{S}$ comes from its second largest singular vector (its largest singular vector is a constant vector) and the most significant cut of $\mathbf{F}\mathbf{F}^T$ comes from its largest singular vector. Let $\text{SVD}(\mathbf{X}, k)$ denote the function that returns the k-th largest singular value of $\mathbf{X}$, γ can be set to:

$$\gamma = \text{SVD}(\mathbf{F}\mathbf{F}^T, 1)/\text{SVD}(\mathbf{S}, 2). \tag{8}$$

This scales the influence of $\mathbf{F}\mathbf{F}^T$ to the same level of the normalized min-cut of $\mathbf{S}$. Similarly, the ratio $1/\beta$ controls the influence of $\mathbf{G}\mathbf{G}^T$ on $\mathbf{S}$ in Eq.(4). Since we want to preserve the given labels in $\mathbf{Y}$, in our implementation, we set β to a large number such that $1/\beta$ will be small, say 0.1.

Algorithm 1: RISGAL

Input: Similarity graph $\mathbf{W}$, input labels $\mathbf{Y} \in \{0,1\}^{n \times c}$, parameters c', β, γ, $\mu = 1$, ρ;
Output: Group indictor matrix $\mathbf{G} \in \mathbb{R}^{n \times c'}$;

```
Normalized the graph kernel: S ← D^{-1/2} W D^{-1/2}, where D is the degree matrix of W;
Compute the normalized Laplacian: Δ ← I_n − S;
Perform c'-way spectral clustering on S and initialize G ∈ {0,1}^{n×c'} as the corresponding group assignment matrix;
repeat
    G' ← G;
    F ← (1−ρ)(I_n − ρ(S + (μ/β) G^t G^{tT}))^{-1} Y;
    H ← G;
    repeat
        H' ← H;
        H ← H' ⊙ sqrt( (S + (μ/γ) F F^T) H' / (H' H'^T (S + (μ/γ) F F^T) H') );
    until ‖H − H'‖ < ε;
    G ← H;
until ‖G − G'‖ < ε;
return G;
```

Setting ρ. $\rho \in (0,1)$ is a tradeoff factor between the graph structure and the input labels. Larger ρ will make $\mathbf{F}$ biased more towards the normalized min-cut of $\mathbf{S} + \frac{\mu}{\beta}\mathbf{G}\mathbf{G}^T$. [3], [10] provided detailed discussion on how to choose ρ. In our implementation, we use a simple heuristic to set ρ:

$$\rho = (1 - \frac{|\mathcal{L}|}{n})a_1 + a_2,\ \forall a_1, a_2 \geq 0,\ a_1 + a_2 < \frac{\beta}{\beta + \mu c'}. \tag{9}$$

Eq.(9) bounds the value of ρ between a_1 and a_2 and the value of ρ will decreases when the number of labeled nodes increases (thus $\mathbf{F}$ must adhere more strictly to $\mathbf{Y}$).

Setting c'. Ideally, $c' > c$ is the true number of risk groups in the patient cohort. c' is usually set by domain experts. If sufficient domain knowledge is lacking, we could set c' in two different ways. One is to set $c' = c+1$, which essentially merges all risk groups into one meta-group. The other is to estimate c' through a regularizer.

Complexity. Inside each iteration, the complexity of our algorithm is dominated by that of LLGC (Eq.(4)) and nonnegative min-cut (Eq.(6)). The complexity of LLGC is dominated by computing the pseudoinverse of an $n \times n$ matrix, which is $O(n^3)$ in the worst case. The complexity of nonnegative normalized min-cut is $O(n^2k)$, where k is the number of iterations needed to converge. An extra $O(n^2c')$ time is needed to initialize $\mathbf{G}$ using c'-way spectral clustering.

Table I
BENCHMARK DATASETS USED IN OUR EXPERIMENTS

Identifier	#Instances	#Classes	#Unseen
Iris	150	3	1
Wine	178	3	1
Soybean	47	4	2
20 Newsgroups	1,759	4	1
USPS Digits	400	4	2
DBLP	421	4	2

IV. EMPIRICAL STUDY ON BENCHMARK DATASETS

In this section we justify the effectiveness of our algorithm on a variety of benchmark datasets with comparison to several existing techniques. The datasets we used in this section are all publicly available.

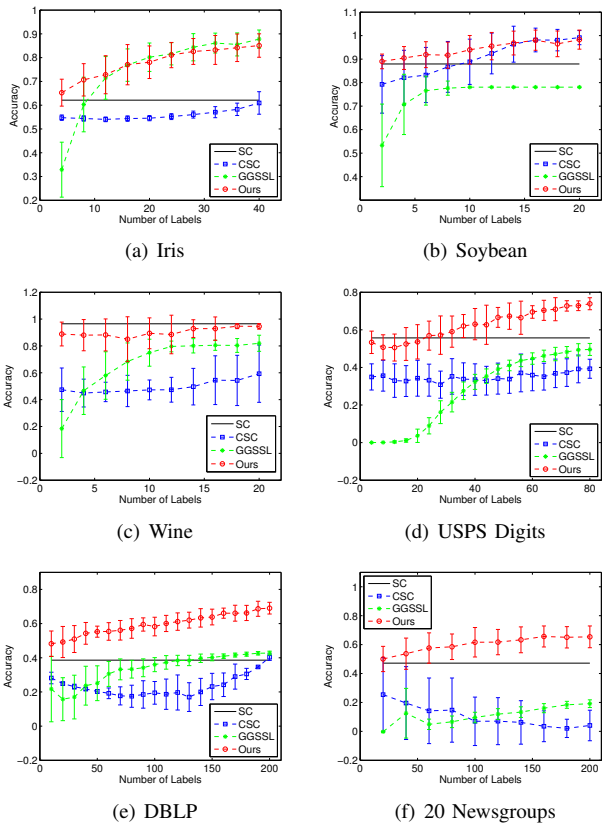
A. Methodology

The datasets we used are summarized in Table I. We used three scientific datasets from the UCI archive, namely Iris, Wine, and Soybean, a subset of the USPS Handwritten Digits, a subset of 20 Newsgroups, and co-author graph constructed from the DBLP dataset. For each dataset, we randomly chose a subset of ground truth labels as the training labels $\mathbf{Y}$. To simulate unseen class, we withheld labels from certain classes. For Iris, we kept the Setosa class hidden; for Wine, we kept Class 3 hidden; for Soybean, we kept D1 and D2 hidden. For the USPS dataset we picked digits 1, 2, 3, and 4 for our experiment and kept digit 1 and 3 hidden from the training labels. For the 20 Newsgroups, which contained 4 high-level topics (rec, comp, sci, and talk), we withheld comp from the training data. For the DBLP dataset, we collected authors and their papers from four areas of computer science, namely data mining (KDD, ICDM, SDM), machine learning (NIPS, ICML), database (SIGMOD, VLDB), and computer vision (CVPR, ICCV). We used the areas as class labels and withheld data mining and database from the training data. For the first five datasets, we used the RBF kernel to construct graphs. The optimal kernel bandwidth was chosen using cross validation. For the DBLP dataset we constructed a co-author graph where $\mathbf{W}_{ij}$ is the number of papers author i and j have co-authored.

We compared our algorithm to three existing techniques:

- **SC**: Spectral Clustering with the true number of classes. It serves as a baseline to show if the training labels have helped to improve the results.
- **CSC**: Constrained Spectral Clustering without label propagation. This is a special case of our framework where $\mathbf{F}\mathbf{F}^T$ in Eq.(6) is replaced with $\mathbf{Y}\mathbf{Y}^T$.
- **GGSSL**: This is the graph-based algorithm proposed in [2], which can deal with only one unseen class.

The parameters of our algorithm were set following the discussion in Section III.



(a) Iris (b) Soybean (c) Wine (d) USPS Digits (e) DBLP (f) 20 Newsgroups

Figure 1. Accuracy on benchmark datasets.

To evaluate the accuracy of prediction, we computed Adjusted Rand Index against the ground truth labels. Higher ARI means higher accuracy, 1 means perfect match between the prediction and the ground truth while 0 means the prediction is as good as random guess.

B. Results and Analysis

In Figure 1 we compare the accuracy of our algorithm to the three baseline algorithms on all six datasets. We report the mean and standard deviation of each technique (except SC) over 50 randomly sampled training label sets. We can see that our algorithm outperformed spectral clustering (SC) on all but one dataset. This indicates our algorithm can effectively utilize given guidance to improve the accuracy of prediction. The only exception is the Wine dataset (c), where SC already achieved close-to-perfect performance and did not leave much room for improvement. Comparing our algorithm to the constrained spectral clustering baseline (C-SC) shows that our algorithm can improve the performance more than CSC using the same amount of guidance. In some cases, the accuracy of CSC was even worse than SC due to the incompleteness of input knowledge. Our algorithm also outperformed the GGSSL algorithm, especially when there were more than one unseen classes (b)(d)(e).

V. Application to Risk Stratifying Congestive Heart Failure Patients

In this section we present the results of applying RISGAL to risk stratify a set of potential Congestive Heart Failure (CHF) patients extracted from a real electronic medical record database. For this dataset, we have $1,296$ patients that are confirmed with CHF using the diagnosis criteria mentioned in [11], who are subsequently referred to as *case* patients. For each case patient, we matched it with a *control* patient, i.e. a patient who does not meet the diagnosis criteria for CHF, but is similar to the case patient in terms of gender, age, and some key clinical characteristics as mentioned in [11]. For all selected patients we extracted medical features in terms of the first three digits of ICD9 (International Classification of Diseases, 9th version), which is also referred to as *diagnosis group codes*. In our database, there are in total 1,230 distinct diagnosis group codes. When constructing the patient graph, we set the edge weights, i.e. pairwise patient similarities, to be the total number of co-occurred comorbidities in terms of diagnosis group codes between the patient pairs.

In our investigation, we combined all case and control patients and segmented them into six risk groups according to our medical experts' suggestion. First we applied unsupervised spectral clustering to discover those six risk groups, and the results are summarized in Table II. For each risk group, we present the number of patients assigned to that group, the five diagnosis group codes with highest in-group frequency as well as the group risk score, which is the percentage of case patients in that group. We also give a name to each risk group in the first column of the table to summarize the medical characteristics. From Table II we can see that unsupervised spectral clustering can already discover some well-known risk factors for CHF, such as *documented heart diseases* and *diabetes*.

When we presented these results to our medical experts, however, they suggested that some important risk factors are missing, such as *kidney disease* and *pulmonary disease*. Thus we treated these two types of diseases as knowledge-driven risk factors and injected them into the RISGAL framework. We selected two specific diagnosis, namely *Severe Chronic Kidney Disease* and *Chronic Obstructive Pulmonary Disease* as the labels and applied our RISGAL framework. The results are summarized in Table III. From the table we can clearly observe that the kidney and pulmonary disease risk groups were discovered, whose risk scores confirmed the guidance from the medical experts that these two groups lead to high risk of CHF onset. In the meanwhile, those data-driven risk groups discovered by unsupervised exploration, such as heart diseases and diabetes, are still retained.

VI. Conclusion

We propose RISGAL, a novel graph-based semi-supervised learning framework for patient risk group exploration. Given some known risk factors according to prior knowledge and the corresponding patient cohort, our method finds the optimal partition over the patient similarity graph guided by incomplete knowledge. The obtained patient groups tend to align with the knowledge-driven risk factors, while revealing additional data-driven risk groups in the patient cohort. We first validated our algorithm on a variety of benchmark datasets with comparison to existing techniques. Then we applied our algorithm to a real medical dataset to identify risk groups from a CHF patient cohort. The empirical results demonstrated the effectiveness of our approach.

Acknowledgment

The authors would like to thank Dr. Robert K. Sorrentino for his valuable inputs as the medical advisor.

References

[1] G. C. Fonarow, J. Kirkwood F. Adams, W. T. Abraham, and et al., "Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis," *JAMA*, vol. 293, no. 5, pp. 572–580, 2005.

[2] F. Nie, S. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing and Applications*, vol. 19, no. 4, pp. 549–555, 2010.

[3] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS*, 2003.

[4] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, 2002, pp. 387–394.

[5] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *NIPS*, 2009, pp. 1410–1418.

[6] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009, pp. 951–958.

[7] D. Zhang, Y. Liu, and L. Si, "Serendipitous learning: learning beyond the predefined label space," in *KDD*, 2011, pp. 1343–1351.

[8] T. Yang, R. Jin, A. K. Jain, Y. Zhou, and W. Tong, "Unsupervised transfer classification: application to text categorization," in *KDD*, 2010, pp. 1159–1168.

[9] C. H. Q. Ding, T. Li, and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding," in *ICDM*, 2008, pp. 183–192.

[10] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," in *ICML*, 2006, pp. 985–992.

[11] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches," *Med Care.*, vol. 48, no. 6 Suppl, pp. S106–113, 2010.

Table II
6 RISK GROUPS IDENTIFIED BY UNSUPERVISED LEARNING (SORTED BY GROUP RISK SCORE)

Risk Group	Top Risk Factors	Proportion %	Group Risk
HEART DISEASE (195)	427 *Cardiac Dysrhythmias*	35.9	0.7365
	V58 *Other and Unspecified Aftercare*	12.0	
	V45 *Other Postsurgical States*	8.4	
	414 *Other Forms of Chronic Ischemic Heart Disease*	7.6	
	424 *Other Diseases of Endocardium*	5.5	
DIABETES RELATED (360)	250 *Diabetes Mellitus*	21.2	0.5712
	414 *Other Forms of Chronic Ischemic Heart Disease*	10.6	
	585 *Chronic Renal Failure*	9.8	
	272 *Disorders of Lipoid Metabolism*	7.3	
	401 *Essential Hypertension*	5.6	
BONES & TISSUES (323)	724 *Other and Unspecified Disorders of Back*	13.9	0.5117
	715 *Osteoarthrosis and Allied Disorders*	13.0	
	719 *Other and Unspecified Disorder of Joint*	9.8	
	722 *Intervertebral Disc Disorders*	8.8	
	729 *Other Disorders of Soft Tissues*	6.9	
MISC (828)	496 *Chronic Airways Obstruction, Not Elsewhere Classified*	5.8	0.4504
	285 *Other and Unspecified Anemias*	4.3	
	599 *Other Disorders of Urethra and Urinary Tract*	4.1	
	244 *Acquired Hypothyroidism*	3.9	
	401 *Essential Hypertension*	3.7	
SKIN (155)	173 *Other Malignant Neoplasm of Skin*	26.3	0.4415
	702 *Other Dermatoses*	25.5	
	238 *Neoplasm of Uncertain Behavior of Other and Unspecified Sites and Tissues*	16.5	
	427 *Cardiac Dysrhythmias*	5.6	
	600 *Hyperplasia of Prostate*	5.1	
EYE (157)	365 *Glaucoma*	19.4	0.3931
	366 *Cataract*	17.2	
	250 *Diabetes Mellitus*	15.8	
	362 *Other Retinal Disorders*	14.6	
	244 *Acquired Hypothyroidism*	11.1	

Table III
6 RISK GROUPS IDENTIFIED BY RISGALWITH GUIDANCE TO LOOK FOR TWO SPECIFIC RISK FACTORS: CHRONIC OBSTRUCTIVE PULMONARY DISEASE AND SEVERE CHRONIC KIDNEY DISEASE

Risk Group	Top Risk Factors	Proportion %	Group Risk
KIDNEY DISEASE (22)	585 *Chronic Renal Failure*	47.4	0.8458
	586 *Renal Failure, Unspecified*	15.0	
	403 *Hypertensive Renal Disease*	11.1	
	250 *Diabetes Mellitus*	10.7	
	584 *Acute Renal Failure*	10.0	
HEART DISEASE (208)	427 *Cardiac Dysrhythmias*	33.7	0.7306
	V58 *Other and Unspecified Aftercare*	11.7	
	V45 *Other Postsurgical States*	9.0	
	414 *Other Forms of Chronic Ischemic Heart Disease*	8.3	
	424 *Other Diseases of Endocardium*	6.0	
PULMONARY DISEASE (255)	496 *Chronic Airways Obstruction, Not Elsewhere Classified*	23.5	0.5591
	491 *Chronic Bronchitis*	8.3	
	493 *Asthma*	6.1	
	250 *Diabetes Mellitus*	4.8	
	427 *Cardiac Dysrhythmias*	4.4	
DIABETES RELATED (394)	250 *Diabetes Mellitus*	24.5	0.4824
	272 *Disorders of Lipoid Metabolism*	8.1	
	414 *Other Forms of Chronic Ischemic Heart Disease*	7.5	
	366 *Cataract*	6.7	
	365 *Glaucoma*	6.4	
MISC (995)	715 *Osteoarthrosis and Allied Disorders*	5.1	0.4525
	719 *Other and Unspecified Disorder of Joint*	4.6	
	724 *Other and Unspecified Disorders of Back*	4.4	
	244 *Acquired Hypothyroidism*	4.0	
	272 *Disorders of Lipoid Metabolism*	4.0	
SKIN (133)	702 *Other Dermatoses*	26.3	0.4452
	173 *Other Malignant Neoplasm of Skin*	25.8	
	238 *Neoplasm of Uncertain Behavior of Other and Unspecified Sites and Tissues*	17.0	
	600 *Hyperplasia of Prostate*	5.5	
	365 *Glaucoma*	5.1	