MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-491

Precision Cohort Finding with Outcome-Driven Similarity Analytics: A Case Study of Patients with Atrial Fibrillation

Haifeng Liu[a], Xiang Li[a], Guotong Xie[a], Xin Du[b], Ping Zhang [c], Chengming Gu[d], Jingyi Hu[d]

[a] IBM Research China, Beijing, China
[b]Anzhen Hospital, Beijing, China
[c]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[d]Pfizer China, Beijing, China

Abstract

Dividing patients into similar groups plays a significant role in implementing personalized care. Clinicians and researchers have been applying patient grouping techniques in disease phenotyping, risk stratification, and personalized medicine. However, the current approaches are either based on pure domain knowledge where the underlying patient similarity cannot be precisely quantified, or based on unsupervised clustering techniques which completely ignore the clinical context of measuring patient similarity. In the study, we propose an outcome-driven approach to identify clinically similar patients which are grouped together as a precision cohort. The approach quantitatively measures the similarity between patients in terms of a particular clinical outcome of interest, thus patients who have a similar clinical outcome tend to be grouped into the same group. We demonstrate the effectiveness of the approach in a real-world case study: from an atrial fibrillation patient cohort that is usually considered to be at high risk for ischemic stroke (IS), according to current clinical guidelines. Our approach successfully identified a precision cohort of patients with truly low risk of IS.

Keywords:

Cluster Analysis; Machine Learning; Precision Medicine

Introduction

In order to understand complex disease conditions and to provide personalized care, it is crucial for clinicians to divide patients into subgroups such that patients in one group are similar to each other. Successful patient grouping is particularly beneficial for the tasks of disease phenotyping, risk stratification and personalized medicine. In current clinical research and/or practice, patient grouping is generally conducted based on some scoring schemes recommended by clinical guidelines, which stratify patients into groups with different levels of risk. The patients with the same level of risk are considered to be similar. This mechanism may lead to some problems: 1) a general clinical guideline may not fit for the local clinical practice or population; 2) the similarity between patients with the same risk level are not quantified because the scoring schema is generated for a population and the detailed patient conditions are not differentiated in the same level.

In recent years, unsupervised clustering has been applied to identify groups of patients with different phenotypes and implement personalized medicine [1;2] in clinical research. Although the method can be adapted to local clinical practice and quantify the similarity between patients using distances between vectors of patient features, it ignores the fact that patient similarities are usually context-based, i.e., the similarity degree between two patients' conditions may vary in terms of particular clinical outcomes of interest. For example, for three patients A, B and C with atrial fibrillation (AF), clinicians would regard that A and B are more similar in terms of stroke-occurrence risk, and A and C are more similar to each other when considering myocardial infarction (MI) risk. This is because the impacts of risk factors on these two outcomes are totally different (e.g., smoking and body mass index play important roles in the risk of MI while their impacts on stroke are relatively smaller).

This study aims to demonstrate how an outcome-driven similarity analytics method can be utilized to address the issues above. With the given clinical outcome of interest, the key idea is to cluster patients into groups based on a learned similarity (distance) metric from patients' clinical records where patients with the same clinical outcome are considered to be similar. One related work [3] has been reported using a learned distance metric to retrieve the K most similar patients and providing the prognosis insight based on the physiological time-series data of similar patients. Another work [4] is to learn the similarity metric from physicians' feedbacks and use patient similarity for decision support. Our work differs from them in using the learned distance metric to divide the patients into groups, followed by identifying the characteristics of similar patients in each group. To further understand the resultant groups, we also discover the discriminating rules between groups. The discovered rules can guide clinicians to easily assign new patients into their similar patient group, and we call such a patient group a precision cohort. Personalized care can then be recommended to the patients based on the insights discovered from this precision cohort. An alternative approach to divide patients into subgroups without similarity metric learning and clustering is to directly build a decision tree to split the patients against the outcome [5]. However, it suffers from two limitations: 1) similar patients could be dispersed in different branches, thereby requiring manual regrouping after the tree is built; 2) there is no distance metric for measuring the exact similarity between two patients.

We validate our approach in a real-world case study where we stratify a population of AF patients with high risk of ischemic stroke (IS) into a few groups and identify a particular group of patients with truly low risk. With the demonstrated effectiveness, we believe that other diseases and scenarios of finding precision cohorts could generally benefit from the proposed approach.

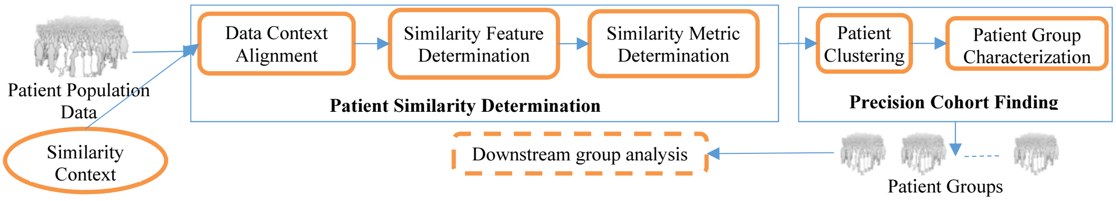
Patient Population Data
Similarity Context
Data Context Alignment
Similarity Feature Determination
Similarity Metric Determination
Patient Similarity Determination
Patient Clustering
Patient Group Characterization
Precision Cohort Finding
Downstream group analysis
Patient Groups

Figure 1 - Methodology of outcome-driven patient similarity analysis

Methods

As precise patient similarity is context-dependent, and varies along the patient conditions, outcomes of interest, and particular clinical scenarios, we first define patient similarity context as follows. A patient similarity context consists of:

- A target patient population of study, e.g., the patients with diagnosis of diabetes type 2 but without any other complications,
- A clinical outcome of interest, e.g., mortality in two years, re-hospitalization in six months, and so on. The assumption is that two patients with the same outcome are considered to be similar, and
- A clinical scenario when the similarity analytics need to be done, e.g., when patients are first diagnosed with type 2 diabetes, when patients are hospitalized, or when patients are registered into a particular study.

Figure 1 illustrates the overall methodology of how we find precision cohorts by outcome-driven patient similarity analytics. With a given clinical data set and a patient similarity context, we adopt a machine learning approach. First, we determine the set of features used to compute the similarity between patients, and determine the exact similarity metric based on the selected features where the outcome of each patient is considered. Then, we segment the patients into groups based on their similarities and characterize the groups with their unique characteristics. Downstream analytics can be performed on each group to discover insights for personalized care, e.g. local risk analysis and treatment efficacy analysis, which are beyond the focus of this paper. We describe the pipeline analytics in detail as follows.

Patient Similarity Determination

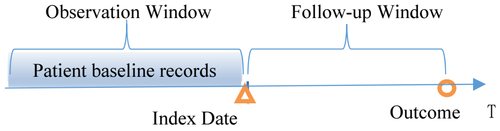

Figure 2 - Alignment of patient data with a specific similarity context

Patient Data Context Alignment

Given a clinical data set and a patient similarity context, we prepare the patient data as illustrated in Figure 2. For each patient, we identify an index date as the time when a clinical scenario is setting, e.g. when the diagnosis of type 2 diabetes is initially made, and use the records before the index date (observation window) to conduct the similarity analysis (those records after the index date and before the outcome date are ignored). The patient is then represented with a vector of features summarized from these clinical records. These features could incorporate the patient's information involving demographics, diagnosis, lab test, medication, and so on. In addition, we label the patient as positive or negative in terms of the outcome of interest (we now focus on the binary outcome only).

Similarity Feature Determination

Not all features are relevant to determine contextual patient similarities, and there are two typical ways to filter the relevant features from the candidates obtained from the previous step: 1) based on established domain knowledge (e.g. clinical guidelines), we identify the set of relevant features known as risk factors regarding the specific outcome; 2) we apply supervised feature selection methods [6] to automatically select the features relevant to the outcome. The analysis scenario determines which method to apply.

Similarity Metric Determination

Since the distance metric can be regarded as a measure of dissimilarity, similarity learning is closely related to metric learning. A few alternatives exist to measure the clinically similarity between two patients represented by vectors of the selected similarity features:

- Using the Euclidean distance between them or between their corresponding vectors with reduced dimensions by PCA (Principal Component Analysis) if there are too many similarity features.
- Using the distance between their predicted risk scores regarding the outcome of interest where we first develop a risk prediction model based on the selected features, and then applying the model to compute the risk score for each patient regarding the outcome. Patients with similar risk scores are considered similar.
- Using a learned Mahalanobis distance between them, which can automatically adjust the importance of each feature against the given outcome of interest. Formally, we represent a patient as a N-dimensional feature vector x where N is the number of identified relevant features. Let S be the set of equivalence constraints denoted by S = {(x_i, x_j)| x_i, x_j belong to the same outcome class} and D be the set of inequivalence constraints denoted by D = {(x_i, x_j)| x_i, x_j belong to the different outcome classes}. Our goal is to learn a generalized Mahalanobis distance between patient x_i and patient x_j defined as:

  $$d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)}$$

  Where A is positive semi-definite matrix and is designed by solving the optimization problem:

  $$\min_{A \in R^{NxN}} \sum_{x_i, x_j \in S} d_A^2(x_i, x_j) \text{ s.t. } \sum_{x_i, x_j \in D} d_A^2(x_i, x_j) \geq 1$$

  As a result of metric learning, we expect to keep pairwise vectors in S close and those in D separated away.

While a few algorithms [7] have been proposed to learn a Mahalanobis distance metric, in this study we have implemented three popular ones including linear discriminant analysis (LDA) [8], which projects the original feature vectors into a subspace that preserves the variance between class labels; large margin nearest neighbor (LMNN) [9], which learns a linear transformation of the input space where k nearest neighbor should have matching labels; and information-theoretic metric learning (ITML) [10], which maximizes the differential entropy of a multivariate Gaussian subject to constraints on the associated Mahalanobis distance.

To determine which metric to use for computing patient similarities, we evaluate them using a nearest-neighbor based method. That is, for a specific metric, we compute the classification performance against the given target outcome using a KNN (K Nearest Neighbor) classifier that is built upon the metric. We consider a metric a better fit if it achieves the best classification performance.

Precision Cohort Finding

Patient Clustering

We apply agglomerative hierarchical clustering to group patients so that the patients within a group are contextually similar. The method starts with singleton clusters and proceeds by successively merging the two "closest" clusters at each stage. We customize the method by using the previously determined distance metric to determine the distances between clusters rather than using the conventional unsupervised distance metrics. With the learned distance metric, we expect that there could be a substantial divergence in the result of the targeted outcome (proportion of patients with a positive outcome) between the resultant groups.

One advantage of hierarchical clustering is the flexibility of determining the number of produced clusters based on the results of one running. For our purpose, we determine the number of reported clusters and evaluate the clustering performance by: 1) an internal clustering performance metric: the silhouette coefficient SH [11] which is defined as $\mathrm{SH} = (\mathrm{b} - \mathrm{a})/\max(\mathrm{a}, \mathrm{b})$ where a is the mean distance between a patient and all other patients in the same group, and b is the mean distance between a patient and all other patients in the next nearest group. A SH near 1 indicates that the sample is far away from the neighboring clusters, and a SH greater than 0.2 is generally considered to be a fair clustering result; 2) An external metric to measure the group outcome disparity (OD) between the resultant groups which is defined as the difference between the maximal and minimal positive outcome rates of the groups. This metric reflects the effect of stratified risks among the groups, and the grouping result with higher OD is better for our purposes.

Patient Group Characterization

After obtaining the patient groups with stratified outcome results, our interest is to identify the characteristics of each group and understand the differences between groups. We address this by two means: 1) we compare the key feature differences between groups using statistical tests (a Kruskal-Wallis test for continuous features and Pearson chi-square test for categorical features.); 2) in order to identify the unique characteristics of groups, we build a decision tree that can differentiate the resultant M clusters where M classes of patients are labeled in accordance with their respective cluster memberships. For easier applicability, we further derive the explicit rules from the tree to interpret the group membership of a patient. In this work, we use C5.0 [12] to build a decision tree and convert it to rules.

Results-Risk Stratification of AF Patients

The CHA2DS2-VASc (CV) [13] score ranging from 0 to 9 has been widely recommended and used to identify AF patients with a high risk of IS (CV≥2) who need to be treated with oral anticoagulant (e.g. Warfarin) or radiofrequency ablation (RFA). However, it is still arguable that the CV score may not precisely capture the risk of particular AF patients from local populations. For example, there is a subgroup of patients with high CV but a low IS-occurrence rate for whom anticoagulation may not be indicated. This is crucial because anticoagulants may have severe side effects, such as warfarin-related bleeding, and RFA incurs additional economic burden on patients. It therefore might be unnecessary to treat those patients with truly low risk. Thus, this study aims to apply our proposed approach to identify such AF patient subgroups, who have truly low risk but are misclassified as high-risk by high CV score.

Data Set

We use a data set from a cohort study for around 18,000 AF patients across China [14]. The collected data includes patients' structured baseline records (i.e. demographics, history, medication history) and clinical records (i.e. interventions, outcomes) during follow-up visits in a 3-year period. We are interested in studying the risk of IS during one year of follow-up by using the baseline features of patients, and also a selected population of 2,907 patients from the population (with an IS-occurrence rate is 4.6%) with the criteria as follows: 1) complete 12-month follow-up records are available for the patient, or follow-up records until the occurrence of an IS event within the 12-month follow-up period; 2) the patient either has no intervention (warfarin or RFA) until the end of 12 months of follow-up, or IS occurred before the intervention was started; 3) the patient's CV was ≥2 (i.e. the patient is considered high-risk). As the raw data has non-standard, missing and dirty values, we apply the same approach as our previous study [14] to automatically clean the data and impute the missing values. In the end, we have a data set with 132 input features and a binary feature for the outcome of having IS occurrence in 12 months of follow-up.

Furthermore, to validate the study results of analytics, we split the data into a derivation patient set (1,743 patients) and a validation patient set (1,164 patients) (60% and 40% of the population, respectively). The splitting strategy is to keep both the IS-occurrence rates and the CV score distributions the same between the two sets. We observe that the lowest IS-occurrence rate for the patients with CV=2 is approximately 2.5%, thus our objective is to discover a subgroup of patient with truly low risk of IS where the IS-occurrence rate should be lower than 2.5%.

Selecting Similarity Features

From the original data set with 132 input features, we first remove those that are relevant to IS occurrence but have strong correlation with known risk factors as defined by the CV score. We then automatically select the other relevant features using a filter-based method (SPSS modeler version 17). In other words, we select the top significant continuous features based on the p-value of using the F statistic and categorical features based on the p-value of using Pearson's χ^2 statistic ($p <= 0.05$). Besides the four known CV features including prior CHF, prior IS, prior vascular diseases, and age, Table 1 lists the other 10 features selected as potential risk factors.

Table 1 – Similarity features selected from the data

Whether having a history of established coronary artery disease (ECAD)
Whether using drugs for ventricular rate control at the baseline (VRCD)
Whether there is a IS within the recent 5 years (IS5)
Whether there is a CHF within the recent 5 years (CHF5)
Whether there is a DM within the recent 10 years (DM10)
Whether statin were used to treat hyperlipidemia (Statin)
Total bilirubin at the baseline (TBIL)
Whether ACEI was used at the baseline (ACEI)

Left ventricular septum thickness on echocardiography at the baseline (IVS)
Left ventricular posterior wall thickness on echocardiography at the baseline (LVPW)

Determining the Similarity Metric

With the derivation data set where each patient is represented by 14 features above, we develop a few distance metrics to measure patient similarity (using scikit-learn 0.17): 1) EUCL. The Euclidean distance between the patient vectors; 2) LR_Score. The distance between the predicted risk scores using a logistic regression (LR) model trained on the derivation set (AUC of the derived LR is 0.72). 3) LDA. The learned Mahalanobis distance using LDA. 4) LMNN. The learned Mahalanobis distance using LMNN 5) ITML. The learned Mahalanobis distance using ITML. Table 2 reported their KNN classification performance on the validation set with the averaged results when KNN classifiers were trained on the derivation set with K set to 1, 3, 5, 7, and 9 respectively. While the overall F1 score is low for all metrics (this is due to our highly imbalance data set), LDA outperforms the other metrics.

Table 2 – KNN classification performance of different metrics.

	EUCL	**LR_Score**	**LDA**	**LMNN**	**ITML**
Average Precision	0.481	0.262	0.594	0.483	0.410
Average Recall	0.124	0.131	0.153	0.122	0.112
Average F1	0.132	0.151	0.194	0.132	0.134

Patient Grouping Results

Table 3 – Clustering results using different similarity metrics

	2 clusters			**3 clusters**		
	IS_occurrence rate	SH	OD	IS_occurrence rate	SH	OD
EUCL	3.8%, 6.6%	0.20	2.8%	3.8%, 5.3%, 9.4%	0.22	5.6%
LR_Score	3.4%, 13.8%	0.56	10.4%	1.6%, 4.6%, 13.8%	0.48	12.2%
LDA	2.9%, 10.2%	0.58	7.3%	**1.5%**,4.1%, 10.2%	**0.52**	**8.7%**
LMNN	0%, 4.7%	0.65	4.7%	0%, 4.1%, 10.6%	0.24	10.6%
ITML	4.5%, 5.4%	0.32	0.9%	0% ,4.5%, 6%	0.30	6%

With all the considerations above, we decide to adopt the three clusters resultant from the clustering with LDA for downstream analysis. As summarized in Table 4, on the one hand, there is a stratified risk of IS among groups where group 1, 2 and 3 corresponds to low, medium, and high risk groups respectively. Particularly the lowest risk rate is 1.5% in group 1 which is even close to the patients with CV=1 (1.4% in our source data set). This implies a group of patients with truly low risk of IS. On the other hand, the IS-occurrence rates increase with the rising of the CV median values of groups. This to some extent verifies the rough validity of CHA2DS2-VASc in this local population.

Table 4 – Summary of the resultant patient groups

	Derivation Set			**Validation Set**		
Group	**1**	**2**	**3**	**1**	**2**	**3**
Propotion	34.7%	41.1%	24.2%	33.0%	49.4%	17.6%
IS- Rate	1.5%	4.1%	10.2%	1.6%	4.8%	9.3%
CV (Median)	3	4	6	3	4	6

Moreover, we validate the resultant clustering model against the validation set where each patient is assigned to one of three clusters based on his similarity with the center of each cluster. In this way, the patients are also divided into three groups as shown in Table 4. We observe that it approximately coincides with the result from the derivation set, and in particular, a precision cohort with truly low risk of IS at 1.6% is successfully identified too. Likewise, we also test the clustering model with LR_Score against the validation set because clustering with LR_score achieves a comparable result with using LDA on the derivation set. However, it fails to get a satisfactory result on the validation set where among the resultant three clusters, the lowest risk rate is 3.4% and there is an extreme small group with only 61 patients.

Group Characterization Results

Table 5 summarizes the key baseline characteristics that are significantly different among the groups from the derivation set where a p value <= 0.05 is considered statistically significant using Pearson chi-square test. Group 1 patients are the youngest and tend to have the lowest rates of comorbidities while group 3 patients are the oldest and have the highest rates of all comorbidities. The conditions of group 2 patients are in between group 1 and group 3 in terms of either comorbidities or medication taken or examination results. These group differences coincide with their varied IS-occurrence rates.

Table 5 – Baseline characteristics of the resultant groups

Group (patient count)	**1 (605)**	**2 (716)**	**3 (422)**	**P value**
Age (median)	71	74	77	<0.001
CHF	12%	34%	65%	<0.001
Prior IS	3%	16%	56%	<0.001
Vascular diseases	17%	22%	42%	<0.001
ECAD	10%	17%	32%	<0.001
CHF5	0%	9%	39%	<0.001
IS5	0%	4%	34%	<0.001
DM10	16%	8%	6%	<0.001
IVS (median)	9.7	9.8	10	<0.001
LVPW (median)	9.4	9.4	10	<0.001
VRCD	46%	80%	85%	<0.001
Statin	9%	25%	43%	<0.001

To further interpret the group characteristics, we build a decision tree using C5.0 (with a classification accuracy of 77% using SPSS modeler version 17) to classify these patients according to their cluster membership and derive a few rules to explicitly differentiate them. The resultant rules capture the unique characteristics of patients in different groups, and are easier to understand. In particular, we are the most interested in the patients in group 1 because they have the lowest IS-occurrence rate, and Table 6 lists the resultant 5 rules to characterize group 1 with high confidences (all are all above 85%). Due to the space limitation, we do not list the other grouping rules.

Table 6 - Rules to characterize the group 1 patients.

No	**Rules**	**Confidence**
1	Statin not used, no CHF in recent 5 years, no prior IS, and LVPW <=7.9mm	90.8%
2	Statin not used, no CHF in recent 5 years, no prior IS, age<75, and 8<=LVPW<=8.9mm	87.7%
3	VRCD not used, no CHF in recent 5 years, no prior IS, and age<75	87.3%
4	Statin not used, no prior CHF, no prior IS, and age<65	86.4%
5	VRCD not used, no prior CHF, no prior IS	85.5%

Table 7 - Validation result from the decision tree.

Group	**1**	**2**	**3**
Size (Percentage)	34.1%	41.2%	24.7%
IS-Occurrence Rate	2.4%	4.9%	6.6%
CV Score (Median)	3	4	6

We validate the developed decision tree by applying it to the validation set. As show in Table 7, the resultant three groups have the same CV median values with those from the clustering. To our interest, group 1 patients still have a low rate of IS-occurrence at 2.4% which is even lower than the patients with CV=2 (2.5% in our source data set). The results above support validity and stability of our approach.

Discussion

One issue for further investigation is if the group 1 patients that we identified with truly low risk are simply a subset of patient with the lowest CV score. To answer that, we compared the breakdown of CV distribution of group 1 patients between the clustering results on the derivation and validation sets, and the result shows that more than 50% of patients are actually have a CV>2 in both sets. This implies that CHA2DS2-VASc may not precisely fit this local population. In fact, prior hypertension and sex are not selected as our similarity features. This suggests a further systematic study on risk factor evaluation which should include all CHA2DS2-VASc factors and the novel feature candidates as in Table 1.

The critical part of our proposed methodology lies in determining an appropriate similarity metric for a specific context. While the outcome-driven learned metrics (including LR_Score and Mahalanobis distance metrics) generally outperform the unsupervised EUCL, it is still challenging to identify the most appropriate one from various outcome-driven metrics. The performance of these metrics may vary depending on the different data sets or different clinical scenarios. Furthemore, in order to cope with a large and sparse data set, we could adopt deep phenotyping [15] (which is not the focus of this study) to learn a set of latent features to compute patient similarity rather than use the seleclted raw features.

Conclusions

In this study, we developed an outcome-driven approach to identify groups of similar patients in terms of a particular clinical outcome. We validated the effectiveness of the approach by grouping AF patients with high risk of IS into three subgroups using a real-world data set and then identifying a precise group of patients with low risk of IS and their unique characteristics. This may help to better inform IS risk stratification in clinical guidelines. Further research would be necessary to verify the utility of novel risk factors identified. Subdividing high-risk patient groups may better target personalized care recommendations and improve patient outcomes. For example, a clinician could prescribe a treatment to a patient based on a treatment effectiveness analysis and comparison of the patient's characteristics against those of a precision cohort which includes similar patients. Our future work includes adopting two-stage clustering to handle a very large data set and incorporating temporal similarity features and other outcomes of interest.

References

[1] S. Ather, L.E. Peterson, V.G. Divakaran,.et al. Digoxin treatment in heart failure--unveiling risk by cluster analysis of DIG data. *Int J Cardiol.* 2011 Aug 4;150(3):264-9.

[2] Ahmad T, Pencina MJ, Schulte PJ,.et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J Am Coll Cardiol.* 2014 Oct 28;64(17):1765-74.

[3] Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting Patient's Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics *AMIA Annual Symposium Proceedings*. 2010;2010:192-196.

[4] Sun J, Wang F, Hu J, Ebadollahi S. Supervised Patient Similarity Measure of Heterogeneous Patient Records. *ACM SIGKDD Explorations Newsletter*, 2012:14(1).

[5] G.C. Fonarow, K.F. Adams, W.T. Abraham, C.W. Yancy, W.J. Boscardin. Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure: classification and regression tree analysis. *JAMA,* Feb 2005, 293 (5).

[6] Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 2003:3:1157-1182.

[7] Kulis B. Metric Learning: A Survey. *Foundation Trends in Machine Learning* 2012:5 287-364.

[8] Hastie T., Tibshirani R. Discriminant adaptive nearest neighbor classification. *IEEE Pattern Analysis and Machine Intelligence*, 1996:18(6):607-16

[9] Weinberger KQ, Saul LK. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research* 10 (2009) 207-244

[10] Davis JV, Kulis B, Jain P, Sra S, Dhillon IS. Information-Theoretic Metric Learning. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[11] Rousseeuw PJ. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics,* 1987 20: 53–65.

[12] Kuhn M, Johnson K. *Applied Predictive Modeling*, 2013

[13] Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest.* ***137** (2):* 263–72.

[14] Li X, Liu H, Du X, .et al. Integrated Machine Learning Approaches for Predicting Ischemic Stroke and Thromboembolism in Atrial Fibrillation. *AMIA Annual Symposium Proceedings.* 2016.

[15] Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH5, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. *J Biomed Inform.* 2015 Dec;58:156-65.

Address for correspondence

Haifeng Liu. Email: liuhf@cn.ibm.com.