Research on cardiovascular disease prediction based on distance metric learning

To cite this article: Zhuang Ni et al 2018 J. Phys.: Conf. Ser. 1004 012032

View the article online for updates and enhancements.
Research on cardiovascular disease prediction based on
distance metric learning

Zhuang Ni1, 2, a, Kui Liu1, b and Guixia Kang1,2, c
1Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and telecommunications. Beijing, China.
2Wuxi BUPT Sensory Technology and Industry Institute CO.LTD

Abstract. Distance metric learning algorithm has been widely applied to medical diagnosis and exhibited its strengths in classification problems. The k-nearest neighbour (KNN) is an efficient method which treats each feature equally. The large margin nearest neighbour classification (LMNN) improves the accuracy of KNN by learning a global distance metric, which did not consider the locality of data distributions. In this paper, we propose a new distance metric algorithm adopting cosine metric and LMNN named COS-SUBLMNN which takes more care about local feature of data to overcome the shortage of LMNN and improve the classification accuracy. The proposed methodology is verified on CVDs patient vector derived from real-world medical data. The Experimental results show that our method provides higher accuracy than KNN and LMNN did, which demonstrates the effectiveness of the Risk predictive model of CVDs based on COS-SUBLMNN.

1. Introduction
From the World Health Organization (WHO) statistics [1], cardiovascular diseases (CVDs) which are the leading cause of death globally, known as a threat of public health across the world, are estimated to cause about 17 million deaths every year. Moreover, an estimated 23 million new probable cases will die of cardiovascular diseases over the next decades. Obviously, the high mortality and death rate in cardiovascular diseases patients is a major public health concern. Cardiovascular diseases are multifactorial which are really hard to make an accurate diagnosis. Efforts should be taken for considering the risk factors of CVDs and finding a global strategy for prediction and prevention of CVDs remains a tough challenge. Making accurate predictions for the risk of cardiovascular diseases is really meaningful for human.

With the emergence and popularity of intelligent technology equipment like medical wearable devices, medical health data adopted by electronic health records (EHRs) grow rapidly and provide various sources of information for medical research. Based on EHRs, many risk prediction models for CVDs have been developed. The Framingham risk score [2], Reynolds risk score [3] and ASSIGN Scottish algorithm [4] are developed to predict the risk of CVDs by using the risk factors of CVDs such as BMI, smoking history, hypertension status diabetes mellitus and some blood biomarkers[5-6]. Some other variables are added to make the prediction more precise [3, 7]. However, these prediction models did not take personalized medicine decision into consideration, making the predictive task not accurate. For these reason, considering patient similarity of pairs of patients based on EHRs will increase the accuracy of prediction model. Fei Wang et al. [8] proposed the SimProX system which
uses low-rank mahalanobis distance with physician’s feedbacks to measure patient dissimilarity and the accuracy of this model is nearly 92% better than the performance of other traditional machine learning algorithm. Gottlieb et al [9] combine types of similarity measure to predict the likely discharge diagnosis for a new cases using information from medical history. Nevertheless, they never care about the patient similarity in local scope, which can capture the discriminative feature better.

KNN is a basic and simple distance metric algorithm, which calculates the distance between vectors to measure the similarity and make prediction, but KNN treats each variable of feature vector equally. However, different features have different weight and make different contributions to label classification. Weinberger et al. [10] proposed the large margin nearest neighbour classification (LMNN) to overcome the weakness of KNN and improves the performance. The aim of LMNN is to minimize the distance between samples belonging to the same class and make vectors with opposite label a large margin by a global distance metric. Considering the direction not just the distance, some studies [12-13] discussed distance metric algorithm based on cosine similarity. All algorithm mentioned above, they did not consider locality of data distributions, which is a significant part to capture nuances of local heterogeneity among patients.

In this paper, we propose the COS-SUBLMNN algorithm adopting cosine metric and large margin nearest neighbour to capture differences between local samples to improve the classification accuracy. We will build an automated predictive system based on COS-SUBLMNN that infers the CVDs diagnosis of patients using patient EHRs derived from routine medical data. Medical data we use includes routine physical examination data, medical questionnaire survey results and a return visit followed 1 month after the physical examination for all the candidates [11]. Every features in medical data list are considered as potential risk factors for CVDs. The diagnostic feedback and personalized suggestion will be sent to all patients and the feedback of patients for the model’s results will be used to optimize algorithm. The system of predicting CVDs is shown in Figure.1.

In this part, we present the details of our COS-SUBLMNN algorithm and the system of predicting CVDs. Firstly, we show the way we construct the representative vector of every candidates based on EHRs.
2.1. Study Subjects

All medical data for our task are extracted from CiMing Health Checkup Center. Here we present the routine physical examination data briefly in Table 1 which is introduced in detail in our last paper [10]. In addition, some other predictor variables displayed are from medical questionnaire survey, such as family medical history, medical history, education level, occupation, dietary law and so on. We use data warehouse, hive to save all medical data. The data table in hive is comprised of some columns: candidate ID, inspection item and its value and all tables in hive are combined by the primary key, candidate ID. The label annotation is still a tough work for us. Some criteria used to define the cases of CVDs is introduced below: (i) The candidate was diagnosed with CVDs, such as cerebral stroke, hypertension, coronary disease, hyperglycaemia, hyperlipoidemia, cerebral infarction and so on in 30-day interval return visit. (ii) The diagnosis of CVDs appeared on the corresponding list of ICD-9 codes. So the label (class variable) in our model is health status based on the criteria of label annotation. If a candidate met the criteria above, the label is 1, otherwise it is 0.

Table 1. Typical variables in physical examination data.

| routine physical examination data | Age | gender |
|----------------------------------|-----|--------|
| Demographic information         | Body Mass Index (BMI) | Systolic Blood Pressure (SBP) |
| Vitals                           |   | Diastolic Blood Pressure (DBP) |
|                                  |   | … |
| Blood routine examination        |   | Red Blood Cell (RBC) |
|                                  |   | Hemoglobin (HGB) |
|                                  |   | … |
| Urine routine examination        |   | Hydrogen (PH) |
|                                  |   | Specific Gravity (SG) |
| Biochemical examination          |   | Creatinine (Cr), |
|                                  |   | Uric Acid (UA) |
| Other risk predictors            |   | Fasting Blood-Glucose (FBG) |
|                                  |   | … |
|                                  | alpha fetoprotein (AFP) and | carcino-embryonic antigen (CEA) |

2.2. Data preprocessing

Incomplete data is a common problem for most tasks in the area of data mining. Missing values is a typical characteristic for real-world medical data. The approach and specific procedure for handling the missing values is the same as [10]. We use different normalization methods for variables according to the feature type. the discrete variables such as age and gender are normalized by one-hot encoding and all the numerical variables like blood pressure are normalized by Z-Score and. The formula of Z-Score is:

\[
Z = \frac{(X - \mu)}{\sigma}
\]

Where X is the raw value of variables. \( \mu \) and \( \sigma \) is the mean and standard deviation, respectively.

As mentioned above, a representative vector for every candidate is constructed, which is a high dimensional vector combined by risk factors from different data sources and label, serving as the input to distance metric algorithm. The purpose and advantage of vector construction is to capture sufficient difference among candidates and distinguish the group of patients used by distance metric algorithm.
Next, our purpose is to design a distance metric algorithm to make fully use of patient feature vector and get the candidate feedback of whether the result of predictive model is right or not.

2.3. **COS-SUBLMNN distance metric algorithm**

In this section, we present a local distance metric learning method adopting cosine metric and LMNN called COS-SUBLMNN, which first use LMNN to get a global distance metric, mapping samples to new feature space, then get K local cluster by k-means and use LMNN in the specific target local cluster which target vector exists in to make prediction.

2.3.1. **Constructing COS-LMNN.** The feature vector of each candidate is denoted by an n-dimensional feature vector \( x \), and \( z \) is the corresponding label of \( x \). So \( X \) presented below is the dataset we use.

\[
X = \{(x_i, z_i) \mid i = 1, 2, ..., m, x_i \in \mathbb{R}^n, z_i \in \{1, 2, ..., c\}\}.
\]

In this part, we introduce COS-LMNN which learns a global matrix using cosine metric:

\[
\text{COS}(x_i, x_j) = \frac{||A(x_i, x_j)||^2}{\sqrt{x_i^TM(x_i) x_j^TM(x_j)}}
\]

(2)

Where \( M \in \mathbb{R}^{nxn} = A^T A \) is a positive Semi-Definite matrix, and \( A \) is the matrix that COS-LMNN needs to learn. We define the homogeneous neighbourhood of \( x_i \), \( S(x_i) \), which has the same label with \( x_i \) and heterogeneous neighbourhood of \( x_i \), \( I(x_i) \), which has the different label with \( x_i \). The size of \( S(x_i) \) is \( K_s \) which should be set by users. \( \text{COS}(x_i, x_j) \) satisfy the property of cosine, so cosine relationship between \( x_k \), the sample of \( I(x_i) \), and input sample \( x_i \) is presented below:

\[
\text{COS}(x_i, x_j) \leq 1 - \text{COS}(x_i, x_k)
\]

(3)

COS-LMNN aims at learning a global matrix to lessen the gap between samples with same label and magnify the distance between examples with opposite label. So the objective function is described as:

\[
\text{Obj}(A) = \sum_i \sum_{j \in S(x_i)} \text{COS}_A(x_i, x_j) + \mu \sum_i \sum_{j \in S(x_i)} \sum_{k \in I(x_i)} [\text{COS}_A(x_i, x_j) + \text{COS}_A(x_i, x_k) - 1]_+
\]

(4)

As we can see, the first term \( \epsilon_{\text{pull}}(A) \) penalizes big gap between every input and its homogeneous neighbourhood.

\[
\epsilon_{\text{pull}}(A) = \sum_i \sum_{j \in S(x_i)} \text{COS}_A(x_i, x_j)
\]

(5)

The second term \( \epsilon_{\text{push}}(A) \) magnify the distance by penalizing the small distance between examples with opposite label. The indicator variable \( y_k = 0 \) if \( z_i = z_k \), otherwise \( y_k = 1 \).

\[
\epsilon_{\text{push}}(A) = \sum_{i \in S(x_i)} \sum_{k \in I(x_i)} y_k [\text{COS}_A(x_i, x_j) + \text{COS}_A(x_i, x_k) - 1]_+
\]

(6)

Where \( [z]_+ = \max(0, z) \) is the hinge loss. The hinge loss indicates that if the input \( x_i \) keep a safe distance with its heterogeneous neighbourhood, then its hinge loss has no influence on the objective function.

So the optimization function combining two penalty term and the slack variable \( \xi_{ijk} \) is shown:

\[
\max \sum_{i,j \in S(x_i)} \text{COS}(x_i, x_j) + \mu \sum_{i,j \in S(x_i)} \sum_{k \in I(x_i)} \xi_{ijk}
\]

s.t. \( \text{COS}(x_i, x_j) + \text{COS}(x_i, x_k) \leq 1 - \xi_{ijk} \)

\( \xi_{ijk} \geq 0 \)
M = A^T A \geq 0

2.3.2. COS-SUBLMNN algorithm. COS-SUBLMNN algorithm has three steps: Firstly, a global matrix A is learned by COS-LMNN, and then K local clusters are determined by k-means and the target cluster which obtains input sample is determined. Finally, a local matrix $A_s$ is learned in the target cluster by COS-LMNN. We introduce COS-SUBLMNN in detail below.

1. The calculation of global distance

Considering input sample $x_i$ dataset $X$ and global matrix $M = A^T A$, we calculate the distances between $x_i$ and each sample in dataset $X$:

$$D(x_i, x_j) = \cos(x_i, x_j) = \|A(x_i, x_j)\|^2 = \frac{x_i^T M(x_i)}{\sqrt{x_i^T M(x_i)} \sqrt{x_j^T M(x_j)}}$$

$D_{ij}$ denotes the distance between $x_i$ and $x_j$ and $D_s = \{D_{i1}, D_{i2}, \ldots, D_{in}\}$ is set of distances.

2. The determination of target cluster

$K_s$ local cluster is created by k-means. Based on $D_s = \{D_{i1}, D_{i2}, \ldots, D_{in}\}$, choose the cluster which is nearest to input sample $x_i$ from $K_s$ local cluster as the target cluster.

3. The classification in target cluster

COS-LMNN is used to get a local matrix $A_s$ in the target cluster $L_s = \{x_{s1}, x_{s2}, \ldots, x_{sm}\}$ and then the distances between $x_i$ and each sample in the target cluster $K_s$ are calculated by cosine distance.

$$\cos(x_i, x_{sl}) = \|A_s(x_i, x_{sl})\|^2 = \frac{x_i^T M(x_{sl})}{\sqrt{x_i^T M(x_{sl})} \sqrt{x_{sl}^T M(x_{sl})}}$$

Finally, the input sample $x_i$ is classified by KNN ($K_b$ is set by users).

### 3. Experiments and discussions

In this part, the details of the experiments of risk predictive model of CVDs on medical data are presented to demonstrate the effectiveness of COS-SUBLMNN algorithm. To get a fair and persuasive conclusion, we compare the performances of COS-SUBLMNN algorithm with these different classifier models (KNN and LMNN) with 5-fold cross validation which is a good solution to the overfitting problem. Some metrics like accuracy, sensitivity and specificity are proposed to assess the performance of risk predictive model of CVDs. All experiments of risk predictive model of CVDs are run on the platform of MATLAB R2014b. The parameters of COS-SUBLMNN algorithm is $K_b$, $K_s$ and $K_p$ introduced in detail in chapter 2. All parameters are selected by grid search.

1) The results of experiments show that the performances and accuracy of COS-SUBLMNN are better when $K_b=8$ and $K_s=7$. So firstly we compare the performances of COS-SUBLMNN algorithm with KNN and LMNN when $K_b$ varies and $K_s=8$ and $K_p=7$.

| $K_b$ | KNN |  |  |  | LMNN |  |  |  | COS-SUBLMNN |  |  |  |
|------|-----|---|---|---|------|---|---|---|-------------|---|---|---|
|      | Sen | Spe | Acc |     | Sen | Spe | Acc |     | Sen | Spe | Acc |     |
| 2    | 0.641 | 0.828 | 0.767 |     | 0.711 | 0.862 | 0.811 |     | 0.720 | 0.849 | 0.807 |     |
| 4    | 0.640 | 0.831 | 0.769 |     | 0.716 | 0.869 | 0.819 |     | 0.733 | 0.848 | 0.813 |     |
| 6    | 0.667 | 0.821 | 0.775 |     | 0.721 | 0.872 | 0.822 |     | 0.719 | 0.865 | 0.817 |     |
| 8    | 0.671 | 0.822 | 0.781 |     | 0.725 | 0.863 | 0.818 |     | 0.726 | 0.875 | 0.825 |     |
| 10   | 0.672 | 0.840 | 0.787 |     | 0.724 | 0.878 | 0.823 |     | 0.724 | 0.877 | 0.830 |     |
| 12   | 0.679 | 0.838 | 0.783 |     | 0.723 | 0.863 | 0.819 |     | 0.735 | 0.881 | 0.833 |     |
| 14   | 0.671 | 0.848 | 0.799 |     | 0.731 | 0.881 | 0.827 |     | 0.746 | 0.878 | 0.835 |     |
From figure 2 and table 2, when K_b=14, the accuracy of COS-SUBLMNN is up to 0.835 higher than KNN and LMNN, demonstrating that COS-SUBLMNN has advantages of identifying all group of patients. When K_S=14, the accuracy of KNN and LMNN reaches the top. As K_b is bigger, the accuracy of COS-SUBLMNN is smaller than LMNN’s and COS-SUBLMNN and LMNN performs better than KNN. The specificity of COS-SUBLMNN and LMNN is up to 0.88 and performs well, indicating that has a strong ability of identifying negative group (normal people).

Table 3. The results of 3 classifiers when K_S varies and K_a=8 and K_b= 14.

| K_S | KNN | LMNN | COS-SUBLMNN |
|-----|-----|------|-------------|
|     | Sen | Spe  | Acc | Sen | Spe  | Acc | Sen | Spe  | Acc |
| 2   | 0.671 | 0.848 | **0.799** | 0.745 | 0.841 | 0.809 | 0.752 | 0.859 | 0.822 |
According to the introduction of COS-SUBLMNN, when $K_s$ is small, noise decreases the performance of algorithm, otherwise, time complexity is too high. The results show that when $K_s=7$, the performances of COS-SUBLMNN and LMNN reach the top. Obviously, KNN is not a function of $K_s$, so the accuracy of KNN stays still no matter how $K_s$ changes.

![Figure 4. The accuracy of 3 classifiers when $K_s$ varies and $K_a=8$ and $K_b=14$.](image)

(3) $K_s=7$ and $K_s=14$ set, we observe the change of the performance of KNN, LMNN and COS-SUBLMNN corresponding to the change of $K_s$ and get the $K_s$ value when COS-SUBLMNN provides the best results.

| $K_s$ | KNN | LMNN | COS-SUBLMNN |
|------|-----|------|-------------|
| 2    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.733 0.855 0.817 |
| 4    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.717 0.871 0.826 |
| 6    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.734 0.872 0.829 |
| 7    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.724 0.877 0.830 |
| 8    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.746 0.878 0.835 |
| 9    | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.734 0.885 0.833 |
| 10   | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.721 0.872 0.822 |
| 11   | 0.682 0.838 0.787 | 0.731 0.881 0.827 | 0.718 0.873 0.820 |

By observing the consequences in table 4 and introduction to COS-SUBLMNN algorithm, higher $K_s$ value reduces the performance of COS-SUBLMNN algorithm with higher time complexity, otherwise, noise dominate the consequence, causing the accuracy of COS-SUBLMNN lower than LMNN. When $K_s=8$, COS-SUBLMNN provides the best accuracy. As we can see from figure 5, the performance of KNN and LMNN stays still when $K_s$ varies.
4. Conclusions
In our research, we propose a distance metric learning algorithm adopting cosine metric and LMNN named COS-SUBLMNN for risk predictive model of CVDs based on the patient profiling derived from EHR medical data. The proposed algorithm capture nuances of differences on local features among patient vector by considering the locality of medical data. The practical benefits of the optimized algorithm have been assessed by experiments. The model built by COS-SUBLMNN achieved much higher accuracy than KNN and LMNN. Predicting the risk of CVDs accurately plays an important role in daily life for each patient. As our results above shown, the application of COS-SUBLMNN based on real-world medical data for CVDs are valuable for both patients and physicians.

Acknowledgments
This work was supported by National Natural Science Foundation of China (61471064) and National Science and Technology Major Project of China (No.2017ZX03001022-005).

References
[1] Otsu N. Discriminant and Least Square Threshold Selection [J]. Proc 4IJCPR, 1978, 596.
Mendis S, Puska P, Norrving B. Global atlas on cardiovascular disease prevention and control [M] World Health Organization, 2011.
[2] Wilson P W F, Levy D, Belanger A M, et al. Prediction of Coronary Heart Disease Using Risk Factor Categories[J]. Circulation, 1998, 97(18):1837-47.
[3] Ridker P M, Buring J E, Rifai N, et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score.[J]. Jama, 2007, 297(6):611.
[4] Woodward M, Brindle P, Tunstall-pedoe H. Adding social deprivation and family history to cardiovascular risk assessment-the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC) [J]. Heart, 2007, 93(2):172.
[5] Baik I, Cho N H, Kim S H, et al. Dietary information improves cardiovascular disease risk prediction models [J]. European Journal of Clinical Nutrition, 2013, 67(1):25.
[6] Björnson E, Borén J, Mardinoglu A. Personalized Cardiovascular Disease Prediction and Treatment-A Review of Existing Strategies and Novel Systems Medicine Tools [J]. Frontiers in Physiology, 2016, 7:2.
[7] Melander O, Newtoncheh C, Almgren P, et al. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. [J]. Jama the Journal of the American Medical Association, 2009, 302(1):49-57.

[8] F Wang, J Hu, J Sun. Medical prognosis based on patient similarity and expert feedback. International Conference on Pattern Recognition, 2012:1799-1802.

[9] Gottlieb A, Stein G Y, Ruppin E, et al. A method for inferring medical diagnoses from patient similarities [J]. Bmc Medicine, 2013, 11(1):194.

[10] Distance metric learning for large margin nearest neighbor classification [J]. Journal of Machine Learning Research, 2009:207-244.

[11] Kang G, Ni Z. Research on early risk predictive model and discriminative feature selection of cancer based on real-world routine physical examination data[C]/ IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2017:1512-1519.

[12] H.V.Nguyen, L. Bai. Cosine similarity metric learning for face verification [C]. Asian Conference on Computer Vision. Springer, 2010, 709-720.

[13] Kumar M P, Torr P H S, Zisserman A. An Invariant Large Margin Nearest Neighbour Classifier[C]/ IEEE, International Conference on Computer Vision. IEEE, 2007:1-8.