Mining Personal Health Satisfaction and Disease Index from Medical Examinations

Chunshan Li(✉), Xiao Guo, Dianhui Chu, Chongyun Xu, and Zhiying Tu

School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China
{lics,guox,chudh,cyxu,zytu}@hit.edu.cn

Abstract. People have regular medical examinations every year, the primary purpose of these is not only to discover a specific disease but to have a clear overview of their current health. Therefore, it is essential to be able to scientifically analyze and evaluate the results of medical examinations, and provide people with comprehensive feedback on physical health satisfaction (PHS) or a specific disease index (DI) feedback. PHS can reflect the current state of a person's body directly. In this paper, we propose a framework for calculating personal PHS and DI, by which the process of assessing people's health is indicated. We use the public dataset of the National Center for Health and Nutrition to conduct a scientific analysis of the population medical examinations data. Finally, based on a comprehensive experiment of 6166 participants' medical examinations data and some experiments with data from participants with heart disease, we verified the effectiveness of the proposed framework.

Keywords: Personal health satisfaction · Personal disease index · Health analysis

1 Introduction

In recent years, data mining has become more and more popular in the field of medical and health, and has been widely used in the areas of auxiliary medicine, disease prediction, online health assessment, etc., and has achieved remarkable results. Despite this, there are still many issues that need to be resolved urgently. As people's quality of life continues to improve, more and more people are paying attention to their health. Therefore, a large number of people go to a health check every year to accurately obtain the current physical health.

Health scoring systems have become increasingly popular over the years, such as Nuffield Health Score [1], but their scoring mechanism is based on pre-defined measurements of expert knowledge and experience. The SAPS acute physiology scoring system proposed by JR Le Gall et al. [2], the data set is derived from a large number of surgical and medical patient samples, and provides a method for converting the score into the probability of hospital death, the primary purpose of which is to study the current relationship between physical condition and mortality. As the number and dimensions

J. Xiong et al. (Eds.): MobiMedia 2021, LNICST 394, pp. 633–646, 2021.
https://doi.org/10.1007/978-3-030-89814-4_46

of health-related data grow, it becomes difficult to rely on expert knowledge. In 2014, Chen Ling et al. [3] implemented an automatic scoring system adapted to change in data attributes. The data set used personal health check data for the age group of 65 years and older in the past five years, and proposed the concept of personal health index (PHI), with the cause of death as a label. However, when people of other ages conduct physical examination analysis, it is difficult to obtain the cause of death of these people and the physical examination data for five consecutive years. Second, in their research, suffered some shortcoming when it came to pre-processing the data. In addition, people not only want to know the current physical condition through medical examination data, but also want to know that under current physical conditions, the probability of suffering from certain diseases, such as middle-aged and older people, will pay more attention.

Inspired by the above problems, in this paper, we have done the following four main tasks.

1. For the health checkup data of people of all ages, we propose the concept of personal health satisfaction (PHS) and construct a PHS calculation framework. Personal fitness satisfaction is used as a model training target. The patient can directly provide this attribute during the physical examination.
2. We train the model through a neural network algorithm to fill in some missing attributes and achieve impressive performance.
3. For the current physical condition and the probability of suffering from a certain disease, we propose the concept of disease index (DI) and use the PHS framework in question 1 to predict the disease index. This paper uses heart disease as an example to verify the effectiveness of the framework.
4. In addition, in order to make the framework reusable, we implement the algorithm and dataset based on an open-source platform (EasyML). For framework calculation process, each step result can be visualization using the EasyML.

In a word, we constructed a PHS computing framework that includes data filling, feature selection, model training, results prediction, and evaluation. Our job is the first to propose that the same set of frameworks can be used to predict personal health satisfaction and disease index.

2 Related Works

In recent years, a number of scoring systems have been introduced to assist in clinical decision making, such as APMCHE, SAPS, and MPM in patients in intensive care units [4]. At present, the most widely used and authoritative scoring method in the world, APACHE [5] (acute physiology and chronic health assessment) scoring system, American scholar Kuaus first proposed APACHE I in 1981 [6]. It consists of two parts. That is the acute physiology score (APS) and the pre-disease chronic health status (CPS) evaluation reflecting the severity of the acute disease. The former includes 34 physiological parameters. APACHE II [7] was launched in 1985, and the APACHE II score consists of three parts: 12 acute physiological variables, age, and chronic health status. The probability of death can be derived by using the disease category and the APACHE II score.

It can be seen that the use of 34 physiological parameters and 12 acute physiological variables in the selection of attributes are based on expert knowledge and experience. For disease prediction, Wilson et al. [8] studied coronary heart disease prediction, based on age, diabetes, smoking, JVC-V blood pressure categories, etc., to develop a gender-specific prediction equation to predict the risk of coronary heart disease, Palaniappan et al. [9] for heart disease prediction, the study developed a prototype of the Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely decision trees, naive Bayes and neural networks. Dangare [10] et al. developed a Heart Disease Prediction System (HDPS) using a neural network. The HDPS system predicts the likelihood of a patient having a heart attack. For prediction, the system uses 13 medical parameters such as gender, blood pressure, and cholesterol. Using medical profiles such as age, gender, blood pressure, and blood glucose can predict the likelihood of a patient suffering from heart disease, and it is easy to see that expert knowledge is also used when selecting attributes. Moreover, data mining in the health care field now uses more inpatient data, incidence rates [11–13], and mortality predictions [14]. The application of data analysis in large medical data sets is mostly based on population analysis of the distribution of some epidemics or high-risk diseases [15–18]. There are very few assessments based on personal health status and individual disease risk assessments, and there are many analyses of specific diseases that are used for classification, mostly in 2 categories, rather than judging the likelihood of illness. Chen Ling et al. [3] proposed the concept of personal health index, which is mainly aimed at the elderly over 65 years old. The framework is not applicable to people aged out this range. The method used in forecasting is a common algorithm in data mining. SVM and fail to compare with other methods.

3 PHS Calculation Framework

The PHS calculation framework (Fig. 1) mainly consists of five main parts: data preprocessing, feature selection, model training, result prediction, and PHS verification. The input is medical examinations dataset containing all attributes, and the output is personal health satisfaction or disease index ranging from 0 to 1.

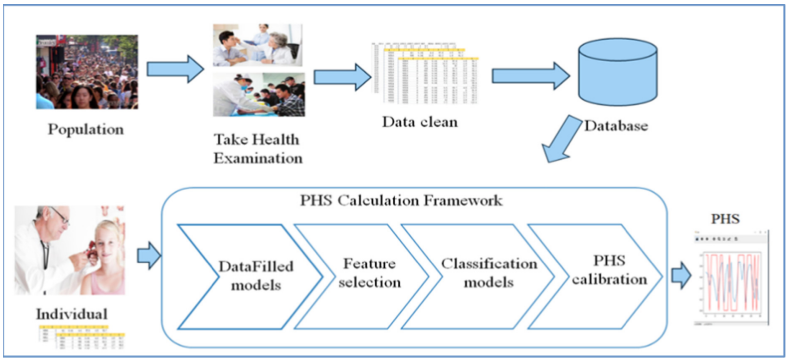


Fig. 1. An overview of the proposed PHS calculation framework

3.1 Data Pre-processing

The attribute values of the medical dataset include numeric values, serial numbers, and text. In the current work, only one text feature will be used as label: the current health satisfaction. The status, the recorded value is also the level information, "very good," "good," etc., we use numerical calculation for level recording. Firstly, we delete some unrecognized symbols and erroneous data. Secondly, with taking a look at missing values, there are many strategies to deal with missing values, a simple mean, regression, direct deletion…, After a few attempts, we kept all the available examples and focused on building model for each feature in order to infer the missing values. It is also the first part of the computational framework. We use simple forward neural networks to build those models. For this, we use three different methods for comparison experiments that are explained below. Finally, we normalize the values of the 30 attributes selected (described in detail in the next section).

We detail below the three neural network structures that we tried to use for reconstructing the missing values. The code is implemented with Tensorflow.

1) A Restricted Boltzman Machine (RBM) to reconstruct the input.

RBM [19] are widely used especially when there are a large amount of unlabeled data as a brick for a bigger network called deep belief network.

Here we use them as an order to mimic the reconstructed data distribution as we trained it with contrastive divergence procedure. We first define an energy function (Hopfield 1982):

$$\mathrm{E(v, h)} = -\sum\nolimits_{i \in visible} a_i v_i - \sum\nolimits_{j \in hidden} b_j h_j - \sum\nolimits_{i,j} v_j h_j w_{ij} \tag{1}$$

v_i, h_j binary states of visible unit i and hidden unit j.

a_i, b_j their biases and w_{ij} is the weight between them.

In a common RBM, the network assigns a probability to every possible pair of a visible and a hidden vector via this energy function:

$$\mathrm{p(v, h)} = \frac{1}{Z} e^{-E(v,h)} \tag{2}$$

$$\mathrm{Z} = \sum\nolimits_{v,h} e^{-E(v,h)} \tag{3}$$

$$\mathrm{p(v)} = \frac{1}{Z} \sum\nolimits_{h} e^{-E(v,h)} \tag{4}$$

It can be proven that when we derive the log probability, we eventually obtain such learning rule.

$$\Delta w_{ij} = lr(<v_i h_j>_{data} - <v_i h_j>_{model}) \tag{5}$$

Where lr is the learning rate.

Except that here our intermediary output is not stochastic but deterministic and their values are continuous. We still assume the learning rule is correct, and we can notice there is still convergence.

2) Auto-encoder

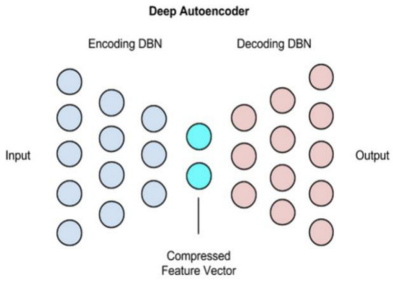


Fig. 2. Auto-encoder principle

In Fig. 2, AE [19] principle, we can notice some similitude with RBM, although this structure allows several hidden layers in both encoding and decoding part. Weights are this time not the same for both phase, it surely allow more freedom. And the objective function is not the same.

$$\text{loss}(\hat{y}, \text{y}) = \frac{1}{n} \sum\nolimits_i |\hat{y}_i - y_i|^a \tag{6}$$

Where $\hat{y}$ is inferred value, y is true value.

There is some freedom to choose geometry or another penalty function, but the simple L2 norm does the job.

Though in the form, there is a bottleneck which can act as a regularizer, in our case, the data dimension is quite low. Consequently, this role is not so useful here.

Corrupt the input by adding some Gaussian noise as it is often the case to limit overfitting and learn more complex pattern has a limited influence, we blame the same reasons.

3) Simple Forward neural Network (FNN)

Once again, the structure is similar to the previous, but there is only one target the weights in the networks are now more specialized, though it makes the network even more sensible to overfitting.

$$\text{loss}(\hat{y}, \text{y}) = \frac{1}{n} \sum\nolimits_i |\hat{y}_i - y_i|^a \tag{6}$$

4) few intermediary results

These figures (Fig. 3, Fig. 4) cannot be seen as proof of effectiveness. At this point, we cannot conclude. Their purpose is mostly to illustrate. Due to a large number of hyper-parameters to optimize, these results will change slightly.

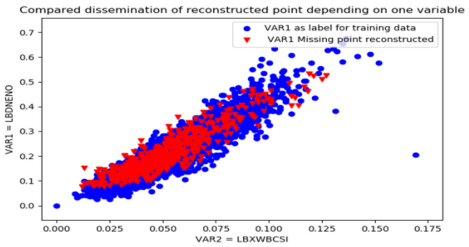


Fig. 3. FNN completion

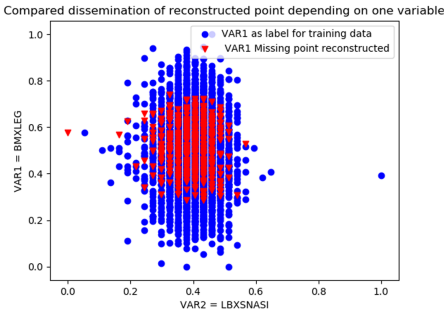


Fig. 4. FNN completion

Light and straightforward network, not too wide and not too deep seems to give a better result. It avoids overfitting, and it prevents activation function in the last layers to be saturated (especially for the simple neural network).

3.2 Feature Selection

There is a large number of attributes of the dataset. We will first go through a regularization phase. In the past, data mining in the field of medical research, attribute selection was mostly based on expert knowledge. In order to be able to pick out the attributes that have a greater impact on the results, we use the Ridge regression [21] algorithm to calculate each attribute and label value separately, and finally select 30 attributes that are better for the prediction.

Ridge regression solves some problems of ordinary least squares by penalizing the size of the coefficient. The minimum of the ridge coefficient is the sum of the squares of the residuals with penalty terms.

$$min||X_w - y||_2^2 + \alpha||w||_2^2 \tag{7}$$

Among them, $\alpha \geq 0$ is a complexity parameter of the amount of shrinkage: the larger the value of α, the larger the reduction.

We also performed a PCA that does not seem to improve the final results.

3.3 PHI and Prediction Models

In this sub-section, we first give the formal definition of PHI:

$\forall x \in \boldsymbol{R^n}$,
PHS: $x \rightarrow 1 - \text{ModelInfer}(x)$
$\boldsymbol{R^n} \rightarrow [0,1]$
x: the n-dimension vector that describe an individual;
n: the dimension that describe an individual;

Model Infer: the complementary function of PHS to 1.

After definition, we can will employ a proper predictor to calculate the PHI. We will ran several cross-validations with numerous learning algorithms. Are common algorithms such as SVM [22], TreeBoost [23], neural networks…, this one seems to outperformed other alternatives with slight differences depending on hyperparameters and kernels for SVM.

3.4 PHS Calibration

In this section, we wonder whether or not we need calibration (Fig. 5). Indeed, to give any individual an understandable and reliable index about its global health we need to be aware of the probability distribution over classes of our models, if the output, a continuous value, of these, is not linearly correlated to the belonging probability to levels, we would need to calibrate it.

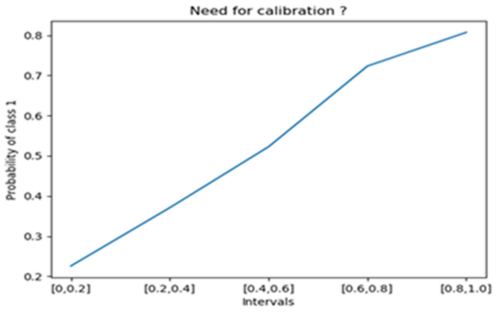


Fig. 5. Belonging probability to class depending on model output. Linear in this case

We can notice here the importance of penalty function among other parameters on the reliability curve. But a simple square error fulfills our expectation.

4 Experimental Evaluation

4.1 Dataset Description

The dataset we used has been released by the National Center for Health and Nutritional Surveys. The participants ranged from a few years to a few decades. The survey center

randomly selected a total of 6,166 participants from all over the United States. The medical examination data is personal privacy, and the public dataset does not have any records that can identify the individual, and the individual is marked in the form of ID. Each person's medical examination data consists of four parts: Demographic, Examinations, Labs tests and Health questions, the specific content is shown in Table 1, due to space constraints, we only listed some attributes. And those are already selected one.

Table 1. Selected attributes by categories

Type	Attribute examples	Number
Demographic	Age, education, marital status, income,	13
Lab tests	Albumin, testosterone, blood lead,blood cadmium, Lymphocyte…	68
Examinations	Blood pressure, Weight, height, arm and leg length…	29
Health questions	Health condition(label), Mental condition(label)	2

4.2 Experiment Setup

For the following experiment, we used "Health condition" as the final target. In the data pre-processing stage, we first select 30 attributes by Ridge regression. In the dataset, there is a questionnaire on personal health satisfaction. The results are divided into 5 grades, 1–5, 1 means very satisfied, 5 means very dissatisfied. Looking at a first confusion matrix, the frontier between predicted classes is not clear, we can't get obvious prediction results. As a consequence, we have processed the label to keep only two classes (Table 2):

Table 2. Label change.

Former label	New label
1, 2	Healthy (0)
3, 4, 5	Not Healthy (1)

We ran cross-validation to select best learning algorithms including Aboost, TreeBoost, DecisionTree [24, 25], etc.

Algorithms are running in a Linux environment. All are written in Python with Scikit-learn package.

4.3 Experiment Results

(1) In the data pre-processing part, firstly, we perform a simple analysis of datasets, there are more than 4,000 records missing about 18 attributes, so the handling of missing values appear to be important.

Secondly, after regularization, as shown in Fig. 6, we can find that the missing values of two preponderant attributes reach 90%. In order to make the result more realistic, when filling in the missing values, the model is not built for these two attributes, only filled with '0', so in the filling model training, the attribute with the missing rate greater than 20% is directly filled with 0, less than 20% attribute will build a model for it.

We want to expose here the influence of the different methods of data blank completion we highlighted in the previous part on the final results (Table 3).

Table 3. Method results

Method	Score
FNN	0.72
RBM	0.70
AutoEncoder	0.72
KNN	0.71
Nothing	0.68

There is not significant difference between the different methods. We mostly blame the quality of the label for that though we decide to use FNN to deploy our code on the platform.

(2) In the feature selection part, we use the Ridge regression algorithm, after performing a grid experiment, the value of α is chosen to be 12, and the experimental result is shown in Fig. 6. An example of training accuracy with the neural network model.

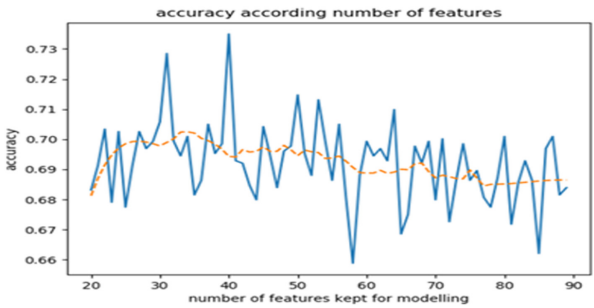


Fig. 6. Attribute number and accuracy

Where the abscissa is the number of selected attributes, and the ordinate is the accuracy of the result. The solid blue line is the specific accuracy obtained, and the yellow dotted line is a smoothed version that shows the trend. We can see that as the number of attributes increases, there is just a slight accuracy difference.

(3) Classification model selection: Firstly, Attempts have been made not to consider the label as they seem really subjective with common algorithms like k-means or Gaussian mixture models, the result in Fig. 7 (a&b). Even by contorting the geometry used, algorithms cannot distinguish that match Healthy/notHealthy group, and the process might be too dependent, too specific to the data.

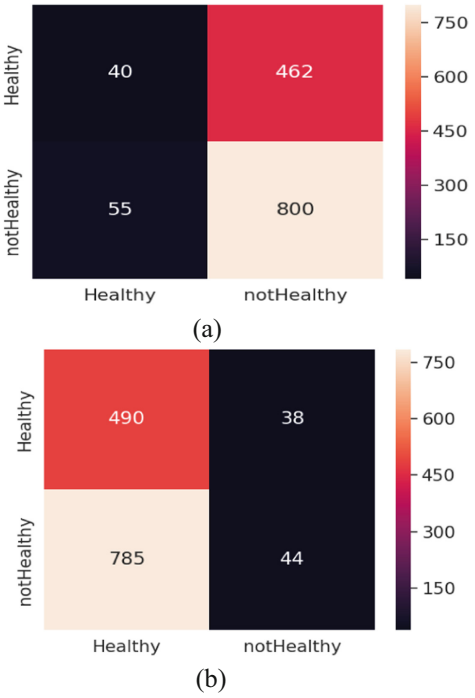
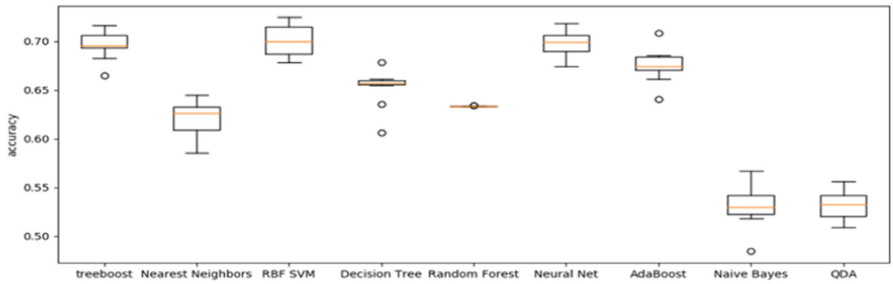


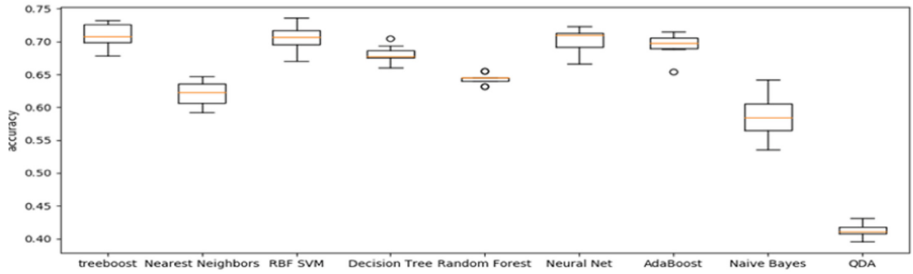
Fig. 7. (a) k-means. (b) Gaussian mixture model (GMM).

We performed cross-validation on quickly processed data normalization, data blank filled with 0 with some popular classification algorithms using 'Health condition' as a label. The comparison is shown in Fig. 8.

(4) (4)The influence of missing values filling algorithms. There is no significant difference between different methods though we decide to use FNN to deploy our code on the platform.
(5) PHS: The final prediction results of PHS, in Fig. 9, in order to make the display of the results more intuitive, we made a comparison between the actual value and the predicted value, and the abscissa in the figure. Indicates the 0-30th person, the ordinate is the prediction result range 0–1, the solid red line is the real value, only 0 and 1, two categories. the solid blue line is predicted value between 0–1.
(6) Calibration: as previously explained we need the output of our model to be linearly correlated to the belonging probability of the class.

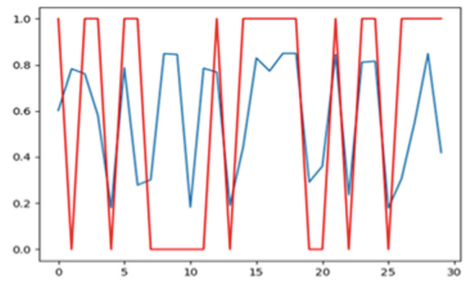


(a) Filled with "0", algorithm comparison

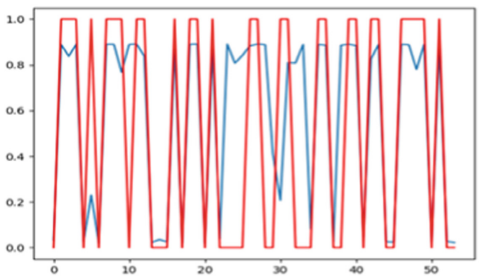


(b) Model filling, algorithm comparison

Fig. 8. Comparison of several classification algorithms



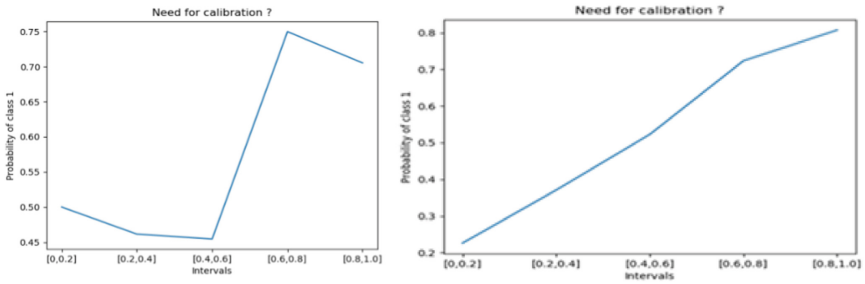
(A)



(B)

Fig. 9. (A) PHS result (B) DI result

We can notice that loss function(Fig. 10(a)) has a significant impact, but classical square loss function(Fig. 10(b)) fit our specification, calibration is in this case not necessary.



(a) L1 Norm (b) L2 Norm

Fig. 10. Class belonging probability according to the model output and the norm used

4.4 Data Processing Platform

In order to make PHS calculation framework and training model more reusable and scalable, we use an open-source machine learning platform that builds a Hadoop cluster that can support the distributed operation of the large dataset and make some modifications to the platform. The algorithm and data can be deployed on the platform through the interface. The advantage is that the algorithms can be moved on the page by dragging. The calculation result of each step of the whole process of the framework can be displayed in real-time so that we can better grasp the result. Moreover, any part of the framework can be optimized in the future, including algorithms and processes.

In Fig. 10, the left panel contains the algorithm, data, and tasks lists. The middle part shows the task running. The right panel let us see the detailed information about the task, real-time monitoring task, and running status. During the running process, the node is yellow, and the running success node is blue. After each algorithm runs successfully, you have access to the result directly from the output node.

5 Conclusion

In the era of big data, we can process a large amount of data. For our study case, we embarked in the healthcare field and tried to build a person's physical health satisfaction and a single disease index. Our work is constrained by the data we have access to indeed label is here highly subjective, although we managed to build a coherent index.

In other studies, PHS was proposed, but the label is based on the cause of death, to predict the current physical condition of the elderly, and it is to infer people's index with out of range. In addition, the missing data is not processed, and the missing data is deleted directly. We use the NN algorithm to train models filling with missing values.

When constructing the classification model, we did a lot of comparison experiments of classification algorithms, including before filling the data and filling the data.

A large number of experiments verify the validity and robustness of the framework. The best predictive PHS accuracy rate is 72.5%, and the disease index DI is 83%, which can better verify the validity and practicability of the model.

References

1. NuffieldHealthScore. http://info.nuffieldhealthscore.com/. Accessed 18 June 2014
2. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA **270**(24), 2957–2963 (1993)
3. Chen, L., Li, X., Wang, S., et al.: Mining personal health index from annual geriatric medical examinations. In: 2014 IEEE International Conference on Data Mining, pp. 761–766. IEEE (2014)
4. Fisher, A., Burke, D.: Critical care scoring systems. In: Brown, S., Hartley, J., Hill, J., Scott, N., Williams, J. (eds.) Contemporary Coloproctology, pp. 513–528. Springer, London (2012). https://doi.org/10.1007/978-0-85729-889-8_35
5. Wong, D.T., Knaus, W.A.: Predicting outcome in critical care: the current status of the APACHE prognostic scoring system. Can. J. Anaesth. **38**(3), 374–383 (1991)
6. Knaus, W.A., Zimmerman, J.E., Wagner, D.P., et al.: APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit. Care Med. **9**(8), 591–597 (1981)
7. Knaus, W.A., Draper, E.A., Wagner, D.P., et al.: APACHE II: a severity of disease classification system. Crit. Care Med. **13**(10), 818–829 (1985)
8. Wilson, P.W.F., D'Agostino, R.B., Levy, D., et al.: Prediction of coronary heart disease using risk factor categories. Circulation **97**(18), 1837–1847 (1998)
9. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 108–115. IEEE (2008)
10. Dangare, C., Apte, S.: A data mining approach for prediction of heart disease using neural networks. Int. J. Comput. Eng. Technol. (IJCET) **3**(3) (2012)
11. Esfandiary, N., Babavalian, M.R., Moghadam, A.-M.E., Tabar, V.K.: Knowledge discovery in medicine: current issue and future trend. Expert Syst. Appl. **41**(9) (2014)
12. Neuvirth, H., et al.: Toward personalized care management of patients at risk: the diabetes case study. In: SIGKDD. California, USA, pp. 395–403. ACM (2011)
13. Tran, T., Phung, D., Luo, W., Harvey, R., Berk, M., Venkatesh, S.: An integrated framework for suicide risk prediction. In: SIGKDD, Chicago, USA, pp. 1410–1418. ACM (2013)
14. Predicting mortality of ICU patients. http://physionet.org/challenge/2012/. Accessed 18 Feb 2014
15. Abbas, K., Mikler, A.R., Gatti, R.: Temporal analysis of infectious diseases: influenza. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 267–271. ACM (2005)
16. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthc. Inf. Manag. **19**(2), 65 (2011)
17. Obenshain, M.K.: Application of data mining techniques to healthcare data. Infect. Control Hosp. Epidemiol. **25**(8), 690–695 (2004)
18. Mellmann, A., Friedrich, A.W., Rosenkötter, N., et al.: Automated DNA sequence-based early warning system for the detection of methicillin-resistant Staphylococcus aureus outbreaks. PLoS Med. **3**(3), e33 (2006)

19. Hinton, G.E.: A practical guide to training restricted Boltzmann machines. In: Montavon, G., Orr, G.B., Müller, K.R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, pp. 599–619. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_32
20. Rifai, S., et al.: Higher order contractive auto-encoder. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS, vol. 6912, pp. 645–660. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23783-6_41
21. Khalaf, G., Shukur, G.: Choosing ridge parameter for regression problems, pp. 1177–1182 (2005)
22. Gupta, S., Kumar, D., Sharma, A.: Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian J. Comput. Sci. Eng. (IJCSE) **2**(2), 188–195 (2011)
23. Jianming, Z., Zhicai, Z., Keyang, C., et al.: Review on development of deep learning. J. Jiangsu Univ. Nat. Sci. Edit. **36**(2), 191–200 (2015)
24. Wu, X., Kumar, V., Quinlan, J.R., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
25. Yang, S., Zou, L., Wang, Z., Yan, J., Wen, J.-R.: Efficiently answering technical questions-a knowledge graph approach. In: AAAI, pp. 3111–3118 (2017)