Journal of Combinatorial Optimization (2021) 42:884–895
https://doi.org/10.1007/s10878-019-00495-x

An intelligent data-driven model for disease diagnosis based on machine learning theory

He Huang[1] · Wei Gao[2] · Chunming Ye[1]

Published online: 29 November 2019

Abstract

In the era of data, major decisions are determined by massive data, especially in the healthcare industry. In this paper, an intelligent data-driven model is proposed based on machine learning theory, specifically, support vector machine (SVM) and random forest (RF). The model is then applied to a case of disease diagnosis, cough variant asthma (CVA). The data of 137 samples with 12 attributes is collected for experiments. The results show that the proposed model achieves better prediction performance than single SVM and single RF. Besides, in order to identify the key medical indicators to enhance diagnosis accuracy and efficiency, the most important factors affecting CVA are generated by the proposed model, including FENO, EOS%, MMEF75/25, FEV1/FVC, PEF, etc. Meanwhile, it is demonstrated that the proposed model could be a user-friendly tool to improve the performance of disease diagnosis.

Keywords Machine learning theory · Healthcare · Random forest (RF) · Support vector machine (SVM) · Disease diagnosis

1 Introduction

With the rapid development of computer and information technologies, the power of data is embodied and emphasized in various domains, such as biology, finance and healthcare. The world is entering the era of big data (Cao et al. 2017).

Meanwhile, data mining technology is generating a huge availability of observations of scientific issues, and machine learning is one of the best-known theories to analyze the data with huge quantity and complex features. Based on machine learning theory, the data-driven models are investigated to enable decisions more reasonable

✉ Wei Gao
gaowei1108@hotmail.com

[1] Business School, University of Shanghai for Science and Technology, Shanghai 200093, China

[2] Shanghai General Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai 200080, China

and accurate, and these models heavily depend on real evidence rather than intuition. However, researchers indicate that limitations might be observed in traditional models with single technology. Contrarily, the hybrid models that integrate two or more technologies would achieve better performance, and remedy the shortages of each technology (Gai and Ji 2019). In fact, random forest (RF) and support vector machine (SVM) are two crucial technologies in machine learning field, and they have been utilized to solve many problems (Bai et al. 2015; Dapogny et al. 2018). Therefore, the data-driven model that combines the advantages of these two technologies would be an effective tool with the value both in academic and practical areas.

In the healthcare industry, medical decisions mostly rely on experts' experiences, it therefore results in inescapable biases and unstandardized results. In fact, the healthcare industry is heavily affected by tremendous amounts of clinical data (Bahri et al. 2018; Zhang et al. 2015). Consequently, it is essential to develop data-driven models to improve medical service level, such as disease diagnosis. The cough variant asthma (CVA) is a disease that shows the symptoms of dry cough without any sputum, and it does not have other traditional asthma symptoms like wheezing and shortness of breath. In reality, there are many factors related to CVA diagnosis. However, the medical staff usually concern with a few factors based on their intuition and experience. This diagnosis procedure overlooks the possibility of other factors that would lead to CVA. This phenomenon motivates this research to develop an intelligent data-driven model for disease diagnosis.

To sum up, an intelligent data-driven model is proposed based on machine learning theory in this research. The model is then applied to cough variant asthma diagnosis to improve accuracy and efficiency. The experimental data is from a real hospital in China. The remainder of the article is organized as follows. In Sect. 2, the literature related to this research is reviewed. The intelligent data-driven model is proposed in Sect. 3. The experiments and results are explained in Sect. 4, and Sect. 5 provides the conclusions.

2 Literature review

By observation of previous papers, it is found that machine learning theory has been widely applied to many industries from different perspectives (Calderoni et al. 2015; Jin et al. 2016; Dongxiao et al. 2017). Meanwhile, healthcare has become a critical research area in recent years, and numerous researchers investigated related issues (Liu et al. 2015; Yang et al. 2015; Zhong and Bai 2019). In reality, the patients' data is extreme vast and chaotic, especially the patient information that consists of demographics, medications, diagnosis codes, etc. The electronic health records (EHRs) is a common tool to record such information. Yadav et al. (2018) realized the value of EHRs data and discussed data mining approaches based on machine learning theory. Papers of this type focus on data mining issues with medical data system, such as EHRs and hospital information system (HIS). Considering the features of specific diseases, Bertsimas et al. (2016) investigated the clinical trials testing combination chemotherapy regimens for Cancer, and then developed machine learning models to suggest regimens to be tested. Alam et al. (2016) proposed an ambient intelligent system of in-home psychiatric care service for emergency psychiatry (EM-psychiatry), and a machine learning algorithm (modified Viterbi) was introduced to generate the most probable psychiatric state sequence. In these papers, different issues in healthcare were optimized or improved by using machine learning methodology. However, the papers related to disease diagnosis, especially for CVA diagnosis, are limited. Among the studies that discuss disease diagnosis, it is found that clustering (such as k-means, mean-shift, DBSCAN, et al.) and classifying (such as decision tree, SVM, neural network, et al.) are two common solutions. In Gao et al. (2019), the decision tree approach was employed that is a classic classification tool; In Khanmohammadi et al. (2017), the diagnosis issue was investigated by an improved overlapping k-means clustering method. From these previous studies, it is clearly showed that machine learning is an important technique in healthcare field, and the performance of disease diagnosis would be further enhanced by using this technique. Besides, in order to achieve better results and take use of the advantages of different approaches, it is essential to explore hybrid models (Thong 2015).

In fact, the disease diagnosis issue could be studied with a hybrid data-driven model, such as RF and SVM. In machine learning, the support vector machine (SVM) is one of the pattern recognition approaches, that is created by Cortes and Vapnik (1995). The basic idea of SVM is to construct a hyperplane or set of hyperplanes used for classification and regression. The SVM not only provides better classification results with less training data, but also deals with the dataset with many features. Dongxiao et al. (2017) presented the wavelet least square support vector machine (w-LSSVM) and the ordinary fruit fly optimization algorithm (FOA) was improved as well, and the model was used to forecast power load. Tsyurmasto et al. (2014) proposed a novel SVM model based on value-at-risk (VaR) measure, and this model outperforms traditional SVM on the datasets with outliers. Apart from SVM, the random forest (RF) is a powerful nonparametric machine learning tool for classification and regression, and it combines the ideas of decision tree, CART algorithm and bagging algorithm. The RF could model complex interactions between predictor variables as well as analyze variable importance. Jin et al. (2016) investigated the issue of 3D fast automatic segmentation of kidney, and a modified RF method was proposed to segment the kidney into four components. Calderoni et al. (2015) studied a new indoor localization system based on the RF classifiers and verified the feasibility of this system in hospital environment. From these previous papers, it is demonstrated that modified SVM and RF could outperform classical models. Meanwhile, SVM and RF are two crucial machine learning techniques for classification problem that is similar to CVA diagnosis with the categories of CVA and non-CVA. As a result, it is essential to develop a new data-driven model by combining these two techniques so as to achieve better performance.

3 Proposed intelligent data-driven model

In this section, an intelligent data-driven model is proposed based on machine learning theory, specifically, random forest (RF) and support vector machine (SVM).

By employing the basic concept of SVM (Gao et al. 2019), it is assumed that there are n samples, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. x is a m-dimensional vector, and x_i denote the explanatory variables, $i = 1, 2, \ldots, n$. In this research, the variables are the factors affecting CVA, such as FENO, EOS%, MMEF75/25, FEV1/FVC, PEF, etc. y_i are the category labels, $y_i \in \{-1, +1\}$. In this sample space, a hyperplane exists, and it is represented as Eq. (1).

$$w^T x + b = 0 \tag{1}$$

Based on this hyperplane, two types of samples are generated. Figure 1 shows that n samples are divided into two categories (the circular signs and the triangular signs). w is the weight vector; b is a constant affecting the position of hyperplane; T is the transposition of vector w. Further, the optimal hyperplane H is defined, that is, the hyperplane divides samples into two categories completely and the gap between categories is as wide as possible.

After normalizing, the distance between sample x_1 and hyperplane is,

$$\frac{|w^T x + b|}{\|w\|} = \frac{1}{\|w\|} \tag{2}$$

The sample x_2 is similar to sample x_1. Therefore, the optimal hyperplane is generated by maximizing $\frac{2}{\|w\|}$, that is the sum of distance from samples to hyperplane, and such distance is denoted by margin. To maximize the margin, the appropriate w and b are identified. Meanwhile, the support vectors are samples on H_1 and H_2 that parallel with H. In fact, maximizing $\frac{2}{\|w\|}$ is equivalent to minimum $\|w\|$ or $\|w\|^2$. Hence, the

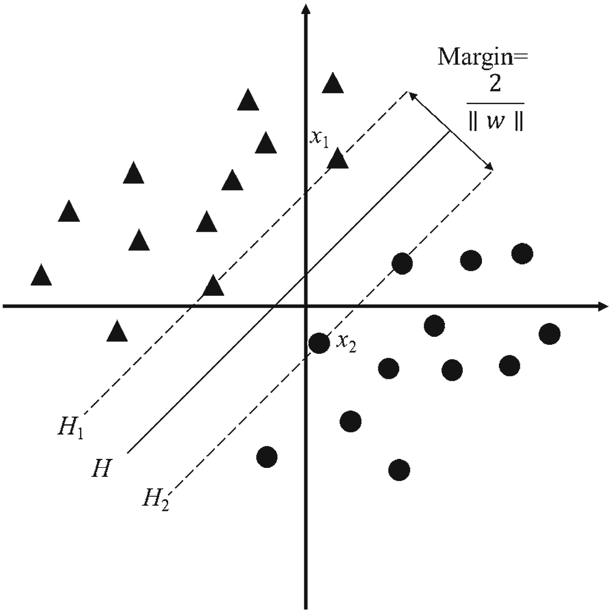

Fig. 1 The schematic diagram of SVM separating samples into two categories

problem is transformed to a quadratic programming problem as follows,

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ s.t. \quad y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \ldots n \end{cases} \tag{3}$$

By introducing the Lagrange function with Lagrange multiplier $a_i (a_i > 0)$, the dual problem corresponding to the above problem is,

$$\begin{cases} \max \left\{ \sum\limits_{i=1}^{n} a_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a_i a_j y_i y_j x_i^T x_j \right\} \\ s.t. \sum\limits_{i=1}^{n} a_i y_i = 0 \quad a_i \geq 0 \quad i = 1, 2, \ldots, n \end{cases} \tag{4}$$

As a result, the judgement function is generated as,

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i y_i x_i^T x + b \right) \tag{5}$$

In terms of the nonlinear classification problem, the data in low-dimensional space is mapped into high-dimensional space by non-linear mapping. Therefore, the sample x is mapped to $\varnothing(x)$, and the kernel function is used for mapping. It is assumed that the kernel function is,

$$k(x_i, x_j) = \varnothing(x_i)^T \varnothing(x_j) \tag{6}$$

Then the problem is expressed as,

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C \sum\limits_{i=1}^{n} \varepsilon_i \\ s.t. \quad y_i(w^T \varnothing(x_i) + b) \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0, i = 1, 2, \ldots n \end{cases} \tag{7}$$

In Eq. (7), the slack variables $\varepsilon_i (\varepsilon_i > 0)$, are introduced to deal with the samples that are not separated to the right category, ε_i therefore indicate the levels of samples deviating from the hyperplane. Meanwhile, the nonnegative constant C is the penalty factor, and the larger the value is, the greater the penalty for the error classification. Further, the corresponding dual problem is,

$$\begin{cases} \max \left\{ \sum\limits_{i=1}^{n} a_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a_i a_j y_i y_j \varnothing(x_i)^T \varnothing(x_j) \right\} \\ s.t. \quad a_i \geq 0, i = 1, 2, \ldots n. \\ \sum\limits_{i=1}^{n} a_i y_i = 0 \end{cases} \tag{8}$$

Then the decision function is,

$$f(x) = sgn\left(\sum_{i=1}^{n} y_i a_i k(x_i, x_j) + b\right) \tag{9}$$

In reality, the people with CVA possibility is a typical two-category problem, with CVA or without CVA. The SVM therefore is an appropriate tool to analyze such problem. In addition, the accuracy rate λ_i is obtained, that evaluates the error between the data predicted by SVM and real data.

Second, the basic idea of RF (Shevchik et al. 2016) is introduced as well. The RF is a typical ensemble learning approach that integrates a number of independent decision trees. The RF holds the view that multiple trees (forest) achieve better predictive performance than utilizing a single tree.

In RF, each tree is built as a binary tree based on the CART (classification and regression tree) algorithm. The training samples are the same as those for SVM. There are n samples with M feature variables, $Z = \{x_i; y_i\}, i = 1, 2 \ldots n, x_i$ denote the explanatory variables and y_i are the category labels. Here, the explanatory variables are also the factors affecting CVA, such as FENO and EOS%.

Figure 2 shows the construction scheme of random forest. According to the concept of bagging algorithm, the bootstrap samples are used for tree training and construction. Specifically, a certain of training samples (number of e) are selected from the training set Z randomly with replacement. This procedure is repeated for many times so as to generate numerous trees as forest. Further, the RF results depend on each tree with the voting mechanism. The one with majority votes determines the final prediction, and each tree has the same voting weight.

Additionally, m variables (denoted by mtry) are randomly chosen from the total M feature variables for constructing trees, $m < M$. At each node, the best variable (split-point) among the mtry is identified to minimize the cost function. Hence, the node is split into two child nodes. For classification problems, the Gini index G is used for measuring the cost. G_l denotes the Gini value for note l,

$$G_l = \sum_{s=1}^{S} p_l^s (1 - p_l^s) \tag{10}$$

where p_l^s is the probability of observation of class $s(s = 1, 2, \ldots, S)$ in node l. This process for each node is repeated until two kinds of species are separated correctly.

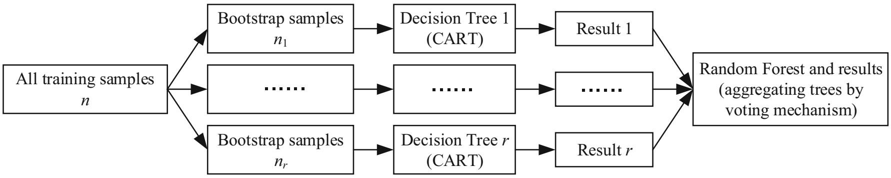

Fig. 2 The construction scheme of random forest

Besides, contrary to traditional CART algorithm, each tree is fully grown and there is no pruning for trees in RF.

Meanwhile, for each tree, the rest of samples that not selected for training data are used for evaluating the accuracy of RF prediction, that is out-of-bag (OOB) samples. For classification problems, the OOB error rate is the misclassification rate, and this rate for tree t is defined as,

$$OOBerr = \frac{1}{|Z_t^C|} \sum_{i \in Z_t^C} I(y_i \neq \hat{y}_i) \tag{11}$$

where Z_t^C is the sample set that not selected for decision tree t construction and constitutes the set of OOB samples. $|Z_t^C|$ indicates the size of corresponding OOB set. y_i is the ith class indicator within the tth OOB set and $\hat{y}_i$ is the class prediction using tree t.

Furthermore, the importance value of each variable is quantified based on OBB error. In this study, diseases are determined by several clinical factors. Hence, the medical decisions usually depend on the factors with higher importance value. It is assumed that $errT_t$ is the error of tree t on its associated OOB sample. Meanwhile, the perturbed sample is employed for which the error of tree t is generated as $\widetilde{errT_t}$. *ntree* is the number of trees in RF. Therefore, for variable x_i, the importance value is expressed as,

$$IMP_{x_i} = \frac{1}{ntree} \sum (\widetilde{errT_t} - errT_t) \tag{12}$$

Similarly, the RF is appropriate for CVA diagnosis as well, and the accuracy rate λ_2 is obtained, that evaluates the error between the data predicted by RF and real data.

Based on the above modules, the accuracy rate λ_1 of SVM module and the accuracy rate λ_2 of RF module are obtained, and the two values reflect the contributions of corresponding modules to the proposed intelligent data-driven model. Therefore, the weight of SVM module and RF module are respectively defined as α_1 and α_2

$$\begin{cases} \alpha_1 = \frac{\lambda_1}{\lambda_1+\lambda_2} \\ \alpha_2 = \frac{\lambda_2}{\lambda_1+\lambda_2} \end{cases} \tag{13}$$

Meanwhile, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the prediction vectors of SVM and RF respectively. Consequently, the final prediction results (vector $\boldsymbol{\eta}$) of the intelligent data-driven model are generated.

$$\boldsymbol{\eta} = \alpha_1 * \boldsymbol{\beta} + \alpha_2 * \boldsymbol{\gamma} \tag{14}$$

The values in vector $\boldsymbol{\eta}$ are the category labels. In this study, the people with CVA is labeled as 1, and the people without CVA is labeled as 0. However, the value η_i in vector η would be between 0 and 1. Hence, the threshold σ is introduced to classify η_i as 0 or 1, see Eq. (15). To this end, the proposed model searches for the most appropriate threshold σ intelligently. The search range is from 0 to 1 with an increasing unit 0.001.

$$\begin{cases} \eta_i = 1 & \eta_i \geq \sigma \\ \eta_i = 0 & \eta_i < \sigma \end{cases} \tag{15}$$

As a result, the intelligent data-driven model is built by integrating SVM and RF. In order to verify the feasibility and reliability of the model, it is then applied to solve a healthcare issue, CVA diagnosis.

4 Experiments and results

4.1 The experimental data

The proposed model is applied to cough variant asthma (CVA) diagnosis. The experimental platform is Intel Core i5-5200U CPU, 4GB RAM and Windows 10 (64 bit). The data is collected from a real hospital in China. This hospital was founded in 1864. There are 48 departments, and 8 departments are national key clinic disciplines, including respiratory department. By cooperating with respiratory department, the research data is obtained.

The data includes 137 samples, specifically, 70 patient samples and 67 non-patient samples. Meanwhile, the data involves 12 attributes for CVA diagnosis, including eosinophils ratio, fractional exhaled nitric oxide, forced vital capacity, etc. The descriptive statistics analysis for the 12 attributes is given in Table 1. Additionally, 0 and 1 are assigned as the category label, the patient with CVA is labeled as 1, and the non-patient without CVA is labeled as 0.

Table 1 The descriptive statistics analysis for the twelve attributes

Attribute	Maximum	Minimum	Mean	Variance
PLT	398.00	81.00	239.93	2944.03
NEUT%	88.90	25.90	58.23	127.28
RBC	5.92	3.59	4.67	0.18
MONO%	20.50	1.60	6.56	6.54
EOS%	17.70	0.00	3.65	12.73
MCH	34.10	24.10	29.96	2.88
FVC	143.90	61.70	98.52	167.60
FEV1%	155.80	34.90	97.71	230.54
FEV1/FVC	99.34	46.29	83.29	58.09
PEF	134.30	28.30	91.32	236.11
MMEF75/25	144.30	12.00	79.59	694.25
FENO	244.00	5.00	48.26	2423.83

The original data of 137 samples is used for calculation

PLT platelet; *NEUT%* neutrophil ratio; *RBC* red blood cell; *MONO%* monocytes ratio; *EOS%* eosinophils ratio; *MCH* mean corpuscular hemoglobin; *FVC* forced vital capacity; *FEV1%* ratio of forced expiratory volume in one second; *FEV1/FVC* ratio of FEV1 and FVC; *PEF* peak expiratory flow; *MMEF75/25* maximum mid-expiratory flow between 25 and 75%; *FENO* fractional exhaled nitric oxide

The 137 samples are divided into two groups randomly, training group and test group. The training group involves 91 samples (47 patient samples and 44 non-patient samples), and 46 samples are in test group (23 patient samples and 23 non-patient samples). Due to the dimensions and units for 12 attributes are not entirely the same, the original data is then normalized as Ep. (16).

$$D_{new} = \frac{D - D_{\min}}{D_{\max} - D_{\min}} \tag{16}$$

where D is the original data; D_{new} is the new data after normalization, $0 \leq D_{new} \leq 1$;$D_{\min}$ is the minimum data of the corresponding attribute; $D_{\max}$ is the maximum data of the corresponding attribute.

4.2 The prediction performance

The new data after normalization is employed as the inputs of the proposed intelligent data-driven model. In SVM module, the training data (91 samples) is used to optimize the SVM module, and the module performance is evaluated with the test data (46 samples). Consequently, 36 test samples are classified into the right category, hence the accuracy rate λ_1 is 78.2609%. The key parameters in this process include: the set type of SVM is nu-SVC; the type of kernel function is POLY (polynomial kernel function); the optimal objective value of the dual SVM problem is 875.0887; the bias term in the decision function is -1.2067; the number of support vectors and bounded support vectors are 49 and 41 respectively; the number of iterations is 42. Additionally, the prediction result of SVM module is denoted as vector $\boldsymbol{\beta}$. Simultaneously, the same training data and test data is utilized for RF module. Hence, 30 test samples are classified into the right category, and the accuracy rate λ_2 is 65.2174%. In this process, 500 decision trees are generated and the number of mtry is 3. Similarly, the prediction result of RF module is recorded as vector $\boldsymbol{\gamma}$.

Based on the value of λ_1 and λ_2, the contributions of these two modules are quantified with Eq. (13). The weight α_1 of SVM module is 0.5455, and the weight α_2 of RF module is 0.4545. Further, the final prediction results (vector $\boldsymbol{\eta}$) of the intelligent data-driven model are generated with Eq. (14). The fractions in $\boldsymbol{\eta}$ hinder the classification with label 0 and 1. Therefore, the classification threshold σ is identified by the proposed model intelligently. The search range is from 0 to 1 with an increasing unit 0.001. It indicates that the iterations for the search process is 1000, and Fig. 3 illustrates the search results.

From Fig. 3, it is clearly found that:

1. The most appropriate threshold σ is 0.546. Therefore, if $\eta_i \geq 0.546$, it is classified as 1, and if $\eta_i < 0.546$, it is classified as 0.
2. Based on the threshold σ (0.546), the accuracy rate λ of the proposed intelligent data-driven model is the best with the value of 80.43%. It should be noticed that the prediction performance of the proposed model ($\lambda = 80.43\%$) outperforms the single SVM approach ($\lambda_1 = 78.26\%$) and single RF approach ($\lambda_2 = 65.22\%$).

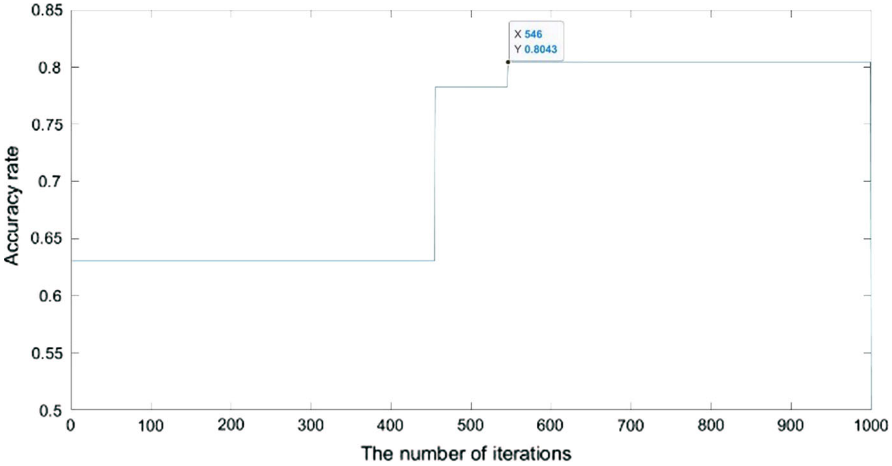

Fig. 3 The searching results for classification threshold

As a result, it is demonstrated that the proposed model could be a user-friendly tool to enhance the diagnosis accuracy for CVA.

4.3 The importance of variables

The impact of the 12 variables on CVA is obviously different, that then affects medical decisions. Therefore, identifying variables importance is essential for healthcare issue. By using the proposed intelligent data-driven model, the importance of the 12 variables is computed with the experimental data, and Fig. 4 shows the results. The top 5 important variables are FENO (fractional exhaled nitric oxide), EOS% (eosinophils ratio), MMEF75/25 (maximum mid-expiratory flow between 25%-75%), FEV1/FVC (ratio of forced expiratory volume in one second and forced vital capacity) and PEF (peak expiratory flow), with the importance values of 14.8831, 11.6639, 7.6610, 4.7220 and 4.6642 respectively. As a result, these attributes are the crucial indicators for

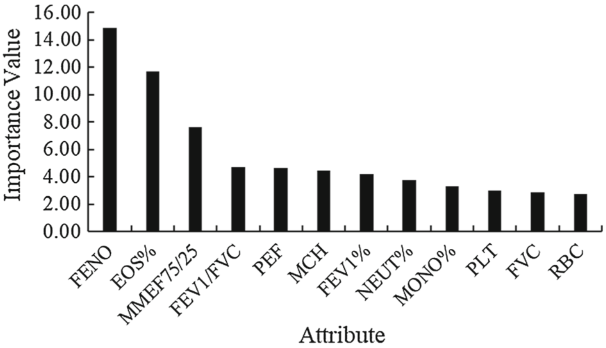

Fig. 4 The importance value of variables

CVA, and the efficiency and accuracy for diagnosis could be improved significantly. Meanwhile, this case also verifies the application value of the proposed model in healthcare field.

5 Conclusions

In this era, various activities are reformed by massive data. In this paper, an intelligent data-driven model is developed for disease diagnosis. This model integrates the advantages of both SVM (support vector machine) and RF (random forest), and it is then applied to the diagnosis of CVA (cough variant asthma). The experimental data is collected from a real hospital in China, that includes 137 samples. According to the experiment results, it is demonstrated that the proposed model outperforms single SVM approach and single RF approach in prediction performance. Meanwhile, the importance levels of factors affecting CVA are identified, and the top 3 important factors are FENO, EOS% and MMEF75/25. Based on the case application, it is verified that the proposed model is an effective tool in healthcare field. This research proposed a hybrid model by combining two machine learning approaches, with theoretical contribution. In terms of application value, the model could effectively improve CVA diagnosis performance. The future study could investigate applications of the model in other areas and related parameters could be optimized further.

Acknowledgements This research is supported by five projects: The project of Shanghai Shenkang hospital development center, clinical science and technology optimization (SHDC12017623); The doctoral start-up project of USST (BSQD201901); National natural science foundation of China (71840003, 71801150); The scientific and technological development project of USST (2018KJFZ043).

References

Alam MGR, Haw R, Kim SS, Azad MAK, Abedin SF, Hong CS (2016) Em-psychiatry: an ambient intelligent system for psychiatric emergency. IEEE Trans Ind Inform 12(6):2321–2330

Bahri S, Zoghlami N, Abed M, Tavares JMR (2018) Big data for healthcare: a survey. IEEE Access 7:7397–7408

Bai Y, Han X, Chen T, Yu H (2015) Quadratic kernel-free least squares support vector machine for target diseases classification. J Combin Optim 30(4):850–870

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016) An analytics approach to designing combination chemotherapy regimens for cancer. Manag Sci 62(5):1511–1531

Calderoni L, Ferrara M, Franco A, Maio D (2015) Indoor localization in a hospital environment using random forest classifiers. Exp Syst Appl 42(1):125–134

Cao X, Liu L, Cheng Y, Shen XS (2017) Towards energy-efficient wireless networking in the big data era: a survey. IEEE Commun Surv Tutor 20(1):303–332

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Dapogny A, Bailly K, Dubuisson S (2018) Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. Int J Comput Vis 126(2–4):255–271

Dongxiao N, Tiannan M, Bingyi L (2017) Power load forecasting by wavelet least squares support vector machine with improved fruit fly optimization algorithm. J Combin Optim 33(3):1122–1143

Gai L, Ji J (2019) An integrated method to solve the healthcare facility layout problem under area constraints. J Combin Optim 37(1):95–113

Gao W, Bao W, Zhou X (2019) Analysis of cough detection index based on decision tree and support vector machine. J Combin Optim 37(1):375–384

Jin C, Shi F, Xiang D, Jiang X, Zhang B, Wang X, Zhu W, Gao E, Chen X (2016) 3D fast automatic segmentation of kidney based on modified AAM and random forest. IEEE Trans Med Imaging 35(6):1395–1407

Khanmohammadi S, Adibeig N, Shanehbandy S (2017) An improved overlapping k-means clustering method for medical applications. Expert Syst Appl 67:12–18

Liu L, Tang G, Fan B, Wang X (2015) Two-person cooperative games on scheduling problems in outpatient pharmacy dispensing process. J Combin Optim 30(4):938–948

Shevchik SA, Saeidi F, Meylan B, Wasmer K (2016) Prediction of failure in lubricated surfaces using acoustic time-frequency features and random forest algorithm. IEEE Trans Ind Inform 13(4):1541–1553

Thong NT et al (2015) Hifcf: an effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. Expert Syst Appl 42(7):3682–3701

Tsyurmasto P, Zabarankin M, Uryasev S (2014) Value-at-risk support vector machine: stability to outliers. J Combin Optim 28(1):218–232

Yadav P, Steinbach M, Kumar V, Simon G (2018) Mining electronic health records (EHRS): a survey. ACM Comput Surv (CSUR) 50(6):85

Yang Y, Shen B, Gao W, Liu Y, Zhong L (2015) A surgical scheduling method considering surgeons' preferences. J Combin Optim 30(4):1016–1026

Zhang Y, Qiu M, Tsai CW, Hassan MM, Alamri A (2015) Health-CPS: healthcare cyber-physical system assisted by cloud and big data. IEEE Syst J 11(1):88–95

Zhong L, Bai Y (2019) Three-sided stable matching problem with two of them as cooperative partners. J Combin Optim 37(1):286–292

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.