COPD Multidimensional Feature Extraction and Integrated Diagnostic Method Postprint

Authors: Fang Youli, Wang Hong, Di Ruitong, Wang Lutong Song Yongqiang

Date: 2018-06-19T00:00:00+00:00

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a chronic pulmonary disorder that causes progressive deterioration of respiratory function in patients, requiring the assistance of big data analytics and algorithms to facilitate more accurate disease diagnosis by clinicians. Current COPD research exhibits limitations: firstly, studies merely utilize data analysis to examine the impact of single features on the disease; secondly, research outcomes only validate case data through simplistic algorithmic models. Therefore, this paper proposes a COPD multi-dimensional feature extraction and ensemble diagnosis methodology. Initially, a Maximum Dependency MDF-RS algorithm is introduced to extract optimal combinations of multi-dimensional features. Subsequently, a DSA-SVM ensemble model is proposed to construct classifiers for diagnostic and predictive purposes. Finally, cross-validation methods are employed to verify various performance metrics including accuracy. The effectiveness of the proposed algorithm is validated through comparative experimental verification.

Full Text

Preamble

Multidimensional Feature Extraction and Integrated Diagnosis Method for COPD

Fang Youli[1,1,2], Wang Hong[1,1,2†], Di Ruitong[1,1,2], Wang Lutong[1,1,2], Song Yongqiang[1,1,2]

(1. a. School of Information Science & Engineering; b. College of Life Science, Shandong Normal University, Jinan 250358, China; 2. Shandong Provincial Key Laboratory of Distributed Computing Software, Jinan 250358, China)

Abstract: Chronic obstructive pulmonary disease (COPD) is a chronic lung disease that leads to progressive decline in respiratory function, requiring big

data analysis and algorithms to assist physicians in more accurate diagnosis. Current COPD research suffers from limitations: on one hand, studies only analyze the impact of single features on disease using data; on the other hand, research merely validates case data through simple algorithm models. This paper proposes a COPD multidimensional feature extraction and integrated diagnosis method. First, we introduce the Maximum Dependency Degree MDF-RS algorithm to extract optimal combinations of multidimensional features. Second, we propose the DSA-SVM integrated model to construct classifiers for diagnosis and prediction. Finally, cross-validation methods verify accuracy and other performance metrics. Experimental comparisons demonstrate the effectiveness of the proposed algorithm.

Keywords: COPD; multidimensional features; integration methods; cross validation

0 Introduction

Chronic obstructive pulmonary disease (COPD) is a chronic lung condition causing progressive respiratory function decline, ranking as the fourth leading cause of death globally with over 170 million patients worldwide. COPD progression is gradual: in early stages, symptoms are subtle—primarily cough and sputum production—making it difficult for patients to recognize, yet this represents the optimal treatment window. In intermediate stages, worsening conditions may cause dyspnea on exertion, aggravated airway obstruction, and irreversible loss of lung elasticity, rendering most medications ineffective. Advanced stages may develop complications such as pulmonary heart disease and respiratory failure, severely impacting quality of life and physical/mental health if treatment is delayed. Therefore, early COPD detection is crucial, requiring long-term, stable patient management. Without prevention and management, disease progression—particularly acute exacerbations—poses greater risks. Acute exacerbations involve sudden deterioration of cough, sputum, dyspnea, chest tightness, and wheezing, potentially requiring treatment modifications.

With advances in computer data mining technology, COPD analysis has become a research hotspot. Data mining is now widely applied in COPD pathological analysis and clinical diagnosis. Current research focuses on two aspects: (a) using existing analytical tools to mine electronic medical records and analyze the impact of single features on disease, and (b) validating COPD patient prognosis risk through simple models. Researchers have conducted extensive work on analyzing feature influencing factors and risk prediction at different disease stages with promising results. Himes et al. extracted features and demographic information from asthma patient records to build a COPD prediction model, achieving 0.83 accuracy using eight features including age, gender, ethnicity, and smoking history to predict COPD risk in independent asthma patients. Hoogendoorn et al. identified key predictive factors including cough, wheezing, sputum, 6-minute walk test, inhaled corticosteroid use, and oxygen saturation, though low BMI, cardiovascular disease, and emphysema were also important

predictors for secondary care hospitalization. Guo Huimin et al. used R language for model identification, parameter estimation, and testing, constructing an ARIMA model from monthly admission data that effectively fitted COPD admissions and provided short-term predictions, showing 2016 admissions would increase and offering valuable guidance for medical resource allocation.

Other researchers have analyzed COPD risk prediction at different stages. Literature [6–8] examined prerequisites for risk stratification to improve diagnosis and treatment, avoiding higher health-related quality of life impacts, longer lifespans, and lower medical costs. Literature [9,10] experimentally validated risk-stratified treatment through prediction models. Mega et al. evaluated the BODE index (a multidimensional grading system for predicting mortality) to assess its ability to predict COPD patient conditions, concluding it was a better predictor of acute exacerbation frequency and severity.

This paper proposes a hybrid decision-making method based on DSA-SVM algorithm for COPD diagnosis, constructing a classifier that extracts multidimensional features through the Maximum Dependency Degree MDA-RS algorithm and validates accuracy metrics via cross-validation.

2 COPD Methods

This paper' s main contributions include: (a) proposing the MDA-RS algorithm to extract optimal COPD feature subsets for better classification results; (b) proposing the DSA-SVM hybrid model for COPD classification and prediction; and (c) conducting extensive experiments to demonstrate method effectiveness. Our objective is to analyze chronic obstructive pulmonary disease patient data, extract feature subsets from multidimensional features, and diagnose/predict disease using the hybrid decision model DSA-SVM algorithm. To achieve this, we address four problems through the following steps: data preprocessing, multidimensional feature selection using MDA-RS algorithm, DSA parameter optimization, and construction of the DSA-SVM classifier hybrid decision model.

2.1 Data Preprocessing

Data preprocessing aims to improve data quality, making data mining more effective and easier while enhancing result quality. Preprocessing targets noisy and missing data, primarily using data cleaning, correlation analysis, and data transformation. This paper applies appropriate preprocessing to raw data:

a) Converting categorical attributes to numerical values. Each category is represented numerically (e.g., smoking = 1, non-smoking = 0).

b) Handling missing values. The COPD dataset' s 8th and 16th features contain 37 and 23 missing values respectively, which can be filled using the mode of each attribute.

c) Data normalization. For example, the forced expiratory volume in one second to forced vital capacity ratio (FEV1/FVC) can be normalized to

(0~1) range. Normalization facilitates computation and improves precision. The normalization formula is shown in Equation (1), where X_{norm} is normalized data, X is original data, and X_{max} and X_{min} are maximum and minimum values of original data.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2.2 Multidimensional Feature Selection

Advances in image processing, information retrieval, and bioinformatics have generated ultra-large-scale multidimensional datasets. Effectively extracting or selecting useful feature information and patterns for classification has become a fundamental challenge. Feature selection chooses an optimal feature subset from the original feature set based on evaluation criteria, enabling classification or regression models built on this subset to achieve similar or better predictive accuracy than using all features. RS (Rough Set) is a mathematical method for feature selection, extraction, reduction, and decision rule extraction, particularly effective with uncertain and incomplete data. This paper proposes the Maximum Dependency Degree Feature Selection algorithm (MDF-RS) based on rough set theory, using likelihood ratio tests for validation.

2.2.1 RS Rough Set Theory RS is an effective data processing method with strong classification capability that can reduce knowledge (features) while maintaining classification consistency. In literature [17], a knowledge system is defined as $S = (U, A, V, f)$, where U is a non-empty object set; A is a non-empty feature set; $V = \bigcup_{a \in A} V_a$ is the value domain of feature a; and $f : U \times A \to V$ is a knowledge function specifying feature values for each object in U.

Definition 1 Let $S = (U, A, V, f)$ be a knowledge system, $B \subseteq A$ be any subset of features. The indiscernibility relation with respect to B is denoted as $IND(B)$. Clearly, every subset of A can derive a unique indiscernibility relation (also called equivalence relation), which in turn derives a unique clustering. The equivalence classes of U derived by B are denoted as U/B.

Definition 2 In knowledge system $S = (U, A, V, f)$, let $X \subseteq U$ be any subset of objects and $B \subseteq A$ be any subset of features. The lower approximation of X with respect to B is denoted as $\underline{B}X$, and the upper approximation as $\overline{B}X$. The boundary region can be represented by the lower approximation of X' s complement, as shown in Equation (3). The approximation precision of any subset $X \subseteq U$ with respect to B is expressed in Equation (4).

$$\underline{B}X = \{x \in U \mid [x]_B \subseteq X\}$$
$$\overline{B}X = \{x \in U \mid [x]_B \cap X \neq \emptyset\}$$

where $[x]_B$ denotes the equivalence class of x under relation B. For empty set $\emptyset$, $\underline{B}\emptyset = \overline{B}\emptyset = \emptyset$. If X is a union of some equivalence classes of B, then $\underline{B}X = \overline{B}X = X$, and we say X is precise with respect to B. Conversely, if X is not a union of some equivalence classes, then $\underline{B}X \neq \overline{B}X$, and we say X is not precise with respect to B. Higher approximation precision $\alpha_B(X)$ indicates a more precise subset.

2.2.2 Feature Dependency Degree In rough set theory, feature importance can be understood as follows: given a knowledge system $S = (U, A, V, f)$ and a feature subset $B \subseteq A$, if adding feature a to B improves the system' s ability to distinguish objects, this improvement degree represents feature importance. Greater improvement indicates higher importance of a to B. Feature dependency degree reveals intrinsic relationships: dependency between important features is small, while dependency between important and secondary features is strong, and dependency between unimportant features and important/secondary features is small. Thus, we can remove unimportant classification features or extract important ones through dependency degree.

Definition 3 In knowledge system $S = (U, A, V, f)$, let C and D be any subsets of feature set A. If every value in C can be precisely associated with a value in D, then D is functionally dependent on C, denoted as $C \Rightarrow D$. The dependency degree is defined in Equation (5): let k be the dependency degree, D depends on C to degree k, denoted as $C \Rightarrow_k D$.

$$k = \gamma_C(D) = \frac{|POS_C(D)|}{|U|}$$

where $POS_C(D) = \bigcup_{X \in U/D} \underline{C}X$ is the positive region. If $k = 1$, D completely depends on C; if $0 < k < 1$, D partially depends on C. The coefficient k describes the ratio of elements in U that can be correctly classified into blocks of U/D using feature C. When $k = 1$, all or part of U' s elements can be partitioned into equivalence classes of C. When $k = 0$, no element in U can be partitioned using C. Thus, higher inter-feature dependency degree has greater impact on classification decisions.

2.2.3 Feature Maximum Dependency Degree Algorithm (MDF-RS) Since higher feature dependency degree indicates greater importance and stronger influence on classification decisions, the goal of the Maximum Dependency Degree algorithm is to select features with maximum dependency as classification attributes. The algorithm steps are:

a) Calculate equivalence classes for each feature using indiscernibility relations;
b) Compute feature dependency degree using Equation (5);
c) Select the maximum dependency degree for each feature;

d) Choose attributes with maximum dependency degree as classification features.

Example of Maximum Dependency Degree Selection: Assume four attributes A, B, C, D with dependency degrees shown in . Comparing all dependency degrees k, the maximum $k = 1$ appears on attributes A and B. Then comparing other dependencies of A and B, the maximum $k = 0.4$ appears on attribute B (A→B = 0.4). Thus, B is selected as the classification feature attribute.

2.2.4 Feature Selection Based on MDF-RS This paper applies the proposed MDF-RS algorithm for COPD multidimensional feature selection:

a) Feature clustering. Clustering groups functionally similar features together. To extract low-redundancy features, K-means clustering analyzes original data features, using Euclidean distance to measure inter-point distances and Sum of Squared Errors (SSE) as the clustering objective function to minimize SSE, as shown in Equations (6) and (7), where c_i represents k cluster centers.

$$d(x, c_i) = \sqrt{(x_1 - c_{i1})^2 + (x_2 - c_{i2})^2}$$

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} \text{dist}(x, c_i)^2$$

b) Main feature selection. After clustering, each group contains functionally similar features, so we select primary features to represent each category and combine them into a feature set. The COPD feature selection method is described in Algorithm 1.

Algorithm 1: Feature Selection of COPD

```
Input: Sample set S, feature set A = {A , A , ..., A }
Output: Feature subset G
1. Cluster features using K-means to get k clusters G = {g , g , ..., g }
2. FOR each cluster g  (i = 1 to k)
3. {
4.   Calculate sample equivalence classes
5.   Calculate feature dependency degree
6.   Compare and select main features from category
7.   Add selected features to G
8. }
9. return G
```

2.3 Direct Search Simulated Annealing Algorithm DSA

DSA (Direct Search Simulated Annealing) is an improvement over SA (Simulated Annealing). DSA differs from SA in two key aspects: SA maintains only

one current optimal point, searching near a single point, which may lead to local convergence; DSA maintains a set of working points, searching near this set to effectively escape local optima. The improved DSA algorithm is shown in Algorithm 2.

Algorithm 2: A Direct Search Variant of the Simulated Annealing Algorithm

```
Input: Initial state G , precision P(s)
Output: Optimal solution G_best, optimal score p_best
1. G = G , p = P(s)  // Initial state, precision
2. G_best = G, p_best = p
3. k = 0; k_max = Constant Value
4. MaxScore = A constant Value  // Evaluation count
5. while (k < k_max && p <= MaxScore)
6. {  // While time left & not good enough
7.   G_new = Neighbor(G)
8.   p_new = P(G_new)
9.   if exp(p_new - p) > Random()
10.  {
11.     G = G_new  // Yes, change state
12.     p = p_new
13.  }
14.  if p_new > p_best
15.  {
16.     G_best = G_new  // Save 'new neighboring' 'best found'
17.     p_best = p_new
18.  }
19.  k = k + 1
20.}
21. return G_best, p_best  // Return the best solution found
```

2.4 Model Establishment

The classifier model construction flowchart is shown in [Figure 1: see original paper]. The model first outputs optimal (C,) values, then builds the classifier. After initializing DSA parameters, SVM parameters (C,) are randomly initialized. Neighbors are selected and adjusted via DSA search. Different (C,) combinations are compared using cross-validation to continuously optimize parameters. To further adjust kernel function parameters, we construct a virtual window around the best local (C,) until parameters stabilize within an acceptable range. Finally, the optimal (C,) parameters are used to establish the DSA-SVM model and test the dataset. The parameter interval is set to (2 , 2^1) for C and (2^{-1} , 2) for , with all possible combinations evaluated via cross-validation.

Cross-validation is used in DSA to optimize parameters through k-fold cross-

validation:

a) Randomly divide sample set S into k disjoint subsets, each containing m/k samples, denoted as S , S , ···, S ;
b) Use k-1 subsets as training set S_train and remaining subset as validation set S_validation to train model and compute generalization error;
c) Calculate average generalization error for each model and select model with minimum error according to Equation (8).

The (C,) combinations obtained through cross-validation are shown in Equation (9).

After obtaining optimal (C,), we build a Pairwise Coupling (PWC) probabilistic classifier. PWC is a popular multi-level classification method that combines all pairwise class comparisons, constructing r = k(k−1)/2 binary classifiers for 1 i k, 1 j i. Classification decisions are made by aggregating classifier outputs. Binary classifiers estimate pairwise class probabilities, with estimated by training on classes i and j. We use Du' s method [20] to compute these probabilities, then estimate p(y = i | x) for i = 1···K using all r . During testing, each classifier estimates classification result probability as shown in Equation (10).

3.1 COPD Dataset

The COPD dataset was extracted from electronic medical records of partner healthcare systems, selecting patients with at least 5 years of observation data. The goal is to predict COPD presence through various medical test results and symptom presentations. The dataset contains 1200 samples across two categories: 750 COPD patients (62.5%) and 450 non-COPD patients with similar symptoms (37.5%). We extracted 26 original features from electronic medical records, described in .

3.2 COPD Feature Selection Results

Extracting high-dimensional features from the original dataset is critical for model accuracy. Traditional learning methods cannot extract optimal feature combinations, so we propose the MDF-RS algorithm to obtain feature subsets.

a) Feature clustering. Using K-means clustering with k = 7 based on original data characteristics, features are clustered into 7 groups, with results shown in [Figure 2: see original paper].

b) Main feature selection. From the 7 clustered feature groups, MDF-RS selects primary features to form subsets. shows feature combinations comprising 9 to 19-dimensional subsets, yielding 14 feature combinations (R1-R14) via MDF-RS. After feature weight normalization, features are sorted by weight as shown in [Figure 3: see original paper]. The optimal feature subset combination is used as input for the DSA-SVM model. Likelihood

ratio test results are shown in , with test statistic formula in Equation (11).

$$\mathrm{LR} = -2\ln\left(\frac{L_0}{L_1}\right)$$

All 19 test statistics exceed $\chi^2_{0.05,1} = 3.84$, $p < 0.05$, indicating statistical significance, consistent with feature combination R13 selected by MDF-RS. Results demonstrate that each variable significantly impacts the model when other 18 variables remain constant, confirming these 19 multidimensional features are meaningful for COPD diagnosis.

3.3 Experimental Results DSA-SVM

After feature selection via MDF-RS, DSA-SVM classifies the dataset. Parameter combination (C,) is crucial for model accuracy, so we use direct search simulated annealing to optimize SVM parameters. A virtual window is built around local optimal parameters with threshold settings until parameters stabilize within acceptable range. Cross-validation finds the optimal (C,) combination. We represent parameter combinations and corresponding accuracy in a 3D plot in [Figure 4: see original paper], where the boxed point shows optimal parameters (C = 14.5, = 0.352) achieving 94.6% accuracy. Classification accuracy for different feature combinations R is shown in .

3.4 Experimental Comparison

Many researchers have diagnosed COPD using single and hybrid methods, but limitations exist in handling missing values and model parameters. Our MDF-RS feature extraction and DSA-SVM classification model achieve good results.

First, we compare our method with previous machine learning models. The proposed DSA-SVM algorithm achieves good performance in accuracy, recall, and F1-score, as shown in .

TABLE:6 Method Comparison | Method | Accuracy | Recall | F1-Score | | ———-|——-|—-|——-| | Logistic | 90.3% | 85.65% | 89.2% | | Decision Tree | 92.26% | 83.7% | 87.8% | | XGBoost | 93.76% | 87.43% | 90.4% | | DSA-SVM | 94.6% | 93.2% | 92.9% |

Second, we compare with literature [3,5]. Himes used Bayesian network models for COPD prediction with 83.3% accuracy, while Guo et al. used ARIMA models achieving 90.2% accuracy. Our method improves accuracy and F1-score, with AUC reaching 0.94, demonstrating effective results. Comparison charts for accuracy, F1-score, and ROC are shown in [Figure 5: see original paper]-[Figure 7: see original paper].

4 Conclusion

This paper proposes a DSA-SVM model and MDF-RS multidimensional feature selection algorithm for COPD diagnosis and prediction. Comparisons across various metrics show the DSA-SVM diagnostic algorithm achieves good results for COPD. The proposed DSA-SVM diagnostic algorithm can assist physicians in making final decisions, helping diagnose COPD and reducing misdiagnosis rates. Future COPD diagnosis research will explore different feature extraction methods and other learning approaches to improve diagnostic system accuracy.

References

[1] Lópezcampos J L, Tan W, Soriano J B. Global burden of COPD [J]. Journal of Applied Mathematics, 2016, 21(1): 14.

[2] Celik M U, Sharma G, Tekalp A M. Lossless watermarking for image authentication: a new framework and an implementation [J]. IEEE Trans on Image Processing, 2006, 15(4): 1042-1049.

[3] Himes B E, Dai Y, Kohane I S, et al. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records [J]. Journal of the American Medical Informatics Association, 2009, 16(3): 371-379.

[4] Hoogendoorn M, Feenstra T L, Boland M, et al. Prediction models for exacerbations in different COPD patient populations: comparing results of five large data sources [J]. International Journal of Chronic Obstructive Pulmonary Disease, 2017, 12(5): 3183-3194.

[5] Guo Huimin, Du Jun, Huang Lufei. Application of ARIMA model based on R language in predicting incidence of patients with acute exacerbation of chronic obstructive pulmonary disease [J]. Chinese Health Statistics, 2017, 34(2): 288-289.

[6] Steyerberg E W. Clinical Prediction Models [M]. Springer US, 2010.

[7] Moons K G M, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? [J]. Bmj, 2009, 338(7706): b375-b383.

[8] Steyerberg E W, Moons K G M, Windt D A V D, et al. Prognosis research strategy (progress) 3: prognostic model research [J]. PLoS Medicine, 2013, 10(2): e1001381-e1001389.

[9] Moons K G M, Kengne A P, Grobbee D E, et al. Risk prediction models: II. External validation, model updating, and impact assessment [J]. Heart, 2012, 98(9): 691-699.

[10] Bleeker S E, Moll H A, Steyerberg E W, et al. External validation is necessary in prediction research: a clinical example [J]. Journal of Clinical Epidemiology, 2003, 56(9): 826-832.

[11] Mega J L, Braunwald E, Wiviott S D, et al. Rivaroxaban in patients with a recent acute coronary syndrome [J]. New England Journal of Medicine, 2012, 366(1): 9-18.

[12] Cheung N. Machine learning techniques for medical analysis [J]. Journal of Clinical Epidemiology, 2017, 26(4): 126-132.

[13] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease [J]. Applied Soft Computing Journal, 2013, 13(8): 3429-3438.

[14] Kaneiwa K. A rough set approach to multiple dataset analysis [M]. Elsevier Science Publishers B. V. 2011.

[15] Chen Y, Miao D, Wang R. A rough set approach to feature selection based on ant colony optimization [J]. Pattern Recognition Letters, 2010, 31(3): 226-233.

[16] Shafiq M, Atif M, Viertl R. Generalized likelihood ratio test and cox' s f-test based on fuzzy lifetime data [M]. John Wiley & Sons, Inc. 2017.

[17] Thangavel K, Pethalakshmi A. Dimensionality reduction based on rough set theory: a review [J]. Applied Soft Computing, 2009, 9(1): 1-12.

[18] Ali M M, Törn A, Viitanen S. A direct search variant of the simulated annealing algorithm for optimization involving continuous variables [J]. Computers & Operations Research, 2002, 29(1): 87-102.

[19] Sartakhti J S, Zangooei M H, Mozafari K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA) [J]. Computer Methods & Programs in Biomedicine, 2012, 108(2): 570-579.

[20] Du Zhanlong, Li Xiaomin, Xi Leiping, et al. Multi-class probabilistic extreme learning machine and its application in remaining useful life prediction [J]. Systems Engineering and Electronics, 2015, 37(12): 2777-2784.

Note: Figure translations are in progress. See original paper for figures.