An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation

Mahdi Abdollahi[1], Xiaoying Gao[1], Yi Mei[1], Shameek Ghosh[2], Jinyan Li[3]
[1]*School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand*,
[2]*Founder & CTO, Medius Health, Sydney, Australia*,
[3]*Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia*
{mahdi.abdollahi, xiaoying.gao, yi.mei}@ecs.vuw.ac.nz, Shameek.ghosh@mediushealth.org, jinyan.li@uts.edu.au

***Abstract*—Document classification (DC) is the task of assigning pre-defined labels to unseen documents by utilizing a model trained on the available labeled documents. DC has attracted much attention in medical fields recently because many issues can be formulated as a classification problem. It can assist doctors in decision making and correct decisions can reduce the medical expenses. Medical documents have special attributes that distinguish them from other texts and make them difficult to analyze. For example, many acronyms and abbreviations, and short expressions make it more challenging to extract information. The classification accuracy of the current medical DC methods is not satisfactory. The goal of this work is to enhance the input feature sets of the DC method to improve the accuracy. To approach this goal, a novel two-stage approach is proposed. In the first stage, a domain-specific dictionary, namely the Unified Medical Language System (UMLS), is employed to extract the key features belonging to the most relevant concepts such as diseases or symptoms. In the second stage, PSO is applied to select more related features from the extracted features in the first stage. The performance of the proposed approach is evaluated on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set which is a widely used medical text dataset. The experimental results show substantial improvement by the proposed method on the accuracy of classification.**

***Index Terms*—Medical Text Classification, Particle Swarm Optimization, Feature Selection, Conceptualization, Ontology.**

I. INTRODUCTION

Text mining is one of the important topics in artificial intelligence which deals with analyzing different types of unstructured text to extract useful knowledge. There are many tasks in text mining such as text classification, text clustering, entity extraction, document summarization and semantic analysis. Text classification is one of the extensively studied natural language processing tasks. In text classification, the goal is to automatically classify text documents into one or more predefined classes. For example, detecting spam and non-spam emails, automatically tagging client queries and categorizing news articles are some applications of text classification. The main steps of text classification consist of preprocessing, text representation, feature weighting, feature selection, training, testing and interpretation.

In text classification, the bag-of-words model using all of the unique words in the documents as features is the simplest way of text representation [1]. The first problem with this method is that the number of features is big and the second issue is the existence of noisy features. The massive data can reduce the learning speed and increase the time cost. The noisy data can have negative effects on the learned model which lead to poor accuracy of label prediction. Feature selection can improve the performance of classification by selecting meaningful features and at the same time reducing the number of noisy features.

The early feature selection algorithms in analyzing text is single feature ranking [2], which is a filter technique of choosing m features as a sub set from the n features by considering the top m features based on their rank. The significance of each feature is defined by its contribution to the classification task, and some basic measures of relevance are: Logistic Regression [3], Statistical Testing [4], Pearsons correlation [5], and distinctive data hypothesis based measures [6]. Single feature ranking is a simple method with a low computation cost which encourages many researchers to apply it in their research, however, it does not consider interactions between features. As a matter of fact, most filter methods assess features separately and cannot distinguish interactions between features [5].

A large search space is one of the difficulties in feature selection problem and the number of feasible feature subsets will grow by increasing the number of the original features. An effective global search method is needed to address the issues of feature selection. Evolutionary computation (EC) methods have been utilized in solving the feature selection problem because they have robust search abilities [7]. Among EC methods, Particle Swarm Optimization (PSO) [8] is a powerful technique. PSO is a population-based EC algorithm which each individual is a particle. Particles are a set of solutions [9], which represent feature subsets in our case. The convergence of PSO is quick and it has only a few parameters to set [7]. Based on the mentioned properties of PSO, it is a suitable candidate for feature selection. PSO has achieved good performances in solving feature selection in different fields [10]–[12]. Some preliminary researches have used PSO for medical image classification and promising outcomes have been accomplished [13]. However, the use of

PSO is still generally new in medical document classification, it has potential for further research in clinical discharge note classification.

The majority of previous researches on text classification utilizes only one method to carry out feature selection, and have problems because of the extremely large search space [14]. Specifically, given an extremely large feature set, a single feature selection method such as PSO can still result in a large number of selected features, which limits the effectiveness of the feature selection. To overcome this drawback and improve the effectiveness of feature selection in a very high dimensional feature space, this paper targets to develop a two-stage method to extract and select meaningful features for text classification. The first stage detects meaningful phrases from the original documents as features and then extracts their concepts by considering the domain of documents and labels. In this approach, we use a tool to extract concepts to reduce the number of extracted features. We targeted on clinical notes classification, so we used a domain specific ontology for feature extraction. After extracting features from the raw notes, particle swarm intelligence (PSO) will be applied for feature selection to find a more meaningful subset. In this paper, we aim to investigate the following research questions:

1) Whether the proposed method can extract meaningful information from the document set;
2) Whether the proposed method can reduce the number of features and keep the meaningful features; and
3) Whether the proposed method can increase the classification accuracy in the aimed clinical notes classification.

The rest of the paper is organized as follows: section 2 describes the problem and summaries the related work. Our method is presented in section 3. The obtained experimental results are shown in section 4. At the end, the conclusions and future work are presented in section 5.

II. BACKGROUND

A. Clinical Text Mining

In clinical text mining, the text describes a set of clinical events within a narrative, with the goal of producing an explanation as precise and comprehensive as possible when describing the health status of a patient. Generally, such texts include a heavy use of domain specific terminology and the frequent inclusion of acronyms, which makes clinical text analysis very different from standard text mining. Specially, a discriminative combination of domain-specific medical events reported within a clinical note can be highly indicative of a patient's condition.

There has been research that applies text classification on clinical text. Previously, Pratt et al [15] employed words, medical phrases, and their combinations as features for medical document classification. Multi-label classification performance based on an associative classifier is examined on medical articles [16]. In another study, Hidden Markov models were used for classification [17]. In a recent study, an approach using support vector machines and latent semantic indexing was applied to some data sets including the ones consisting of medical abstracts [18]. The performances of classifiers on medical document classification was analyzed for two cases where stemming was applied and not applied [19]. The impact of different text representations of biomedical texts on the performance of classification were analyzed [20]. Feature selection methods using Gini Index were employed along with models like Bayesian networks and decision trees to improve medline document classification [21].

Besides, there exist a number of studies in the literature where ontology-based classification approaches have been applied [22], [23]. The use of ontologies like Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine (SNOMED), and Medical Subject Headings (MeSH) have proved very useful for improving classification performance [24]–[26]. Our approach applies an ontology as a feature selection method for text classification and our target is to identify Coronary Artery Disease (CAD) disease. The UMLS is employed for conceptualization. It is chosen because it is more comprehensive than other tools.

In addition, some work has used clinical records for prominent tasks such as finding risk factors for diabetic patients [27], extracting Framingham risk score (FRF) for target population [28], using rule-based and dictionary-based methods to identify heart disease risk factors [29], and applying a rule-based method by combining with regular expression and UMLS to spot risk of heart disease [30]. The majority of the previous works have developed statistic rule-based systems which need expert assist when the model should be updated with new features. These kind of systems are not scalable and when the labels of the problem changes, the new rules should be set by experts.

B. Text Mining and Text Classification

Data mining, as another sub field of computer science, uses methods that have intersection with machine learning, statistics and database systems. Six common areas of data mining are: Anomaly detection, Association rule learning, Clustering, Classification, Regression, and Summarization. During the past decades, machine learning algorithms like classification has been developed and many classifiers such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machines (SVM), Naive Bayes (NB), and Neural Networks have been proposed.

Text classification is the task of assigning label l_i to document d_j, where $l_i \in L = \{l_1, \cdots, l_{|L|}\}$ and $d_j \in D = \{d_1, \cdots, d_{|D|}\}$, using a function F:

$$F : D \rightarrow L \tag{1}$$

In formula (1), function F is a classifier which gets documents (D) as input and allocates labels (L) as output to each of the input documents. In this paper, we focus on binary classification. Hence, the set L is $\{0, 1\}$, where 1 is for clinical notes with Coronary Artery Disease (CAD), and 0 for clinical notes without CAD.

Text classification is one of the broadly investigated natural language processing tasks. The goal of text classification is to learn a model from available training data set with predefined classes to predict the classes of the unseen documents. For instance, filtering spam emails, labeling client queries and tagging patient reports are a number of the document classification applications. There is a pipeline in text mining to classify the unlabeled documents which includes preprocessing, representing features, selecting features, classifying and evaluating. Fig. 1 shows the stages of document classification. The feature extraction and selection step is one of the important tasks which can have significant effects on the quality of the classification. Since the data in text classification often appears in raw form such as medical discharge notes, hence, extracting meaningful information to use as features in document classification is a substantial task.

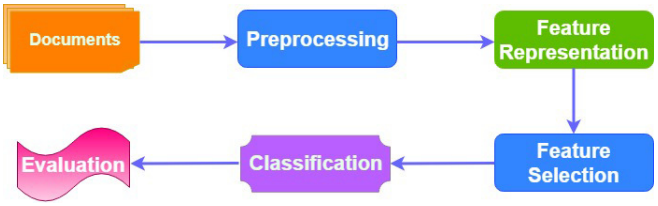

Fig. 1. Stages of document classification

C. Feature Selection

The extraction and selection of features for document classification problems has received a lot of interest in the past. Typically, a lot of these algorithms rank features using statistics from the distribution of features in the given corpus [25], [31]. Existing methods have employed metrics associated with word frequency, information gain, mutual information, term frequency-inverse document frequency (tf-idf) for extracting textual features [32]. However, the aforementioned techniques tend to treat each feature separately, i.e they ignore the dependencies between features.

Feature selection [33] is a NP-hard problem. Feature selection methods can be divided into three groups with respect to the applied feature evaluation method: filter methods, wrapper methods and embedded [34], [35]. These methods are different in evaluating the features. Filter methods assess feature subsets apart from classification approaches. They are fast in computation, but they do not consider dependency between features. On the other hand, wrapper methods utilize classification approaches to evaluate feature subsets. They consider dependencies between features, however, they are slow in computation. As an alternative, embedded methods incorporate feature selection task into the training step of the classifier. They are faster than wrapper approaches, but they make actions which depend on classifier and for this reason they may not act with other classifiers [36]. These feature reduction methods have been used broadly to document classification problems, but there are very limited research in medical text classification field.

D. Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is a subset of computational intelligence in the field of swarm intelligence which is proposed by Kennedy and Eberhart in 1995 [37]. In PSO, all of the particles look for the best point and each time they move, the particles calculate their own value of the fitness function. Each particle who has the best value for the fitness function and closer to the responds, tell it to others, therefore other particles moving toward it. This movement continues until all particles come together at the best point. This algorithm has a kind of memory and the knowledge of good solutions is maintained by each particle as a social sharing characteristic.

PSO optimizes a problem with a population of candidate solutions. The particles are similar, and it moves these particles into a search space by simple mathematical formulas to calculate the position and velocity of each particle. The motion of each particle is influenced by the best known local position, which leads to the best known positions throughout the search space ,that are found by finding better situations by particle. This process leads the particles to the best of solutions totally. PSO is a pattern search method that does not use gradient optimization. This means that PSO, unlike classic optimization methods, such as downside gradient methods and quasi-Newton method, does not need to be differentiable. Therefore, it can be used for optimization problems that have to somewhat arbitrary, noise, variable with time and so on.

PSO is used to predicting and analyzing different diseases in medical field. For example, Eberhart and Hu [38] utilized PSO to checkup human tremor. They targeted two different human tremor: Parkinson's disease and important tremor. PSO is used to improve a neural network that makes a distinction between normal people and those have tremor. Another study utilized a PSO-based approach which utilizes a Radial Basis Function Neural Network (RBFNN) to diagnosis Parkinsonian tremors [39]. Moreover, Fong et al. [40] employed PSO method to find optimum feature subsets. They used PSO beside three different classifiers: navies bayes, decision tree, and pattern network. This study presented a high classification accuracy in two different experimental clinical datasets: the Micro Mass and the Arrhythmia datasets.

Li et al. [41] employed a hybrid evolutionary algorithm by using GA and PSO for selecting gene. They targeted three datasets (i.e. leukemia, colon and breast cancer) to test their method. They introduced a hybrid algorithm to reduce the dimension of the dataset and increase the accuracy of classification. Nazir et al. [42] also utilized PSO-GA method for selecting optimal feature subsets by analyzing face and cloth objects of each person to classify people based on their gender. They reduced the dimension of data by considering only two mentioned objectives.

The combination of PSO and SVM methods is used broadly in medical area and it achieved good performances. The hybrid method is employed for selecting gene and classifying tumor. They utilized the PSO method for gene selection, and SVM as a classifier. Then, the proposed hybrid algorithm was tested

on microarray dataset and it improved the classification accuracy [43]. In another research, Jiang et al. [44] utilized a new hybrid method based on PSO and SVM approaches to discern liver cancer issue. PSO is employed to determine the parameters of SVM. Therefore, it is able to select the parameters impartially for SVM. Furthermore, Mandal [45] introduced another PSO-SVM approach to deal with feature selection problem by employing machine learning ensembles to achieve better ensemble's accuracy. Liu and Fu [46] proposed a new method which combines three different techniques (PSO, SVM and Cuckoo Search (CS)). The suggested approach includes two phases. In the first phase, an improved cuckoo search (CS) is utilized to enhance the parameters of SVM to set appropriate initial parameters for the SVM kernel function, and then in the second phase, PSO is employed in training step of the SVM classifier to identify the biggest parameters of SVM.

Most of the existing research have utilized PSO for selecting features and parameter tuning in medical area to increase the classification accuracy and reduce the feature dimensionality. However, as the clinical discharge notes contain information which is hidden and need to be extracted, doing only feature selection is not sufficient and some new knowledge-based methods are needed to analyze medical text and extract meaningful information to use as features.

E. Two-Stage Feature Selection Approaches

The majority of existing works do feature selection in one step. However, some data sets might be very large and include a large number of features. Using all of the features in feature selection precess to select most important features lead to large search space and might decrease the effectiveness and accuracy of the learned model. Hence, feature selection process can be done in two stages. In the first stage, we employ an efficient method to extract informative features and eliminate obviously redundant and unrelated features, and in the second stage, we conduct feature selection on the much smaller feature set obtained from the first stage. Recently, some two-stage feature selection approaches have been suggested. Uguz [47] applied a two-stage feature selection by utilizing information gain (IG) in the fist stage for selecting informative features and deliver to Principal Component Analysis (PCA) and Genetic Algorithm (GA) to gain the final subset of features. Xue et al. [48] proposed another two-stage method which using PSO for feature selection in the first step, then, in the second step, the obtained features are considered in other fitness function to further reduce the number of features and increase the classification efficiency. Bello et al. [49] introduced a method which performs feature selection by applying PSO in two steps, where the obtained primary results in the first step can be utilized to make the initial population for the second step. Furthermore, Bai et al. [14] presented another two-stage approach which uses four different ranking method to select features in the first stage and applies PSO on the selected features to reduce the number of them more.

All the mentioned two-stage methods accomplish good classification results on the candidate data sets. Meanwhile, the size of selected features by two-stage algorithms is smaller than that gained by single stage algorithms. Furthermore, the smaller size of features reduce the training time to make a classifier model. However, there is not much research in the medical area. Clinical discharge notes are different from other data sets and include meaningful information which are hidden. Our approach differs from others in that it uses a domain specific ontology to do feature extraction in the first stage.

III. THE PROPOSED TWO-STAGE FEATURE SELECTION ALGORITHM

In this section, the proposed two-stage algorithm and the used tools for extracting concepts of phrases are described in details. Fig. 2 shows the flowchart of the proposed two-stage method.

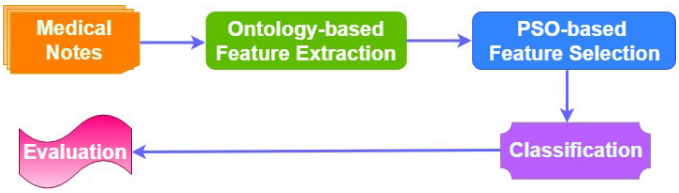

Fig. 2. The proposed two-stage method

The input of the proposed approach is a set of clinic texts. Firstly, the approach detects all of the meaningful expressions in the documents and then applies MetaMap tool to extract their concepts from the UMLS. After deleting redundant features in the first step, PSO is employed to select a feature subset from the extracted features in the first stage. The output is a classifier along with the selected features that predicts the label of a text. First step reduces the size of search space for PSO and assists it to better search. It is expected that the suggested method extracts meaningful features and selects more informative subset of them and maintains or enhances the classification performance.

A. Feature Extraction Method

The UMLS is an abstract of various vocabularies in the biomedical field. It provides an ontology structure of medical vocabulary concepts. In this research the input of UMLS is our documents and the output of it is concepts of the detected meaningful phrases. In the first step, the MetaMap tool is employed to send all of the documents to the UMLS to extract all of the concepts of the detected meaningful phrases. Next, a simple idea is applied in concept section step. Since the label of the candidate problem is a disease, only two concepts are targeted to select: "Disease or Syndrome" and "Sign or Symptom". Generally, these two concepts are closely related to some diseases. Hence, the method keeps these concepts and eliminates the rest of the concepts. Finally, the extracted concepts which are features will be transformed to a vector by using the tf-idf measure. This step reduces the number of features significantly and keeps the informative features too.

Fig. 3 presents the outline of the feature extraction method. The two main selected tools are detailed as follows.

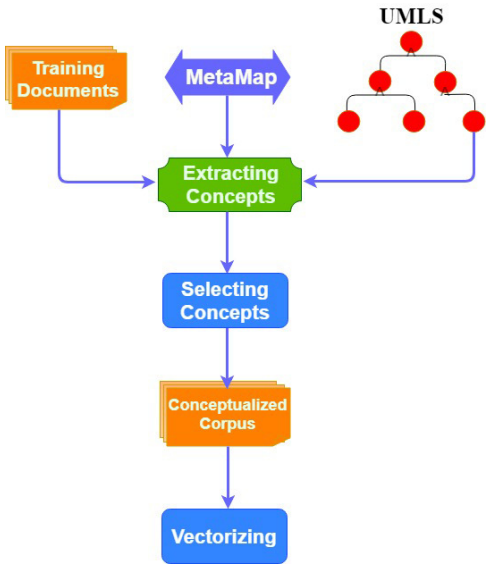

Fig. 3. Feature extraction method

1) Unified Medical Language System (UMLS): The Unified Medical Language System (UMLS) [50] was introduced for modeling the language of health and biomedicine. UMLS is a source of knowledge which improves the performance of information systems in the biomedical area. It provides three main resources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The largest component of UMLS is the Metathesaurus. It gives services such as finding biomedical concepts of phrases and relationships between concepts (e.g. SNOMED-CT, Mesh, etc.). The Semantic Network includes a collection of extensive topic classes, and different types of Semantics, which cover a matchable classification of concepts provided in the UMLS Metathesaurus, and a category of relationships and Semantic Relations between Semantic Types. The SPECIALIST lexicon includes a specialised English vocabulary of biomedical words.

2) MetaMap Tool: MetaMap [51] is an exceptionally configurable program created by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM). It maps text to the UMLS Metathesaurus to find Metathesaurus concepts in text. MetaMap is a knowledge concentrated approach that utilizes computational-linguistic, natural-language processing (NLP) and symbolic methods. MetaMap is applied broadly in Information Retrival (IR), data mining applications. Furthermore, it is utilized for automatically biomedical literature indexing at the U.S. National Library of Medicine (NLM). It allows the mapping between text content and related concepts in UMLS. To achieve this goal, MetaMap breaks the content into expressions and after that, for each expression, it selects the mapping alternatives based on the ranking of mapping quality.

B. PSO-based Algorithm in the Second Stage

In this stage, PSO is employed to further eliminate the irrelevant and unnecessary features from the extracted features in the first stage. The value of particles are initialized randomly by numbers in [-1, 1]. Each particle in PSO corresponds to a feature subset and is coded as a vector. For example, a positive number indicates the corresponding feature is selected and a negative number means the feature is not selected. The dimension of a vector is d and consists of real numbers. In other words, d represents the dimension of the search space which is equal to the size of the primary features which obtained by the first step. A random value is initialized for position and velocity of each particle. Next, PSO moves particles by updating their $pbest$ (best position has found so far) and $gbest$ (the best position). Toward the end of the process, $gbest$ is obtained based on particles' fitness value and the gained best particle will be figure out to achieve the selected feature subset. Algorithm 1 presents the pseudocode for PSO for feature selection in the second stage.

During the algorithm (line 5), the fitness value of each particle is evaluated based on the classification accuracy:

$$Fitness(S) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where S represents the feature subset, TP (True Positive) is the number of correctly identified documents, FP (False Positive) is the number of incorrectly identified documents, TN (True Negative) is the number of correctly rejected documents and FN (False Negative) is the number of incorrectly rejected documents. Our approach is a wrapped-based method. Hence, a classifier is employed to run with PSO to evaluate value of fitness function parameters (TP, FP, TN and FN).

Algorithm 1: Pseudo-code of PSO to select best feature subset

Input : Training instances
Output: The best feature subset ($gbest$)
1: Keep only the features that are extracted in the first stage;
2: Randomly initialize the position and velocity of particles;
3: $gen \leftarrow 0$
4: **while** $gen < maxGen$ **do**
5: **Evaluation**: Evaluate fitness of particles based on classification accuracy on the training set;
6: **for** i = *1 to* $|Particle|$ **do**
7: Update $pbest$ and $gbest$ for particle i;
8: **end**
9: **for** i = *1 to* $|Particle|$ **do**
10: **for** d = *1 to dimension* **do**
11: Update the velocity of particle i
12: Update the position of particle i
13: **end**
14: **end**
15: $gen \leftarrow gen + 1$
16: **end**
17: **return** the position of $gbest$;

Fig. 4 shows the flowchart of how we calculate the fitness function value for each particle. All the training data set is entered as input for PSO to do feature selection. 10-fold cross validation is used to compute a particle's fitness value. The training data is divided into 10 subsets. Nine training

subsets are used as input for PSO and one test subset is used for calculating the fitness of each particle. The average of calculated 10 classification accuracies will be the fitness value of a particle. Please note that the test data set is not used in this PSO feature selection process. The test set is only used in the final evaluation where the final classification accuracy is calculated for the selected best feature subsets. From the aspect of time, PSO takes more time to select best feature subset in the suggested method, but it takes the same time in the test step.

IV. EXPERIMENTAL DESIGN

A. Dataset and Feature Extraction

The 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set is used to analyze the performance of the proposed two-stage approach. The labels of the candidate dataset are CAD (Coronary Artery Disease) and non-CAD which forms a binary classification problem. The total number of documents for the 2010 i2b2 data set is 426.

The features of the 2010 i2b2 documents are extracted by employing the MetaMap tool and using the UMLS. Then, the following preprocessing steps are applied on the obtained results:

The features of the 2010 i2b2 documents are extracted by applying the following preprocessing steps:

- Keep only words and ignore punctuation, numbers, etc. Change all words to lowercase.
- Eliminate words which are less than 3 letters long. For instance, deleting "we" but keeping "our".
- Eliminate the 524 SMART stopwords.
- Extract stems of the remained words.

Next, the TF-IDF method is applied to transform the extracted features to vectors and create a sparse vector matrix.

B. Parameter Settings

We formulize our task as a binary classification problem. The 2010 i2b2 data set consists of 426 documents with 7554 attributes exhibiting various terms which 170 documents belong to training set and 256 documents belong to test set. Five different classifiers (Naive Bayes (NB), Linear Support Vector Machine (LSVM), K-Nearest Neighbor (KNN), Decision Tree (DT) and Logistic Regression (LR)) are utilized for the experimental comparison. The performance of the classifiers are evaluated based on classification accuracy.

Table I shows the set parameters of PSO which are suggested in [14]. We initialise the values using numbers in [-1, 1], and the threshold (θ) is adjusted to zero, so we select roughly 50% of the features. Some documents will not be represented if we select fever than 50% of features.

Some of the classifiers parameters are turned to achieve better results. Hence, the number of the neighbors in KNN is set to 28 for the "n_neighbors" parameter. In Decision Tree classifier, the random number generator and the maximum depth of the tree are set to 11 for the "random_state" and 14 for the "max_depth" parameters, respectively. Value "1e1" is set to the "C" parameter which is the inverse of regularization strength in the Logistic Regression. Additionally, early stopping rule is selected to prevent overfitting in training Logistic Regression and Linear SVM classifiers. Default values are kept for the rest of the classifiers' parameters.

TABLE I
PSO PARAMETERS SETTING

PSO Parameters	**Value**
Population Size	30
Maximum Number of Iteration	100
Dimension	7554
Velocity	[-3, 3]
Threshold (θ)	0
Acceleration Coefficients	2.0
Run Times	40

V. RESULTS AND DISCUSSIONS

The performance of the proposed two-stage method is assessed on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set. Five different classifier are employed to evaluate the suggested approach. The performance of the classifiers are evaluated based on classification accuracy.

Table II shows the number of selected features by four different methods. In the first one, all of the features are selected for training the classifiers and in the second method, MetaMap tool is applied for feature extraction and selection and only 10.33% of the original features are selected. In the third case, PSO is employed to select features from the original feature set and on average 50% of the original feature set is selected. In the 4th case, our two-stage method significantly reduces the number of the selected features to around 5% of the original feature set. The smallest number of features are highlighted in the table. In 3rd and 4th cases, mean and standard deviation of 40 independent PSO runs are presented based on the selected features. Furthermore, the number of selected features for the best subset is showed.

TABLE II
NUMBER OF SELECTED FEATURES

Methods	All	UMLS	All+PSO		Two-Stage	
Classifiers	(100%)	(10.33%)	Ave±Std,	Best(%)	Ave±Std,	Best(%)
NB	7554	780	3779.35±38.01,	3671(48.60)	**387.20±14.61, 396(5.24)**	
LSVM	7554	780	3768.75±48.22,	3827(50.66)	**386.08±14.79, 371(4.91)**	
KNN	7554	780	3774.13±39.36,	3732(49.40)	**394.35±10.68, 397(5.26)**	
DT	7554	780	3775.25±43.04,	3716(49.19)	**388.60±15.14, 394(5.22)**	
LR	7554	780	3767.65±32.77,	3803(50.34)	**388.25±12.31, 374(4.95)**	

The proposed two-stage method is applied on the training set using 40 independent PSO runs. Then, the obtained best feature subsets from each run is used on test set to evaluate the quality of the selected feature subsets. After determining the classification accuracies for the 40 selected feature subsets, the experimental results are calculated. Table III compares the statistical results for four methods. The average and standard deviation of accuracies are provided for each classifier and the significance test is done using the experiment results of the 40 runs over the i2b2 data set. The Wilcoxon signed ranks test [52] with significance level of 0.05 is applied to examine whether the proposed method has made significant difference in classification accuracy. According to Table III, "T" column

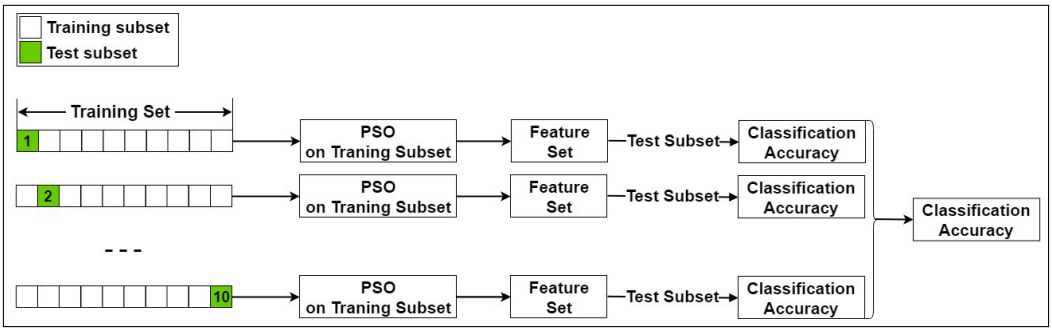

Fig. 4. PSO for feature selection using 10 fold cross validation

shows the significance test of the suggested method against the other three methods, where "+" implies the proposed two-stage technique is significantly more accurate, "=" implies no significant difference, and "-" implies significantly less accurate. The best results are highlighted in the table.

From Table III, it can be concluded that the proposed method has achieved considerably higher classification accuracy than other methods for Naive Bayes, Linear SVM and Logistic Regression classifiers. The average accuracy of KNN classifier is worse than the UMLS method [53], but still the two-stage method is able to achieve similar classification accuracy by using only 5.26% features which is 50 percent less than UMLS method [53]. Our approach gains significantly better classification accuracy in most of the cases.

VI. CONCLUSION AND FUTURE WORK

This paper introduces a two-stage approach to investigate domain concepts and determine which concepts are discriminative to a classification problem. It is able to extract meaningful features from the document set and reduce the number of the features. Moreover, the two-stage approach improves the classification accuracy in the majority of the candidate classifiers by using a small size of feature subset. Experimental and statistical results illustrate that the proposed method can achieve significantly better classification accuracy.

This paper presents the potential of utilizing a two-stage feature extraction and selection approach in medical text classification, but it still requires further investigations to improve the classification performance and reduce the number of features. We will study other ways to extract features for the first stage, and investigate and analyze the features. At the same time, we will target to improve the PSO method by using different fitness functions. Also, our system should provide the ability to allow a domain user to interactively change the concepts and auto-build machine learning models for Coronary Artery Disease (CAD) investigation. We also will consider more datasets and other methods to explore more about other diseases.

REFERENCES

[1] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*. ACM, 1998, pp. 148–155.

[2] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 212–217.

[3] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[4] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, pp. 1–37, 2014.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[6] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2016.

[7] A. Engelbrecht, "Artificial intelligence: an introduction," *England: John Wiley and Sons Limited*, 2007.

[8] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*. IEEE, 1998, pp. 69–73.

[9] H. B. Nguyen, B. Xue, and P. Andreae, "Mutual information estimation for filter based feature selection using particle swarm optimization," in *European Conference on the Applications of Evolutionary Computation*. Springer, 2016, pp. 719–736.

[10] E. Alba, J. Garcia-Nieto, L. Jourdan, and E.-G. Talbi, "Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*. IEEE, 2007, pp. 284–290.

[11] S.-W. Lin and S.-C. Chen, "Psolda: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis," *Applied Soft Computing*, vol. 9, no. 3, pp. 1008–1015, 2009.

[12] B. Tran, B. Xue, M. Zhang, and S. Nguyen, "Investigation on particle swarm optimisation for feature selection on high-dimensional data: Local search and selection bias," *Connection Science*, vol. 28, no. 3, pp. 270–294, 2016.

[13] H.-P. Chan, B. Sahiner, K. L. Lam, N. Petrick, M. A. Helvie, M. M. Goodsitt, and D. D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces," *Medical physics*, vol. 25, no. 10, pp. 2007–2019, 1998.

[14] X. Bai, X. Gao, and B. Xue, "Particle swarm optimization based two-stage feature selection in text mining," in *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2018, pp. 1–8.

[15] M. Yetisgen-Yildiz and W. Pratt, "The effect of feature representation on medline document classification," in *AMIA annual symposium proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 849.

[16] R. Rak, L. A. Kurgan, and M. Reformat, "Multilabel associative classification categorization of medline articles into mesh keywords," *IEEE engineering in medicine and biology magazine*, vol. 26, no. 2, p. 47, 2007.

[17] K. Yi and J. Beheshti, "A hidden markov model-based text classification of medical documents," *Journal of Information Science*, vol. 35, no. 1, pp. 67–81, 2009.

[18] A. K. Uysal and S. Gunal, "Text classification using genetic algorithm oriented latent semantic features," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5938–5947, 2014.

TABLE III

COMPARISON OF CLASSIFICATION ACCURACY AND STANDARD DEVIATION AVERAGES USING 40 INDEPENDENT RUNS. THE HIGHLIGHTED ENTRIES ARE SIGNIFICANTLY BETTER(WILCOXON TEST, $\alpha = 0.05$)

Methods	Two-Stage Stochastic		All [53] Deterministic		UMLS [53] Deterministic		ALL+PSO Stochastic		
Classifiers	Accuracy Ave±Std	Accuracy Best(Lowest)	Accuracy	T	Accuracy	T	Accuracy Ave±Std	Accuracy Best(Lowest)	T
NB	**83.50**±0.018	**86.96**(80.43)	77.47	+	79.57	+	77.58±0.007	78.66(75.89)	+
LSVM	**92.87**±0.007	**93.91**(91.30)	87.35	+	92.61	+	87.22±0.008	88.93(84.98)	+
KNN	93.61±0.005	**94.78**(92.61)	84.98	+	**94.78**	-	86.80±0.014	89.33(82.21)	+
DT	88.71±0.021	91.30(82.61)	85.77	+	87.39	+	**90.09**±0.011	**92.25**(86.96)	-
LR	**93.27**±0.007	**94.35**(91.74)	86.96	+	92.61	+	87.62±0.008	89.33(86.17)	+

[19] B. Parlak and A. K. Uysal, "Classification of medical documents according to diseases," in *Signal Processing and Communications Applications Conference (SIU), 2015 23th*. IEEE, 2015, pp. 1635–1638.

[20] A. J. J. Yepes, L. Plaza, J. Carrillo-de Albornoz, J. G. Mork, and A. R. Aronson, "Feature engineering for medline citation categorization with mesh," *BMC bioinformatics*, vol. 16, no. 1, p. 113, 2015.

[21] B. Parlak and A. K. Uysal, "On feature weighting and selection for medical document classification," in *Developments and Advances in Intelligent Systems and Applications*. Springer, 2018, pp. 269–282.

[22] F. Camous, S. Blott, and A. F. Smeaton, "Ontology-based medline document classification," in *Bioinformatics Research and Development*. Springer, 2007, pp. 439–452.

[23] R. B. Dollah and M. Aono, "Ontology based approach for classifying biomedical text abstracts," *International Journal of Data Engineering (IJDE)*, vol. 2, no. 1, pp. 1–15, 2011.

[24] V. N. Garla and C. Brandt, "Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 882–886, 2012.

[25] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on*. IEEE, 2016, pp. 2264–2268.

[26] K. Buchan, M. Filannino, and Ö. Uzuner, "Automatic prediction of coronary artery disease from clinical narratives," *Journal of biomedical informatics*, vol. 72, pp. 23–32, 2017.

[27] A. Khalifa and S. Meystre, "Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes," *Journal of biomedical informatics*, vol. 58, pp. S128–S132, 2015.

[28] J. Jonnagaddala, S.-T. Liaw, P. Ray, M. Kumar, N.-W. Chang, and H.-J. Dai, "Coronary artery disease risk assessment from unstructured electronic health records using text mining," *Journal of biomedical informatics*, vol. 58, pp. S203–S210, 2015.

[29] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of biomedical informatics*, vol. 58, pp. S171–S182, 2015.

[30] C. Shivade, P. Malewadkar, E. Fosler-Lussier, and A. M. Lai, "Comparison of umls terminologies to identify risk of heart disease using clinical notes," *Journal of biomedical informatics*, vol. 58, pp. S103–S110, 2015.

[31] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[32] Y. Ko, "A study of term weighting schemes using class information for text classification," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 1029–1030.

[33] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1-2, pp. 237–260, 1998.

[34] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems*, vol. 36, pp. 226–235, 2012.

[35] P. Kumbhar and M. Mali, "A survey on feature selection techniques and classification algorithms for efficient text classification," *International Journal of Science and Research*, vol. 5, no. 5, p. 9, 2016.

[36] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 2015, 2015.

[37] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*. IEEE, 1995, pp. 39–43.

[38] R. C. Eberhart and X. Hu, "Human tremor analysis using particle swarm optimization," in *Proceedings of the congress on evolutionary computation*. IEEE Press Piscataway, NJ, 1999, pp. 1927–1930.

[39] D. Wu, K. Warwick, Z. Ma, M. N. Gasson, J. G. Burgess, S. Pan, and T. Z. Aziz, "Prediction of parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization," *International journal of neural systems*, vol. 20, no. 02, pp. 109–116, 2010.

[40] S. Fong, S. Deb, X.-S. Yang, and J. Li, "Feature selection in life science classification: metaheuristic swarm search," *IT Professional*, no. 4, pp. 24–29, 2014.

[41] S. Li, X. Wu, and M. Tan, "Gene selection using hybrid particle swarm optimization and genetic algorithm," *Soft Computing*, vol. 12, no. 11, pp. 1039–1048, 2008.

[42] M. Nazir, A. Majid-Mirza, and S. Ali-Khan, "Pso-ga based optimized feature selection using facial and clothing information for gender classification," *Journal of applied research and technology*, vol. 12, no. 1, pp. 145–152, 2014.

[43] Q. Shen, W.-M. Shi, W. Kong, and B.-X. Ye, "A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification," *Talanta*, vol. 71, no. 4, pp. 1679–1683, 2007.

[44] H. Jiang, F. Tang, and X. Zhang, "Liver cancer identification based on pso-svm model," in *Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*. IEEE, 2010, pp. 2519–2523.

[45] I. Mandal, "Svm-pso based feature selection for improving medical diagnosis reliability using machine learning ensembles," *Computer Science & Information Technology (CS & IT)*, vol. 2012, pp. 267–276, 2012.

[46] X. Liu and H. Fu, "Pso-based support vector machine with cuckoo search technique for clinical disease diagnoses," *The Scientific World Journal*, vol. 2014, pp. 1–7, 2014.

[47] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[48] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[49] R. Bello, Y. Gomez, A. Nowe, and M. M. Garcia, "Two-step particle swarm optimization to solve the feature selection problem," in *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*. IEEE, 2007, pp. 691–696.

[50] "Unified medical language system (umls®), http://www.nlm.nih.gov/research/umls/. last updated 20 apr 2016."

[51] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[52] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[53] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li, "Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2018, pp. 104–110.