IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Stage Prediction of Lung Tumor Identification: A Decision Tree Model for Particle Swarm Optimization Algorithm

P.Jyotsna[1*], P. Govindarajulu P[2]

[1,2]Department of Computer Science, Sri Venkateswara University, Tirupati, India

**Corresponding Author: jyotsna.naidu@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 20/Jul/2018, Published: 31/Jul/2018

Abstract- Text mining has become a basic methodology for computational applications in the field of medical Reports. Text mining can be applied in the medical field for diagnosis of organs like Lung tumor, Head and neck, diabetes and other related diseases. Lung tumor is the most common disease, with more than one million cases being reported worldwide each year. The most effective way to reduce lung tumor deaths is by early diagnosis. This study aims to determine the lung tumor TNM Staging diagnosis. This research data uses National Cancer Institute (NCI) from UCI machine learning. Historical Medical Text Reports constitute a rich and varied source of information, which is today readily accessible, due to large-scale digitization efforts. But in spite of such large scale digitization efforts, stage data in Tumor (cancer) registries is often incomplete, inaccurate, or simply not collected. This paper describes a classification that automatically extracts Tumor staging information from the medical reports that identify malignant cases that are well suited for TNM staging using one class. This is because the decision is unaffected by the outliers and the form of the data fits more precisely. The system uses text classification techniques to extract elements of the stage listed in Tumor staging guidelines. When processing new reports, it classifies relevant sentences to help reach the staging decision and consequently, assigns the most likely stage. This Staging decision is appropriate for medical Text mining.

Key words: Text data mining, Tumor staging, Decision tree and PSO**.**

I. INTRODUCTION

Extensive efforts to computerize voluminous data pertaining to historical text are making it increasingly easier for researchers to search, explore and gain access to a wealth of information that was previously available only in a printed form. The wide time spans covered by the computerized texts provide significant scope to study historical change in the reports. The vast amount of data earlier available in a printed form made it very difficult to process and analyze it manually. But with the advent of computers this task has been simplified helping researchers to analyze the data easily. The computerization of medical Manuscripts and public health reports presents new opportunities for medical historians, e.g., to analyze data pertaining to extended periods of time. Tumor stage is assigned according to standard criteria such as the tumor-node- metastases (TNM) staging standard. Tumor stage annotation could add significant value to existing incidence and mortality data collected by personnel of the department of oncology. They collect the data and prepare population-based registries. These registries form a basis for planning clinical management. They are also a major pre-processing factor in the analysis of outcomes across a population.

To gather data related to TNM stages, the cancer case summaries created by the Basavatarakam Indo American Cancer Hospital & Research Institute Pathologists (BCHP) is proving useful. These case summaries are used as synoptic checklists containing tumor site-specific items including cancer staging information. The value of tumor stage along with other key characteristics in the BCHP cancer checklists has been recognized by the Medical experts. The documentation of check listed items in pathology reports is now mandated as a minimum requirement for finding Tumor.

This is an important step in standardiasing the collection of TNM stages. But Data Reliability continues to depend on the skill of the clinician documenting the stage.

II.MEDICAL BACKGROUND

Tumor staging is a critical step in critical step in diagnosis. Its objectives are many and they include 1) helping the clinician to recommend a treatment plan; 2) giving some indication of prognosis; 3) aiding in the evaluation of the results of treatment; 4) facilitating the exchange of information between treatment centers and 5) contributing to the continuing investigation of uncontrolled growth and spread of human cells. The T category describes

the size and extent of the primary tumor. The N category describes the extent of involvement of regional lymph nodes. The M category describes the presence or absence of distant metastatic spread. The addition of numbers to these categories describes the extent of the cancer. All possible combinations of the T, N, and M categories are then used to create TNM subsets (Table 1). TNM subsets with similar prognoses are then combined into stage groupings. NSCLC stages range from one to four (I through IV). The lower the stage, the less the spread of the tumor. SCLC is defined using two stages: Limited (confined to the hemi thorax of origin, the mediastinum, or the supraclavicular lymph nodes) and extensive (spread beyond the supraclavicular areas) [4].

Table 1 Lung Tumor staging, Tumor, Node, Metastasis staging

Staging		
T	**Primary Tumor**	
	T_x	Cannot be assessed; Tumor in sputum/bronchial washings not in imaging/bronchoscopy
	T_0	No evidence
	Tis	Carcinoma in situ
	T_1	≤ 3 cm surrounded by lung/visceral pleura, not involving main bronchus
	T1a(mi)	Minimally invasive adenocarcinoma
	T1a	≤ 1 cm
	T1b	> 1 to ≤ 2 cm
	T1c	> 2 to ≤ 3 cm
	T2	> 3 to ≤ 5 cm or Involves main bronchus without carina involvement or Visceral pleural invasion or atelectasis/post obstructive pneumonitis extending to hilum
	T2a	> 3 to ≤ 4 cm
	T2b	> 4 to ≤ 5 cm
	T3	> 5 to ≤ 7 cm or Separate tumor in same lobe or Direct invasion of chest wall (includes parietal pleura and superior sulcus)/parietal pericardium/phrenic nerve
	T4	> 7 cm or Separate tumor in different lobe of ipsilateral lung or Invasion of heart/ great vessels/diaphragm/ mediastinum/ trachea/carina/esophagus/ recurrent laryngeal nerve/ vertebral body
N		Regional lymph node
	Nx	Cannot be assessed
	N_0	No involvement
	N_1	Ipsilateralperibronchial and/or hilar nodes and intrapulmonary nodes
	N_2	Ipsilateralmediastinal and/or subcarinal nodes
	N_3	Contralateral mediastinal or hilar; ipsilateral / contralateral scalene/ supraclavicular
M		Distant metastasis
	M_0	No distant metastasis
	M_1	Distant metastasis is present
	M_{1a}	Tumor (s) in contralateral lung; pleural/ pericardial nodule/malignant effusion
	M_{1b}	Single extrathoracic metastasis
	M_{1c}	Multiple extrathoracic metastases, in one/more organs

source : International Association for the Study of Lung Cancer (IALSC)

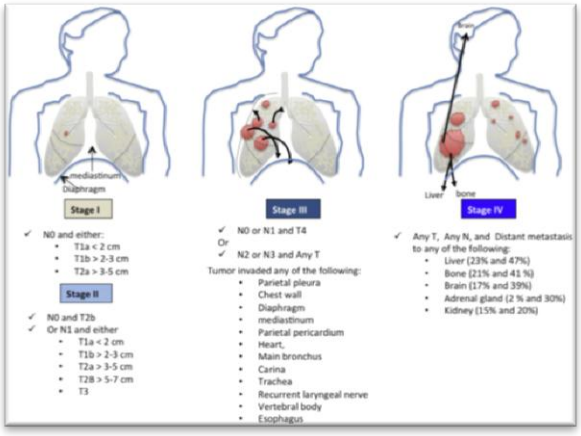

Fig. 1. Schematic illustration of the non-small lung cancer(tumor) (NSCLC) staging[9]

The problem of reducing the number of irrelevant documents that are included within search results is relatively complicated by the fact that many words can have various meanings. An additional issue is that keyword-based search cannot be used effectively to restrict search results to just those in which concepts of interest are only mentioned in the context of a relevant relationship of interest. As an example, consider that a researcher is interested in finding concepts that correspond to causes of Tumor. Just as there are various possible ways in which Tumor could be mentioned in text, there are also many means of expressing interconnection, including words and phrases such as cause, due to, result of, etc. Although a researcher could try to formulate a query incorporating several variant terminologies for both Tumor and interconnection, keyword-based queries do not allow the specification of how different query terms should be linked

to each other. Hence in the Medical Reports retrieved, there is no guarantee that search terms will even occur within the same sentence and if they do, the nature of the relationship may not be the one that is required. For example, retrieved documents may talk about things caused by Tumors rather than causes of Tumors Stages.

Lung Tumor staging is the process by which the extent of the primary tumor and the extent of tumor spread within the body are established. The TNM staging system guides patient management, provides information regarding prediction and eligibility for clinical trials, and allows international comparisons. TNM staging is based on the characteristics of the primary tumor (T), the degree of lymph node involvement (N) and the presence or absence of metastasis (M). The combination of T, N and M descriptors are then used to give the tumor an overall stage (I - IV), with the objective of grouping patient reports into stages with similar predictions. Treatment options also vary from stage to stage.

III. RELATED WORK

Lung Tumor is considered to be the main cause of Tumor death in worldwide, and it is difficult to detect in its early stages because symptoms appear only in the advanced stages causing the mortality rate to be the highest among all other types of tumor. In this paper, we use Decision tree Classification approaches using PSO for improving sensitivity for the differentiation of malignant and benign tumor. For developing the model, the different malignant and benign features, which are identifiable using bronchoscope, needle biopsy, CT scan and MRI modalities are used. The identification of malignant cases can be well suited for TNM staging using one class because the decision is unaffected by the outliers and the form of the data fit more precisely.

S. Tsim , C.A. O'Dowd, R. Milroy , S. Davidson[23] Present "Staging of non-small cell lung cancer (NSCLC):A review" .It Describes the review of TNM Staging System for lung tumors. The most commonly used cancer staging system is the tumor, node, metastasis (TNM) staging system, which is maintained by the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC). In 1998, the International Association for the Study of Lung Cancer (IASLC) established The Lung Cancer Staging Project. It collected data on over 100,000 patients diagnosed with lung cancer between 1990-2000 worldwide, in order to revise the 6th edition TNM staging system for non-small cell lung cancer (NSCLC).The 7^{th} edition was published in late 2009. This review of staging in NSCLC, includes a summary of the different staging techniques Currently available and the 7th edition TNM staging system for NSCLC.

One of the most powerful and widely used techniques for classification and prediction is decision tree [16]. Decision tree is frequently used classification algorithm and has a simple structure as well as easy to be interpreted[13]. Decision tree transforms a very large fact into a decision tree presenting the rules. The C4.5 algorithm[14] proves its performance in predicting with best results in terms of accuracy and minimum execution time. Many researchers have tried to apply the machine learning algorithm to diagnose Lung cancer (tumor).

Decision Tree Algorithm is one of the classification algorithm. It is frequently used by the researchers to classify the data. The decision tree is very popular because it is easy to build and require less domain knowledge. Also the decision tree method is scalable for large database. The decision tree algorithm has weaknesses in handling large data, including: (1) empty branch, nodes with zero value or near zero value do not contribute to generate rules or help to build classes for classification tasks but make bigger and more complex tree sizes, (2) insignificant branch not only reduce the usefulness of the decision tree but also bring over fitting problems, (3) Over fitting occurs when the algorithm model takes data with unusual characteristics (noise).

Data quality such as noise and over fitting data can affect the performance of classification algorithms. Feature selection is commonly used in machine learning when it involves attributes of high-dimensional and noise datasets. Feature Selection is the process of selecting relevant features, or a subset of feature and then process the data with the selected features to the learning model. Feature selection search locally. Metaheuristic optimization can find the solutions in full search space and use global search capabilities that significantly improve the ability to find high-quality solutions within a reasonable timeframe. Improved algorithmic accuracy is required. For example through the application of Discretization and Bagging Techniques to Improve Classification Accuracy in decision tree Algorithm [12].

One of metaheuristic optimization for feature selection is Particle Swarm Optimization (PSO). PSO has proven to be more competitive than genetic algorithms in some cases, especially in the area of optimization [11]. In this study, a combination of PSO-based decision tree algorithms is proposed to improve the accuracy of Lung tumor diagnoses and to overcome weaknesses in the decision tree algorithm using PSO metaheuristic optimization for feature selection and to optimize decision tree algorithm accuracy. Based on the above description, it is necessary to improve the method of diagnosing Lung tumor accurately.

IV. METHODOLOGY

The Method adopted here relies on analysis of comparison and fusion of two classification methods of Text mining. The method uses the decision tree algorithm and

particle swarm optimization. The first step in this research is to measure the accuracy of decision tree algorithm. The next step is to measure the accuracy of decision tree algorithm based on PSO. PSO is a feature selection-based optimization process. It measures the accuracy of decision tree algorithm. Then it compares other algorithms and finds out which algorithm gives better accuracy.

Decision tree algorithm

A decision tree is a decision support tool that uses a tree-like structure or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way of displaying an algorithm.

A decision tree model is in TNM stages. It is a structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails). Each branch of the tree represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes)[4]. The paths from root to leaf represent classification rules.

Basic steps used for decision tree algorithm:

Input: Data partition D, which is a set of training tuples and their associated class labels;

Attribute list: This list includes the set of candidate attributes; attribute selection method and a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes.

This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree

1. Create a node N;
2. If tuples in D are all of the same class, it is then C
3Return N as a leaf node labelled with class C;
4.If attribute list is empty then
5.Return N as a leaf node labelled with the majority class in D; //majority voting //
6.ApplyAttribute_Selection_Method(D, attribute list) to find the "best" splitting criterion.
7. Label node N with Splitting_Criterion;
8. If splitting attribute is discrete-valued and splits are allowed then // not restricted to binary trees//
9.attribute_list divide-attribute // remove splitting_attribue//
10. for each outcome j of splitting _criterion //partition the tuples and grow subtrees for each partition//
11. Let Dj be the set of data tuples in D satisfying the outcome j; // a partition//
12. If Dj is empty then
13. attach a leaf labelled with the majority class in D to node N;
14. else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N endfor
15. return N;

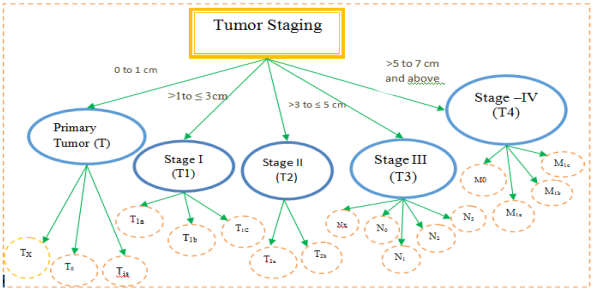

Fig 2: Decision Tree model with TNM Staging system

A) Data cleaning

At this stage cleaning was done on incomplete, empty, or null data, data containing noise, and inconsistent data. There were 32 missing values on bare nuclei attribute. There are several ways of handling missing value data. These ways include ignoring tuples, filling missing value manually, using global constants to fill missing value, using measures of central tendency for attributes (eg, mean or median), using mean or median attributes for all samples included in the class which is the same as the tuple given, and using the value that is most likely to be filled in the lost value . Handling of missing value using average in this study reduced the level of fitness. Therefore, the handling of missing value in this study is done by reducing the data object so that the amount of BCHP Lung Tumor dataset which was originally 100 records became 68 records. The details of data to be cleaned are shown in Table 1.

Table No .1 Data cleaning

Data type	No. of lung tumor data	Remarks
Data collection	100	Randomly pick data for oncology dept.
Missing data	32	Exclude the unstructured data
No. of Cleaned data	68	Data set

B) Data selection

At this stage data selection would be done to reduce irrelevant and redundant data. The NCI dataset is used here to eliminate the attribute of sample code number due to the attribute included into nominal or ordinal feature that is of a categorical type and has qualitative value. This value was actually a symbolic value. It was impossible to perform arithmetical operations as in numerical type so that only 7 attributes were used with 3 attributes as predictor variables and 1 attribute as destination / target variable. The significant pattern attribute details are as follows.

Significant Pattern mined using Decision tree conditions

1. Tumor(T) – X-ray/CT Scan - Carcinoma - Primary tumor (T) – T1 lung tumor measuring ≤ 3 cm into T1a, T1b, T1c, lesion based on specific size cutoffs. Type of tumor : small cell tumor Weightage : 100
2. Tumor (T)- MRI –Invasive adenocarcinoma – Primary tumor (T) – T2 lung tumor measuring >3 to ≤ 5 cm of tumor size or obstructive pneumonities extending to hilum : Non small cell tumor, weightage : 200
3. Tumor (T) – Bronchoscopy – tumor in same lobe – Primary Tumor (T) –T3 lung tumor measuring > 5 to ≤ 7 cms of tumor size – parietal pericardium/ phrenic nerve, Non small cell tumor, weightage : 250
4. Tumor (T) - FNAC test – Tumor in different lobe of ipsilateral lung –Primary Tumor (T) –T4 lung tumor measuring > 7 cm of tumor size – recurrent laryngeal nerve / vertebral body. Non small cell tumor, weightage : 270
5. Regional Lymph node (N) – Biopsy – Tumor in different hailer nodes and intrapulmonary nodes Non small cell tumor, weightage : 280
6. Regional Lymph node (N1, to N3) – Biopsy – ipsilateralMediastinal and sub carinal nodes Non small cell tumor, weightage : 290
7. Metastasis (M) – Biopys – tumor in contralateral lung; - Single extrathoracic metastasis – multiple extrathoracic metastases in one or more organs. Non small cell tumor, weightage : 295

Weightage for Significant Pattern

The weightage is calculated for every frequent pattern, based on the attributes, to analyse the pattern's impact on the output. The pattern that is frequently mined, which satisfies the condition mentioned, is taken as the most significant Frequent Pattern.

$Sw(i)= \Sigma (Wi*Fi)$

(1).Where Wi is the weightage of each attribute and Fi represents number of frequency for each rule. And significant Frequent Pattern is selected by using the following Equation (2) $SFP=Sw(n) \geq \varphi$ for all values of n (2). Where SFP denotes significant frequent pattern and φ denotes significant weightage.

Table No.2 Risk scores for the attributes that represent the significant patterns.

Attributes	**Values**	**Risk score**
Habits	Smoking Non smoking	6 5
Type of tumor	Small cell lung tumor Non-small cell lung tumor	4 5
Stage 0	Primary Tumor(T) Tx,T0,Tis,	2 3
Stage I	Tumor(T) - stageT T1 -- Tumor size is >1 to <3cm T1a,(<1cm) T1b,(>1 to <2cm) T1c,(>2 to < 3cm)	 3 4 5
Stage II	T Stage T2 --Tumor Size is >3 to < 5cm T2a, (>3 to <4cm) T2b,(> to <5cm)	5 4 5
Stage III	N Stage T3 --Tumor size is >5 to <7cm Nx,N0,N1,N2,N3 N0 or N1and T4 Or N2 or N3 and Any T T2a,T2b,T3	3 5
Stage IV	M Stage T4 – Tumor size is > cm M0,M1 M1a, M1b,M1c T2b>5-7cm, N0 or N1 and T4 N2 or N3 and Any T ,Any N, and Distant Metastasis M	3 6

Class-labeled training tuples from Tumor staging patients

Rules for Decision Tree

TUMOR is stage 0 =>Primary (T0 to T1) and risk score = x < 20 then result = you don't have danger tumor, tests = do simple clinical tests to confirm.

TUMOR is Stage I =>(T1 to T2) and risk score = 20 <x < 40 then result = you may have Tumor present, tests = sputum test and x ray to confirm.

TUMOR is Stage II =>(T2 to T3) related to Lung and risk score = x > 50 then result = you have TUMOR, tests = CT or MRI

TUMOR is Stage III =>(T2 to T3) Regional lymph node (N) related to lung with nodes and risk score =x > 60 then result = you have TUMOR with lymph node, type= Non small cells lung tumor, tests= FNAC.

TUMOR is Stage IV =>(T3 to T4) Distant Metastasis (M) related to chest and shoulder and risk score =x > 70 then result = you have tumor with distant metastasis, type = Non small cells lung tumor, tests = bronchoscopy (biopsy).

Based on the above mentioned rules and the calculated risk scores, the severity of lung tumor is known. Some clinical tests are prescribed to confirm the presence of tumor.

C) Data Transformation

At this stage transformation of data takes place. The data of class value had formats 1 and 3. Among These formats , format 1 is used for benign tumors and 3 formats for malignant ones. After the pre-processing stage is

completed, the data is divided based on tenfold cross validation. Tenfold cross-validation divids data into 10 sets. The size of data set is divided by 10. Then 9 sets of data are used for training and 1 set of data for testing. Then the step is repeated up to 10 times iteration. Training data is used to build the model while testing data is used to validate the model. Later, data training is used for the modeling of decision tree algorithm based particle swarm optimization. Particle swarm optimization gives weight to each attribute and produces the best solution (fitness). Then the calculation of decision tree algorithm is done. The steps to generate fitness are as follows.

Steps for Constructing a Decision-Tree with PSO

This section introduces decision tree with PSO algorithm and its variant, the stage PSO(SGPSO). Performance criteria of the algorithm are discussed by considering the effect of algorithmic parameters. Particle organization and knowledge sharing through decision tree are also discussed.

The PSO algorithm is initialized with the population of individuals being randomly placed in the search space and by searching for an optimal solution by updating individual generations. At each iteration, the velocity and the position of each particle are updated according to its previous best position (p_{best}, i,j) and the best position found by informants (g_{best},i,j). In the original continuous version[2]. Each particle's velocity and position are adjusted by the following formula:

Particle Velocity .

$$v_{id}^{t+1} = W * v_{iN}^{t-1} + c_1 r_1 (p_{iN}^{t} - x_{iN}^{t}) + c_2 r_2 (p_{gN}^{t} - x_{iN}^{t}), \quad ---(1)$$

The new position

$$x_{iN}^{t+1} = x_{iN}^{t} + v_{iN}^{t+1}, \quad ----(2)$$

Whereas,

i is the index of a particle in the swarm (i=1,2,….. n), N is the index of position in the particle (N=1,2,……m), t represents the iteration number, v_{id}^{t} is the velocity of the i^{th} particle, and x_{iN}^{t} is the position. Note that r1 and r2 are random numbers uniformly distributed between 0 and 1. c1 and c2 are called acceleration coefficients and w is risk score weight. These are the directions of pbest and gbest positions respectively.

In the classification problem, accuracy or mixture of sensitivity and specificity is usually used as the fitness function. Like other evolutionary computation algorithms, the PSO requires some parameters as inputs. The size of the population, weight (i.e., w), acceleration coefficients (i.e., c1 and c2), and the maximum number of iterations are parameters. Among these parameters, weight and acceleration coefficients have a large impact on the algorithm's performance. SGPSO is one variant of PSO. SGPSO which includes TNM stages of lung tumor, improves the search efficiency by adjusting w, c1 and c2 adaptively.

Proposed algorithm

As mentioned before, most decision tree algorithms are structure standardizations and genetic approach function of mutation and cross over. The fitness of protocol identification is carried out; using the C4.5 algorithm which is an improvement of IDE3 algorithm, developed by Quinlan Ross (1993)[20]. It is based on Hunt's algorithm and like the IDE3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node, thereby reducing the error rate.

The PSO algorithm was introduced by Kennedy and Eberhart [7] to modify the staging Particle Swarm Optimization algorithm to operate in TNM Staging problem spaces. It uses the concept of velocity as a probability that a bit (position) takes on one or zero. In the SGPSO, Eq.(1) for updating the position is redefined by the rule[8].

$$x_{iN}^{t+1} - \begin{cases} 0, & if\ r(\ ) \geq S(V_{iN}^{(t+1)}) \\ 1, & if\ r(\ ) \leq S(v_{iN}^{(t+1)}) \end{cases} \text{-----(3)}$$

Where S is the sigmoid function for transforming the velocity to the probability as the following expression:

$$S(v_{iN}^{(t+1)}) = \frac{1}{1+e^{-v_{iN}^{(t+1)}}} \quad ----(4)$$

and r ( ) is the pseudo random number selected from uniform distribution over (0, 1). It should be noted that the SGPSO is susceptible to sigmoid function saturation, which occurs when velocity values are either too large or too small. In such cases, the probability of a change in bit value approaches zero and thereby, limits exploration. For a velocity of 0, the sigmoid function returns a probability of 0.5, implying that there is a 50% chance for the bit to flip. However, velocity clamping will delay the occurrence of the sigmoid function saturation.

The risk score parameter w, controls the influence of the previous velocity on the current velocity(4). It can improve the manner in which the PSO algorithm converges to a solution by smoothing particle optimization. Their relation is shown in the following equation with an intermediate parameter φ,

$$\begin{cases} w = \frac{1}{\varphi - 1 + \sqrt{\varphi^2 - 2\varphi}} \\ c_1 = c_2 = \varphi w \end{cases} \quad \text{-------- (5)}$$

For φ the above formula imposes values greater than 2. The risk score weight, w, has to be a real number. In addition, experimental results show that $c_1$and c_2 must be greater than 1. These two remarks lead us to changing φ in the interval (0,1).

Maximum velocity V_{max}

A restriction may be placed on the individual velocity component values to enforce the limitation that a particle does not exceed a certain acceleration. This

constraint, V_{max}, reduces the chance that a particle may accelerate uncontrollably and explode off the bounds of the search space. C4.5 builds decision trees from a set of training data in the same way as IDE3, using the concept of information entropy. The training data is a set $S = \{s_1, s_2 \ldots\}$ of already classified samples. Each sample s_{iN} consists of a p-dimensional vector $(x_1,(t)\}$, $x_2,(t)$, ...,xp,(t) , where the xj represent attributes or features of the sample, as well as the class in which s_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs in the smaller sub lists.

Modified TNM staging particle swarm optimization With Decision Tree Algorithm

In this section, our proposed scheme, the modified SGPSO, is described using the C4.5 decision tree. When this happens, it simply creates a leaf node for the decision tree,to choose that class. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using expected value of the class and modified SGPSO, respectively. The main difference between the PSO and the SGPSO is in the position update function. Specifically, the position update of the SGPSO does not use the information of the current position. In other words, the next position of the SGPSO is not influenced by the current position but influenced by the velocity only. This implies that when updating a position in the SGPSO, it is meaningless to know where a particle is currently located in the staging search space. Because of this fact, it seems that the updated velocity of a particle is its maximum velocity, even though the TNM staging position already exists in the SGPSO. Thus, we suggest the concept that the velocity and the position of the original SGPSO be taken as a particle and a solution transformed by the sigmoid function. Let the velocity of the original SGPSO be a continuous search space, and then a staging position can be decoded by the sigmoid function. Then, the updated functions of (1), (2), (3) and (4)of the original SGPSO are changed as the following expressions:

$$v_{id}^{t+1} = W * V_{iN}(t) + c_1 r_1 \left(P_{best\ iN} - X_{p\ iN}(t)\right) + c_2 r_2 \left(g_{best\ iN} - X_{PiN}(t)\right), \text{-(6)}$$

$$X_{giN}(t+1) = X_{giN}(t) + V_{iN}(t+1), \text{---(7)}$$

$$X_{PiN}(t+1) = \begin{cases} 0 & if\ r(\ ) \geq S(X_{giN}(t+1)) \\ 1 & if\ r(\ ) \leq S(X_{giN}(t+1)) \end{cases} \text{---(8)}$$

Where

$$S(X_{giN}(t+1)) = \frac{1}{1+e^{X_{gij}(t+1)}} \text{---(9)}$$

In updated functions of the modified TNM Staging PSO, we emphasize the following: instance of previously unseen class encountered. Again C4.5 creates a decision node higher up the tree using the expected value. Create a decision node that splits on a_best Recurrence on the subsists obtained by splitting the best search space, and add those nodes as children of node.

Mathematical expression for calculating the fitness values of Particles:

Entropy H(S) is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S)

$$H(S) = -\sum_{x \in X} p(x) log_2\, p(x) \text{----------(10)}$$

Where, S is the current (data) set for which entropy is being calculated (changes every iteration of the IDE3 algorithm)

X is set of classes in S

P(x) is the proportion of the number of elements in class x to the number of elements in set S When H(S) = 0, then the set S is perfectly classified (i.e. all elements in S are of the same class)

This process is backed up by the C4.5 Fitness function, where the entropy value calculated provides the algorithm with a probabilistic value so as to choose what particle, for the remaining particles to follow. Every iteration also updates the velocity of the particles, the velocity specifies to choose a random number, leaving behind a finite number of numbers behind. As the velocity keeps on updating, the efficiency of the algorithm further improves, and the target is reached in a much lesser time period. After the target is reached, the records that have a value closest to the target are returned as output. The people with these records are having a higher chance of getting the disease, based on the history of patients.

Algorithm for optimized Decision Tree With PSO Staging

Input

c1, c2, r1, r2, w, v_{min} ,$v_{max.}$

Step 1 : Initializing position, Velocity, c_1, c_2

Step 2 : for I =1 to p

Obtaining the velocity for particle 1 by using equation 1

Calculate the fitness (objective) function for particle 1 by using equation (entropy)

Step 3: initialize population

While (max iteration or convergence criteria is not met)

do

for i= 1 to numbers of particles

Evaluate fitness value of the particle by C 4.5

for stage 0 = Primary tumor,

If size is 0 to < 1cm Then

Classification is Tx, T0, Tis

Else If Stage I = T Then

If size is < 1 cm to > 3 cm Then

Classification is T_{1a}, T_{1b}, T_{1c} T_2
Else If Stage II = T Then
If size is > 3 cm to < 5 cm Then
Classification is T_{2a}, T_{2b}
Else If stage III = N Then
If size is > 5 cm to < 7cm Then
Classification is T_3 N_x ,N_0, N_1, N_2, N_3
Else If Stage IV = M Then
If size is >7 cm Then
Classification is T_4 M_0, M_1, M_{1a}, M_{1b}, M_{1c}
Step 4:If the fitness value of Xi is greater than that of PBi
Then PBi = Xi
If the fitness value of Xi is greater than that of PBi
Then GB = Xi
End if
for d = 1 to no of particles

$$v_{id}^{t+1} = W * V_{iN}(t) + c_1 r_1 \left(P_{best\ iN} - X_{p\ iN}(t)\right) + c_2 r_2 \left(g_{best\ iN} - X_{PiN}(t)\right)$$

If V_{iN}>V_{max}Then $v_{id}^{t+1} = V\max$
If V_{iN}<V_{max}Then $v_{id}^{t+1} = V\min$
If *sigmoid* (v_{id}^{t+1}) >*U(0,1)*
then
x_{id} *(t)* = 1
else
x_{id} *(t)* = 0
end if
next *d*
next*i*
end if
end while
return it new subset if text features .

C4.5 made a number of improvements to IDE3.Some are: Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those who attribute value is above the threshold and those that are less than or equal to it. Handling training data with missing attributes values C4.5 allows missing attribute values are simply not used in gain and entropy calculations. Handling attributes with differing costs. Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

V. EXPERIMENTS FOR TEST DATASET

To validate the proposed algorithm, numerical experiments are performed with training data and real data. Each data set is separated into two sets: the two thirds of the observations are used as a training set and the remaining one third is set aside as the test set. The number of particles and the maximum number of iterations are set to 68 and 100, respectively in the SGPSO. The initial values of (w, c1, c2) are set to (0.9, 0.1, 2).We compare our results against the IDE3 results using 10-foldcross validation to see the performance improvement through the proposed optimization method. In the proposed method, we use the classification accuracy as the fitness function.

Experiments using different objective functions

In order to analyze the algorithm parameters influence on classifying effectiveness, classifying results are compared using different objective functions. The test datasets originate from reference (related dataset). This is concerned with lung tumor diagnosis. There are 7 attributes and lung tumor class is decision attribute which has three stages (The first one is primary tumor, the second is nodes and the third is malignant). We choose one hundred data cases as the test data. The number of primary tumors are 56, Regional Lymph node are 8 and 4 Distant Metastasis. Other data is similar to the reference[11]. The fitness functions are set Euclidean distance, and Murkowski distance (cube root). The classification results are using the Swarm optimization-based algorithm as Table 3 shows. It can be seen that different objective functions have distinct influences on the classification results which demonstrate that the worst is Murkowski distance and Mahalanobis distance function. The best is Euclidean distance function which falls between these functions.

Table 3 Classification results based on other distant function

Function	Lung tumor	T(Primary tumor)	N(Regional lymph nodes)	M(distant metastatic)
Euclidean distance	T	56	0	0
	N	0	8	0
	M	0	1	3
Murkowski distance	T	54	1	1
	N	1	5	2
	M	0	1	3
Mahalanobis distance	T	50	4	2
	N	0	8	0
	M	2	1	1

To further confirm this improved SGPSO-based classifier's performance, we select NCI dataset to make a study. There're 3 values for decision attributes : tumor(T), Regional Lymph node(N) and Distant Metastasis(N); The parameters in our algorithm are set as follows. Maximum iteration number is set to 50; maximum weighted inertia coefficient is 0.9; minimum weighted inertia coefficient is 0.1; learning factors c1 and c2 are set to 2. The classification results are compared with other algorithms such as J48 in Weka and REP tree in the table presented below.

Table 4.Classification results based on different classifier algorithms

Classifier Algorithm		T	N	M
J48	T	55	1	0
	N	0	7	1
	M	1	1	2
REPTree	T	56	0	0
	N	1	6	1
	M	2	0	2
SGPSO	T	56	0	0
	N	0	8	0
	M	0	0	4

As can be seen from Table 4, although different classifiers have different classification results, SGPSO-based classifier is the best one.

VI. CONCLUSION

Instance-based learning is one of important classifier design approaches in text mining. As a widely used evolutionary algorithm, PSO is receiving more and more attention in Text mining processes. Aimed at classifier design for instance-based learning, an improved PSO-based algorithm is proposed to optimize multiple objective fitness functions in classification decision problem with multiple attribute values. Decision tree is widely used as text mining tool for classification and prediction. Since the classifying rule in forms of 'IF-THEN' rule is provided by decision tree, analysts can understand the result of decision tree easily. However, most decision tree algorithms consider one variable at a time to search splitting variables. This approach has a risk of reaching a local optimal solution. Still, the SGPSO optimizes only threshold values of the selected variables. Once several variables are selected by IDE3, other variables cannot be considered in further analysis. It makes computational cost reduced, but the decision tree may be optimized locally. Nevertheless, the SGPSO can improve the prediction without the depth growing of decision tree. Then comparison analysis is carried out using test cases and results demonstrate that this PSO-based algorithm performs better compared to existing classifiers. Furthermore, other feature selection methods can be adopted when generating the preliminary tree instead of IDE3. If we can get a good set of features which are critical to predict classes, then thresholds of splitting variables can be achieved by the SGPSO. In this sense, combining random forest and the SGPSO may be a good alternative.

REFERENCES

[1] K.Arutchelvan,2dr.Ponperiasamy"evaluation of staging classification in lung cancer"
International journal of advanced research in computer science and software engineering volume 6, issue 8, august 2016

[2] Alexandreszabo and leandronunes de castro" a constructive data classification version of the particle swarm optimization algorithm" hindawi publishing corporation Mathematical Problems in engineering volume 2013, article id 459503, 13pageshttp://dx.doi.org/10.1155/2013/459503

[3] Chi-hyuckjun†, yun-jucho and hyeseon lee present "improving tree-based classification rules using a particle swarm optimization" springer, ifip advances In information and communication technology, aict-398 (part ii), pp.9-16, 2013

[4] David e. Midthun "early detection of lung cancer [version 1; referees: 3 approved]" 25 April 2016, 5(f1000 faculty rev):739 (doi:10.12688/f1000research.7313.1)

[5] David e. Rumelhart, geoffrey e. Hinton, ronald j. Williams, "learning representations by backpropagating errors", letter to nature, 1986.

[6] Divya chauhan, varunjaiswal" development of computational tool for lung cancer prediction using data mining" international journal of computer applications technology and research volume 5– issue 7, 417 - 421, 2016

[7] Eberhart, R.C. and Kennedy, J. "A new optimizer using particle swarm theory", Proceeding of Sixth International Symposium on Micromachine and Human Science, pp. 39-43, 1995.

[8] Kennedy, J., Eberhart, R.C. and Shi, Y. "Swarm intelligence", Morgan Kaufmann Publishers, San Francisco, 2001.

[9] Hassanlemjabbar-alaoui , omeruihassan, yi-wei yang, petrabuchanan" lung cancer: biology and treatment options" © 2015 elsevier

[10] Iain a. Mccowan, phd, darren c. Moore, meng, anthony n. Nguyen, phd, Rayleen v. Bowman, phd, belinda e. Clarke, phd, edwina e. Duhig, mary-jane fry" collection of cancer stage data by classifying free-textMedical reports" j am med inform assoc. 2007;14:736 –745. Doi 10.1197/jamia.m2130

[11] kun-huangchen, kung-jengwang, min-lung tsai, kung-min wang, angeliamelani adrian1, wei-chungcheng, tzu-sen yang, nai-chia teng, kuo-pin tan and ku-shangchang: "gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm". Chen et al. Bmc bioinformatics 2014.

[12] y.-j. Lee, o. L. Mangasarian, and w. H. Wolberg, "breast cancer survival and chemotherapy: a support vector machine analysis", data mining institute, computer sciences department, university of wisconsin, 2000.

[13] C.J. Mantas, J. Abellán, Credal-c4.5: Decision tree based on imprecise probabilities to classify noisy data, Expert Systems with Applications, 41 (2014) 4625-4637.

[14] Mamuslim, shrukmana , e sugiharti1 , b prasetiyo1 and s alimah2" optimization of c4.5 algorithm-based particle swarm optimization for breast cancer diagnosis" international conference on mathematics, science and education 2017 (icmse2017)

[15] Paul thompson, rizatheresa batista-navarro, georgios kontonatsios,Jacob carter, elizabeth toon, john mcnaught, carsten timmermann,Michael worboys, sophia ananiadou "Text mining the history of medicine" plos one | doi:10.1371/journal.pone.0144717 January 6, 2016

[16] Perveen S 2016 Performance Analysis of Data Mining Classification Techniques to Predict Diabetes publisher: *Procedia Comp. Sci.* 82 115-121

[17] Pranavtejagarikapati , naveenkumarpenki, sashankgogineni "improvised gene selection using particle swarm optimization with decision tree as classifier" international journal of new technology

and research (ijntr) issn:2454-4116, volume-3, issue-9, september 2017 pages 80-86
[18] j.r, quinlan, "induction of decision trees", journal of machine learning, volume 1, number 1, march, 1986.(10)
[19] N.V. Ramana Murthy and Prof. M.S. Prasad babu" a critical study of classification algorithms for lungcancer disease detection and diagnosis" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1041-1048
[20] J. Ross Quinlan(1993) C4.5: Programs for Machine Learning by. Morgan Kaufmann Publishers, Inc.,
[21] Sujatha, dr.k.usha rani, "evaluation of decision tree classifiers on tumor data sets", ijettcs, vol2, issue4, july-aug2013, pg.no:418-423
[22] Thangaraju p, barkavi g "lung cancer early diagnosis using some data mining classification techniques: a survey" an international journal of advanced computer technology, 3 (6), june-2014 (volume-iii, issue-vi)
[23] S. Tsim , c.a. O'dowd, r. Milroy , s. Davidson "staging of non-small cell lung cancer (nsclc):a review" 2010 elsevier ltd
[24] Vanaja, s. And k. Rameshkumar presents "performance analysis of classification algorithms on medical diagnoses-a survey" journal of computer science,2014.
[25] International Journal of Scientific Research in Computer Sciences and Engineering (ISSN: 2320-7639)
[26] International Journal of Scientific Research in Network Security and Communication (ISSN: 2321-3256).

Author's profile

P.JYOTSNA received Master of Computer Applications degree from Sri.Venkateswara University, Tirupati, AP and. Pursuing Ph.D in the department of Computer Science, Sri Venkateswara University, Tirupati. Her research area are Databases and Data Mining, Her research focus is on Text Mining Techniques for Detection of Tumors Using Ontology Based Particle Swarm Optimization with Clustering Approaches.

P. GOVINDARAJULU, Professor, Department of Computer Science, Sri Venkateswara University, Tirupathi, AP, India. He received his M. Tech., from IIT Madras (Chennai), Ph. D from IIT Bombay (Mumbai), His area of research are Databases, Data Mining, Image processing, Intelligent Systems and Software Engineering.

