International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Lung Cancer with Prediction Using Dbscan

Rokkala Sowjanya & Dr. Kunjam Nageswara Rao
[1]PG Scholar, Department of CS & SE, Andhra University, Visakhapatnam

[2]Associate Professor, Department of CS & SE, Andhra University, Visakhapatnam

Abstract- *Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in the diagnosis and treatment of diseases. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes, stroke, cancer and heart disease. Lung cancer, also known as lung carcinoma, is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. This growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas. The two main types are small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC). symptoms of lung cancer: thready pluse, wiry pulse, soft pulse, thin fur, yellow fur, white fur, obesity, red tongue, deep red tongue, pale tongue, internal bleeding, abdominal distension, stomachache, constipation, loose stool, cough, serious cough, occasional cough, phlegm, little phlegm, phlegm that is hard to cough up, dry mouth, chest pain, oppression in chest, sweating, fever, weakness, weak legs, rash, oral ulcer. As there is a big growth in large volume of data now days, this will create a need for extracting meaningful data from the information. From the various biomedical datasets, cancer is the widest disease that has killed human life over 7 million every year and lung cancer among them is nearly 17% of moralities. Previous research works show that survival rate of patients affected with cancer is larger and higher, when compared to the diagnosed at the initial stage, Lung cancer is the most historic data and dependent disease in for early diagnosis. This has created the researcher to use data mining technique for early diagnosis of lung cancer in stage 1.There has been an increase in survival rate to about 70% at the early stage of detection, when tumor is not spread. Pre- existing techniques The five year survival rate increases to 70% with the early detection at stage 1, when the tumor has not yet spread. Existing medical techniques like X-Ray, Computed Tomography (CT) scan, sputum cytology analysis and other imaging techniques not only require complex equipment and high cost but is also proven to be efficient only in stage 4, when the tumor has metastasized to other parts of the body. Our proposed work involves the uses of data mining technique used in classification of lung cancer patients and the categorization of stage to which it belong positive. The work is based on early diagnosis of prediction of lung cancer which suggests the doctors in treating the patients for increasing the survival rate of the human. For Early stage of prediction there are so many data mining techniques are there. I studied prediction using DBSCAN Algorithm*

Keywords: Data Mining, DBSCAN, Lung Cancer, Clustering, prediction.

1. INTRODUCTION

Lung cancer is the most rapidly increase disease which is causing human death world-wide due to respiratory problems. This cancer disease has exceeded the death rate compared to breast cancer. This disease has characterized based on growth of uncontrolled cells. If this disease is not diagnosed at the early stages and cured before the second stage, it will increase the death percentage in human. This tissue will be spread rapidly to other parts of the body like brain, heart, bones, glands and liver. As from early research, there is no such tool for early detection of lung cancer disease in human. We come across two types of lung cancers, one is SCLS and NSCLS. There are two major types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Over 85 percent of all lung cancers are non-small cell lung cancers, while about 13 percent contribute small cell lung cancers [2]. Staging lung cancer is based on whether the cancer is local or has spread from the lungs to the lymph nodes or other organs. Because the lungs are big, tumors can grow in them for a long time before they are found. Even when symptoms like coughing and fatigue occurs; people think that they are due to other causes. Deaths in both women and men. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases. [1]. Treatment and prognosis depend on the histological type of cancer, the

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

stage (degree of spread), and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy Survival depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis.

Symptoms that may suggest lung cancer include:

- Dyspnea (shortness of breath with activity),
- Hemoptysis (coughing up blood),
- Chronic coughing or change in regular coughing pattern,
- Wheezing,
- Chest pain or pain in the abdomen,
- Cachexia (weight loss, fatigue, and loss of appetite),
- Dysphonia (hoarse voice),
- Clubbing of the fingernails(uncommon),
- Dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm,
- Bronchitis or pneumonia,
- Decline in health and unexplained weight loss.

Mortality and morbidity due to tobacco use is very high. Usually lung cancer develops within the wall or epithelium of the bronchial tree. But it can start anywhere in the lungs and affect any part of the respiratory system. Lung cancer mostly affects people between the ages of 55 and 65 and often takes many years to develop [2]. There are two major types of lung cancer. They are Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) or oat cell cancer. Each type of lung cancer grows and spreads in different ways, and is treated differently. If the cancer has features of both types, it is called mixed small cell/large cell cancer. Non-small cell lung cancer is more common than SCLC and it generally grows and spreads more slowly. SCLC is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. They often start in the bronchi near the center of the chest. Lung cancer death rate is related to total amount of cigarette smoked [3]. Smoking cessation, diet modification, and chemoprevention are primary prevention activities. Screening is a form of secondary prevention. Our method of finding the possible Lung cancer patients is based on the systematic study of symptoms and risk factors. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Environmental factors have an important role in human cancer. Many carcinogens are present in the air we breathe, the food we eat, and the water we drink. The constant and sometimes unavoidable exposure to environmental carcinogens complicates the investigation of cancer causes in human beings. The complexity of human cancer causes is especially challenging for cancers with long latency, which are. Researchers have been investigating on applying various data mining techniques on lung cancer dataset for early diagnosis of lung cancer. This paper proposes a model for measuring if applying data mining techniques to lung cancer dataset can provide reliable performance in the detection of lung cancer at Stage I. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction.

Clustering methods can be broadly Density based clustering algorithm has played a vital role in finding non-linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity. Comparing the performance of the DBSCAN algorithm with a proven segmentation algorithm that utilizes k-means clustering demonstrated that the DBSCAN algorithm had a higher sensitivity and correctly segmented more swallows. Comparing its performance with a threshold-based algorithm that utilized the quadratic variation of the signal showed that the DBSCAN algorithm offered no direct increase in performance.

2. Literature Survey

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy company strength. Once these things are satisfied, then next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book, previous work done on the project or from websites.

According to the World Health Organization (WHO), cancer is the number one cause of mortality worldwide, accounting for 7.6 million deaths, or 13% of all deaths, across the globe in 2008. The toll is expected to rise continuously to over 11 million in 2030. Environmental factors are believed to be a primary contributor to the pathogenesis. The hazards range from physical agents such as ionizing and ultraviolet radiations, chemical agents such as dioxins and arsenic, to biological agents such as human papillomavirus and hepatitis B virus. Other risk factors include smoking, alcoholism/diet, obesity, ageing and genetics. Cancer cells, with mutated genomes, share three characteristics: uncontrolled cell multiplication, invasion of adjacent tissues, and migration to non-adjacent sites.

Treatment of cancer in modern western medicine includes surgery, radiotherapy and chemotherapy. Treatment modality depends on the site of cancer origin and stage of cancer progression, and typically involves a combination of modalities, for example a surgical removal followed by radiation or chemotherapy. Radiotherapy kills cells by breaking the DNA with X-, gamma-rays or charged particles and the free radicals generated in the radiation. Cytotoxic chemotherapy tames cells by stopping their division with small molecules that stop cell cycle or DNA synthesis. The two therapies seldom eliminate all cancer cells as, at increasing dosages, more nearby healthy and normal fast growing cells are compromised.

The outcome of the "war on cancer", initiated 40 years ago by the United States administration, has been debated in public media. According to a latest analysis by the American Cancer Society, the overall age-adjusted death rate of all cancers in men (women) dropped by 11% (6%) to 2.2 (1.5) per 1000 in the United States in 2006, compared to that in 1970. The declines were however largely attributed to better prevention and early detection including reduced smoking and increased mammogram and Pap tests, suggesting room for further development of new, complementary and alternative (CAM), as well as integrative cancer therapies..

What cancers are in TCM is best revealed from TCM prescriptions to cancers. Toward a systematic and scientific investigation of TCM treatment of cancers, the elements in TCM prescriptions have to be established and standardized, similar to the consideration of genes as the fundamental elements in genomics. In a typical TCM prescription to a patient can be found, for example, two TCM formulas and four TCM herbs. TCM formulas are believed to evolve from synergistic combinations of multiple TCM herbs. Many TCM formulas from authoritative TCM classics with specified ingredient herbs and relative weights, stand the test of time and are still highly received today. The chemical composition within an herb can change depending on the harvest times/regions and processing methods. Minimizing the variability in the chemical profiles of the herbs/formulas can be relegated to certified manufacturers of TCM medicinals. A certification system involving government, for example, mandatory GMP compliance, also helps eliminate the concern of herbal toxicity due to heavy metal, pesticide and microbiological contaminations. A TCM prescription p is then represented by p?=?a1 m1+a2 m2+…+aN mN, where mi can be either a TCM herb or a classical TCM formula, ai a numeric value between 0 and 1 for the percentage weight of mi in the prescription, and N the number of mi in the prescription so that a1+a2+…+aN?=?1.

TCM has developed its own system of diagnostics or TCM syndrome differentiation via for example tongue and pulse readings. Outcomes of TCM diagnoses in TCM terms could lack of consistency among TCM doctors. As modern western medicine has become an integral part of the core curriculum of the TCM education in Taiwan, diagnosis made in western terms, i.e., the International Classification of Diseases codes (ICD-9), has been enforced for the reimbursement of TCM prescriptions to the public health insurance program in Taiwan. With a large quantity of diagnosis and prescription data, we are then able to statistically associate an ICD-9 coded cancer to a TCM prescription: p(cancer ICD9 code)?=?a1 m1+a2 m2+…+aN mN. ICD-9 codes on one side of the association and ai's on.

3. KNOWLEDGE DISCOVERY AND DATA MINING

This section provides an introduction to knowledge discovery and data mining. We list the various analysis tasks that can be goals of a discovery process and lists methods and research areas that are promising in solving these analysis tasks.

A. Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following figure. 1 shows data mining as a step in an

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

iterative knowledge discovery process. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [5].

The iterative process consists of the following steps:

(1) Data cleaning: also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.

(2) Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source

(3) Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

(4) Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

(**5) Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

(6) Pattern evaluation: this step, strictly interesting patterns representing knowledge are identified based on given measures.

(7) Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

B. Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms. Data mining applications can use different kind of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).Data mining involves some of the following key steps [6]-

(1) Problem definition: The first step is to identify goals. Based on the defined goal, the correct series of tools can be Applied to the data to build the corresponding behavioural model.

(2) Data exploration: If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be Treated consistently.

(3) Data preparation: The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.

(4) Modelling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighbourhoods and clustering but also next generation techniques such as decision trees, networks and rule based Algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.

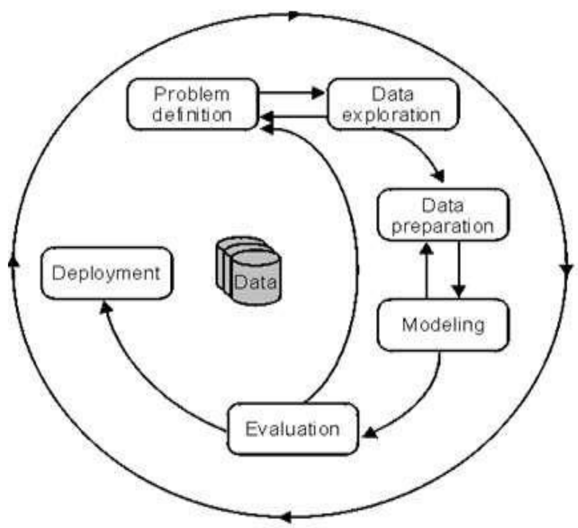


Figure 2. Data Mining Process Representation

(5) Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

4. LITERATURE FOR LUNG CANCER

The approach that is being followed here for the prediction technique is based on systematic study of the statistical factors, symptoms and risk factors associated with Lung cancer. Non-clinical symptoms and risk factors are some of the generic indicators of the cancer diseases. Initially the parameters for the pre-diagnosis are collected by interacting with the pathological, clinical and medical oncologists (Domain experts).

. Statistical Incidence Factors:

- Age-adjusted rate (ARR)
- Primary histology
- Area-related incidence chance
- Crude incidence rate
- B. Lung cancer symptoms:

The following are the generic lung cancer symptoms [14].

i. A cough that does not go away and gets worse over time
ii. Coughing up blood (heamoptysis) or bloody mucus.
iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
iv. Weight loss and loss of appetite
v. Increase in volume of sputum
vi. Wheezing
vii. Shortness of breath
viii. Repeated respiratory infections, such as bronchitis or pneumonia
ix. Repeated problems with pneumonia or bronchitis x. Fatigue and weakness
x. New onset of wheezing
xi. Swelling of the neck and face
xii. Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
xiii. Paraneoplastic syndromes which are caused by biologically active substances that are secreted by the tumor.
xiv. Fever
xv. Hoarseness of voice
xvi. Puffiness of face
xvii. Xviii.Loss of appetite
xviii. Xix.Nausea and vomiting

C. Lung cancer risk factors:

a. Smoking:

i. Beedi
ii.
Cigarette
iii. Hukka

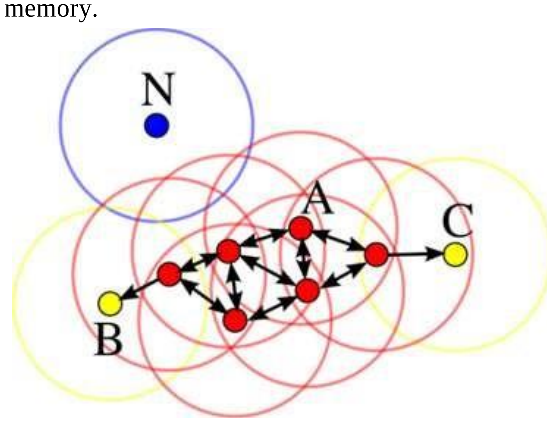
b. Second-hand smoke
c. High dose of ionizing
radiation d. Radon exposure
e. Occupational exposure to mustard gas chloromethyl
Ether, inorganic arsenic, chromium, nickel, vinyl
Chloride, radon asbestos
f. Air pollution
g. Insufficient consumption of fruits & vegetables
h. Suffering with other types of malignancy

5. DBSCAN ALGORITHM

The DBSCAN algorithm can be abstracted into the following steps:

1. Find the ε (eps) neighbors of every point, and identify the core points with more than min Pts neighbors.
2. Find the connected components of core points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assign it to Noise.

A naive implementation of this requires storing the neighbourhoods in step 1, thus requiring substantial memory. The original DBSCAN algorithm does not require this by performing these steps for one point at a time. DBSCAN visits each point of the database, possibly multiple times (e.g., as candidates to different clusters). For practical considerations, however, the time complexity is mostly governed by the number of region Query invocations. DBSCAN executes exactly one such query for each point, and if an indexing structure is used that executes a neighborhood query in O(log n), an overall average runtime complexity of O(n log n) is obtained (if parameter ε is chosen in a meaningful way, i.e. such that on average only O(log n) points are returned). Without the use of an accelerating index structure, or on degenerated data (e.g. all points within a distance less than ε), the worst case run time complexity remains $O(n^2)$. The distance matrix of size $(n^2- n)/2$ can be materialized to avoid distance recomputations, but this needs $O(n^2)$ memory, whereas a non-matrix based implementation of DBSCAN only needs O(n) memory.



DBSCAN(DB, dist, eps, minPts) {
C = 0 /* Cluster counter */
for each point P in database DB {
if label(P) ≠ undefined then continue /* Previously processed in inner loop */
Neighbors N = RangeQuery(DB, dist, P, eps) /* Find neighbors */

```
    if |N| < minPts then {                /* Density check */
        label(P) = Noise                   /* Label as Noise */
        continue
    }
    C = C + 1                              /* next cluster label */
    label(P) = C                           /* Label initial point */
    Seed set S = N \ {P}                   /* Neighbors to expand */
    for each point Q in S {                /* Process every seed point */
        if label(Q) = Noise then label(Q) = C      /* Change Noise to border point */
        if label(Q) ≠ undefined then continue      /* Previously processed */
        label(Q) = C                       /* Label neighbor */
        Neighbors N = RangeQuery(DB, dist, Q, eps)   /* Find neighbors */
        if |N| ≥ minPts then {             /* Density check */
            S = S ∪ N                      /* Add new neighbors to seed set */
        }
    }
  }
}
RangeQuery(DB, dist, Q, eps) {
  Neighbors = empty list
  for each point P in database DB {        /* Scan all points in the database */
    if dist(Q, P) ≤ eps then {             /* Compute distance and check epsilon */
        Neighbors = Neighbors ∪ {P}        /* Add to result */
    }
  }
  return Neighbors
}
```

6. RELATED WORK

In the last few years, the digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases containing rich medical information and made available through the Internet for Health services globally. Data mining techniques applied on these databases discover relationships and patterns that are helpful in the progression of disease [17].

The steps involved in data mining are:

1) Data Integration: Heterogeneous data from various health organizations which are in different forms are collected from multiple sources and made into a common source.

2) Data Selection: The dataset collected from various sources contain all sorts of data. Some of the data may be irrelevant for the mining process and also, some data contain a lot of missing information. Such data are discarded. Only those data relevant to the mining process are considered.

3) Data cleaning: Some patient record contains errors, noise or missing information. Certain data are corrected and those that cannot be corrected are discarded. Fuzzy Self Organising Maps can also be used to filling the missing values some of the attributes identified.

4) Data Transformation: The acquired data from the process of cleaning will not be ready for mining. This data has to be converted into suitable form for data mining. Larger values have to be normalized for fast calculation. This is achieved by using the formula.

5) Data Clustering and prediction

Different type of patterns is discovered in this data mining for clustering. Various types of techniques are there. It is been identify that DBSCAN will give better results when compared to other methods. For this reason, this method can be used for the diagnoses of lung cancer at an early stage. With modification done to the learning rate and the neighborhood distance and the weight updation formula, the model can yield a better performance result.

7. RESULTS AND DISCUSSIONS

The aim of this survey is to evaluate the technique for extracting knowledge and notify the existing lung cancer data profile. Various types of techniques of data mining are applied on the cancer data but we used only DBSCAN Method. A survey work has been done in this area related to data mining algorithms applied on lung cancer data. Data cleaning is the major challenging processed involved in this area, because of data extracting from various sources base done required attributes. As there is an increase in training data, performance is also increase and improved

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

CONCLUSION

An incremental density-based clustering(DBSCAN) algorithm is introduced to incrementally build and update clusters in datasets. The algorithm incrementally partitions the dataset to reduce the search space to each partition instead of scanning the whole dataset. After that the algorithm incrementally forms and updates dense regions in each partition. Following identifying possible dense regions in each partition, the algorithm uses an inter-connectivity measure to merge dense regions to form the final number of clusters. Experimental results show that the proposed algorithm has a comparable accuracy compared to related incremental clustering algorithms. However, the proposed algorithm has significant improvements on the runtime with a speedup factor of 3.2. The proposed algorithm is also proved to perform better in large datasets with higher dimensions compared to related algorithms.

Other possible enhancements to the proposed algorithm are planned for future work. Probably a major enhancement is designing the algorithm to work in a parallel manner. Given the independence of the partitions, incremental DBSCAN for each partition can be applied in parallel. It is expected that the parallel version of the proposed algorithm will achieve better performance with comparable accuracy.

REFERENCES

[1] K.P.Soman, ShyamDiwakar, and V.Ajay, “Insight into Data Mining- Theory and Practice”, Eastern Economy Edition, Prentice Hall of India Pvt.Ltd, New Delhi, 2006

[2] Margaret H.Dunham “Data Mining Introductory and Advance Topics”, Low price Edition – Pearson Education, Delhi, 2003.

[3] "Euclidean Distance” in http://people.revoledu.comkardi/tutorial1/Similarity /EuclideanDistance.html.

[4] "Cluster analysis” in http://en.wikipedia.org/wiki/Cluster_ analysis.

[5] Hui-Chuan Lin “Survey and Implementation of Clustering Algorithms” an Unpublished master's thesis for master's degree, Hsinchu, Taiwan, Republic of China (2009).

[6] Schilham A, Prokop M, van Ginneken B. Computer analysis of computed tomography scans of the lung: a survey. IEEE TransMedicalImaging. 25(4):38405doi:10.1109/TMI.2005.862753.; 2006.

[7] Jiang WY. Therapeutic wisdom in traditional Chinese medicine: A perspective from modern science. Trends Pharmacol Sci 2005;26:558-63.

[8] Qiu J. Traditional medicine: A culture in the balance. Nature 2007;448:126-8.

[9] Li S. Mapping ancient remedies: Applying a network approach to traditional Chinese medicine. Science 2015;350:S72-4.

[10] Li S. Possible Relationship Between Traditional Chinese Medicine ZHENG and Molecular Networks. The First Academic Annual Meeting of the China Association for Science and Technology; 1999.

[11] Li S, Zhang ZQ, Wu LJ, Zhang XG, Li YD, Wang YY, et al. Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network. IET Syst Biol 2007;1:51-60.

[12] Fan X, Zhao X, Jin Y, Shen X, Liu C. Network toxicology and its application to traditional Chinese medicine. Zhongguo Zhong Yao Za Zhi 2011;36:2920-2.

[13] Li S, Zhang B. Traditional Chinese medicine network pharmacology: Theory, methodology and application. Chin J Nat Med 2013;11:110-20.

[14] Fan XH, Cheng YY, Zhang BL. Network formulaology: A new strategy for modern research of traditional Chinese medicine formulae. Zhongguo Zhong Yao Za Zhi 2015;40:1-6.

[15] Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics 2005;21:2076-82.