PalArch's Journal of Archaeology of Egypt / Egyptology

PARTICLE SWARM OPTIMIZATION ASSISTED SUPPORT VECTOR MACHINE BASED DIAGNOSTIC SYSTEM FOR LUNG CANCER PREDICTION AT THE EARLY STAGE

Jyotsna Seth[1], Dr. Parma Nand[2], Prabhdeep Singh[3], Dr. Rajbir Kaur[4]

[1*]Dept. of Computer Science and Engineering, Sharda University, India,

[2] Dept. of Computer Science and Engineering, Sharda University, India ,

[3] Dept. of Computer Science and Engineering, Punjabi University, Patiala, India,

[4] Dept. Of Electronics & communication, Punjabi University, Patiala, India

Email: [1]jyotsna.seth@sharda.ac.in, [2]Parma.nand@sharda.ac.in, [3]ssingh.prabhdeep@gmail.com,[4]rajbir277@yahoo.in

ABSTRACT

Cancer is one of the countries' deadliest illnesses and it will cure if diagnosed early. Lung cancer is the main cause of death in humans, as the signs of lung cancer occur in advanced stages, so it is difficult to diagnose and contributes to high mortality among other forms of cancer. Therefore, early prediction of lung cancer is mandatory for the diagnostic process and offers better odds of effective therapy. Researchers focus on healthcare to diagnose and avoid lung cancer early. Medical data has achieved its full capacity by offering large data sets to researchers. . Machine learning is a division of artificial intelligence that utilizes a range of mathematical, probabilistic, and optimization techniques that allow computers to "learn" from past examples and identify correlations that are difficult to distinguish from big, noisy, or complex data sets. Machine Learning is commonly utilized in the detection and prognosis of lung cancer. In this paper, Particle Swarm Optimization assisted Support Vector Machine based Diagnostic System for Lung Cancer prediction at an early stage is proposed. The primary objective of this paper is to evaluate the effect of the PSO and SVM for mining the lung cancer

dataset The aim of this paper is to improve the accuracy of the machine learning algorithm. The proposed technique was also verified by using the various standard lung cancer classification data sets. The comparison is drawn among the proposed and the existing technique based upon the various standard quality of service parameters. Experimental results indicate that the proposed algorithm is more efficient than existing techniques..

1.   Introduction

Lung cancer, like other cancers, is unregulated irregular lung tissue development. During formation, cells migrate across the lung by metastasizing through surrounding tissue or other areas of the body. This cell development gradually creates a tumor mass. Three major forms of lung cancer include non-small cell lung cancer, small cell lung cancer, and lung cancer. NSCLC is the most prevalent form of lung cancer, around 85% of all lung cancers. Lung cancer is the primary cause of cancer-related deaths worldwide nearly 1.6 million individuals are expected to die of lung cancer in 2012. Overall, a man's lifetime risk of contracting lung cancer is around 1 in 14; for a woman, it's around 1 in 17. Survival stats The primary source of lung cancer (85%) is long-term cigarette use. Data indicate a clear association between cigarette consumption and lung cancer. Lung cancer rises and declines are strongly parallel to tobacco use. As intake decreases, the occurrence of lung cancer rises. Tobacco use, passive smoking, radon radiation, genetic causes, and air quality also cause lung cancer.

Healthcare services can never be isolated from machine learning technology growth. Machine learning technology plays a rising function in all stages of lung cancer such as lung field segmentation, bone suppression, and irregular tissue identification. Machine learning is a cross-discipline multi-domain. machine learning's key objects are items, specific algorithms that boost their experience efficiency. .Here characteristics are variables used to characterize entity trends. Extracting features requires identifying suitable variables to characterize artifacts. Feature classification is the method of making objective judgments based on factors used to classify artifacts. Machine learning is ideal for automated object detection. Machine-learning technology learns trends from training data to detect artifacts, and the algorithm can continuously enhance its efficiency before convergence is accomplished through repetitive learning.

Healthcare is an important aspect of human life and frequent analysis of crucial criteria since therapies are the fundamental feature of healthcare. These processes become much more important in addressing psychiatric illnesses that need qualified medical professionals to control their state. At distinct times, the current procedure for monitoring certain essential health metrics of healthcare is assessed by doctors or medical professionals. Furthermore, self-reported subjective health measurements are important to determine care efficacy. These methods also contribute to the lack of critical details on specific treatments during the day or night

These kinds of styles can be observed when a form of a summary of your insight information and facts, and also can be employed throughout even more investigation and also, to get a good example, throughout appliance discovering as well as predictive analytics.

To give an example, the info mining phase could discover various teams through your data, which can be used to obtain more genuine prediction benefits utilizing a decision, help system. Or your current information selection, information and facts prep, neither outcome meaning as well as reporting will be the principle information pursuit phase but accomplish remain in the whole KDD method when even more steps.
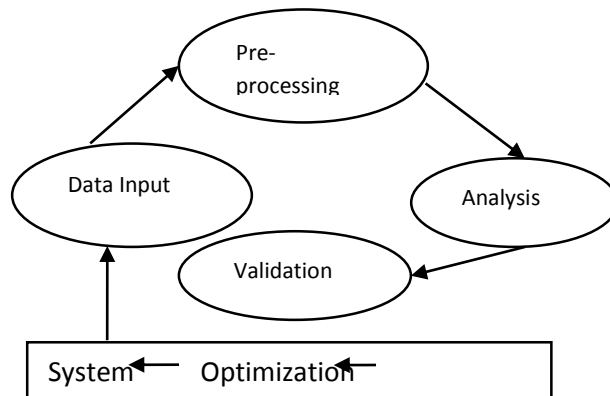
Fig 1 *Machine Learning Process*

Details exploration is the period often utilized in pc science. Oahu is the approach used to acquire the actual helpful details via big data fixed making use of various techniques. A variety of data exploration algorithms are utilized to acquire details via the data fixed including Group, Clustering, Aggregation, and many more. The entire aim involving the data exploration approach is mostly to be able to acquire details via the data fixed as well as change the item into a superior easy to Understand variety which they can use further. Merely, getting helpful details via the data is called data mining.

A. **Abbreviations and Acronyms Data Mining In Healthcare**

The medical industry currently builds massive amounts of complex info regarding clients, medical centers resources, ailment a diagnosis, electric client files, healthcare devices, etc. These large amounts of information are an essential learning resource being processed as well as analyzed about knowledge removal allowing aid concerning cost-savings as well as making decisions. Data mining gives some sort of tools and methods that may be relevant to this particular ready-made info to find out secret patterns that include health pros an extra source of knowledge to create decisions. The actual decisions relax using medical professionals.

The rest of the paper is organized as follows. Section II contains the introduction of Particle Swarm optimization, Section III contains the related work, section IV explains the methodology with a flow chart, Section V describes results and discussion, Section VIII concludes research work with future directions.

2.   Particle Swarm Optimization

The Particle swarm optimization (PSO) is a Fuzzy C-Mean clustering formula that is usually ways to show exactly how facts might be labeled in addition to the group with a group or even in any program [13]. This began by Dunn [14].In this kind of paper, utilizing Fuzzy c-means clustering formula background front things usually are segmented from your picture or even frames. This kind of formula mostly helps to portion your p whether it belongs to track record or even foreground. The sheer numbers of groups are generated a good range of things from the frames. Using this wooly d indicates clustering formula centroid will probably be selected. First, your centroid is usually picked out randomly good suggest in the pixels. The right centroid will probably be worked out immediately after obtaining the quality of pixel utilizing many iterations. Within this paper wooly c-means clustering strategy is employed for choosing your centroid good p along with the found sides utilizing the fresh border detectors formula [11]. The pursuing formula displays how a wooly c-mean clustering procedure could be used to portion your front thing from your offered image/frame.

3.   Related Work

CIOs et al. [1] deal with the particular exclusive top features of information mining by using healthcare data. He's discussed a variety of honest plus legalized areas linked to healthcare information mining just like information management, fear of legal cases, forecasted benefits, plus exclusive supervision issues. With this paper, your dog stated which the numerical knowledge of approximation plus theory development in healthcare data is mainly not the same as all other information variety routines. Mitra et al. [2] come with market research on the obtainable materials on information mining utilizing gentle processing. A categorization offers been recently supplied predicated on various gentle computing devices and hybridizations put on, the data mining operate put on, plus the desire criterion identified by way of the unit. Inherited algorithms provide a successful look for algorithms to pick out one, via mixed advertising information, predicated on several choice criterion objective function. Hard items are usually suitable to handle various types of concerns with data. Bellazzi et al.[3] provides discussed which the prevalent option of brand new computational techniques plus resources about information exam plus predictive modeling necessitates healthcare informatics gurus plus experts to help systematically select the most appropriate approach to cope with controlled forecast problems. An enormous assortment of these methods needs common plus essential pointers that can help experts within the appropriate variety of knowledge mining resources, progression plus agreement with predictive types, combined with distribution with predictive patterns within just clinical environment. Palaniappan et al.[4] researched that the health care industry collects huge amounts of medical care information. These studies operate have created one Smart Soul Disease Prediction System (IHDPS) implementing information mining strategies, Choice Timber, Unaware Bayes

plus Sensation problems Network. Success express that many approaches have exceptional muscle with spotting the particular ambitions on the acknowledged mining goals. Making use of healthcare single profiles just like grow older, gender, blood pressure level plus sugar levels perhaps it will predict the particular likelihood of people finding a coronary heart disease. Marengo, Fumbeya, et al.[5] introduced an information exploration program made to support the particular quick development of data-derived NTCP models. Prestashop exploits the normal healthcare workflow and information encoded by using a normal ontology. Mcdougal stated which the system referred to is a piece of helpful information on the advance with irradiation oncology information exploration types especially plus local-level LHS components with the general. Gholap, et al.[6] offers consist of your collaborative files mining procedure to offer multi-level evaluation coming from health test out data. The aim would be to examine success by simply collaboratively implementing various files mining techniques such as classification, clustering, along with connection procedure mining. General, the method seeks from getting information coming from health test out files to be able to increase model of health checks by simply creating peace of mind inside the final results implementing multi-level evaluation Nie, Liqiang, et al.[7] offers your story technique so that you can code health data files by simply collectively employing local mining along with worldwide studying approaches, that are tightly hooked up along with mutually reinforced. Neighborhood mining efforts so that you can procedure the individual medical record by simply independently receiving the health strategies from the medical record itself and then maps these phones' authenticated terminologies. Duan, K. B.et al.[8] offers consist of the latest feature choice procedure which relies on a backward elimination treatment a lot like which put in place in assist vector equipment recursive feature elimination (SVM-RFE). As opposed to an SVM-RFE procedure, at each step, a consist of approach computes a feature ranking score coming from a mathematical analysis connected with bodyweight vectors connected with various straight line SVMs skilled with subsamples connected with the very first coaching data. Bramer.M et al. [9] Offers talked about a pair of methods of speeding connected with the anatomical encoding approach. First, an example may be the usage of a proficient algorithm that eliminates code. Subsequent one particular is a demotic approach to just about parallelizing the system one processor.GP efficiency with health classification troubles can be as opposed to coming from a benchmark data source having success received by simply nerve organ networks. Success demonstrates that GP functions equally in classification along with generalization. Prather J.C et al.[10] offers used the approaches of data mining (also called Understanding Uncovering in databases) to find interactions inside of a large professional medical database. They will illustrate methods linked to mining your professional medical data source including file warehousing, file query& clean-up along file analysis.

4. Methodology

This section discusses the proposed algorithm.

Steps	Description
Algorithm parameters	A: Population of agents p_i: Position of agent a_i in the solution space f: Objective function v_i: Velocity of agents a_i $V(a_i)$: Neighborhood of agent a_i (fixed) The neighborhood concept in PSO is not the same as the one used in another meta-heuristics search, since in PSO each particle's neighborhood never changes (is fixed).
Required attributes with values	Number of particles usually between 10 and 50 C_1 is the importance of personal best value C_2 is the importance of neighborhood best value Usually $C_1 + C_2 = 4$ (empirically chosen value) If velocity is too low → algorithm too slow If velocity is too high → algorithm too unstable
Particle update rule	$p = p + v$ with $v = v + c1 * rand * (pBest - p)$ $+ c2 * rand * (gBest - p)$ where: particle's position, v: path direction, c_1: the weight of local information, c_2: the weight of global information, $pBest$: the best position of the particle, $gBest$: the best position of the swarm, and $rand$: random variable
PSO Algorithm	$[x *] = PSO()$ $P = Particle\_Initialization();$ $For\ i = 1\ to\ it\_max$ $For\ each\ particle\ p\ in\ P\ do$ $fp = f(p);$ $If\ fp\ is\ better\ than\ f(pBest)$ $pBest = p;$ end $endgBest = best\ p\ in\ P;$ $For\ each\ particle\ p\ in\ P\ do$ $v = v + c1 * rand * (pBest - p)$ $+ c2 * rand * (gBest - p);$ $p = p + v;\ endend$
PSD based K-means algorithm for training	1. Create a 'population' of attributes (particles) from developed clusters of data set X using the K-means algorithm. 2. Evaluate each particle's position according to the objective function i.e. minimum root mean squared error (RMSE). 3. If a particle's current position is better than its previous best position, update it. 4. Determine the best particle (according to the particle's previous best positions). 5. Return optimized trained model

5. Results and Discussion

For research and implementation, the proposed technique is appraised using the WEKA tool. The evaluation of the proposed method is done on the origin of the following parameters as Accuracy, Mean square error, and Sensitivity. The subsequent data demonstrates the comparison regarding response to diverse parameters. The result demonstrates the proposed solution provides improvement over active approaches. After the results, we compared the proposed solution against the current procedures.
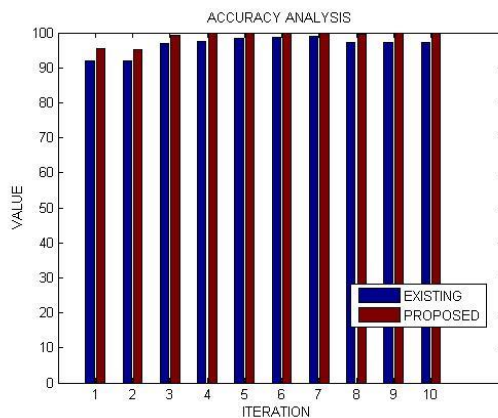
B. **Accuracy**

Accuracy is one metric for assessing model classification. Classification quality is what we generally say by utilizing accuracy. It is the ratio of the number of accurate predictions to total input samples. It is calculated as-

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 1. *Accuracy*

ITERATION	EXISTING	PROPOSED
1	91.9271	95.4479
2	91.9271	95.3177
3	97.0052	99.3958
4	97.3958	99.9648
5	98.4979	99.8590
6	98.5979	99.9593
7	98.9583	99.9794
8	97.2188	99.6592
9	97.2188	99.8698
10	97.2188	99.8698

*Fig 2.**Accuracy Evaluation*

C. **Mean square error**

The mean squared error says the closeness of a regression line across points. It achieves this by taking and squaring distances from points on regression rows. The squaring must eradicate all derogatory indications. It also offers greater discrepancies in weight. The mean squared error is considered to find the average collection of errors.

$$\text{MSE} = \frac{1}{ab}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2$$

Where Aij and Bij are the image pixel value of the reference image.

Table 2. *Mean Square Error*

ITERATION	EXISTING	PROPOSED
1	0.0807	0.0755
2	0.0788	0.0723
3	0.079	0.0725
4	0.0765	0.0739
5	0.0771	0.0732
6	0.0773	0.0738
7	0.077	0.0742
8	0.0778	0.0739
9	0.0786	0.0729
10	0.079	0.0721
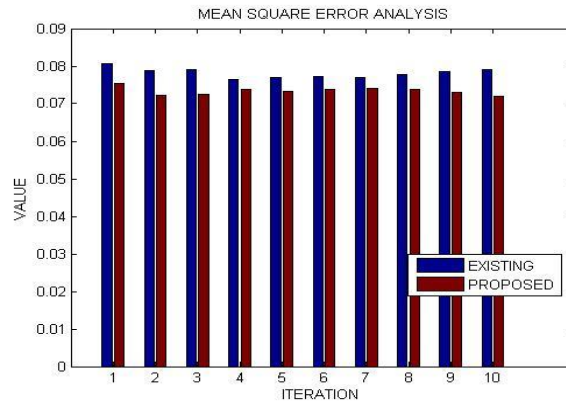
Fig 3.*Mean Square Error Evaluation*
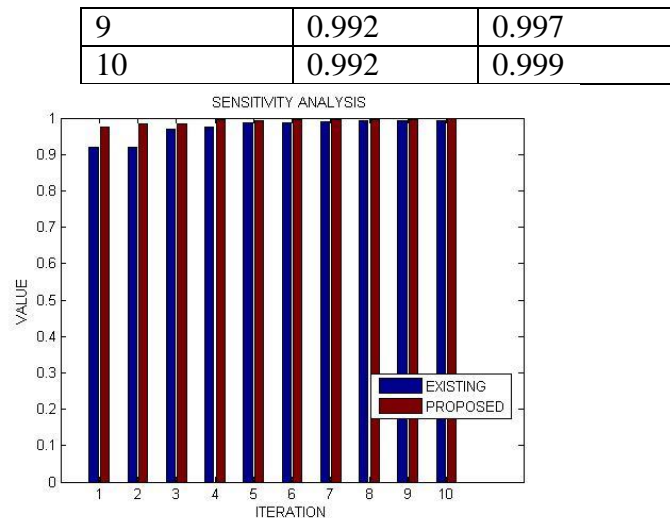
D. **Sensitivity**

Sensitivity is a calculation of the proportion of real positive cases expected positive. Often named Recall, awareness. This means that there would be another proportion of positive events, which will be wrongly predicted as negative. This may also be represented as a false negative rate. The addition of sensitivity and the false-negative rate is 1. It lies between 0-1. Value of sensitivity near 1 signifies efficient results.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Table.3. *Sensitivity*

ITERATION	EXISTING	PROPOSED
1	0.919	0.974
2	0.919	0.983
3	0.97	0.984
4	0.974	0.997
5	0.987	0.993
6	0.987	0.996
7	0.99	0.996
8	0.992	0.995

| 9 | 0.992 | 0.997 |
| 10 | 0.992 | 0.999 |

Fig 3 *Sensitivity Evaluation*

6.   Conclusion and Future Scope

Lung cancer is one of the world's biggest causes of death. Many lung cancer diagnosis and identification is conducted using different analyses and classification techniques. Although the origin of lung cancer remains obscure, avoidance becomes difficult, but early diagnosis of lung tumor is the only way to cure lung cancer. In this paper, the integration of Particle swarm optimization (PSO) with a support vector machine (SVM) is proposed to detect lung cancer at a very early stage. The proposed technique is verified by using the various standard lung cancer classification data sets. The comparison is drawn among the proposed and the existing technique based upon the various standard quality metrics of the data mining. Experimental results indicate that the proposed algorithm is more efficient than existing techniques. In the future, the suggested method would help recommend diagnosing cancer in various human organs. The proposed strategy for particular cancer, i.e. community of diseases that help to minimize the development of unhealthy cells or spread to other areas of the body

Reference

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W.E.: Medical data mining: knowledge discovery in a clinical data warehouse. In Proceedings of the AMIA annual fall symposium (p. 101). American Medical Informatics Association(1997).

Parpinelli, R. S., Lopes, H. S., &Freitas, A. A.: An ant colony based system for data mining: applications to medical data. In Proceedings of the genetic and evolutionary computation conference (GECCO-2001) (pp. 791-797)(2001,July).

Ghazavi, S. N., & Liao, T. W.: Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine, 43(3), 195-206(2008).

Delen, D., Walker, G., &Kadam, A.:Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127(2005).

Moses, D.:A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, 42(2)(2015).

Ji, S., Wang, Z., Liu, Q., & Liu, X.:Classification Algorithms for Privacy Preserving in Data Mining: A Survey. In International Conference on Computer Science and its Applications (pp. 312-322). Springer Singapore(2016,December).

Rani, G., Gladis, D., &Mammen, J.Classification and Prediction of Breast Cancer Data derived Using Natural Language Processing. In Proceedings of the Third International Symposium on Women in Computing and Informatics (pp. 250-255). ACM(2015,August).

Das, T. K.:A customer classification prediction model based on machine learning techniques. In 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 321-326). IEEE(2015,October).

Yu Zhou, Xiaokang Yang, Yongzheng Zhang, Xiang Xu, Yipeng Wang, Xiujuan Chai, Weiyao Lin, Unsupervised adaptive sign language recognition based on hypothesis comparison guided cross validation and linguistic prior filtering, Neurocomputing, Volume 149, Part C, 3 February 2015, Pages 1604-1612.

WalidMagdy, Tamer Elsayed, Unsupervised adaptive microblog filtering for broad dynamic topics, Information Processing & Management, Volume 52, Issue 4, July 2016, Pages 513-528.

Shiqiang Du, Yide Ma, Shouliang Li, Yurun Ma, Robust unsupervised feature selection via matrix factorization, Neurocomputing, Volume 241, 7 June 2017, Pages 115-127.

Daniel Carlos GuimarãesPedronette, Ricardo da S. Torres, Unsupervised Rank Diffusion for Content-Based Image Retrieval, Neurocomputing, Available online 16 May 2017.

Herman Kamper, Aren Jansen, Sharon Goldwater, A segmental framework for fully-unsupervised large-vocabulary speech recognition, Computer Speech & Language, Available online 18 May 2017.

Wei He, Xiaofeng Zhu, Debo Cheng, Rongyao Hu, Shichao Zhang, Unsupervised feature selection for visual classification via feature-representation property, Neuro-computing, Volume 236, 2 May 2017, Pages 5-13.

ZouhairMbarki, HasseneSeddik, Ezzedine Ben Braiek, A rapid hybrid algorithm for image restoration combining parametric Wiener filtering and wave atom transform, Journal of Visual Communication and Image Representation, Volume 40, Part B, October 2016, Pages 694-707.

Gloria Re Calegari, EmanuelaCarlino, Diego Peroni, Irene Celino, Filtering and windowing mobile traffic time series for territorial land use

classification, Computer Communications, Volume 95, 1 December 2016, Pages 15-28.

Mostafa Mohammadpourfard, Ashkan Sami, AlirezaSeifi, A statistical unsupervised method against false data injection attacks: A visualization-based approach, Expert Systems with Applications, Volume 84, 30 October 2017, Pages 242-261.

EliahuKhalastchi, Meir Kalech, LiorRokach, A hybrid approach for improving unsupervised fault detection for robotic systems, Expert Systems with Applications, Volume 81, 15 September 2017, Pages 372-383.

Chong Yang, Xiaohui Yu, Yang Liu, YanpingNie, Yuanhong Wang, Collaborative filtering with weighted opinion aspects, Neurocomputing, Volume 210, 19 October 2016, Pages 185-196.

Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, WangmengZuo, Subspace clustering guided unsupervised feature selection, Pattern Recognition, Volume 66, June 2017, Pages 364-374.

Angurala, Mohit, et al. "An internet of things assisted drone based approach to reduce rapid spread of COVID-19." Journal of Safety Science and Resilience 1.1 (2020): 31-35.

Gupta, Vivek, et al. "An energy efficient fog-cloud based architecture for healthcare." Journal of Statistics and Management Systems 21.4 (2018): 529-537.

Singh, Prabhdeep, and Rajbir Kaur. "An integrated fog and Artificial Intelligence smart health framework to predict and prevent COVID-19." Global transitions 2 (2020): 283-292.

Singh, Narinder, Prabhdeep Singh, and Rajbir Kaur. "Design and Development a Hybrid Classifier to Improve Lung Cancer Diagnosis." Journal of the Gujarat Research Society 21.15 (2019): 323-328.