Indo-Iranian Journal of Scientific Research (IIJSR)

(Peer Reviewed International Journal), Volume 2, Issue 3, Pages 343-352, July-September 2018

A SURVEY ON DATA CLASSIFICATION IN DATA MINING

R.NANDHAKUMAR[1] AND ANTONY SELVADOSS THANAMANI[2]

[1]Research Scholar-Department of Computer Science, NGM College, Pollachi-642001, India.

[2]Associate Prof & Head, Department of Computer Science, NGM College, Pollachi-642001, India.

E-Mail- nkumarram@gmail.com, selvdoss@gmail.com

ABSTRACT: One of the main leading causes of death for human is the cancer, which plays as severe threat to the human society as the change of weather, urbanization and chosen of fast food culture that raises the happening of cancer in current age. The methods that are applicable in data mining offer a significant contribution to the ground of medical diagnostics for the accurate prediction of the disease. The predictive analysis methods of data mining frame a knowledge prediction paradigm by analyzing the present history of patients so as to analyse the future data. This paper reviews the results of different classification technique that have been established in the research articles from the year 2014-2017 and makes observation among presented outputs of the previous works. Furthermore, experimentation is also conducted to validate the authenticity of the best clustering algorithms as an evidence of concept.

Keywords*:* *Classification, prediction models, predictive analysis.*

INTRODUCTION:

Machine learning is not latest to cancer research. Artificial Neural Networks (ANNs), decision trees (DTs), support vector machines, and etc have been used in cancer prediction and diagnosis for almost 20 years. Today machine learning techniques are being used in a wide range of applications ranging from predicting and classifying tumors through X-ray and CRT images to the classification of virulence from proteomic and genomic (microarray) as says according to the latest PubMed statistics, more than 1500 papers have been published on the subject of machine learning and cancer. Nevertheless, the vast majority of these papers are concerned with using machine learning methods to identify, classify, detect, or distinguish tumors and other malignancies. In the other words machine learning has been used fundamentally as a help to cancer diagnosis and prediction. It has been used been relatively recently that cancer researchers have tried to apply machine learning close to cancer prediction and prognosis. As an impact the body review in the ground of machine learning and cancer prediction or prognosis is relatively small.

Certainly, a cancer diagnosis typically associates multiple physicians from various specialties using different subsets of biomarkers and many clinical factors, including the age and general health of the patient, the location and type of cancer also the grade and size of the tumor. Usually histological clinical and demographic information must all the carefully integrated by the attending physician to come up with a reasonable prognosis. Even for the most skilled clinician, this is not easy to do. Similar disputed also exist for both physicians and patients alike when it comes to the problems of cancer prevention and cancer susceptibility prediction. Family history, age, diet, weight (obesity), high-risk habits (chain smoking, heavy alcohol drinking), and exposure to environmental carcinogens (UV radiation, radon, asbestos, PCB) all play a role in predicting an individual's risk for developing cancer. Sadly these traditional macro-scale clinical environmental and behavioral parameters mainly do not provide enough information to make robust predictions or prognoses. Specifically what is required is some very particular molecular details about either the cancer or the patient's possess genetic make-up.With the accelerated development of genomic (DNA sequencing, microarrays), proteomic (protein chips, tissue arrays, immune-histology) and imaging (fMRI, PET, micro-CT) technologies, this kind of molecular-scale information about patients or tumors can now be readily obtained. Still, as the number of parameters we measure grows, so too does the challenge of trying to make sense of all this information.

REVIEW OF LITERATURE :

[1] A.Priyanga.et.al presented data mining based cancer prediction system (DMBCPS) that estimates the risks of the breast, skin, and lung cancers. The Authors validated by comparing its predicted results with patient's prior medical information and it was analysed by using weka tool. The major focus of this model is to provide the early warning to the users and also cost efficient to the user. The comparison is accomplished with Naïve Bayes, J48, and ID3 algorithms. This system is available with online so that people can easily check their risk level and take proper medicine based on their risk status.

[2] P.Ramachandran.et.al proposed a novel multi layered method combining clustering and decision tree techniques to build a cancer risk prediction system is proposed here which predicts lung, breast, oral, cervix, stomach and blood cancers as well as user friendly, time, and cost saving. The using methodology is data mining's classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed, fed into the database and classified to yield significant patterns using decision tree algorithm. The data is clustered using K-Means algorithm to separate data as cancer suffered patients and non-cancer suffered patient. Moreover the cancer cluster is subdivided into six clusters. Finally a

prediction system is developed to analyse risk levels which aid in prognosis. The research is helps in finding of a person's cancer predisposition for cancer before going for clinical and lab experiments which is cost and time consuming.

[3] Konstantinakourou.et.al proposed a machine learning applications in cancer prognosis and prediction. This paper discussed applying machine learning techniques to improve the understanding of cancer progression and an appropriate level of validation for the methods to be considered in the everyday clinical practice. Machine learning is a branch of artificial intelligence that relates the problem of learning from data samples to the general concept of inference. The predictive models discusses based on different supervised ML techniques as well as on various input features and data samples. Given the growing trend on the application of ML methods in cancer research and presented most recent publications that employ some techniques as an aim to model cancer risk patient outcomes.

[4] Meng-Yun Wu.et.al proposed a gene selection and cancer classification algorithm that is the Laplace naïve Bayes model with mean shrinkage where the regularization parameter perhaps over sighted by the predetermined number of the chosen biomarkers. Because the objective function of the author's method is a bitwise linear function with concern the mean of every class, its ramification is relative low, which has been theoretically analyzed. The Laplace distribution, which is utilized as the conditional distribution of the samples made this method less sensitive to outliers. LNB-MS has taken into account of the group effects in terms of two views that are gene expression profiles and GO annotation. Publicly available cancer data sets shown that LNB-MS achieved good classification even on unbalanced data sets such as DLBCL2 and Pros3.

[5] Benjamin Harvey.et.al presented a novel methodology that uses a CSDP divisible single-dimensional method for wavelet based transformation for initializing a threshold which will hold significantly expressed genes by way of the denoising process for robust classification of cancer patients. Furthermore, the complete study was implemented and surrounded by CSDP atmosphere. The application of cloud computing and wavelet-based thresholding in order to denoising was utilized for the classification of samples with in the Global Cancer Map (CGM), Cancer Cell Line Encyclopedia (CCLE and the Cancer Genome Atlas (TCGA). For the wavelet-based thresholding of the GCM data, the divisible single-dimensional technique took about 400.29 seconds to be estimated in the cloud atmosphere and 1000.53 seconds in the CSDP atmosphere. The existing techniques like SVM, NN, and K-NN

classification methods have given accuracy up to 88%. The result of the one-dimensional denoising method achieved accuracy up to 97%.

[6] Shu-Lin Wang.et.al proposed a robust classification method of tumor subtype by using correlation filters to recognize the complete pattern of tumor subtype hidden in differentially expressed genes. The work has introduced two concrete correlation filters that are Minimum Average Correlation Energy (MACE) and Optimal tradeoff synthetic discriminant function (OTSDF) which is a test sample suits the templates synthesized for every subclass. The proposed method applied on six publicly available data sets which is robust to noise, and more efficiently avoid the impacts of dimensionality curse. According to this paper, correlation filter-based method achieved better performance when compared to many model based methods. The balanced training sets are exploited to synthesize the templates especially, this method could detect the resemblance of a complete pattern when ignoring mismatches between the test sample and the synthesized template. The work has performs well even if only a less training samples are available.

[6] Ji-xin LIU.et.al presented a newly compressed sensing (CS) for mass spectrum data processing. MS sensing data is utilized to realize the prior MS analysis over Compressed Sensing Recognition (CSR) method. It can extract the valuable prior information from MS sensing data for pathological diagnosis such as SD can guarantee validity recovery quality for MS analysis with much lower computational cost. The framework results shown the efficiency and feasibility of MS data processing via CSR and SD and it seems more effective one.

[7] Isabelle guyon.et.al addressed the difficulties of the selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA microarrays. The authors have used available training examples from cancer and normal patients and built a classifier suitable for genetic diagnosis also drug discovery. The proposed new method of gene selection using Support Vector Machine (SVM) method depends on Recursive Feature Elimination (RFE). It has eliminated gene verbosity automatically and yielded better and more compact gene subsets. The method has achieved 98% accuracy.

[8] Yan-jun Hong.et.al presented a discrimination has been investigated for the probabilistic classification between healthy and ovarian cancer serum samples utilizing proteomics data from mass spectrometry (MS). The method applies data normalization, clustering and a linear discriminant analysis

of surface-enhanced laser desorption ionization (SELDI) time-of-flight MS data. The probabilistic classification approach estimates the optimal linear discriminant using the intricate human blood serum SELDI spectra. Cross-Validation and training/testing data split tests are over sighted to verify the optimal discriminant and determined the accuracy and robustness of the approach. The results were seems excellence in performance up to 97% in sensitivity, specificity and positive predictive values on ovarian cancer. The discrimination analysis assists the molecular recognizes of differentially expressed proteins and peptides the health versus ovarian patients.

[9] Lilien RH.et.al have developed an algorithm called Q5 for probabilistic classification between healthy and disease serum samples utilizing mass spectrometry. The algorithms apply LDA on SELDI-TOF mass spectrometry (MS) data determined on four real data sets from complete, complex SELDI spectra of human blood serum. The solution focuses on computationally effective such as it is no iterative and computes the optimal linear discriminant using closed form equations. The Q5 approach performs out earlier full spectrum complex sample spectral classification techniques and can provide indications as to the molecular identities differentially expressed proteins and peptides.

[10] Conrad T.et.al proposed a high-throughput proteomics techniques, being mass spectrometry (MS)-based method, generate very high-dimensional data sets. Spectra from healthy patients versus spectra from patients having a specific disease. The Machine learning algorithm is required to identify these discriminating characteristics and classify obscure spectra depends on this feature set. Because the acquired data are generally noisy, the algorithms should be robust against noise and outliers, while the recognized feature set should be small as possible. The presented algorithm Sparse Proteomics Analysis (SPA) based on the theory of compressed sensing that permits us to recognize a minimal discriminating set of characteristics from mass spectrometry data sets. The results shown that how this method performed on artificial and real world data sets and competitive with standard algorithms for analyzing proteomics data and robust against random and systematic noise.

[11] Shulin Wang et.al proposed a novel method for robust tumor classification based on a MACE filter to resolve the problems of too much noise and the curse of dimensionality that the number of genes far exceeds the size of the samples in tumor data sets, tumor classification by selecting a small set of gene subset from the thousands of genes becomes a dispute. The authors have given an approach that a novel, high accurate method, which used the global scheme of differentially expressed genes matching to every

tumor subtype, which is resolved by tumor-related genes, to classify tumor samples by utilizing Minimum Average Correlation Energy (MACE) filter method to computing the resemblance degree between an experiment sample with unknown labels in the test set and the template constructed with the training set. It was a method, which gives an efficient and robust in classification performance.

[12] Prachi Kawalkar et.al discussed various techniques for pre-processing, segmentation, feature extraction, and classification of biomedical images to predict and classify glands in human tissues. Furthermore, the focus on resolving biomedical images and to provide a solution for that problems. The study was about introducing different methods like polar conversion, along with object based segmentation and SVM classification. The given solutions provide an honest base for additional analysis within the space of cytological image segmentation.

[13] Shikha Agrawal et.al survey clearly displays the efficiency of neural network technologies in the detection of cancer. Maximum neural network shows the overwhelming result to classify tumor cells accurately. Particularly, MLP (Multi-Layer Perception) has given 97.1% accuracy and PNN (Probabilistic Neural Network) which has given 96% accuracy, perceptron with 93% and also ART1 shown 92% accuracy. The removal of missing values from the data set improved the test results. The neural network structures could provide an enhanced result also classification rate and training time is very high.

[14] Benjamin Harvey et.al presented a novel methodology that uses a CSDP divisible single-dimensional method for wavelet based transformation for initializing a threshold which will hold significantly expressed genes by way of the denoising process for robust classification of cancer patients. Furthermore, the complete study was implemented and surrounded by CSDP atmosphere. The application of cloud computing and wavelet-based thresholding in order to denoising was utilized for the classification of samples with in the Global Cancer Map (CGM), Cancer Cell Line Encyclopedia (CCLE and the Cancer Genome Atlas (TCGA). For the wavelet-based thresholding of the GCM data, the divisible single-dimensional technique took about 400.29 seconds to be estimated in the cloud atmosphere and 1000.53 seconds in the CSDP atmosphere. The existing techniques like SVM, NN, and K-NN classification methods have given accuracy up to 88%. The result of the one-dimensional denoising method achieved accuracy up to 97%.

[15] Meng-Yun Wu et.al proposed a gene selection and cancer classification algorithm that is the Laplace naïve Bayes model with mean shrinkage where the regularization parameter perhaps over sighted by the predetermined number of the chosen biomarkers. Because the objective function of the author's method is a bitwise linear function with concern the mean of every class, its ramification is relative low, which has been theoretically analyzed. The Laplace distribution, which is utilized as the conditional distribution of the samples made this method less sensitive to outliers. LNB-MS has taken into account of the group effects in terms of two views that are gene expression profiles and GO annotation. Publicly available cancer data sets shown that LNB-MS achieved good classification even on unbalanced data sets such as DLBCL2 and Pros3.

[16] Shu-Lin Wang et.al proposed a robust classification method of tumor subtype by using correlation filters to recognize the complete pattern of tumor subtype hidden in differentially expressed genes. The work has introduced two concrete correlation filters that are Minimum Average Correlation Energy (MACE) and Optimal tradeoff synthetic discriminant function (OTSDF) which is a test sample suits the templates synthesized for every subclass. The proposed method applied on six publicly available data sets which is robust to noise, and more efficiently avoid the impacts of dimensionality curse. According to this paper, correlation filter-based method achieved better performance when compared to many model based methods. The balanced training sets are exploited to synthesize the templates especially, this method could detect the resemblance of a complete pattern when ignoring mismatches between the test sample and the synthesized template. The work has performs well even if only a less training samples are available.

[17] Ji-xin LIU et.al presented a newly compressed sensing (CS) for mass spectrum data processing. MS sensing data is utilized to realize the prior MS analysis over Compressed Sensing Recognition (CSR) method. It can extract the valuable prior information from MS sensing data for pathological diagnosis such as SD can guarantee validity recovery quality for MS analysis with much lower computational cost. The framework results shown the efficiency and feasibility of MS data processing via CSR and SD and it seems more effective one.

[18] Isabelle guyon et.al addressed the difficulties of the selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA microarrays. The authors have used available training examples from cancer and normal patients and built a classifier suitable for genetic diagnosis also drug

discovery. The proposed new method of gene selection using Support Vector Machine (SVM) method depends on Recursive Feature Elimination (RFE). It has eliminated gene verbosity automatically and yielded better and more compact gene subsets. The method has achieved 98% accuracy.

[19] Shulin Wang et.al proposed a novel method for robust tumor classification based on a MACE filter to resolve the problems of too much noise and the curse of dimensionality that the number of genes far exceeds the size of the samples in tumor data sets, tumor classification by selecting a small set of gene subset from the thousands of genes becomes a dispute. The authors have given an approach that a novel, high accurate method, which used the global scheme of differentially expressed genes matching to every tumor subtype, which is resolved by tumor-related genes, to classify tumor samples by utilizing Minimum Average Correlation Energy (MACE) filter method to computing the resemblance degree between an experiment sample with unknown labels in the test set and the template constructed with the training set. It was a method, which gives an efficient and robust in classification performance.

[20] Prachi Kawalkar et.al discussed various techniques for pre-processing, segmentation, feature extraction, and classification of biomedical images to predict and classify glands in human tissues. Furthermore, the focus on resolving biomedical images and to provide a solution for that problems. The study was about introducing different methods like polar conversion, along with object based segmentation and SVM classification. The given solutions provide an honest base for additional analysis within the space of cytological image segmentation.

CONCLUSION:

Thus the literature study aids to identify data mining techniques to predict cancer disease at precious stage. Various researchers have presented different techniques to detect the cancer disorder and various kind of accuracy levels as per used techniques. These technique helps to reduce the irrelevant data of patient's from the databases in healthcare center. Algorithms namely decision, neural networks, PNN, SVM, Naïve bayes, Mass spectrometry, correlation, KNN and etc consider for the review. These techniques gave different result based on time span, accuracy, performance and cost. Also these effective classification data helps to detect the treatment to patient. In future a best method to predict the cancer disease can be found out with improvements in existing methods.

REFERENCES:

[1] Joseph A.Cruz and David S.Wishart "Applications of machine learning in cancer prediction and prognosis" cancer informatics, Vol 2, PP 59-77.

[2] A.Priyanga, S.Prakasam, "Effectiveness of Data Mining –based cancer prediction system", International Journal of Computer Applications", (0975 -8887) Vol 83, PP 10, 2013.

[3]P.Ramachandran, N.Girija, T. Bhuvaneshwari, Early detection and prevention of cancer using data mining techniques" international journal of computer applications (0975-8887) Vol 97-No.13, 2014.

[4] Konstantinakourou, themis P.Exarchos, Konstantinos P.Exarchos Michalis, V.Karamouzis Dimitrios I fotiadis, "Machine learning applications in cancer prognosis and prediction", computational and structural biotechnology journal, Vol 13, PP 8-17, 2015.

[5] Meng-yun wu, Dao-Qing Dai, Yu Shi, Hong Yan, Xiao-Fei Zhang "Biomarker identification and cancer classification based on Micro data using Laplace Naïve Bayes Model with mean shrinkage", IEEE/ACM transactions on computational biology and bioinformatics, Vol 9, issue,6, PP 1649-1662, 2012.

[6] Shu-Lin Wang, Yi-Hai Zhu, Wei Jia, De-Shuang Huang,"Robust Classification method of tumor subtype by using correlation filters", IEEE/ACM Transaction on computational biology and bioinformatics, Volume 9, issue 2, PP 580-591, 2012.

[7] Isabelle Guyon+, Jason Weston, Stephen Barnhill, M.D. and Vladimir Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", Machine learning, PP 1-39.

[8] Yan-jun Hong, Xiao-dan Wang, David Shen, Su Zeng, "Discrimination analysis of mass spectrometry proteomics for ovarian cancer detection", Acta Pharmacologica Sinica, Vol 29, No 10 (October 2008)

[9] Lilien RH1, Farid H, Donald BR,"Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum", Journal of computational biology, 2003;10(6):925-46.

[10] Shu-Lin Wang , Yi-Hai Zhu, Wei Jia, De-Shuang Huang,"Robust classification method of tumor subtype by using correlation filters", computational biology and bioinformatics, IEE computer society, 2012 vol. 9 Issue No. 02 - March/April.

[11] Aha D. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies. 1992;36:267–287.

[12] Ando T, Suguro M, Hanai T, et al. Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn J Cancer Res. 2002;93:1207–12. [PubMed]

[13] Atlas L, Cole R, Connor J, et al. Performance comparisons between backpropagation networks and classification trees on three real-world applications. Advances in Neural Inf. Process. Systems. 1990;2:622–629.

[14] Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003;95:470–8. [PubMed]

[15] Baldus SE, Engelmann K, Hanisch FG. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. Crit Rev Clin Lab Sci. 2004;41:189–231. [PubMed]

[16] Benjamin Harvey-IEEE Member and Soo-Yeon Ji-IEEE Member," Cloud-Scale Genomic Signals Processing for Robust Large-Scale Cancer Genomic Microarray Data Analysis", 2168-2194 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.See http://www.ieee.org/publications_standards/publications/rights/index.html for more information

[17]Meng-Yun Wu, Dao-Qing Dai, Yu Shi, Hong Yan, and Xiao-Fei Zhang," Biomarker Identification and Cancer Classification Based on Microarray Data Using Laplace Naive Bayes Model with Mean Shrinkage", IEEE/ACM Transactions On Computational Biology And Bioinformatics, vol. 9, no. 6, November/December 2012

[18] Shu-Lin Wang, Yi-Hai Zhu, Wei Jia, and De-Shuang Huang,"Robust Classification Method of Tumor Subtype by Using Correlation Filters", IEEE/ACM Transactions On Computational Biology And Bioinformatics, vol. 9, no. 2, March/April 2012.

[19] Ji-xin LIU, Quan-sen SUN, "Mass spectrum data processing based on compressed sensing recognition and sparse difference recovery", 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).

[20] Isabelle guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik," Gene selection for cancer classification support vector machines", Machine learning, Springer, January 2002, Volume 46, Issue 1-3, pp 389-422.