International Journal of Latest Trends in Engineering and Technology
Special Issue IC3NS 2018, pp. 062-064
e-ISSN:2278-621X

ANALYSIS OF MACHINE LEARNING AND STATISTICS TOOL BOX (MATLAB R2016) OVER NOVEL BENCHMARK CERVICAL CANCER DATABASE.

Mehbob Ali[1], Abid Sarwar[2], Vinod Sharma[3]

Abstract:-Uterine Cervix Cancer is one of the leading Cancer names effecting the female population worldwide [1] [2]. Incidence of Cervical Cancer can be reduced by 80% through a routine Pap smear test. Pap smear test requires skilled cytologists and is always prone to inaccurate and inconsistent diagnosis due to manual error. Automated systems for easy recognition and proper staging of the cancerous cells can assists the medical professionals in correct diagnosis and planning of the proper treatment modality [3]. In this research 23 well-known machine learning algorithms available in MatlabR2016 are extensively analyzed for their classification potential of Pap smear cases. To Train and Test the algorithms a huge database is created containing 8091 cervical cell images pertaining to 200 clinical cases collected from three medical institutes of northern India. The raw cases of cervical cancer in form of Pap smear slides were photographed under a multi-headed digital microscope. After profiling the cells were vigilantly assigned classes by multiple cytotechnicians and histopathologists [4]. Cervical cases have seven classes of diagnosis [4].Quadratic SVM performed best among the 23 algorithms applied.

Keywords Machine learning, Neural networks, Cervical Cancer, Pap smear test.

1 INTRODUCTION.

1.1 Cervical Cancer.

Cervical cancer is the second most common form of cancers affecting the female population after breast cancer. This malignant cancer affects the cervix uteri or cervical area of the female reproductive organs by uncontrolled cell division and growth. Human papillomavirus (HPV) an icosahedral DNA virus, non-enveloped with a diameter of 52-55nm is the main agent for the pathogenesis of cervical cancer [5]. More than 120 types of HPV types are acknowledged today [6], among them only 15 are classified as high-risk types [7], 3 as probable-high-risk, and 12 as low-risk. The cells over the surface of the cervix affected by HPV shows precancerous developments called CIN which passes through various stages CIN1, CIN2, CIN3 and finally invasive cervical carcinoma (ICC). This progression takes over a period of two to three decades [8]. The most important part for any therapy is therefore to detect and wipe out local CIN3 lesions before it progresses to ICC [9]. According to WHO system the growth of CIN can be divided into three grades 1,2 and 3 and at least two-thirds of CIN1, half of the CIN2 and one third of CIN3 has the chance to regress back to normal[9]. A new system called Bethesda system categorizes cervical epithelial precursor lesions into two classes: the Low-grade Squamous Intraepithelial Lesion (LSIL) and High-grade Squamous Intraepithelial Lesion (HSIL). The LSIL corresponds to CIN1, while the HSIL includes CIN2 and CIN3 [10].

1.2 Machine Learning.

Machine Learning a branch of Artificial intelligence produces computer programs that learns from data samples without being explicitly programmed thus it relates learning from data to common concept of inference[11][12][13] . In biomedical field machine learning with its different techniques and algorithms has proven its ability of reaching to an acceptable generalization by searching through an n-dimensional space of complex bio-medical datasets [14]. Machine learning algorithms are trained by two methods 1) Supervised learning and 2) unsupervised learning. A Machine learning algorithm provide a data sample with less dimension produces better results as compared to data samples with large dimensionality[15]. Reducing dimension/Feature selection is done though methods called embedded, filter and wrapper approaches [15].The models in machine learning are usually trained to classify the data items into one of several predefined classes. A good classification model is rated on the basis of classification and generalization errors. Machine learning has a large no of algorithms able to learn the intricate relationships existing in complex multi-dimensional datasets e.g. ANN, KNN, SVM, Decision tress etc.

2. METHODS.

Decision treeare tree structured classifiers where an attribute is tested at internal nodes, each outgoing branch from an internal node represents one of the possible values of the test. Each test instance after tracing a particular path from the root node

[1] Department of Computer science and IT university of Jammu.
[2] Department of Computer science and IT university of Jammu.
[3] Department of Computer science and IT university of Jammu.

through the internal nodes based upon the test results, will halt at aleaf node holdingclass label for the test example. Decision trees are trained by ID3,C4.5 techniques and CART. SVM classifies instances of different classes by constructing set of hyperplanes in a high dimensional space. The hyperplane that largely separates (maximum margin hyperplane) classes is chosen for constructing classifier. KNN is a non-parametric and instance based method for classification. KNN assigns an instance to a class most common among its K nearest neighbors. Ensemble system of classification engages number of independent trained classifiers to propose the class label for a testing instance. Ensemble system produces much greater classification accuracy than independent classifiers. Artificial neural network (ANN) acts as a gold standard method in number of classification tasks and non-linear analysis of complex data [16] [17] [18]. ANN architecture consists of number of independent nodes/processing units arranged in input, hidden and output layers, connected by weighted connections called weights. The no of nodes in input layer corresponds to number of clinical variables in the data sample, nodes in hidden layer receives the weighted signals from the input nodes and calculates its output by passing the sum of weighted input values through an activation function. The output nodes then produce the output of the network by passing the sum of weighted signals received from the hidden nodes through activation function.

3. LITERATURE REVIEW.

[19] designed an automated cervical cell segmentation and classification system. The system using fuzzy c-means clustering technique (FCM) segmented each cervical cell into cytoplasm and nucleus regions. Five machine learning algorithms KNN, ANN, SVM, LDA and Bayesian classifier were implemented to classify the segmented cells in to their respective class of diagnosis. [20] Accessed the capability of artificial neural network to clearly distinguish malignant from benign breast cancer cases and also to predict the probability of breast cancer for individual patients. A large dataset consisting of 62,129 mammography findings are used to train a three layer feed forward network. [3] Proposed an innovative method ensemble of ensembles technique called hybrid ensemble method to increase the classification efficiency of AI based automated screening models. [21] Surveyed the applicability of recent machine learning techniques in cancer prognosis and prediction. A variety of machine learning techniques including ANN, SVM, Decision trees, Bayesian Networks have been widely used in the development of automated predictive models.

4. DATASETS FOR ANALYSIS.

In this research a huge database of cells of cervix obtained from slides of Pap smear test has been used. The database consist of about 8091 cell images pertaining to 200 clinical cases which had been reported in [4]. These cases are collected from three leading medical institutes of northern India. The database is designed according to the 2001 Bethesda system of Pap smear classification. Each of the 200 Pap smear slides were analyzed under NIKON microscope (Nikon Eclipse E400 DS-F12 microscope) attached with a digital camera and a computer to capture the image of the slide.

5. RESULTS AND DISCUSSION.

An accurate and precise automated diagnostic system for cervical cancer requires the correct classification of the Pap smear images to their respective classes of diagnosis [3]. In this research the screening potential of 23 machine learning algorithms had been extensively tested over a database of 8091 Pap smear images, against four performance metrics classification accuracy, Precision, Sensitivity and F-measure. The classification results of all classifiers 10 fold cross validation are tabulated in table 4. Quadratic SVM with a classification accuracy of 78.25% and F-value 0.69490 was the best classifier. The digital database developed along with potential machine learning algorithms especially quadratic SVM can play pivotal role in designing automated cervical cancer detection tool for efficient and timely detection of cancer.

S.no	Machine Learning Algorithm		Classification Accuracy	Precision	Sensitivity	F-Value
1	Decision Trees	Complex Tree	73.06%	0.66454	0.63366	0.648733
2		Medium Tree	72.84%	0.62366	0.57787	0.599892
3		Simple Tree	70.50%	0.60213	0.47229	0.529365
4	Support Vector Machines	Linear SVM	77.40%	0.69848	0.62257	0.658344
5		Quadratic SVM	78.25%	0.72323	0.66871	0.694902
6		Cubic SVM	74.78%	0.70764	0.68876	0.698072
7		Fine Gaussian SVM	60.82%	0.74568	0.42368	0.540346
8		Medium Gaussian SVM	78.02	0.78562	0.65447	0.714073
9		Coarse Gaussian SVM	76.01%	0.71833	0.55492	0.626139
10	Nearest Neighbor Classifier	Fine KNN	67.63%	0.64815	0.62745	0.637632
11		Medium KNN	71.84%	0.71262	0.57511	0.636523
12		Coarse KNN	72.22%	0.6766	0.5402	0.600755
13		Cosine KNN	69.71%	0.67136	0.56846	0.61564
14		Cubic KNN	69.61%	0.72095	0.54871	0.623147

15		Weighted KNN	72.63%	0.72047	0.66315	0.690623
16	Ensemble Classifiers	Boosted Trees	75.55%	0.68472	0.62401	0.652957
17		Bagged Trees	78.14%	0.78213	0.72147	0.750576
18		Sub Space Discriminant	74.85%	0.68504	0.61716	0.649331
19		Sub Space KNN	70.66%	0.65356	0.60415	0.627884
20		RU Boosted Trees	73.11%	0.61624	0.73012	0.668364
21	Discriminant Analysis	Linear Discriminant	65.80%	0.52151	0.5753	0.547086
22		Quadratic Discriminant	67.15%	0.50804	0.60565	0.552567
23	Monolithic Neural Networks	Feed forward network with 'trainlm'	76.40%	0.69422	0.62401	0.657246
		Feed forward network with 'trainscg'	75.20%	0.67566	0.612322	0.642434
		Feed forward network with 'trainbr'	76.40%	0.69422	0.62401	0.62445

Table 1. Classification results of all the 23 machine learning algorithms over 10 cross validations for the Novel benchmark database.

6. REFERENCES.

[1] Parkin DM, Bray FI, Devesa SS (2001) Cancer burden in the year 2000: the global picture. Eur J Cancer 37:S4–S66

[2] Goldie SJ, Kuhn L, Denny L, Pollack A, Wright T (2001) Policy analysis of cervical cancer screening strategies in low-resource setting: clinical benefits and cost effectiveness. J Am Med Assoc 285:3107–3115]

[3] Sarwar A, Sharma V, Gupta R (2015) Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis. Personalized Medicine Universe , Elsevier,4:54–62. doi:10.1016/j.pmu.2014.10.001.

[4] Abid Sarwar, Jyotsna Suri, Mehbob Ali, and Vinod Sharma, "Novel Benchmark database of digitized and calibrated cervical cells for artificial intelligence based screening of cervical cancer", Journal of Ambient intelligence and Humanized computing, Springer Verlag-Berlin Heidelberg 2016, DOI 10.1007/s12652-016-0353-8

[5] X. Castellsagué, S. de Sanjosé, T. Aguado, K.S. Louie, L. Bruni, J. Muñoz, M. Diaz, K. Irwin, M. Gacic, O. Beauvais, G. Albero, E. Ferrer, S. Byrne, F.X. Bosch, "HPV and Cervical Cancer in the World 2007 Report" ,Vaccine, Elsevier

[6] Chaturvedi Anil, Gillison Maura L. Human Papillomavirus and head and neckcancer. Epidemiology, pathogenesis, and prevention of head and neck cancer.2010. Pp. 87-116.

[7] Munoz Nubia, Xavier Bosch F, de Sanjose Silvia, Herrero Rolando, Castellsague Xavier, Shah Keerti V, et al. Epidemiologic classification of human Papillomavirus types associated with cervical Cancer. N Engl J Med 2003;348: 518e27. February 6, 2003.

[8] Cronjé HS. Screening for cervical cancer in the developing world. Best Practice and Research: Clinical Obstetrics and Gynaecology. 2005;19(4):517–529.

[9] Delgado G, Bundy B, Zaino R, Sevin BU, Creasman WT, Major F (1990) Prospective surgical—pathological study of disease-free interval in patients with stage Ib squamous cell carcinoma of the cervix: a gynecologic oncology group study. Gynecol Oncol 38:352–357.

[10] Frankel K, Sidawy MK. Formal proposal to combine the papanicolaou numerical system with Bethesda terminology for reporting cervical/vaginal cytologic diagnoses. Diagnostic Cytopathology. 1994;10(4):395–396.

[11] C.M. Bishop Pattern recognition and machine learning Springer, New York (2006).

[12] The discipline of machine learning: Carnegie Mellon University Carnegie Mellon University, School of Computer Science, Machine Learning Department (2006)

[13] I.H. Witten, E. Frank Data mining: practical machine learning tools and techniques Morgan Kaufmann (2005).

[14] A Niknejad, D. Petrovic Introduction to computational intelligence techniques and areas of their applications in medicine Med Appl Artif Intell, 51 (2013)

[15] Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining; 2006.

[16] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited. Cancer 2010; 116: 3310–21.

[17] Baxt WG (1995) Application of artificial neural networks to clinical medicine. Lancet 346:1135–1138

[18] Lundin J (1998) Artificial neural networks in outcome prediction. Anns Chir Gynaecol 87:128–130.

[19] Thanatip chankong, Nipon Theera-umpon and Sansanee Auephanwiriyakul , "Automatic cervical cell segmentation and classification in pap smears", computer methods and programs in biomedicine,Elsevier, pp. 539-556 year 2013

[20] Turgay Ayer, Oguzhan Alagoz, Jagpreet Chhatwal, Jude W. Shavlik, Charles E. Kahn, and Elizabeth S. Burnside , "Breast Cancer Risk Estimation with Artificial Neural Networks Revisited: Discrimination and Calibration", Cancer, 116, p. 3310–21

[21] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadis, "Machine learning applications in cancer prognosis and prediction", Computational and structural biotechnology journal 13, 8-17.