I.J. Education and Management Engineering, **2020, 2, 28-37**
Published Online April 2020 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2020.02.04

Available online at ***http://www.mecs-press.net/ijem***

Evaluation of Data Mining Categorization Algorithms on Aspirates Nucleus Features for Breast Cancer Prediction and Detection

Gajendra Sharma

School of Engineering, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Kavre, Nepal

Received: 10 January 2020; Accepted: 06 March 2020; Published: 08 April 2020

Abstract

With the development of technology the use of Computer Aided Diagnosis has become a key for breast cancer diagnosis. It is important to increase the accuracy and effective of such systems. The concept of data mining can be applied on the data gathered through such systems for prediction and prevention of breast cancer. In this research, we have conducted the comparison between seven classification algorithms with the help of WEKA (The Waikato Environment for Knowledge Analysis) tool on the 569 instances (10 nucleus attributes) of data with two classes Malignant(M) and Benign (B) of breast cancer aspirate cells. Furthermore the influence of each attribute on prediction was evaluated. The accuracy of these algorithms was above 91% with the highest value of 94.02% for random forest and the predictive power of conclave points was highest whereas lowest was of Fractal Dimension.

Index Terms: Data Mining, Breast Cancer, Classification Techniques, Prediction, Diagnosis,WEKA

* Corresponding author.
E-mail address: gajendra.sharma@ku.edu.np

1. Introduction

Breast cancer has been one of the leading cancers for women as compared to other cancers worldwide; it is as well forecasted to be the leading cause of death over the next few decades [1]. According to the data published by World Health Organization (WHO) breast cancer accounted for 15% of all cancer deaths in 2018 and is estimated that it is impacting 2.1 million women each year [2]. Breast cancers are a Malignant (M) or Benign (B) tumor, inside breast, where in cells divide and grow without control [3]. Fine needle aspirations (FNAs) are one of the processes followed for the diagnosis of such cells and the diagnosis is highly dependent upon the skill and experience of the physician [4].

TABLE 1.TUMORS IN THE BREAST

Tumor Type	Characteristics
Benign (B)	Not cancer Are not harmful Don't spread to other parts of body Can be removed and don't grow
Malignant (M)	Cancer Spread to other body Can be removed but there is probability of growing back

Table 1 shows different types of tumors found in the breast.Though there is no effective way for its prevention, efficient diagnosis in early stages can be incredibly supportive. The development of Computer-aided detection or diagnosis (CAD) systems can play a key role in its screening and diagnosis [5].But it is necessary to make these systems more accurate and effective.In the recent years with the development of data mining and availability of data has been able to provide various methods and procedures for the early stage prediction and detection of the breast cancers cells. The classification of breast cancer data can be of great use for this task to predict the outcome of the disease and discovery of the genetic behavior of the tumors. There are many techniques for the prediction and classification of breast cancer pattern. So, this paper empirically compares the performance of classification techniques of data mining along with the detection of influence of several cancer nuclear cell factors affecting the predictive power of the algorithms.

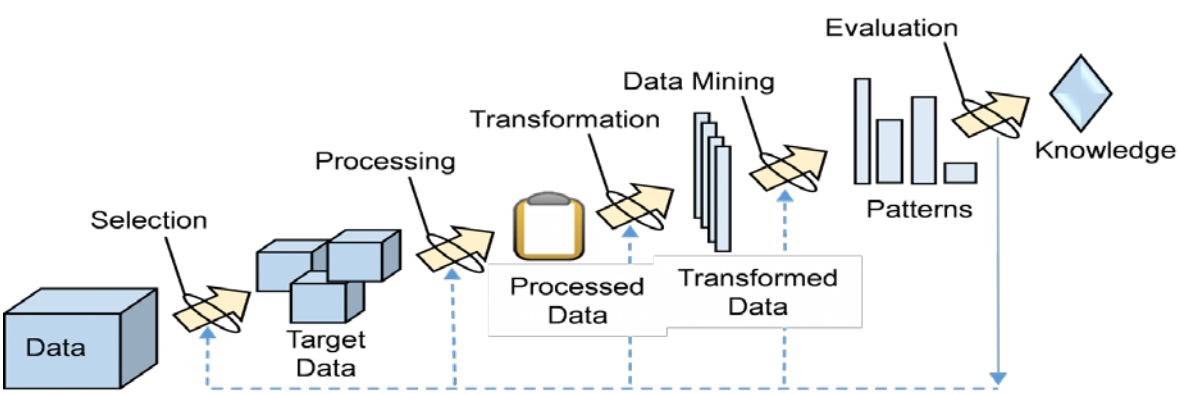

Fig. 1:Process of Knowledge Discovery in Databases [7]

Data Mining (DM) is the procedure, which deals with the extraction and analysis of data from large sets of data to explore hidden and unknown patterns [6]. It is a core step of the *Knowledge Discovery in Databases (KDD). KDD* is the process to extract knowledge from data in the context of large data sets. The KDD process is iterative and interactive process which starts with understanding the domain and ends with discovering the knowledge from the patterns generated using data mining methods. The discovered knowledge from KDD process is used for different purposes like: understanding data's behavior, assist users, improve and evaluates and systems/procedures.

The process of KDD is conducted through the following steps [8]:

1. Data Cleaning: Raw, noisy and inconsistent data is cleaned and irrelevant data are removed from the collection of the datasets.
2. Data Integration: Data from heterogeneous sources are combined in a coherent data store.
3. Data Selection: At this step relevant data required for the analysis is extracted from the database.
4. Data Transformation: In this phase selected data is transformed into the forms appropriate for mining.
5. Data Mining: Data patterns are extracted using different intelligent methods such as machine learning, artificial intelligence etc.
6. Knowledge representation: Here the discovered knowledge is visually represented and presented. In this step visualization techniques are used to help users understand and interpret the data mining results.

Huan Liu and Lei Yu [12] investigated active feature selection that promotes the idea to actively selecting instances for feature selection. Vanaja et al. [13] found that each feature selection methodology has advantages and disadvantages inclusion of larger attribute causes the reduction of accuracy. Dong-Sheng Cao's [14] evaluated a new decision tree based method combined with feature selection method backward elimination strategy with bagging to find the structure activity relationships in chemo metrics. Liu Ya-Qin's [15] discovered on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability. Medhat et al. [16] introduced decision tree with minimum error rate and maximum average gain.

Although a large number of researches have been carried out to identify the use of data mining in prediction and diagnosis of breast cancer a few of them have focused on comparative study of those techniques.

2. Algorithms Used

The following classification algorithms will be used in this study for effective analysis of results:

K-Nearest Neighbors(k-NN):k-NN, also known as a distance based classifier is based on analogy learning. It compares the given test datasets with training datasets that are similar to it[9]. It is also called lazy learning as it does not try to build a general internal model but simply stores instances of the training data.

Decision Tree (DT): A decision tree is a flow-charting like structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label [review2 5]. These are simple to understand and visualize.To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached.

Random Forest(RF): Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model. One of the most important advantage of using random forest is it deduce the problem of over-fitting.

Support Vector Machine(SVM):Also known as support vector networks these are supervised learning models that classify by finding the hyperplane that maximizes the margin between two classes. It is Effective in high dimensional spaces and uses a subset of training points in the decision function so it is also memory efficient.

Logistic Regression(LR):In this technique the probabilities relating the possible outcomes of a particular examination are modeled using a logistic function.

Naïve Bayes(NB): Based on Bayes theorem it is a classification technique which assumes that the presence of a particular feature in a class is unrelated to the presence of other feature.It is fast and requires small amount of training data.

Stochastic Gradient Descet(SGD):SGD is a simple and very efficient approach to fit linear models. It is particularly useful when the number of samples is very large. It supports different loss functions and penalties for classification.

3. Breast Cancer Dataset Summary and Nucleus Features

The breast cancer databases used in this researchis obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg, which are available on the UCI Machine Learning Repository. The data sets contain569 instances of data differentiated into2 classes M and B with 10 nucleus features. There are 212(37.26%) M class and 357(62.74%) B class distribution. Table I gives the brief description of attributes of the breast cancer dataset. All of these were generated from diagnosis based on fine-needle aspirates [4].

TABLE2. ATTRIBUTES OF BREAST CANCER DATASET

Name	Description
Radius	Individual nucleus radius.
Perimeter	Nuclear perimeter.
Area	Nuclear area.
Compactness	$Perimeter^2/Area$.
Smoothness	Difference between the length of a radial line and mean length of the lines surrounding.
Concavity	Severity of concavities/indentation in a cell nucleus.
Concave Points	Similar to Concavity but measures only the number, rather than the magnitude, of contour concavities
Symmetry	Length difference between lines perpendicular to the major axis to the cell boundary in both directions.
Fractal Dimension	Fractal dimension of a cell.
Texture	Nucleus texture.

4. Experiment Setup

To carry out the necessary experiment in this research we have used WEKA toolkit. WEKA is powerful open source software which contains supervised learning as well unsupervised learning methods. It contains Classification, Clustering, Association Mining, Feature Selection, Data Visualization, etc. which helps us to implement and compare DM techniques easily and effectively [1].The data set was divided into two sets one used for training sets and other for testing purpose. The training sets contained about 60% of the total data.10 iterations were donewith each classifier (K-Nearest Neighbors (IBK), Decision Tree (J48), Random Forest, Support Vector Machine (SMO), Logistic Regression, Naïve Bayes, Stochastic Gradient Descet). The results of the experiment are discussed in the next section.

5. Experimental Results and Discussion

5.1.Data Visualization

Fig. 2 provides the in glance visualization of the distribution data and different attributes.From the visualization we can see that lower the radius the cell has a higher probability to fall under M class so is the case with the concave points.

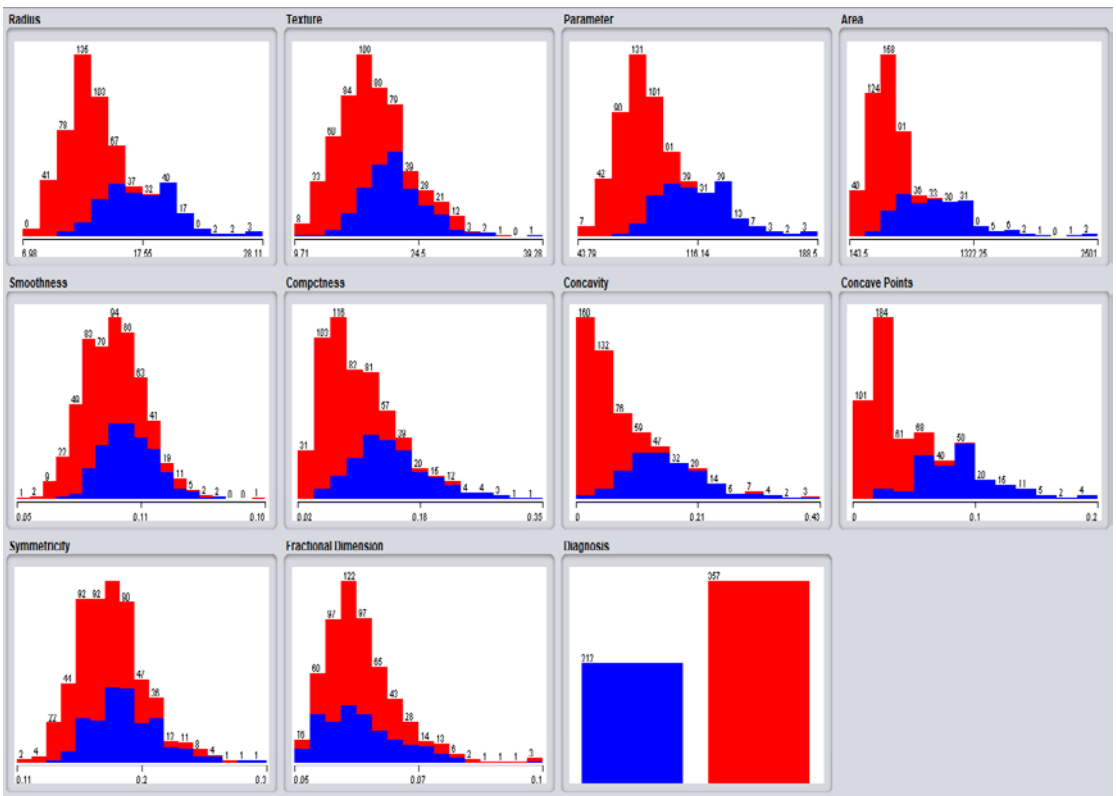

Fig. 2. Visual form of Breast Cancer Cell Data Using All Attributes (red: M and blue: B class).

5.2.Performance study of Algorithms

Table 3 consists performance value of different classification algorithms along with the Kappa statistic, mean absolute error ,root mean squared error and relative absolute error. As the table illustrates the lowest time taken is 0.01 seconds for three of the algorithms K-NN,DT,SVM and NB. Whereas the highest time taken was 0.10 seconds by RF. In case of accuracy the random forest stands first with 94.02 % and KNN and Naïve Bayes have the least accuracy percentage 91.52 and 91.61 respectively.

TABLE 3. PERFORMANCE OF THE CLASSIFIERS

Algorithms Implemented	Time Taken (Sec)	Correctly classified Instances (%)	Incorrectly classified instances (%)	Keppa Statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)
K-Nearest Neighbors	0.01	91.52	8.48	0.82	0.09	0.29	18.65
Decision Tree	0.01	92.36	7.64	0.84	0.09	0.26	19.15
Random Forest	0.10	94.02	5.98	0.87	0.10	0.21	20.56
Support Vector Machine	0.01	93.59	6.41	0.86	0.06	0.25	12.71
Logistic Regression	0.03	93.37	6.63	0.86	0.09	0.22	18.24
Naïve Bayes	0.01	91.61	8.39	0.82	0.09	0.27	18.24
Stochastic Gradient Descet	0.03	93.67	6.33	0.86	0.06	0.25	13.52

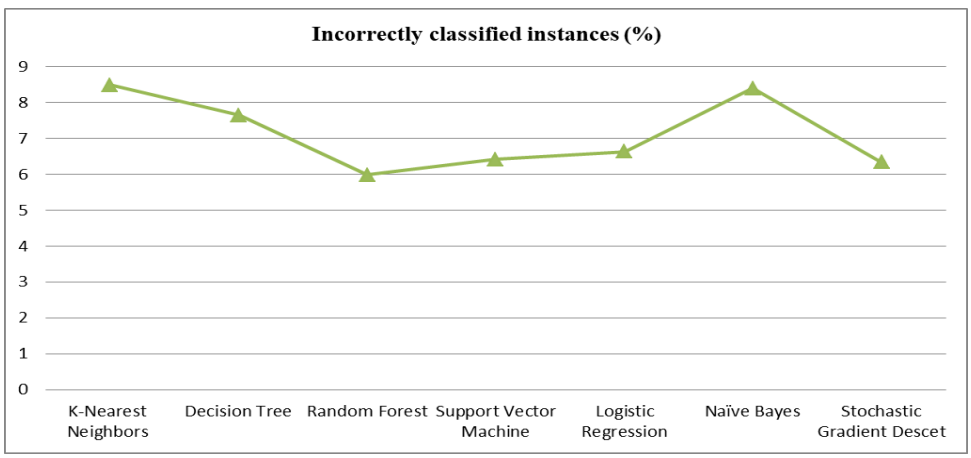

Fig. 3. Incorrectly Classified Instances Graph

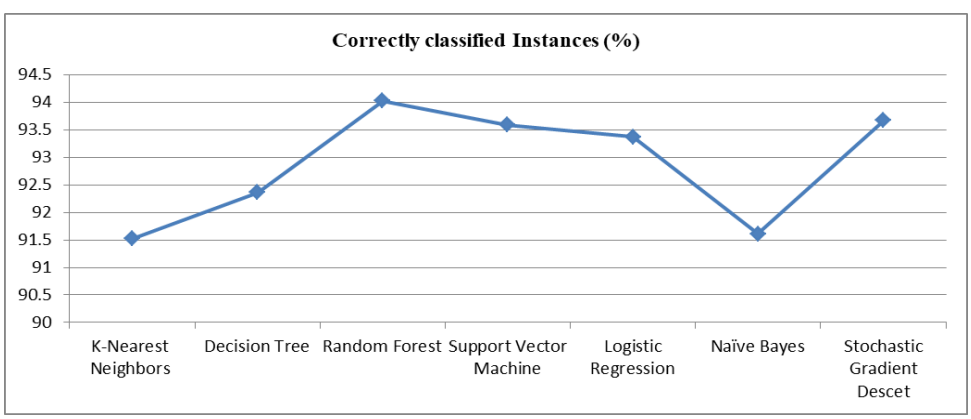

Figure 4. Correctly Classified Instances Graph

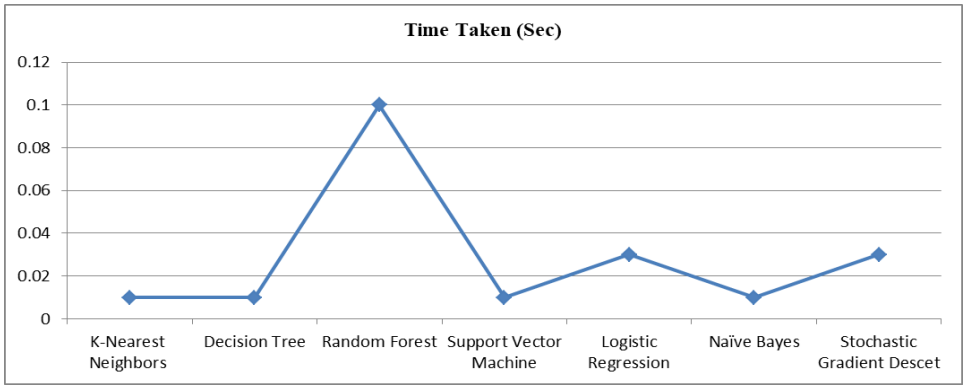

Fig. 5. Time Comparison of Classifiers

Whenever a prediction is made it can have four possible outcomes:
True positive (TP) = Predicting M as M.
False negative (FN) = Predicting M as B.
False positive (FP) = Predicting B as M.
True negative (TN) = Predicting B as B.

Table 4 shows the TP rate FP rate, precision recall value for different classifiers.

TABLE 4. COMPARISON ON ACCURACY MEASURES

Classifier	TP	FP	Precision	Recall	Class
K-Nearest Neighbors(IBK)	0.901	0.080	0.882	0.901	M
	0.920	0.009	0.933	0.920	B
Decision Tree(J48)	0.890	0.036	0.942	0.890	M
	0.964	0.110	0.930	0.964	B
Random Forest	0.912	0.036	0.943	0.912	M
	0.964	0.008	0.943	0.964	B
Support Vector Machine	0.857	0.029	0.857	0.902	M
	0.971	0.143	0.911	0.840	B
Logistic Regression	0.923	0.073	0.894	0.923	M
	0.927	0.077	0.948	0.927	B
Naïve Bayes	0.868	0.058	0.908	0.868	M
	0.942	0.132	0.915	0.942	B
Stochastic Gradient Descet	0.901	0.044	0.932	0.901	M
	0.934	0.077	0.934	0.934	B

TABLE 5. CONFUSION MATRIX

Classifier	M	B	Class
K-Nearest Neighbors	82	9	M
	11	126	B
Decision Tree	81	10	M
	5	132	B
Random Forest	83	8	M
	5	132	B
Support Vector Machine	78	13	M
	4	133	B
Logistic Regression	84	7	M
	10	127	B
Naïve Bayes	79	12	M
	8	129	B
Stochastic Gradient Descet	82	9	M
	6	131	B

Table 5 shows the confusion matrix which displays the frequency of correct and incorrect predictions made by the developed model.

TABLE 6.RESULT OF TEST AND AVERAGE RANK

Attribute	**Chi-Squared**	**Info Gain**	**Gain Ratio**	**Average Rank**
Radius	324.272	0.541	0.303	325.116
Perimeter	334.818	0.564	0.311	335.693
Area	327.45	0.543	0.317	328.31
Compactness	201.845	0.309	0.211	202.365
Smoothness	63.83	0.098	0.107	64.035
Concavity	319.379	0.517	0.33	320.226
Concave Points	368.155	0.636	0.332	369.123
Symmetry	65.83	0.095	0.071	65.996
Fractal Dimension	2.239	0.009	0.007	2.255
Texture	113.324	0.171	0.153	113.648

Table 6 shows the sensitivity (predictive power) of individual attributes. For this purpose three test were conducted (chi-squares, info gain and gain ratio). Average rank is calculated based on these values to determine the most and least indicative attribute. Figure 6illustrates the bar graph representation of the predictive power of each attributes from which we can see that the concave points have the highest predictive power followed by perimeter, area and radius. Whereas, the fractal dimension has the least predictive power.

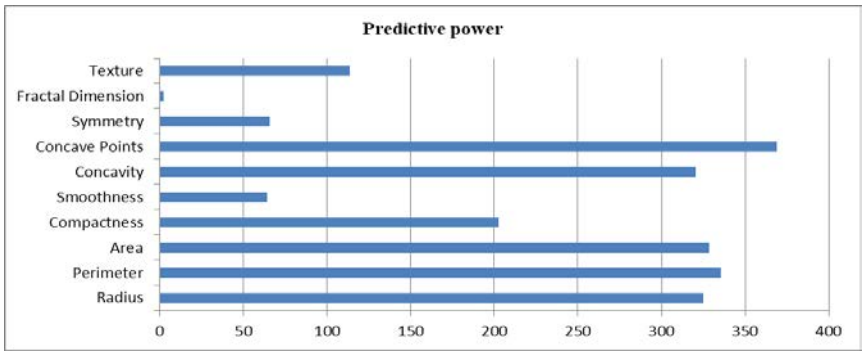

Fig. 6. Predictive power of Attributes

6. Conclusion

With the increase in availability of health data, different data mining classification techniques can be applied for the identification and prevention of breast cancer among the patients. This paper discussed about seven classification techniques and made a comparative analysis based on various parameters. Also, the attributes which are more indicative was identified. Our studies filtered the classifiers based on lowest computing time and accuracy and the results shows that the Random Forestis superior algorithm compared to other based on accuracy. Also it is found that the concave points attribute has the highest influence on the prediction where as Fractal Dimension has the least significance. Future study is needed for further elaboration of findings of this study.

References

[1] Jinshan Tang, R. Rangayyan, Jun Xu, I. El Naqa, Yongyi Yang (2019). "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances". IEEE Transactions on Information Technology in Biomedicine. 13(2), Pp. 236-251. Available online: 10.1109/titb.2008.2009441 [cited 2019 20 May].

[2] "Breast cancer" (2019). World Health Organization.Available online: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/. [cited 2019 20 May].

[3] Karthikeyani, V., I. Parvin, K. Tajudin, I. Shahina Begam. “Comparative of data mining classification algorithm in Diabetes disease prediction”. International journal of computer application.

[4] Street, W., W. Wolberg, O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis" (1993). Biomedical Image Processing and Biomedical Visualization,. Available: 10.1117/12.148698 [cited 2019 29 July].

[5] Sampat, M. P., M. K. Markey, A. C. Bovik (2005), “Computer-aided detectionand diagnosis in mammography,” in Handbook of Image and VideoProcessing, A.C. Bovik, Ed., 2nd ed. New York: Academic, Pp. 1195–1217.

[6] Maimon, O., L. Rokach (2019), "Decomposition Methodology for Knowledge Discovery and Data Mining", Data Mining and Knowledge Discovery Handbook, Pp. 981-1003. Available: 10.1007/0-387-25465-x_46.

[7] "Data Mining Occupant Behavior Research at LBNL BTUS" (2019), Behavior.lbl.gov, 2019. Available online: https://behavior.lbl.gov/?q=node/11. [cited: 2019 25 May].

[8] Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. Available online: webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf.
[9] Jiawei Han and Micheline Kamber (2012), "Data Mining Concepts and Techniques", third edition, Morgan Kaufmann Publishers an imprint of Elsevier.
[10] Wolberg, W.H., W.N. Street, O.L. Mangasarian (1994). "Machine learning techniques to diagnose breast cancer from fine-needle aspirates", Cancer Letters 77.Pp163-171.
[11] Chaurasia, Vikas & Pal, Saurabh. "A novel approach for breast cancer detection using data mining techniques".International Journal of Innovative Research in Computer and Communication Engineering. 3297. 2320-9801.
[12] Liu, Huan, and Lei Yu (2005). "Toward integrating feature selection algorithms for classification and clustering." IEEE Transactions on knowledge and data engineering. 17(4) Pp. 491-502.
[13] Vanaja, S., and K. Ramesh Kumar (2014). "Analysis of feature selection algorithms on classification: a survey." International Journal of Computer Applications. 96 (17).
[14] Cao, Dong-Sheng (2010). "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity." Chemometrics and Intelligent Laboratory Systems . 103 (2), Pp. 129-136.
[15] Ya-Qin, Liu, Wang Cheng, and Zhang Lu (2009). "Decision tree based predictive models for breast cancer survivability on imbalanced data." 2009 3rd International Conference on Bioinformatics and Biomedical Engineering.
[16] Abdelaal, Medhat Mohamed Ahmed (2010). "Using data mining for assessing diagnosis of breast cancer." Computer Science and Information Technology (IMCSIT). IEEE Proceedings of the 2010 International Multiconference.

Author's Profile

Gajendra Sharma completed doctoral degree in Information Systems Engineering from Harbin Institute of Technology, China. He received the degree of Masters of Engineering in Electronics and Communication in 1997 from Moscow Technical University of Telecommunication and Informatics, Russia. His research and teaching interest is focused on information systems, e-commerce (including e-business), strategic management of information technology (IT), IT adoption, design and evaluation of IT infrastructure, strategic management of IT as well as e-governance and ethics. He published research papers in some of the top-tier information systems and IT journals such as Information Systems Frontiers, Internet Research, Information Technology and People, Telecommunications Policy, International Journal of Web Based Communities and Electronic Commerce Research. He is a reviewer and technical editor of a number of peer review journals relating to information systems and IT. He worked in Liaoning Technical University, China at the department of Information Systems as an Associate Professor from 2011-2014. He completed postdoctoral research on technology philosophy (e-government and ethics) from Dalian University of Technology, China coordinating with Delft University of Technology, Netherlands. In the meantime, he has been working in Kathmandu University, Nepal.

How to cite this paper: Gajendra Sharma. "Evaluation of Data Mining Categorization Algorithms on Aspirates Nucleus Features for Breast Cancer Prediction and Detection", International Journal of Education and Management Engineering(IJEME), Vol.10, No.2, pp.28-37, 2020.DOI: 10.5815/ijeme.2020.02.04