Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Probabilistic Computational Model for Microarray Data Classification

Barnali Sahu, Ishara Priyadarsani
Department of CSE, Siksha 'O' Anusandhan University, Bhubaneswar,
Odisha, India

DOI: 10.23956/ijarcsse/SV7I5/0221

Abstract— For medical classification problems, it is often desirable to have a probability associated with each class to predict the disease. Various probabilistic computational models are currently used for classifying microarray data. A few classification models which are used for microarray classification are Naive Bayes, logistic regression and probabilistic neural network. Probabilistic classifiers have received relatively little attention in the literature of less number of sample sizes and a large number of gene sizes in microarray data and microarray data exhibit a high degree of noise. Most of the time probabilistic model does not adequately address the problem of dimensionality and noise and not giving the good accuracy. For achieving good accuracy feature selection techniques are used to reduce the high dimensional data and remove noisy data. This paper presents various probabilistic computational models for microarray data classification, and also reviews the state-of-the-art probabilistic computational model by grouping the literatures into three categories: Naïve Bayes, probabilistic neural network, logistic regression.

Keywords— Microarray data classification, Feature selection, Feature discretization, Probabilistic model.

I. INTRODUCTION

In machine learning, a probabilistic classifier is a classifier that is capable of predicting a given input sample and a probability distribution over a set of classes, rather than only outputting the most likely class that the sample should belong to. Probabilistic classifiers support classification with a degree of certainty, which can be very useful in its own right or when combining classifiers into ensembles. Binary probabilistic classifiers are also called binomial regression models in statistics. In econometrics, probabilistic classification is known as discrete choice. Also, some classification models, such as Naive Bayes, logistic regression and multilayer perceptrons (when trained to an appropriate loss function) are naturally probabilistic. As well as other models such as support vector machines are not, but their methods exist to turn them into probabilistic classifiers [11]. For improvement of gene expression microarray technology, there are many types of classifiers are developed for tumour sample classification. Classification rules are commonly accomplished using thousands of attributes with just only hundreds of samples. For medicinal decision making, the probabilistic classifier is very beneficial and also to know clearly which cases belong to which class. Medicinal decision making is very complicated. Misclassification costs are usually dissymmetric, tedious to compute as well as differ between physicians and patients. In this situation, probabilistic classifiers that provide an estimate of the probability of membership in each class for new cases are much useful than classification rules that simply allows cases to a class. Probabilistic classifiers give tools for a different audience of users who can also use the probabilities in alliance with other information such as treatment options and patient preferences for making the complicated combination of medical decisions. Based on the study of gene expression microarray data appropriate probabilistic classification is much more complicated than deterministic classification [24]. Over fitting and multi-collinearity are the most common problems that arise in high-dimensional data when applying statistical classification methods. These issues make statistical microarray classification methods very difficult [46]. Based on probabilistic model for microarray data there are few review paper. That is why we have chosen this topic for survey.

The roadmap of this survey is as follows: Section 2, discusses the overview of probabilistic classifier, Advantages and disadvantages of probabilistic modifier. Section 3, describes the microarray data, types of microarray data, their application and microarray data classification. In Section 4, Probabilistic computational models for microarray data and advantages and disadvantages of probabilistic model in dealing with microarray data are discussed. Section 5, includes Feature selection and literature detail of different probabilistic model. Section 6; contains discussion part of this survey paper. Section 7contains the summary.

II. OVERVIEW OF PROBABILISTIC CLASSIFIER

A. Naïve Bayes classifier

In machine learning, naive Bayes classifier is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features. Naive Bayes is a simple technique for creating classifiers models. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all Naive

Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable [30].

B. Probabilistic Neural Network (PNN):

A probabilistic neural network (PNN) is a feed forward neural network, which is broadly used in classification and also for pattern recognition problems [31]. It is a supervised neural network that is generally used in the area of pattern recognition, nonlinear mapping, and estimation of the probability of class membership and likelihood ratios [38]. In a PNN, the operations are organized into a multilayered feed forward network with four layers and they are:

1) Input layer
2) Hidden layer
3) Pattern layer/Summation layer
4) Output layer

C. Logistic regression

In statistics, logistic regression is also known as logit regression or logit model is a regression model where the dependent variable is categorical. Logistic regression only solves the binary dependent variable which can take only two values i.e. "0" and "1", outcomes are represented like pass/fail, win/lose, alive/dead or as healthy/sick. When the dependent variable has more than two outcome categories they can be analysed in multinomial logistic regression if the multiple categories are ordered in ordinal logistic regression [32]. In the language of economics, logistic regression is an example of a qualitative response or as a discrete choice model [33].

Logistics regression is discrimination tool like discriminant analysis, than prediction though called regression. In discriminant analysis consideration of normal distribution are strict than logistic regression. But discriminant analysis is more efficient than logistic regression. But empirically logistic regression is used widely than discriminant analysis because normality assumption is hard to follow every time for everything.

Table 1 Advantages And Disadvantages Of Different Computational Models

Model Name	**Advantages**	**Disadvantages**
Naïve bayes	• Very simple representation doesn't allow for rich hypotheses. • Affords fast, highly scalable model building as well as scoring. • It is used for both binary and multiclass classification problems.	• Makes a very strong assumption on the shape of your data distribution, i.e. any two features are independent given the output class. • Problem happens due to data scarcity • Problem arises for continuous features
Probabilistic neural network	• PNNs are much faster than multilayer perceptron networks. • PNNs can be more accurate than multilayer perceptron networks. • PNN networks are relatively insensitive to outliers. • PNN networks generate accurately predicted target probability scores. • PNNs approach Bayes optimal classification.	• PNN are slower than multilayer perceptron networks at classifying new cases. • PNN requires more memory space to store the model.
Logistic regression	• It does not require that the independents easily understood and also integrated with other domain knowledge. • More robust: the independent variables don't have to be normally distributed, or have equal variance in each group. • It can also handle nonlinear effects. • It does not require that the independents be an interval is unbounded.	• Logistic regression attempts to predict the outcomes based on a set of independent variables, but if researchers include the wrong independent variables, the model will have little to no predictive value. • It cannot predict continuous outcomes.

III. MICROARRAY DATA

A microarray database is an archive containing microarray gene expression data. The basic uses of a microarray database are to store the density data, handle searchable indicants, and also to make the data available to other applications for analysis and clarification. Microarray data present especially very useful to analyse the system in cancer diagnosis as well as produce one of the most effective tools to discover the genetic system makes the failure of cell cycle control. Recently, certain different methods to handle microarray data as a diagnostic tool in cancer classification also have been developed. These procedures take changes in the expression of particular genes into account but do not consider a break into certain gene interplay create by the tumour. It is probable that some genes participating in tumour

development do not change their expression level desperately [25]. The microarray data are images, which have to be converting into gene expression matrices in which rows represent genes and columns show various samples such as tissues or experiential surrounding and also numbers in each cell defines the expression level of a particular gene in the particular sample. Microarray-based disease classification system takes labelled gene expression data samples and also generates a classifier model that classifies new data samples into different predefined diseases. Microarray data classification is a supervised learning task that predicts the diagnostic category of a sample from its expression array phenotype [26].

A. Types of microarrays

Microarray experiments can be categorized in three ways:

1) Microarray Expression Analysis: In this experimental system, the cDNA borrowed from the mRNA of known genes is disabled. The sample has taken from both the normal as well as from the diseased tissues genes. If the gene is over reveal in the diseased condition then the spots with more anxiety are obtained for diseased tissue gene. This reveals pattern is then compared to the revealed pattern of a gene then we have to know that responsible for which disease.

2) Microarray for Mutation Analysis: For mutation analysis, the researchers use gDNA. The genes might vary from each other by as less as a single nucleotide base. A single base difference between two sequences is known as Single Nucleotide Polymorphism (SNP) and detecting them is known as SNP detection.

3) Comparative Genomic Hybridization: It is used for the recognition of the increase or decrease of the important chromosomal fragments harbouring genes involved in a disease.

B. Applications of Microarrays

1) Gene Discovery: DNA Microarray is a technology which helps to identify the new genes, and also to know about their functioning as well as expression levels under different conditions.

2) Disease Diagnosis: DNA Microarray technology helps researchers to learn more about particular diseases like heart diseases, mental illness, infectious disease and especially the study of cancer.

3) Drug Discovery: Microarray technology has extensive application in Pharmacogenomics. Pharmacogenomics is the study of correlations between therapeutic responses to the drugs and the genetic profiles of the patients.

4) Toxicological Research: Microarray technology provides a strong stage for the research of the brunt of the toxins on the cells and they are fleeting on to the progeny.

5) GEO: In the recent, microarray technology has been broadly used by the scientific community. There has been a lot of creation of data linked to gene expression. This data is being distributed which is not easily accessible for public use.

C. Microarray data classification:

Classifications of microarray data classification are usually presented in matrix. In figure 1 represent the structured representation of microarray data classification N rows represent the attributes of microarray data and M columns represent the samples of microarray data. After having the matrix we have to choose a classifier. The most common way in building classifiers using microarray data is to start with gene selection to select a subset of genes which is expected to contain the most relevant genes. Gene selection is usually applied over the sample-gene matrix directly and it tackles two main issues; discriminative genes and redundant genes. Discriminative genes are those genes whose profiles have strong statistical differences between different classes, so they are good genes to be used to differentiate between samples that belong to different classes. Redundant genes are those genes which have close profiles. Then we have to train the classifier. After training the classifier with the samples using the selected subset of genes the classifier needs to be tested and assigned a numeric performance value. The most common metric to measure classifiers performance is the classification accuracy, which is the percentage of the correctly classified test samples of the entire test sample set. Many methods have been introduced in the literature for classification and validation for the classifiers.

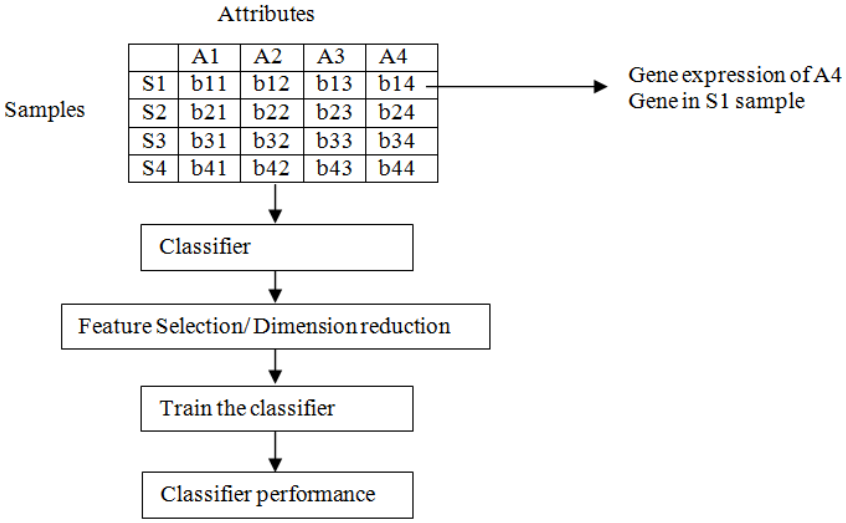


Fig. 1 Representation of N*M matrix for microarray data classification

IV. PROBABILISTIC COMPUTATIONAL MODEL FOR MICROARRAY DATA

Probabilistic classifiers have accepted nearly slight consideration in the literature of small n large p classification problems where the number of candidate attributes exceeds the number of cases available for the model development [27]. Medical decision making is very complicated in high dimensional data. Misclassification outlay is usually asymmetric, tough to compute and also differ with the doctor and the sufferer. In this situation, probabilistic classifiers that give an estimate of the probability of membership in each class for new cases are much more useful than classification rules that simply allow cases to a class [24]. Probabilistic classifiers also provide tools for a diverse audience of users who may also use the probabilities in conjunction with other information such as treatment options and patient preferences for making complex integrated clinical decisions [23]. The probabilistic classifiers can be easily extended to handle more than 2 classes at a time. Proper probabilistic classification is much more difficult than deterministic classification based on simulation studies and also analysis of gene expression microarray data. It is important to provide that a probabilistic classifier is well calibrated or at least not "anticonservative" using the methods generated over here. In figure 2 describe the stages of probabilistic process.

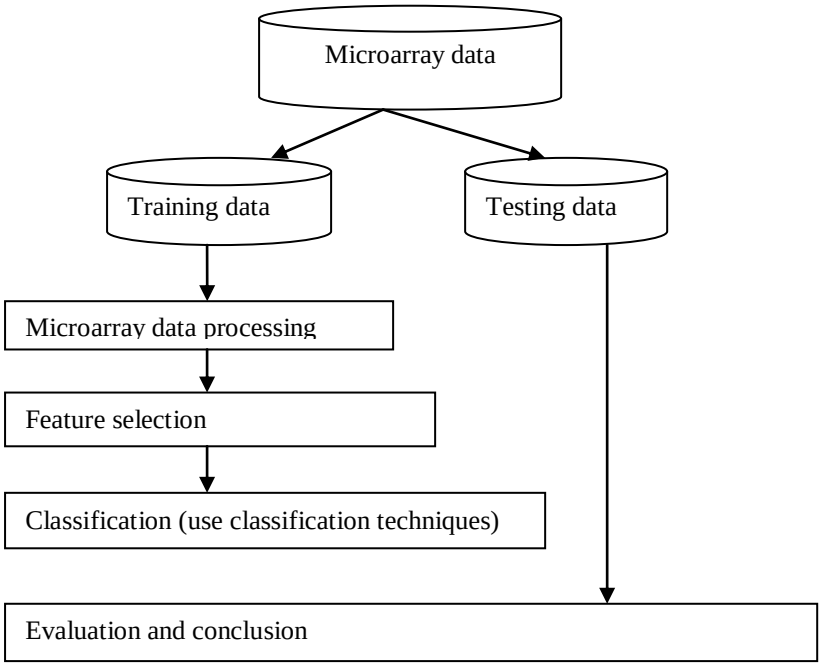


Fig. 2 Stages in probabilistic classifier process

Table 2 Advantages And Disadvantages Of Probabilistic Model In Dealing With Microarray Data

Advantages	**Disadvantages**
• Provide a principled way to deal with the missing data and also to learn simultaneously from other types of data. • Probabilistic classifiers provide tools for a diverse audience of users who may also use the probabilities in conjunction with other information such as treatment options and patient preferences for making complex integrated clinical decisions.	• Proper probabilistic classification is much more difficult than deterministic classification. • Probabilistic classifiers have received relatively little attention in the literature of small n large p classification problems. • The number of candidate variables exceeds the number of cases available for the model development.

V. FEATURE SELECTION

In the field of machine learning, pattern recognition, statistics, and in data mining feature selection is a technique is commonly used for dimensionality reduction. This technique goal to select a subset of relevant features from the original set of features according to some of the criteria [35]. Feature selection usually is used in the domains where the datasets comprise of thousands of features but with relatively small sample size (e.g., gene expression data). Feature selection that is applied to the gene expression data is also known as the gene selection [39]. Gene selection is necessary as the data usually contains many irrelevant, redundant, and noisy expressions, and it is also effective for early tumor detection and cancer discovery as it leads to a more reliable cancer diagnosis or prognosis and a better clinical treatment [40]. The feature relevance score is calculated, and low scoring features are removed. The top ranked genes are used to build the classifier [42].

In Table 3 summarizes the literature detail for three probabilistic model using different feature selection method and validation techniques are given. The majority of researchers focus on the microarray data classification using feature selection. Many feature selection algorithms are designed for reducing high dimensional data and the validation technique giving high accuracy.

Table 3 Literature Details For Naïve Bayes Classifier

Model name	**Feature selection technique**	**Microarray data**	**Validation technique**	**Accuracy**
Naive Bayes classifier [1]	• Class-conditional Independent component analysis (ICA) • Stepwise regression	• Leukemia-ALLAML [13] • Leukemia-MLL [14] • Colon Tumor [15] • Lung Cancer I [16] • Lung Cancer II [17]	• Leave-one-out • Hold-out	95.8 83.3 82.3 82.3 98.3
Naive Bayes classifier [41]	• Information gain (IG) • Relief Algorithm (RA) • T-statistics (TA) • Feature extraction method based on PCA	• Lung_Cancer [32] • Prostate_Cancer [33] • Breast_Cancer [34] • Lukemia [13] • SRBCT [18] • Brain_Tumor [35] • Colon_Cancer [36] • Ovarian_Cancer [37] • 14_Tumors [38]	• 4-fold cross validation	77.0 88.9 79.1 80.4 88.2 83.4 90.0 92.1 82.6
Naive Bayes classifier [2]	• Independent component analysis (ICA) • Class-conditional • Independent component analysis (CC-ICA) • Partition-conditional (PC-ICA)	• Leukemia-MLL [14] • Lung Cancer I [16]	Hold-out	-
Naive Bayes classifier [3]	• Principal component analysis (PCA)	• 12 Synthetic datasets (other) • Real world datasets (other)	10 fold cross validation	-
Naive Bayes classifier [4]	• Wrapper feature subset selection • Incremental algorithms	• Arcene • Madelon • Dorothea (other) • Dexter (other) • Gisette (other) • warpPIE10P (other) • warpAR10P (other) • pixraw10P (other) • orlraws10P (other) • TOX-171 SMK-CAN-187 • GLI-85 • GLA-BRA-180 • CLL-SUB-111 • pcmac • basehock	10 fold cross validation	72 60.5 92.9 83 94.1 91.4 70.8 95 90 71.9 71.1 78.8 67.8 67.6 80.3 89.5

Table 4 Literature Details For Probabilistic Neural Network

Model name	**Feature selection technique**	**Microarray data**	**Validation technique**	**accuracy**
Probabilistic Neural Network (PNN) [5]	filter methods and wrapper methods	• Leukemia [18] • Embryonal CNS tumor [19] • Medulloblastoma Morphology [20]	10-Fold cross-validation	95.4 86.1 89.8

		• Medulloblastoma treatment outcome [20]		69.2
Probabilistic Neural Network (PNN) [6]	Principal component analysis (PCA)	SRBCT data set	n cross-validation	75.3
Probabilistic Neural Network (PNN)		• hepatic cancer diagnosis	10 fold- cross validation	86.56
Probabilistic Neural Network (PNN) [8]	filter and wrapper	• Oral cancer related data ENT (Ear Nose-Throat) and Head-Neck [37]	leave-one-out	73.76
Probabilistic Neural Network (PNN) [48]	Principal component analysis (PCA)	• Colon Tumor • DLBCL • Leukemia		95.83 96.67 95.83

Table 5 Literature Details For Logistic Regression

Model name	**Feature selection technique**	**Microarray data**	**Validation technique**	**Accuracy**
Logistic regression [9]	Recursive feature elimination (RFE)	• Breast Cancer • Central Nervous System • Colon Tumor • Acute Leukemia • Lung Cancer • Ovarian Cancer • Prostate Cancer	leave-one-out cross-validation	99.8 99.8 99.3 100 100 100 96.3
Logistic regression [10]	-	• leukemia ALL [22] • leukemia AML [22]	cross-validation	-
Logistic regression [11]	Combinatorial feature selection	• Thrombin data set [23] • Leukemia data set [18]	-	-
Logistic regression [12]	• Univariate ranking • Recursive feature elimination	• Leukemia data [18] • SRBCT data [19] • Ramaswamy data [20]	• 8-fold CV • 10-fold CV	-
Logistic regression [43]	• principal components analysis	• Breast data [44] • Colon data set [15] • Leukemia data set [18]	• Cross validation	-
Logistic regression[45]	-	• Colon [15] • Prostate [46] • DLBCL [47]	• Cross validation	95 95 95

In Table 3,4,5 summarizes the work on naïve Bayes classifier, probabilistic neural network and logistic regression for microarray data classification using different feature selection methods and different validation techniques. In the above literature review we focus on accuracy. Same probabilistic model using different validation techniques give different accuracy for same data.

A. Dataset Analysis

Based on Tables 3, 4, and 5, the seven most commonly used gene microarray expression datasets in the literatures are Lung cancer, Colon, Leukemia, Prostate_Cancer, Breast Cancer, SRBCT, Ovarian Cancer, The comparison accuracy based on probabilistic models are shown in Table 6. The comparison results indicate that same dataset using different probabilistic model giving different accuracy and finding which model is giving better accuracy for same dataset.

VI. DISCUSSION

This paper provides a review on the probabilistic computational model for microarray data classification. It also discusses the about the feature selection and validation method. Probabilistic classification is important for medical decision making. Probabilistic classifiers have accepted small n large p classification problems where the number of candidate attributes exceeds the number of cases available for model development. To overcome this problem in each stage of feature selection are important to reduce the dimensionality. This paper implies that probabilistic model dealing with microarray data. Many probabilistic classifier methods are used for microarray data classification by researchers. A plenitude of a probabilistic model for microarray data classification approaches have been designed by researchers, yet this paper implies that there are still many open opportunities for further improvement. Most studies treat the highest

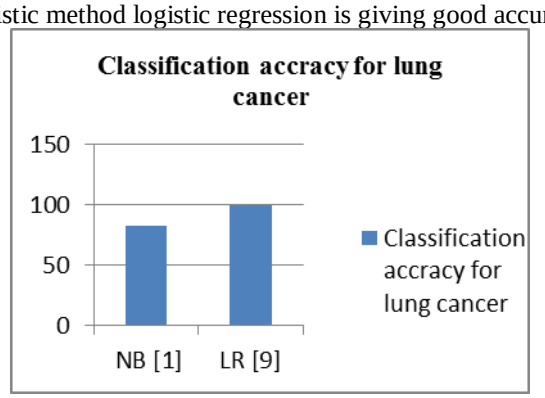
classification accuracy as the ultimate goal. So, more research efforts in evaluation and validation of feature selection, like measurement of specificity, sensitivity, similarity, and stability.

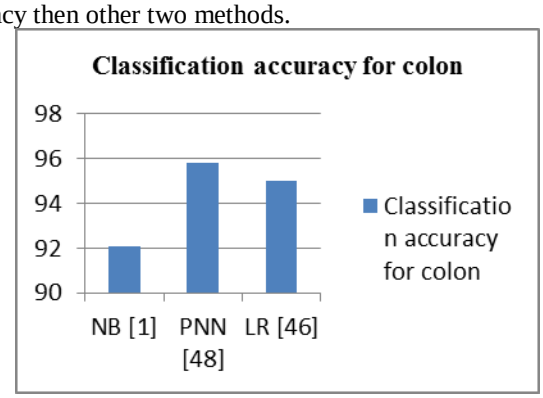
Table 6 Classification Accuracy For Different Dataset Using Different Probabilistic Models

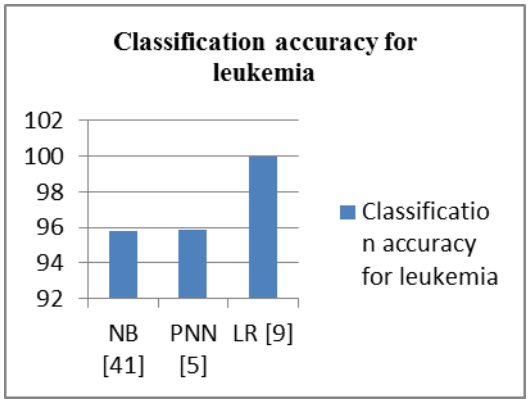
Dataset	**Classifier**	**Accuracy**
Lung cancer	Naïve Bayes [1]	82.3
	Logistic Regression [9]	100
Colon	Naïve Bayes [1]	92.1
	Probabilistic neural network [48]	95.83
	Logistic Regression [45]	95
Leukemia	Naïve Bayes [41]	95.8
	Probabilistic neural network [5]	95.83
	Logistic Regression [9]	100
Prostate_Cancer	Naïve Bayes [41]	88.9
	Logistic Regression [9]	96.3
Breast Cancer	Naïve Bayes [41]	79.1
	Logistic Regression [9]	99.8
sSRBCT	Naïve Bayes [41]	88.2
	Probabilistic neural network [6]	75.3
Ovarian Cancer	Naïve Bayes [41]	82.6
	Logistic Regression [9]	100

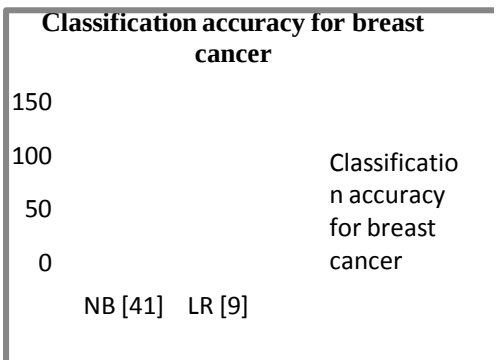
In Table 6 shows the accuracy of same microarray data using different probabilistic model. In this literature we are using three probabilistic methods. Different method gives different accuracy for same data set. For example classification accuracy of colon using different probabilistic method, naïve is giving 92.1% accuracy using feature extraction method based on principal component analysis and 4-fold cross validation method [41]. Probabilistic neural network give 95.83% accuracy using principal component analysis [48]. Logistic regression gives 95% accuracy using using cross validation techniques [45]. Here Probabilistic neural network is giving better accuracy then other probabilistic model and in some other data other model give better accuracy So, the researchers are still doing research on this particular topic that which method is good for all these microarray data. In Fig 3 shows the graph representation for different datasets accuracy using different probabilistic methods.

In general, we observe that many researchers put huge and fruitful efforts in classification of microarray data using different probabilistic models. In Table 6 the comparison results on a few common microarray data using different probabilistic method logistic regression is giving good accuracy then other two methods.









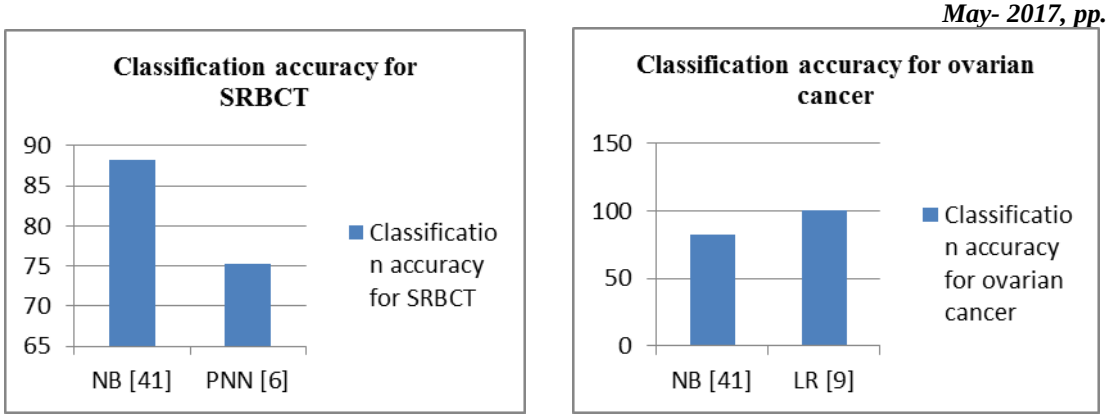


Fig. 3 Classification accuracy for different microarray data set

VII. SUMMARY

This survey presented the different approaches of the probabilistic classifier to solve the problem of microarray data classification. The literature surveyed covers the different probabilistic computational model are used for microarray data classification. This paper does not aim in any way to provide numerical evaluations of probabilistic classifier in terms of which one is the best, but to gather as much as possible domain knowledge about this particular topic. This includes naïve Bayes classifier; Probabilistic Neural Network (PNN) and Logistic regression are used for classification. Several feature selection methods are used in this paper for reducing the high dimensional data and different validation techniques are also used. These methods and algorithm have given a good accuracy results for high dimensional data.

REFERENCES

[1] Fan L., Poh K. L., Zhou P., "A sequential feature extraction approach for naïve bayes classification of microarray data", Expert Systems with Applications, vol. 36, Issue 6, pp. 9919–9923, 2009.

[2] Fan L., Poh K. L., Zhou P., "Partition-conditional ICA for Bayesian classification of microarray data", Expert Systems with Applications, vol. 37, pp. 8188–8192, 2010.

[3] Zhang M. L., José M. Peña, Victor Robles, "Feature selection for multi-label naive Bayes classification", Information Sciences, vol. 179, pp. 3218–3229, 2009.

[4] Bermejo P., Gámez J. A., Puerta J.M., "Speeding up incremental wrapper feature subset selection with Naïve Bayes classifier", Knowledge-Based Systems, vol. 55, pp. 140–147, 2014.

[5] Huang C.J., Liao W. C., "Application of Probabilistic Neural Networks to the Class Prediction of Leukemia and Embryonal Tumor of Central Nervous System", Neural Processing Letters, vol. 19, pp.211–226, 2004.

[6] Xuand R., Wunsch D.C., "Probabilistic Neural Networks for Multi-class Tissue Discrimination with Gene Expression Data", pp. 1696-1701, 2001.

[7] Gorunescu F., Gorunescu M., El-Darzi E., Gorunescu S., "An Evolutionary Computational Approach to Probabilistic Neural Network with Application to Hepatic Cancer Diagnosis", ISSN: 1063-7125, 2005.

[8] Sharma N., Om H., "Usage of Probabilistic and General Regression Neural Network for Early Detection and Prevention of Oral Cancer", The Scientific World Journal, pp. 1-11, 2015.

[9] Shen L., Tan E. C., "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, vol.2, Issue 2, pp. 166-175, 2005.

[10] LiaoJ. G., Chin K. V., "Logistic regression for disease classification using microarray data: model selection in a large p and small n case", BIOINFORMATICS, vol. 23, Issue 15, pp. 1945–1951, 2007.

[11] Bertola P., Felicia G., Festab P., Lanciac G., "Logic classification and feature selection for biomedical data", Computers and Mathematics with Applications, vol. 55, pp. 889–899, 2008.

[12] JIZHU, "Classification of gene microarrays by penalized logistic regression", Biostatistics, vol.5, Issue 3, pp. 427–443, 2004.

[13] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol.286, pp. 531–537, 1999.

[14] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia", Nature Genetics, vol.30, pp.41–47, 2002.

[15] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., " Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proceedings of the National Academy Sciences of the United States of America, vol.96, pp.6745–6750, 1999.

[16] Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., " Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses", Proceedings of the National Academy Sciences of the United States of America, vol.98, pp.13790–13795, 2001.

[17] Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., et al. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research, 62, 4963–4967.

[18] GolubT. R., Slonim D. K., TamayoP. T., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M. L., Downing J R., Caligiuri M. A., Bloomeld C. D., Lander E. S., "Molecular classification of cancer Class discovery and class prediction by gene expression monitoring",, Science 286 (1999), pp. 531–537, 2002.

[19] KHAN J., WEI, J., RINGNER M., SAAL L., LADANYI M., WESTERMANN F., BERTHOLD, F., SCHWAB M., ANTONESCU C., PETERSON C., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine , vol. 7, pp. 673–679, 2001.

[20] HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). The Elements of Statistical Learning. Berlin: Springer.

[21] Chu S., DeRisi J., Eisen M., Mulholland J., Boslein D., Brown P. O., Herskowitz I., "The transcriptional program of sporulation in budding yeast,"Science, vo1.282, pp. 699-705,October 1998.

[22] Golub,T., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, vol. 286, pp.531–536, 1999.

[23] https://en.wikipedia.org/wiki/Probabilistic_classification.

[24] Kim K. I., Simon R., "Probabilistic classifiers with high-dimensional data", Biostatistics, vol. 12, Issue 3, pp.399-412, 2011.

[25] https://en.wikipedia.org/wiki/Microarray_databases.

[26] Lavanya C., Nandihini M., Niranjana R., Gunavathi C., "Classification of Microarray Data Based On Feature Selection Method", International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, Issue 1, ISSN. 2319 – 8753, 2014.

[27] Kim K. I., Simon R., "Probabilistic classifiers with high-dimensional data", Biostatistics Advance, vol. 17, pp. 1–14, 2010.

[28] https://www.quora.com/What-are-the-disadvantages-of-using-a-naive-bayes-for-classification.

[29] https://en.wikipedia.org/wiki/Probabilistic_neural_network#Advantages.

[30] https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[31] https://en.wikipedia.org/wiki/Probabilistic_neural_network.

[32] https://en.wikipedia.org/wiki/Logistic_regression.

[33] Walker SH., Duncan., "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54: 167–178, 1967.

[34] Ang J. C., Mirzal A., Haron H., Hamed H. N. A., "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, vol.13, Issue 5, pp. 971-988, 2016.

[35] Sheikhpoura R., Sarrama M. A., Gharaghanib S., Chahookia M. A. Z., "A Survey on semi-supervised feature selection methods", Pattern Recognition, vol. 64, pp. 141–158, 2017.

[36] Nishanth K. J., Ravi V., "Probabilistic neural network based categorical data imputation", vol. 218, pp. 17–25, 2016.

[37] Department of three Tertiary Care Hospitals of Pune, Maharashtra, India.

[38] Georgiou V. L., Pavlidis N. G., Parsopoulos K. E., Ph.D. Alevizos, Vrahatis M. N., "Optimizing the Performance of Probabilistic Neural Networks in a Bioinformatics Task", eunite, pp. 34-40, 2004.

[39] Yang K., Cai Z., Li J., Lin G., "A stable gene selection in microarray data analysis," BMC Bioinf, vol. 7, Issue 1, p. 228, Apr. 2006.

[40] Boulesteix A. L., Strobl C., Augustin T., Daumer M., "Evaluating microarray-based classifiers: An Overview," Cancer Informat., vol. 6, pp. 77–97, Feb. 2008.

[41] Osareh A., Shadgar B., "Microarray data analysis for cancer classification", Health Informatics and Bioinformatics (HIBIT), DOI: 10.1109/HIBIT.2010.5478893, 2010.

[42] Chitode K., Nagori M., "A Comparative Study of Microarray Data Analysis for Cancer Classification", International Journal of Computer Applications, vol. 81, Issue 15, pp. 14-18, 2013.

[43] Bielza C., Robles V., Larrañaga P., "Regularized logistic regression without a penalty term: An application to cancer classification with microarray data", Expert Systems with Applications, vol.38, Issue 5, pp. 5110–5118, 2011.

[44] West M., Blanchette C., Dressman H., Huang E., Ishida S., Spang R., "Predicting the clinical status of human breast cancer by using gene expression profiles", Proceedings of the National Academy of Sciences USA, vol.98, Issue 20, pp.11462–11467, 2001.

[45] Algamal Z. Y., Lee M. H., "Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification", Expert Systems With Applications, vol. 42, Issue 23, pp. 9326–9332, 2015.

[46] Shipp M. A., Ross K. N., Tamayo P., Weng A. P., Kutok J. L., Aguiar R. C. T., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning", Nature Medicine, vol. 8, pp. 68–74, 2002.

[47] Singh D., Febbo P. G., Ross K., Jackson D. G., Manola J., Ladd C., "Gene expression correlates of clinical prostate cancer behavior. Cancer Cell", vol.1, pp.203–209, 2002.

[48] Sreepada R. S., Vipsita S., Mohapatra P., "An efficient approach for classification of gene expression microarray data", Fourth International Conference of Emerging Applications of Information Technology, pp. 344 - 348, DOI: 10.1109/EAIT.2014.46, 2014.

