Machine Learning Algorithms with Different Gene Expression Datasets

M Ramachandro, Ravi Bhramaramba

Abstract*: Two classification techniques have been compared using microarray dataset. For example SVM, Logistic Regression strategies have been utilized in the process. The usefulness of these techniques has been determined with precision, accuracy, recall and F1-Scores. These techniques Here Prostate tumors, Lung Cancer, were analyzed. For each situation these strategies were applied to two distinctive microarray datasets with two classes. Finally performance analysis was done

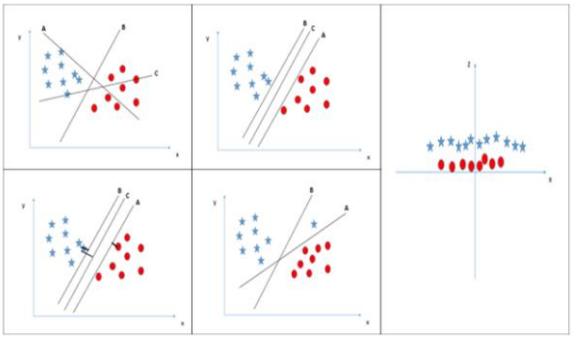
***Keywords* : About four key words or phrases in alphabetical order, separated by commas.**

I. INTRODUCTION

1.1 Support Vector Machine

"Support Vector Machine" (SVM) is an arranged AI technique which can be used for both portrayal and fall away from the confidence burdens. Incidentally, it is usually used in get-together issues. In this figure, we plot each data as a point in n-dimensional space with the estimation of each area being the estimation of a specific sort. We find the hyper-plane that perfectly separates the two classes.

Working of Support Vector Machine:



Recognize the benefit hyper-plane (Scenario-1): Here we have three (A, B and C) hyper-plans. See the hyper-plane advantage to demand star and circle ahead of time.

- "Choose the hyper plane that is best for the two groups." In this condition, this improvement has been played out amazingly by hyper-plane "B."
- Identify the hyper-plane leeway bit (Scenario-2): here we have three hyper-plans (A, B and C) and all of them are well-secluded. Obviously, we have the decision to see the bit of leeway hyper-plane
- Here, increase bundles between the nearest data point and the hyper plane enables us to choose the hyper plane leeway. It is called Margin this piece.

Over the top, the hyper-plane C edge is high when presented in contrast with A and B. Then, we call the hyper plane leeway bit C. Another clearing behind the selection of the high edge hyper plane is quality. Reliability. If we select a low edge hyper plane, the likelihood of miss request remains high.

Distinguish the privilege hyper-plane (Scenario-3) Hint: To view the advantage of the hyper-plan, use the standards as assessed in the past.

Some of you might have chosen the hyper plane B, which turns from A to the top. Nevertheless, the trick is here and the SVM chooses the hyper plane which correctly asks for the classes before the edge is increased. Here, hyper-plane B has a requesting mess up and A has arranged all totally. Thusly, the bit of leeway hyper-plane is A.

SVM has an element to eliminate irregularities and find the most significant edge of the hyper plane. We may assume that SVM has been effective in exceptional cases from this point forward. This problem can be addressed by SVM. It handles this issue effectively by displaying additional components.

Another portion of z = x^2+y^2 will be consolidated here. We should finally put pivot x and z around the data centers:

Advantages of Support Vector Machine:

- The clear edge of division works genuinely well
- In high dimensional spaces it is practical.
- In situations that are more familiar than the proportion of the experiments, the number of estimates is possible.
- A subset of ready concentrations is used in quite a range.

Real time Applications of Support Vector Machine:

- Face identification
- Text and hypertext classification
- Classification of pictures
- Bioinformatics
- Handwriting acknowledgment
- Generalized prescient control

1.2. Logistic Regression

Types of Logistic Regression

- Binary Regression of logistics (Pass / Fail)
- Multinomial Logistic Regression(cats, dogs, sheep)
- Ordinal Logistic Regression(Low, Medium, High)

A logistic regression model can be represented by the equation

Revised Manuscript Received on December 13, 2019.
* Correspondence Author
M Ramachandro* Dept of Computer Science &Engg, GMR Institute of Technology,AP, rama00565@gmail.com
Ravi Bhramaramba, Dept of Computer Science &Engg,, GITAM

$$ln(\frac{p}{1-p}) = a_0 + a_1x_1 + a_2x_2 + a_3x_3$$

- p is the likelihood of an occasion to happen which you are attempting to anticipate
- x1, x2 and x3 are the autonomous factors which decide the event of an occasion for example p
- a0 is the consistent term which will be the likelihood of the occasion happening when no different variables are considered.

R.H.S speaks to the connection work which will assist us with determining a non-direct connection in a straight manner wherein (p/1-p) is the odd proportion. At whatever point the log of the chances proportion is seen as positive, the likelihood of achievement is in every case over half.

Advantages of Logistic Regression

- It's profoundly interpretable
- It doesn't require information highlights to be scaled
- It doesn't require any tuning
- It's anything but difficult to regularize, and it yields well-adjusted anticipated probabilities.
- Can use continuous matching variable
- Can utilize multivariate coordinating
- Works well with diagonal
- Does not give undue weight to correlated features

Applications of Logistic Regression

- Regression analyses
- Widely utilized managed AI procedure
- It is perhaps the best apparatus utilized by analysts

1.3 Introductions to datasets

DATASET

There are different openly accessible microarray datasets from malignant growth quality articulation ponders, including leukemia disease, prostate tumor, colon disease, lymphoma, bosom disease, NCI60, and lung malignancy datasets. Among them two datasets (prostate and lung) are utilized in this paper. All the malignant growth datasets have been gathered from the storehouse of Artificial Intelligence Lab, Ljubljana.

Prostate dataset comprises of 40 examples: 50 examples of ordinary tissue and 52 examples of prostate tumor. Prostate malignant growth is most regular heterogeneous illness among people, regarding profoundly dissimilar clinical and histological results. Each example contains 54675 quality articulation levels.

Lung malignancy dataset comprises of 84 samples: The tumor reaction to neoadjuvant treatment was surveyed after the samples.were got before treatment. Each example contains 12600 quality articulation levels.

II. LITERATURE STUDY

Stephan Dreiseitla, Lucila Ohno-Machado (2002)

The models of judgment are operational regression and artificial neural networks .In this survey, we a bridge the distinctions and similarities of these models from a specialized perspective, and contrast them and other AI calculations. We provide helpful considerations for a basic assessment of the nature of the models and their results. Finally, in an example of therapeutic writing papers, we compile our findings on the fulfilment of quality criteria for strategic relapse and falsified neural system models.

Shuzhou Wang, Bo Meng (2011) Support Vector Machine (SVM) is another displaying technique. In many fields and generally in neural systems, it demonstrated great performance. Before preparing SVM, the parameter should be determined. The SVM parameters were selected for the adjusted particle swarm optimization techniques.

Ersoy Öz, Hüseyin Kaya (2013) Vector support machines are a two-class characterization method, one of the non-parametric controlled classifiers, which is presented to minimize measurable learning pathogens. Support vector machines Variable supports were essentially divided into two categories as explicit variable support machines and nonlinear vector assistance machines. Vector supporting machines are intended to organize by projecting data to that higher dimension information space by constructing an aircraft in the field. In essence, this solution requires solving a quadratic programming problem. The quality control of DNA sequencing data is used in this analysis which has a growing usage level in the model recognition field. The reliability of the whole DNA sequence data is therefore, of course, characterized as "top-notch / low-quality."

Chao-Ying Joanne Peng et al., (2013)

This article shows the favored example for the use of calculated strategies with an outline of strategic relapse applied to an informational collection in testing an examination theory. Proposals are likewise offered for proper detailing arrangements of calculated relapse results and the base perception to-indicator proportion. The creators assessed the utilization and translation of calculated relapse.

Padmavathi Janardhanan et al., (2015) the possibility of restorative information mining is to separate covered information in medicinal field utilizing information mining procedures. One of the positive angles is to find the significant examples. This paper discusses the suitability of SVM, the most famous datasets. The presentation of the Naïve Bayes classification, RBF system and SVM classification is examined in this paper. Various medicinal data sets divide the exhibition of prophetic models.

The datasets were parallel and each database had a variety of features. The datasets include cardiovascular databases, malignant development and diabetes. The SVM classifier provides better classification reliability. The research has been adapted to WEKA standards and has obtained findings that show that the SVM category of restorative data sets is the most efficient and desirable.

Aditi anil ghive, D. R. Patil (2015) the medicinal field is one of the field in which new research is done at a quicker rate. In a medical field automation is picking up force. From the restorative information valuable data can be removed and made helpful for creating programming or MEDLINE applications that can help specialists in the treatment. In this paper, we need to recognize ailments and their medicines in short content utilizing SVM and NB arrangement calculations.

Trial results show that SVM gives preferred characterization result over credulous Bayes. These calculations are assessed utilizing four criteria: Accuracy, Precision, Recall and F-measure.

Janmenjoy Nayak et al., (2015) over the past two decades the Support Vector Machine's use of various information processing undertakings has been expected to provide a significant measure for research efforts. Information Mining is a leading and exciting exploration area because of its huge areas of application and native people. The Support vector machine (SVM) is a strategic task, as it provides strategies with an effective and decent value path that are particularly appropriate. In this article, we discuss SVM's work in various data mining approaches such as grouping, bunching, anticipation and other applications. We also analyzed the quantity of work outputs that have been applied for information-mining applications in several widely speculated publications and have also indicated a possible no. of SVM problems. This paper focuses mainly on extrapolating the various zones of SVM with an understanding of the system and a thorough study while providing analyzers with a modernized picture of the depth and width of the hypothesis and applications.

Chih-Feng Chao, Ming-Huwi Horng (2015)

Parameters for precision and efficiency in the supporting vector machines (SVM) are important. In this article, all SVM parameters including the retribution parameter, the smoothness parameter, and the lagrangians were equipped with firefly measurement. The technique introduced is called the firefly-based SVM. In particular, the single class SVM is not suitable for application in a multiclass characterization together with best part optimal.

Krupal S et al., (2016) Support Vector Machine (SVM) has been extremely well known as a huge edge classifier due its hearty numerical hypothesis. It has numerous down to earth applications in various fields, for example, in bioinformatics, in therapeutic science for analysis of ailments, in different building applications for expectation of model; it is broadly utilized in restorative science due to its incredible learning capacity in grouping. It can arrange profoundly nonlinear information utilizing bit work. Skin patients are not dependent upon appropriate determination bringing about abuse. We think SVM is a decent apparatus for appropriate finding. This paper utilizes different pieces for characterization and accomplishing the best precision of 95.39 %.

K.Sharmila, S.A.Vethamanickam (2016)

The Apache Hadoop is now a big data range. Two phases are included in the above blend template. The K-implies the bunching of the main phase is used to find the misarranged occasion and get rid of it. In the next stage, the Support Vector Machine (SVM) is adjusted using the privileged grouped event at the earliest stage. The result shows that our solution to this convergence design is successful in predicting diabetic patients that, in fact, are at risk for a cardio-vascular disease, nephropathy and retinopathy and, at the same time, guarantees the patients ' good therapy at the right time.

Amit Yadav et al., (2016)

Persistent information had been gathered from the medical clinic of Nepal with the assistance of emergency clinic organization, specialists and patient collaboration. Information examination endeavors to shows the huge connection among infection and components causal of illness. Research investigates the utility of multinomial strategic relapse system in wellbeing area and its most advantageous use for unmitigated information. Papers attempt to display different variables which bring about occurring of wellbeing issue and feature utilization of information mining method in human services. It is felt that this research gives greater specificity and unquestionable value in identifying the causal factors of disease, misrepresentation, the help for all human service activities, the cost reduction, time reduction and the method of diagnosis.

Yanke Zhu, Jiqian Fang et al., (2016)

The grouping tree is an important strategy for prescient displaying and information mining. Not with standing, Current existing patterns are not concerned with the possibility of having a sub-set of individuals who cannot be grouped all round based on data from a given set of predictor variables and who can be mixed with a higher fault rate; in each step, the vast majority of the established order trees do not use a combination.

S.Vanitha, P.Balamurugan (2017)

A structure of arrangements using a neural network and vector support machines for the ordering of therapeutic information with a range of features and incidences that includes two kind of information and tries in particular to disseminate healthier and non-sick details from the collection of information and identification of beneficial and noxious from cleveland coronary disease The test results clearly show that the classifier of the neural network is important in attempts to organize therapeutic data.

Wenjing Pian1,Christopher SG Khoo (2017)

The point of this examination was to research the probability of distinguishing the three client wellbeing data settings (looking for self, scanning for other people, or perusing in light of no specific medical problem) utilizing just hyperlink clicking conduct; utilizing eye-following data; and utilizing a blend of eye-following, statistic, and criticalness data. Prescient models are created utilizing multinomial calculated relapse.

III. PROPOSED SYSTEM

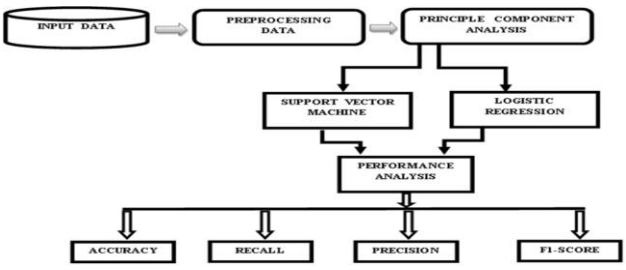


Figure 1: Proposed model architecture

Figure 1: In the above diagram initially taking the input data and assigning to preprocessing module then applying the PCA after getting the result of PCA, assigning to SVM & LR methodologies. After getting the results from SVM & LR methodologies giving to the performance analysis.

IV. RESULTS & GRAPHS

TABLE (1)

ALGORITHM	DAT SET USED	PARAMETERS	WITHOUT PCA				WITH PCA			
			WITH OUT SCALING		WITH SCALING		WITH OUT SCALING		WITH SCALING	
			TRAINING	TESTING	TRAINING	TESTING	TRAINING	TESTING	TRAINING	TESTING
Support Vector Machine (SVM)	Prostate Cancer	Accuracy	1	0.62	0.94	0.62	0.62	0.5	0.66	0.38
		Precision	1	1	1	1	0.62	0.5	0.73	0
		Recall	1	0.25	0.88	0.25	0.62	0.25	0.5	0
		F1-score	1	0.4	0.93	0.4	0.62	0.33	0.59	0
	Lung Cancer	Accuracy	1	0.7	1	0.65	0.53	0.53	0.91	0.55
		Precision	0	0.69	1	0.75	0.53	0.58	0.9	0.6
		Recall	1	0.82	1	0.55	1	0	0.92	0.55
		F1-score	1	0.75	1	0.63	0.69	0.73	0.91	0.57

ALGORITHM	DAT SET USED	PARAMETERS	WITHOUT PCA				WITH PCA			
			WITH OUT SCALING		WITH SCALING		WITH OUT SCALING		WITH SCALING	
			TRAINING	TESTING	TRAINING	TESTING	TRAINING	TESTING	TRAINING	TESTING
Logistic Regression	Prostate Cancer	Accuracy	1	0.62	1	0.5	0.62	0.5	0.62	0.5
		Precision	1	1	1	0.5	0.64	0.5	0.67	0.5
		Recall	1	0.25	1	0.25	0.56	0.5	0.5	0.25
		F1-score	1	0.4	1	0.33	0.6	0.5	0.57	0.33
	Lung Cancer	Accuracy	1	0.75	1	0.65	0.53	0.53	0.94	0.65
		Precision	0.57	0.71	1	0.75	0.53	0.58	0.93	0.7
		Recall	1	0.91	1	0.55	1	0	0.95	0.64
		F1-score	1	0.8	1	0.63	0.69	0.73	0.94	0.67

In this study of Support Vector machine and Logistic Regression we are considering two datasets related to Gene expression micro array data. With trained and tested data with scaling techniques. Again we are taking the measures here with principle component analysis and without principle component analysis. Based on these we are taking four performance analysis like accuracy, precision, recall and f1-score. Here the training and tested data regarding taking 80% of data for training and 20% for testing.

Results Comparison of Training and Testing Data of SVM and Logistic Regression

Table 1: Here considering the training data and testing data results with SVM. Here we are considering the different data sets like Prostate and Lung cancer with different measures and performance analysis parameters taken into consideration to obtain the results.

Table 2: Here considering the training data and testing data results with Logistic Regression. Here we are considering the different data sets like Prostate and Lung cancer with different measures and performance analysis parameters taken into consideration to obtain the results.

Here the abbreviations of Measures:

1.WOP+WOS = Without Principal Component Analysis + Without Scaling
2.WOP+WIS = Without Principal Component Analysis + With Scaling
3.WIP+WOS = With Principal Component Analysis + Without Scaling
4.WIP+WIS= Without Principal Component Analysis + With Scaling

Result Discussions

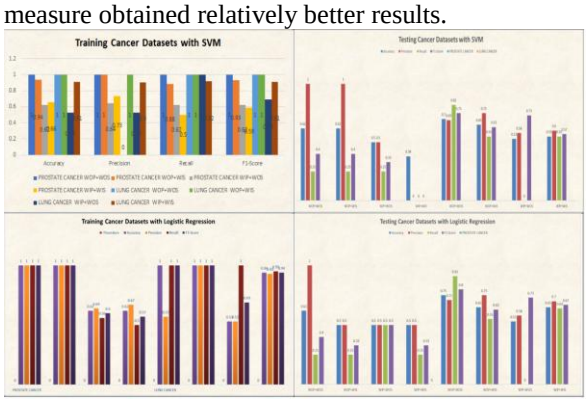
Machine learning techniques are SVM and LR with four preprocessing with WOP+WOS, WOP+WIS, WIP+WOS, WIP+WIS applied into different cancer data sets prostate and lung cancer.

SUPPORT VECTOR MACHINE SUMMARY RESULTS:

1. Training prostate cancer data with **SVM WOP+WOS** this measure obtained best accuracy, precision, recall and F1-Score.
2. Testing prostate cancer data with **SVM WOP+WOS** this measure obtained best precision.
3. Training lung cancer data with **SVM WOP+WOS** this measure obtained best accuracy, recall and F1-Score.
4. Testing lung cancer data with **SVM WOP+WOS** this measure obtained relatively better results.

LOGISTIC REGRESSION SUMMARY RESULTS:

1.Training prostate cancer data with **LR WOP+WOS** this measure obtained best accuracy, precision, recall and F1-Score.
2. Training prostate cancer data with **LR WOP+WIS** this measure obtained best accuracy, precision, recall and F1-Score.
3. Training lung cancer data with **LR WOP+WIS** this measure obtained best accuracy, recall and F1-Score.
4. Testing prostate cancer data with **LR WOP+WOS** This measure obtained best precision.
5.Testing lung cancer data with **LR WOP+WOS** this measure obtained relatively better results.



Final comparisons with machine learning algorithms

V. CONCLUSION

Dataset pertaining to two different types of cancer namely prostate, Lung cancers were taken. Preprocessing of data was done using Principal component analysis without scaling and with scaling using two techniques namely supports vector machine and Logistic regression. Finally Performance analysis was done.

REFERENCES

1. Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics, 35(5–6), 352–359.
2. Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. Journal of Educational Research, 96(1), 3–14.
3. Öz, E., & Kaya, H. (2013). Support vector machines for quality control of DNA sequencing. Journal of Inequalities and Applications, 2013, 1–9.

4. Janardhanan, P., Heena, L., & Sabika, F. (2015). Effectiveness of support vector machines in medical data mining. Journal of Communications Software and Systems, 11(1), 25–30.
5. Nayak, J., Naik, B., & Behera, H. S. (2015). A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. International Journal of Database Theory and Application, 8(1), 169–186.
6. Patil, D. R. (2015). Implementation of Svm and Nb Algorithms for. (4), 141–145.
7. Chao, C. F., & Horng, M. H. (2015). The construction of support vector machine classifier using the firefly algorithm. Computational Intelligence and Neuroscience, 2015.
8. Wang, S., & Meng, B. (Parikh, K. S., & Shah, T. P. (2016). Support Vector Machine – A Large Margin Classifier to Diagnose Skin Illnesses. Procedia Technology, 23, 369–375.
9. Yadav, A., Hui, L., Ali, M., & Anis, M. (2016). Analysis of Healthcare Data of Nepal Hospital using Multinomial Logistic Regression Model. International Journal of Management & Information Technology, 11(2), 2720–2730.
10. Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., … Liu, D. (2016). Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Systems, 96, 61–75.
11. .Zhu, Y., & Fang, J. (2016). Logistic Regression-Based Trichotomous Classification Tree and Its Application in Medical Diagnosis. Medical Decision Making, 36(8), 973–989.
12. B.Santhosh Kumar, P.Kanaga Ranjitham, K.R Karthekk, J Gokila, "Survey on various small file handling strategies on Hadoop", IEEE International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016, DOI:10.1109/ CESYS. 2016. 7889882, Pages 1-4
13. Pian, W., Khoo, C. S. G., & Chi, J. (2017). Automatic classification of users' health information need context: Logistic regression analysis of mouse-click and eye-tracker data. Journal of Medical Internet Research, 19(12).
14. Parameter selection algorithm for Support Vector Machine. Procedia Environmental Sciences, 11, 538–544.
15. Vanitha, S., & Balamurugan, P. (n.d.). Medical Data Classification Using Svm and Neural Network Classifier-a Study. 139–145.
16. 15.B.Santhosh Kumar, S. Karthik, V. P. Arunachalam "Upkeeping secrecy in information extraction using 'k' division graph based postulates", **DOI:**10.1007/s10586-018-1705-2, *Journal of Cluster Computing* **(Springer)**, Volume 22, pp 57–63, January 2019, Pages:1-7

AUTHORS PROFILE

M.Ramachandra is M.Tech in Computer Science from NIT, Tiruchi, India. Since 2008 he has been working as Assistant Professor in the department of CSE, GMR Institute of Technology, Rajam, A.P, and India**.** His area of research includes Data mining & Bio Informatics

Dr. Bhramaramba Ravi obtained her Ph.D from JNTUH in the year 2011. She has about 19 years of teaching experience and is currently Professor in the Dept of Computer Science and Engineering GIT, GITAM , and Visakhapatnam. She has about 32 publications in reputed Journals. Her area of interest is Data Mining and Bioinformatics