International Journal of Future Generation Communication and Networking
Vol. 13, No. 1, (2020), pp. 1453-1459

Auto Encoder Based Feature Learning and Up-sampling to Enhance Cancer Prediction

M.Akkalakshmi
Dept of CSE,
GITAM (Deemed to be University) Hyderabad
lakshmi9.muddana@gmail.com

Y.Md.Riyazuddin
Dept of CSE,
GITAM (Deemed to be University) Hyderabad
riyazymd@gmail.com

V.Revathi
Dept of CSE ,
GITAM (Deemed to be University) Hyderabad
vrevathi530@gmail.com

Abhisek Pal
Dept of Pharmacy,
GITAM (Deemed to be University) Hyderabad
abhisek.cology@gmail.com

Abstract

Early detection of cancer is vital in long term survival of the patient. Cancer detection requires skilled doctors and analysis of patients data in different form like images, clinical data and gene expressions. AI can help in analysing large and complex data of different forms and still achieve good accuracy on par with specialist doctors. Semi-supervised learning techniques of AI can deal with even scarce and incomplete datasets. This paper deals with up-sampling the unbalanced breast cancer datasets and proposes a semi-supervised learning approach for latent features learning using autoencoders to improve the prediction accuracy which helps in cancer diagnosis and treatment.

Keywords

Latent features, Semi-supervised learning, Autoencoders, Up-sampling

Introduction

As per WHO reports, cancer is the second leading cause of death globally, accounting for an estimated 9.6 million deaths. Breast, cervical, thyroid being most common among women and stomach, liver, prostate among men.As per the reports, breast cancer death rates were more than double the rates of any other cause of cancers deaths among women in the age group of 20-50.It is increasing in developing countries where the majority of cases are diagnosed in advanced stages.

Cancer detection and diagnosis is a challenging task and accurate prediction in the early stage is crucial for effectivetreatment. Machine learning plays an increasingly important role in cancer diagnosis and the accurate prediction. Application of traditional machine learning techniques faces the challenges like i) data not recorded and stored for access to analysts ii) unavailability of specialists for labelling the data iii) limited data availability iv) not many malignant cases information available in the datasets v)non-Integration of different forms of data like images, clinical data, gene expression data.

Unlabeled data which is inexpensive is available in large volumes. This precious data can be utilized using semi-supervised learning techniques of AI for improving the performance of prediction models by learning important features from the dataset.

ISSN: 2233-7857 IJFGCN

Related work

Sharifah Hafizah Sy Ahmad Ubaidillaha et.al[1]made comparative study of different machine learning algorithms like SVM,ANN classifiers on datasets of different cancer types.**KonstantinaKourou,.et.al**[2] reviewed the machine learning approaches of cancer prediction on different input features and data samples.**Francisco Azuaje** [3] discussed the key challenges like high dimensionality, insufficient data and need of hybrid models for AI in precision oncology.**Jun Chin Ang et.al**[4] used a semi-supervised SVM-based feature selection (S^3VM-FS) to reduce the dimensionality that resulted in higher accuracy model for lung cancer prediction.**Shi M, Zhang B[**5] uses Low density separation technique based on clustering assumption to handle high dimensional gene and small size data sets.**Rasool Fakoor. et.al[**6] uses data from different types of cancer to learn the features using sparse auto encoders to improve model accuracy.**Padideh Danaee.et.al** [7] used stacked Denoising auto encoder to extract features of high dimensional gene expression datasets before applying classifiers for better accuracy.**N. V. Chawla.et.al**[8] used up sampling and down sampling methods for imbalanced datasets before applying the classifier.

Proposed model

A semi-supervised technique is used in building the model. First latent feature learning using autoencoder followed by balancing the datasets using upsampling and then building the classifier.

Twopublic datasets were taken having featurescharacterizing cell nuclei of the tumour. Dataset1 (Wisconsin) has30 features,dataset2 with 9 features and label that indicate benign or malignancy of the tumour. Both the datasets have imbalanced classes.

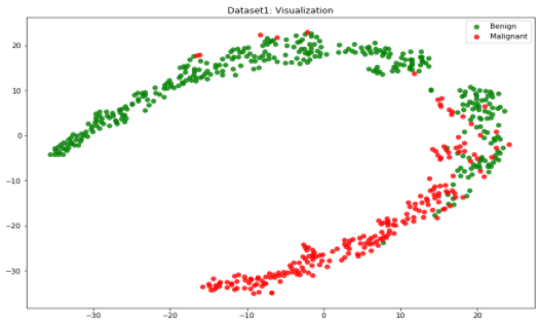
Datasets	No. of Features	No. of Samples with Malignancy	No. of Normal Samples	Total number of Samples
Dataset1	**30**	**212**	**357**	**569**
Dataset2	**9**	**241**	**458**	**699**

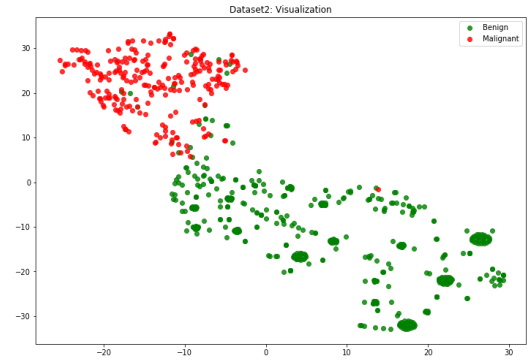
Pre-processing:

Dataset2 contained 16 null values in bare. Nuclei feature and was replaced by mean value of that feature. As the range of values of the features varies, the features were normalized in both the datasets.

Visualization of the datasets

ISSN: 2233-7857 IJFGCN





Linear regression (LR), Support Vector Machine (SVM) using rbf kernel models were built and measured accuracy using F1 score.

DATASETS	LR		SVM	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Dataset1(Wisconsin)	95.39	98.11	97.74	97.19
Dataset2	95.08	94.64	96.53	94.73

As the datasets show Imbalance of classes,techniques can be applied to balance classes. Different methods are available for balancing the imbalanced datasets.

i). Up-sampling

ii). Down-sampling.

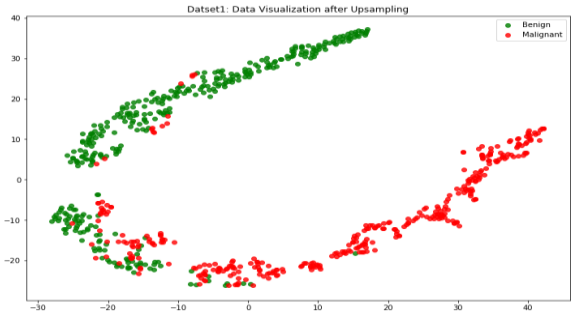
As the dataset size is small, up-sampling isapplied using SMOTE algorithm. The algorithm synthesises the samples by taking k-nearest neighbours of the randomly picked minority samples.

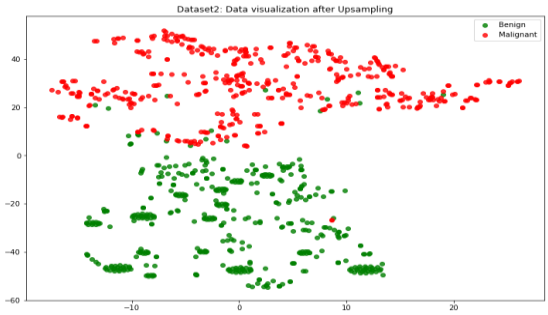
The resultant datasets after applying SMOTE algorithm is as follows

Datasets	**No. of Samples with Malignancy**	**No. of Normal Samples**	**Total number of Samples**
Dataset1	**357**	**357**	**714**

Dataset2	**458**	**458**	**916**

Visualization of the datasets after up sampling.



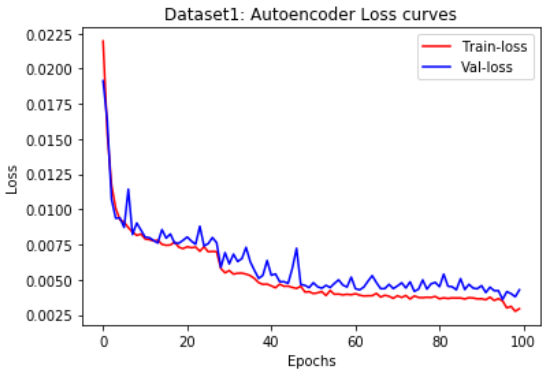


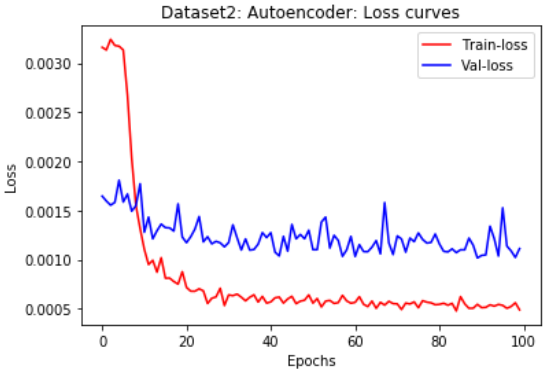
LR and SVM modelsare rebuilt on the up-sampled datasets and the results show improvement in accuracy and variance.

Datasets	LR		SVM	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Dataset1	97.51	98.39	97.89	98.39
Dataset2	96.96	97.29	98.15	97.34

To improve the performance further, a semi supervised learning technique can be adopted. Feature learning was applied using autoencoders to determine latent features. Autoencoder was tuned with different optimizers and mini batch sizes.Low loss was obtained with RMSPROP optimizer with batch size of 16 and trained for 100 epochs. 40 latent features were learnt from 30 features of Dataset1 and 25 latent features from dataset2.

The following loss curves show good performance of the autoencoder.





On the learned features of up-sampled data, Feed forward Neural network classifier model was built. Neural network was tuned with different optimizers, batch sizes and number of hidden layers. Good train accuracy, test accuracy and low variance was achieved with ADAM optimizer, nine hidden layers and mini batch size of 16. As the dataset sizes are small 25% data is used as test sample.

Following is the accuracymeasures on train and test data.

Datasets	Feed Forward Neural Network	
	Train Accuracy	Test Accuracy
Dataset1	99.81	98.88
Dataset2	98.98	99.12

Semi supervised learning using autoencoders for latent feature learning on Up-sampled data and neural network model for binary classifier has shown good performance. Good train, test accuracy and could reduce the variance to less than 1% in both the datasets.

Conclusion

LR and SVM models were built on two breast cancer datasets and analyzed the classification accuracy using F1 score. As there is scope for improving the accuracy, two techniques were used.Due to imbalance in the datasets, SMOTE technique was applied for up-sampling. To further enhance the accuracy, auto encoder was used for determining the latent features on which neural network classifier was built. These two techniques of Up-sampling and latent feature learning on the breast cancer datasets have shown good improvement in accuracy and also resulted in low variance.

References

1) Sharifah Hafizah Sy Ahmad Ubaidillaha*, Roselina Sallehuddina , Nor Azizah Ali; Cancer detection using Artificial Neural networks and SVM:A Comparative Study in Jurnal Teknologi · October 2013
2) KonstantinaKourou,Themis P.Exarchos,Konstantinos P.Exarchos, Michalis V.Karamouzis, Dimitrios I.Fotiadis ; Machine learning applications in cancer prognosis and prediction in Computational and Structural Biotechnology Journal in 2015

ISSN: 2233-7857 IJFGCN

3) Francisco Azuaje; Artificial Intelligence for precision oncology / beyond patient stratification Published in partnership with The Hormel Institute, University of Minnesota in 2019.
4) Jun Chin Ang, Habibollah Haron(B), and Haza Nuzly Abdull Hamed;Semi-supervised SVM-based Feature Selection for Cancer Classification using Microarray Gene Expression Data in Springer International Publishing Switzerland 2015 pp. 468–477, 2015.
5) Shi M[1], Zhang B; Semi Supervised learning improves Gene expression based prediction of cancer recurrences; Bioinformatics. 2011 Nov 1;27(21):3017-23.
6) Rasool Fakoor ,Faisal Ladhak , Azade Nazi , Manfred Huber; Using deep learning enhance cancer diagnosis and classification in 30 th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013
7) Padideh Danaee, Reza Ghaeini, David Hendrix; Deep learning approach for cancer detection and relevant gene identification Computer Science, Medicine, Biology Published in PSB 2017
8) N. V. Chawla,K. W. Bowyer,L. O. Hall,W. P. Kegelmeyer,; "SMOTE: Synthetic Minority Over-sampling technique" JAIR Published**:** Jun 1, 2002

Authors

Dr.M.Akkalakshmi working as Professor, Dept of CSE, School of Technology, GITAM (Deemed to be university), Hyderabad. She has Published 40 no of International Publications (Scopus and Google Scholar). Her expertise in Big Data, AI & Machine Learning.

Dr.Y.Md.Riyazuddin working as Assistant Professor, Dept of CSE, School of Technology, GITAM (Deemed to be university), Hyderabad. He has published 13 No of International Publications (Scopus and Google Scholar). His expertise in IoT, Networks and Security, AI&ML. He has published a Text Book on Computer networks and He is having Patent.

Ms.V.Revathi working as Assistant Professor, Dept of CSE, School of Technology, GITAM (Deemed to be university), Hyderabad.. She has published 3 No of International Publications (Scopus and Google Scholar). Her expertise in Deep Learning, IoT, AI&ML. She has Membership in International Association of Engineers.

Dr.Abhisek Pal working as Assistant Professor, Dept. of Pharmacology, School of Pharmacy, GITAM (Deemed to be university), Hyderabad. He is having 14 Years of teaching Experience. He has Published 50 No of International Publications (Scopus and Google Scholar). Under his Guidance 2 scholars has been awarded PhD Degree. His research interests include AI in health care and drug discovery, Reverse pharmacology & Psychobiotics research. He also perusing different funded projects from DBT,DST etc.

ISSN: 2233-7857 IJFGCN