Combining co-clustering and adaboost ensemble method to breast cancer detection

Sujoy Bhattachary[1,†], Duvvu Avinash[2,†], and Aggunna Sai Teja[3,†]

[1]professor,Vinod Gupta School of management,Kharagpur,721301,India
[2]dual degree student,ocean engineering and naval architecture,Kharagpur,721301,India
[3]dual degree student,ocean engineering and naval architecture,Kharagpur,721301,India
[*]corresponding author(s):Duvvu Avinash (duvvuavinash@gmail.com)
[†]these authors contributed equally to this work

ABSTRACT

purpose

to create a hybrid algorithm of applying biclustering and ensemble learning with adaboost algorithm to outperform the known algorithms on classification problems

Design/methodology/approach

Adaboosting is a weak learner based ensemble algorithm which gives more weights to the weak learners and combining alltogether gets a model sufficient to outperform the known algorithms like neural networks,SVM. from the rules set which we created using biclustering method

Findings

our results are quite promising for the experiments performed on breast cancer data,here the similiarity of trends of behaviour for specific features help for finding the problems where it is applied for such data

Practical implications

the combined algorithm is found to be useful for more data interaction due to the method adopted by us.

originality/value

this paper introduces a method of combining biclustering and adaboost a new hybrid strategy

keywords

biclustering,Adaboosting,malignant rules,benign rules

Paper type

Research type

Please note: Abbreviations should be introduced at the first mention in the main text – no abbreviations lists or tables should be included. Structure of the main text is provided below.

Background & Summary

Breast Cancer is the most commonly diagonised cancer after skin cancer in woman of united states,the fatality rate is high,breast cancer forms in lobules or breast ducts.Lobules are the milk glands and the pathways which take milk from the glands to the nipple.the symptoms of breast cancer are very difficult to be felt in early stages,but an abnormality in mammogram is

visible.there will be lump in the breast but not all lumps are cancerous.some of most common breast cancer symptoms include breast pain,red pitted skin over the region of breast,blood secretion from nipple,swelling or lump under the arm etc.

Types of Breast Cancer

the 2 main categories of breast cancer are "invasive" and "Non-Invasive" or insitu.Although invasive cancer spread to other areas of the breast from the breast canals or glands, noninvasive cancer has not spread from the original tissues. the above 2 categories describes most commonly occuring cancers

Ductal Carcinoma Insitu :-(DCIS) this is a non invasive condition wherein cancer cells stick to ducts in the breast and it is considered as one of the earliest form of breast cancer

Lobular Carcinoma Insitu :-(LCIS) this is quite an uncommon condition wherein the cancer grows in milk producing glands

Invasive Ductal Carcinoma :-IDC starts in milk ducts and invades tissue near the breast.invasive cancer means that the cancer cells have broken away from the lobule and are able to enter the lymph nodes and other regions of the body

Invasive Lobular Carcinoma :-ILC The cancer began in the lobule-line cells and spread to the breast tissue around the lobules.

Methods for identification of Breast Cancer

Breast cancer can be diagnosed through multiple tests, including a mammogram, ultrasound, MRI and biopsy

diagnosis

first step in cancer diagnosis is to do final needle aspirate or extracting some of cells of tumor.at this stage we don't know the cells are malignant or benign.if they are benign the person is safe somehow and cancerous cells are not spreading and if it is malignant the cells are cancerous

Causes and Risk factors of Breast Cancer

family history

gene mutation

late childbearing

early menstruation late Menopause/obesity

increased breast density

prolonged use of oral contraceptives

hormone replacement therapy after menopause

Alcohol intake/Tobacco

1 Previous techniques and comments

The authors of previous work in the 2002-2010 timeframe have carried out a study on the shortcomings of screen-film mammography (SFM) based upon several articles. For instance, this technique produces a large number of false positives [1] , at a rate between 5% and 35%.with the growth of image processing some researchers have mined the severity of cancer with this method [2],[3]. many authors[4]-[5] have used neural networks model and Recurrent Neural Networks(RNN),Deep Neural Networks(DNN) etc.they also underlined the value of breast cancer resolution through groundbreaking computer science techniques.The ability of breast cancer detections with increased breast tissue temperature acquired by infrared thermography was shown (temperature sensitivity of the picture acquired was about 2°C over a couple of minutes)[6]. Texture characteristics are extracted from the matrix and run length matrix. These attributes are consequently fed into an automatic classification of benign and malignant breast conditions by the SVM classifier[7].[8]He used statistical models based on a DT decision table, finding that the patient success rate was 86.52%. In order to resolve the imbalanced problem, they used the under-sampling C5 method using bagging algorithm to boost prediction accuracy on breast cancer.[9]He proposed to find systemic relations in chemometrics a field of pharmaceutical industry .a new decision-tree-based ensemble technique coupled with a backward elimination feature selection strategy with bagging. [10]The results of the Decision Tree for SVM were encouraging and

superior, with low error rates and high average gains. in [11]he compares highly accurate supervised and unsupervised technique whic uses breast thermal images for assisting physicians using a suitable unsupervised learning technique fuzzy c-means clustering to confirm the suspicious areas compared to Adaboost and the mean accuracy is foundd to be 88%.in[12]proposed a hybrid classification technique neuro-fuzzy algorithm to increase the classification accuracy on wiscon breast cancer dataset by 1o fold cross validation.[13]Thermography-based breast cancer diagnosis with a set of statistical features taken from thermograms and evaluating bilateral variations between left and right breast areas which are classified using a fuzzy rule based system.[14]image features which are represented by local nuclei features from convolution neural network models which are trained from imagenet data base.these features are being classified using SVM classifier and belief theory based strategy is eployed to combine the outputs from CNN-SVM classifier with an accuracy of 96.91 % .

Method

1.1 Dataset and preprocessing

this dataset is taken from dataset.the datset has 1062 entries,initially there were 27 features each are being rated on a scale of 5.there were 418 benign cases and 644 malignant cases.we used variance to find BI-RADS features of discriminative abilities.the results showed skin retraction,elastic assessment,nodesinbreast,post surgical fluid,fatnecrosis,mass skin,foreign body show similar features on both benign and malignant cases,these features have a variance below 0.1 so we dopped those features.we use -1 for benign tumor and +1 for malignancy.

1.2 Feature-Extraction

our method of biclustering using cheng and church algorithm helps us in finding biclusters,we use biclustering to find the feautres of each diagnostic rule,we get an overall 18 biclusters,15 are malignant rules and 3 are benign rules,we want to make evaluation more interactive,for that we use feature space dependent diagnostic rule,which helps in predicting the class of training data we feed to prepare the model.the more nearer the training instance is to the classifier,here our weak classifier is formed by using one benign and one malignant rule.

1.3 Biclustering algorithm

block clustering, co-clustering or two-mode clustering is a data mining technique that allows the rows and columns of matrix to be clustered simultaneously.Originally Biclustering was introduced by J. A. Hartigan in 1972[15]In this study, we concentrate in the mining of the biclusters with constant columns, and in this analysis, we focus onIn this study, we concentrate in the mining of the biclusters with constant columns, and in this analysis, we focus on 1972 DirectC.Boris Mirkin[16] later used the term biclustering. This algorithm was not generalised until 2000 when Y. Cheng and G. M. Church proposed a variance-based biclustering algorithm and applied it to data on biological gene expression[17].

Biclustering algorithms cluster the rows and columns of a data matrix simultaneously. Such row and column clusters are known as biclusters. Given a set of m samples represented by an n-dimensional feature vector, the entire dataset can be represented as m rows in n columns (i.e., an m * n matrix). The biclustering algorithm generates biclusters – a subset of rows which exhibit similar behavior across a subset of columns, or vice versa.Biclustering algorithm clusters the rows and columns at the same time.the clustered rows and columns are known as biclusters.here in our study bicluster refers to a local coherent pattern in doctor's perspective.which should be used for obtaining some diagnostic rule.In this study, we concentrate in the mining of the biclusters with constant columns, and in this analysis.Biclusters with a small mean squared residue are of concern to Cheng and Church [18]and is a indicator of biclusters homogeneity.let $\boldsymbol{A}$ be a matrix and $\boldsymbol{(M,N)}$ represent a bicluster.$\boldsymbol{M}$ is set of rows in bicluster,$\boldsymbol{N}$ is set of columns in a bicluster. $\boldsymbol{Aij}$ is a submatrix of $\boldsymbol{A}$.let $\boldsymbol{aij}$ with $i \in M\ j \in N$, be an element of bicluster.

$$a_{iN} = \frac{1}{|\boldsymbol{N}|}\sum_{i} a_{ij}$$

.

$$a_{Mj} = \frac{1}{|\boldsymbol{M}|}\sum_{j} a_{ij}$$

.

$$a_{MN} = \frac{1}{|\boldsymbol{M}||\boldsymbol{N}|}\sum_{i,j} a_{ij}$$

.

these were the means of rows,columns and overall means.the residue of element

$$a_{ij} - a_{iN} - a_{Mj} - a_{MN}$$

.

The mean squared residue of bicluster is

$$H(M,N) = \frac{1}{|\boldsymbol{M}||\boldsymbol{N}|} \sum_{i,j} (a_{ij} - a_{iN} - a_{Mj} - a_{MN})$$

let us see how to find biclusters. the Algorithm: this method starts with a greedy approach,it starts with a single large bicluster after that it removes rows and columns.this greedy removal can be done efficiently,the mean squared residue of any row i and any column j int the bicluster as:

$$d(i) = \frac{1}{|\boldsymbol{N}|} \sum_{j} (a_{ij} - a_{iN} - a_{Mj} - a_{MN})^2$$

$$d(j) = \frac{1}{|\boldsymbol{N}|} \sum_{i} (a_{ij} - a_{iN} - a_{Mj} - a_{MN})^2$$

then single node deletion takes place,in order to speed up multiple rows and columns are removed at once.Multiple and single node deletion stop when

$$H(M,N) <= \delta$$

.This stage often optionally adds rows if their inverses suit the bicluster pattern.the algorithm restarts from the beginning upon node addition and the bicluster is added to resulting list.once bicluster is found entries in original data gets replaced by entries from original uniform random distribution over the range of original dataset.

1.4 modified AdaBoost

, short for Adaptive Boosting, is a meta-algorithm of machine learning formulated by Yoav Freund and Robert Schapire[19] who won the Gödel Prize 2003 for their work. Boosting is a machine learning method based on the concept of constructing a highly accurate predictive rule by combining several fairly weak and imprecise rules. Boosting will be done by averaging the outputs of a weak classifier collections. AdaBoost is one of the most popular boosting algorithm and extremely simple to use and implement.here in our work after getting biclusters we form rules of biclusters,each bicluster may fall under benign or malignant category.the class of bicluster is decided based on nummber of instances of benign or malignant,the greater the instances of benign it comes under benign class vice versa.then average of selected features is taken and formed a benign or malignant rule,each malignant and benign rule are combined to form a weak clasifier,these weak classifiers are trained and adaboost learning algorithm is created

steps followed in adaboost classifier creation:
step1:input malignant and benign rules
step2:form weak classifiers by taking each from 2 groups
step3:feature space dependent normalized distance metric to evaluate similarity between the rule and test instance here t refers to test instance and r refers to rule.the class of new instance is dependent based on the distance from the rule,if it is nearer to benign it goes to benign class else goes to malignant class.

$$FSND(t,r) = \frac{\|v_t - v_r\|}{\|v_{max} - v_{min}\|}$$

the idea of ensemble classifier lies in initializing weight to weak classifiers and updating based on weigted majority,the weights of those correctly classified instances get lower and wrongly classified get higher.subsequent classifers are trained for those hard classifying instances,after T iteration the component classifers form a hypothesis.

Algorithm .1 *modified AdaBoost*

1. Initialize the observation weights $w_i = 1/N$, $i = 1, 2, \ldots, N$
2. For $m = 1$ to M:
   (a) Fit a classifier $G_m(x)$ to the training data using weights w_i.
   (b) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))]$, $i = 1, 2, \ldots, N$.
3. Output $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.

1.5

Example text under a subsection. Bulleted lists may be used where appropriate, e.g.

- First item
- Second item

Experiments and results

the test results base on algorithm show that 73% accuracy for 1062 data points.this can be used for large scale data.

References

1. Köşüş, N., Köşüş, A., Duran, M., Simavlı, S. & Turhan, N. Comparison of standard mammography with digital mammography and digital infrared thermal imaging for breast cancer screening. *J. Turkish Ger. Gynecol. Assoc.* **11**, 152 (2010).

2. Li, H. *et al.* An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. *Sci. reports* **8**, 1–12 (2018).

3. Ribli, D., Horváth, A., Unger, Z., Pollner, P. & Csabai, I. Detecting and classifying lesions in mammograms with deep learning. *Sci. reports* **8**, 1–7 (2018).

4. Osareh, A. & Shadgar, B. Machine learning techniques to diagnose breast cancer. In *2010 5th international symposium on health informatics and bioinformatics*, 114–120 (IEEE, 2010).

5. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).

6. Williams, K. L., Williams, F. L. & Handley, R. Infra-red thermometry in the diagnosis of breast disease. *The Lancet* **278**, 1378–1381 (1961).

7. Acharya, U. R., Ng, E. Y.-K., Tan, J.-H. & Sree, S. V. Thermography based breast cancer detection using texture features and support vector machine. *J. medical systems* **36**, 1503–1510 (2012).

8. Liu, Y.-Q., Wang, C. & Zhang, L. Decision tree based predictive models for breast cancer survivability on imbalanced data. In *2009 3rd international conference on bioinformatics and biomedical engineering*, 1–4 (IEEE, 2009).

9. Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Chen, X. & Li, H.-D. Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity. *Chemom. Intell. Lab. Syst.* **103**, 129–136 (2010).

10. Abdelaal, M. M. A., Abou Sena, H., Farouq, M. W. & Salem, A.-B. M. Using data mining for assessing diagnosis of breast cancer. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 11–17 (IEEE, 2010).

11. Lashkari, A. E. & Firouzmand, M. Early breast cancer detection in thermogram images using adaboost classifier and fuzzy c-means clustering algorithm. (2016).

12. Lee, C. & Jang, M.-G. A modified fixed-threshold smo for 1-slack structural svms. *ETRI journal* **32**, 120–128 (2010).

13. Schaefer, G., Závišek, M. & Nakashima, T. Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recognit.* **42**, 1133–1137 (2009).

14. George, K., Faziludeen, S., Sankaran, P. *et al.* Breast cancer detection from biopsy images using nucleus guided transfer learning and belief based fusion. *Comput. Biol. Medicine* 103954 (2020).

15. Hartigan, J. Direct clustering of a data matrix (1972).

16. Mirkin, B. Mathematical classification and clustering. *J. Oper. Res. Soc.* **48**, 852 (1996).

17. Cheng, Y. & Church, G. M. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 93–103 (AAAI Press, 2000).

18. Cheng, Y. & Church, G. M. Biclustering of expression data. In *Ismb*, vol. 8, 93–103 (2000).

19. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139, 10.1006/jcss.1997.1504 (1997).

Acknowledgements

we thank professor Sujoy Bhattacharya for his timely support and guidance to stay motivated during the overall process of the work and paper .