Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk: ICCCN 2018, NITTTR Chandigarh, India

Chapter · January 2019

DOI: 10.1007/978-981-13-1217-5_57

2 authors, including:

Tawseef Shaikh
Aligarh Muslim University
10 PUBLICATIONS **6** CITATIONS

Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk

Tawseef Ayoub Shaikh and Rashid Ali

Abstract With the advancement in the technological age, the deadly diseases threatening human survival also increase at the same pace. Breast cancer being at number two in causing the deaths among women is equally among the most curable type of cancer if diagnosed prior to time. There is an utmost thirst for diagnosis of breast cancer through an automation system in everyday health applications. This paper uses dimensionality reduction technique offered by Weka tool called WrapperSubsetEval on two benchmark cancer datasets of Wisconsin and Portuguese "Breast Cancer Digital Repository" (BCDR), on top four data mining algorithms available in literature. The final experiments carried in MATLAB and Weka demonstrated that Naive Bayes, J48, k-NN and SVM got an improvement in accuracy from 92.6186, 92.9701, 96.1336, 97.891 to 97.0123, 96.8366, 97.3638, 97.9123% in case of Wisconsin dataset and an improvement from 87.4126, 80.4196, 93.7063, 91.6084 to 89.5105, 90.9091, 97.9021, 95.1049% in case of BCDR-D01_Dataset.

Keywords Machine learning algorithms · Breast cancer · Naïve Bayes · SVM MATLAB · Weka

1 Introduction

Breast cancer is rising at an alarming rate by being the number two among the fatal diseases in the world in women with over 1.5 million foreseen cases spotted in 2010 and triggering a threat to human survival in causing demises to half a million per year, according to a report of World Health Organization [1]. Its effects are dangerously increasing where it is responsible for one in every six deaths among women in the

T. A. Shaikh (✉) · R. Ali
Department of Computer Engineering,
Aligarh Muslim University, Aligarh, Uttar Pradesh, India
e-mail: tawseef37@gmail.com

R. Ali
e-mail: rashidaliamu@rediffmail.com

C. R. Krishna et al. (eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking*, Lecture Notes in Networks and Systems 46, https://doi.org/10.1007/978-981-13-1217-5_57

European Union [2]. A hot research oriented topic in the modern times, breast cancer continuously makes its victims at an alarming rate where in 2012, likely 1.67 million fresh cancer cases were spotted. India is also not opaque from its deadly shadow. With a frequency of almost 1,44,000 fresh cases of breast cancers per annum, it is emerging as number one female cancer in metropolitan India. A likely probability of 100,000 new breast cancer patients is spotted in India per annum [3]. From 5 per 100,000 female population per year in rural areas, its range enlarges to 30 per 100,000 female population per year in urban areas in India [4]. With a trivial share of the world population in developed countries, still it accounts for 50% of breast cancers identified worldwide [5]. This whole scenario compels to come up with new techniques and methods in order to fight against this threatening disease, thus making entry of Information and Communication Technologies (ICT) an ideal choice for the same. Twenty-first century witnessed rapid growth of ICT age and there is hardly any sphere of human life where it has not laid its foot prints. Since the modern healthcare is getting shifted from cure based to care based evidence medicine. The priority is given to early diagnosis and detection of the diseases when they are in initial phase of their development. The latest next generation human genome sequencing is such an example where at early advanced stage could it be possible to see which base pair in the DNA has mutated and can lead to cancer on later stages. The same mutation could be reversed at the initial stage, thus nipping the roots in the bud.

So a woman can get a better chance of complete regaining from the cancer if diagnosed at prior stage, thus firing the need of development of efficient diagnosis techniques. A decrease will occur in the connected disease and mortality rates if timely revealing of breast cancer can become a possibility. Radiologists use the technique of screening mammography as the chief imaging modality for prior breast cancer detection because it has got the credit of being the only method of breast imaging in shrinking mortality rates related to breast cancer [6]. Normally in traditional cases, double-reading (same mammograms are read by two radiologists individually) has been encouraged in decreasing the percentage of overlooked cancers and it is at present the supreme technique incorporated in most of the screening programs instead of the fact that it earns in surplus workload and costs [7]. So a platform for the computer-aided detection/diagnosis (CADe/CADx) systems is established for backing up a single radiologist reading mammograms providing sustenance to her/his decisions [8]. ICT can show impending roles in combating this anti life threat. In fact, big data discussion has made its entry into city's talk nowadays in being a promising dimension that is expected to leave its hall mark on all major fields. Its spectrum in healthcare domain are expanding rapidly because of its increased performance in saving costs, predicting aftermaths, optimal cure within budgets and nurturing quality of health care to protect people's survival.

The suspicious lesions identified by the radiologist are sorted out by the CADx systems and lesions detection is focused by CADe systems [9]. The presented work concentrates on CADx systems. Since CADx systems archetypally has their base on machine learning classifiers (MLC) for affording diagnosis well advanced in time. A combination of forecasters is prerequisite for pronouncing the observation in order to train an MLC for breast cancer diagnosis. In order to make the inference whether

a certain surveillance is from a cancerous finding or not, a high discriminant power should be possessed by the classifier [10]. This not only being an opportunistic but also a challenging theme that has congregated the concentration of research of quite a lot of sciences, from computer vision, artificial intelligence, mathematics and statistics to medicine. Thus, an assembly of related predictors may be used for diagnosis inferring [11].

Remaining paper is fashioned as. Section 2 concentrates on the description of the datasets used in this study. It also throws a brief light on the methodology of extracting the productive feature vector from large high dimensional feature space. The brief information about the mining algorithms which serviced the present work are discussed in Sect. 3. It also brings about the classification parameters used in the evaluation of the classification accuracy of the selected classifiers. Experimental results are drawn both in tabular and graphical form and explored in Sect. 4. The results are discussed in Sect. 5 in discussion part and finally the paper is concluded with Conclusion as Sect. 6.

2 Materials and Methods

This slice designates the assessment procedure of image descriptors for breast cancer verdict. It enlightens about the extraction of the feature vector from the image mamma graphs for training machine learning classifiers to envisage the pinpointing of a lesion (Fig. 1). This section unfolds the data sets upon which the experiments were carried on, shadowed by an ephemeral enlightenment of the image descriptors that were evaluated.

2.1 Data Sets

In this study, two breast cancer datasets are castoff in order to pinpoint the general best method and classifier.

Wisconsin breast cancer diagnosis (WBCD): It is developed by the University of Wisconsin Hospital grounded exclusively on an FNA (Fine Needle Aspiration) test for breast masses finding [12]. A total of 699 clinical instances is present in this dataset, possessing 458 (65.52%) benign and 241 (34.48%) malignant. Every clinical instance is described by a set of 9 attributes having assigned integer values whose range lies from 1 to 10 and one class has a binary value of either 2 or 4 as output as a convenience for representing benign and malignant cases respectively. From this dataset 16 missing occurrences are detached in order to gain high accuracy framing out the final dataset possessing 683 clinical occurrences, with 444 (65.01%) benign and 239 (34.99%). malignant cases.

Breast Cancer Digital Repository (BCDR): It is collected form Portuguese female patients using an average age of 54.4 years old, fluctuating in the range from

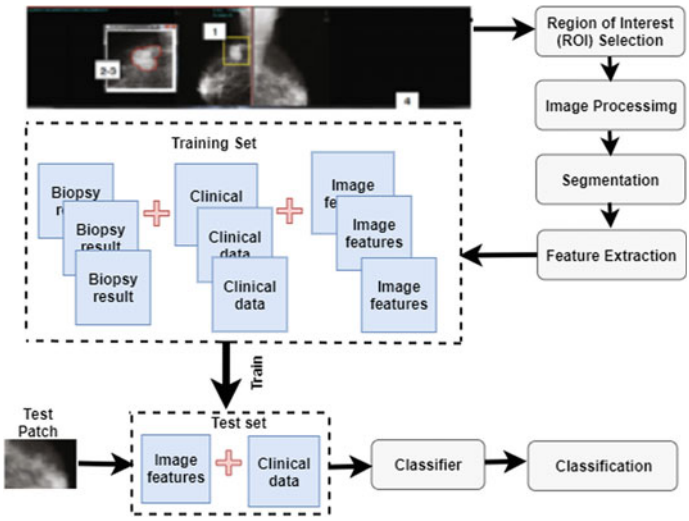

Fig. 1 Extracting image features from the image mammography lesions and training a classifier based upon same features

28 to 82 [13]. BCDR-F01 is the head data set free for public. From 362 segmentation, 187 (51.66%) share is employed by benign findings and the residual 175 (48.35%) go to malignant findings.

3 Evaluation

This chapter deals with the brief defining and working of selected data mining algorithms fruitful for our work in this paper. It also engulfs the metrics which are used as a measuring parameter for calculating the classification accuracy of the selected data mining algorithms on both the datasets. It also adds some more classification measuring parameters even if they were not directly showed in this work.

3.1 Algorithms Used

Four famous and most common used data mining algorithms in studies are used in this paper.

Naive Bayes: It is a special algorithm whose background lies in the famous foundation laid down by Bayes theorem and belongs to probabilistic method of classifiers. Bayes Classifier also known as generative model has its secret in computing class conditional probability in terms of posterior and prior probabilities. Considering a

testing instance possessing 'd' different features and having values $X = <x_1, x_2, \ldots, x_d>$ respectively. For determining posterior probability $P(Y(T) = i|x_1, x_2, \ldots, x_d)$ that the class $Y(T)$ of test instance T is I, the Bayes rule results in:

$$P(Y(T) = i|x_1, x_2, \ldots, x_d) = i).\frac{P(x_1, x_2, \ldots, x_d|Y(T) = i}{P(x_1, x_2, \ldots, x_d)} \tag{1}$$

k-NN: K nearest neighbor is an instance base learning (IBK) which is a type of lazy learning. In it the stage of training model building is often dispensed and test instance has a direct linkage with the training instances for producing a classification model. This approach crafts locally optimized model precise to the test instance.

J48: It is Decision Tree algorithm that uses a split criteria for splitting the data into the corresponding labels. The splitting condition can be single attribute known as Univariate or multiple featured known as Multivariate. The goal here is to recursively make splitting of training data for maximizing the discrimination of different classes over different nodes. Gini-index and Entropy are used to quantify the same. If $p_1, \ldots, p_k$ is the portion of the records fitting to k different classes in a node N, then Gini-index $G(N)$ of a node N is

$$G(N) = 1 - \sum_{i=1}^{k} p_i^2 \tag{2}$$

Corresponding Entropy $E(N)$ is:

$$E(N) = -\sum_{i=1}^{k} p_i \cdot \log(p_i) \tag{3}$$

The objective is to always minimalize the weighted sum of Gini-index or entropy for splitting while developing the training model.

SVM: Support Vector Machine use linear conditions to make the classification. An SVM classifier is equivalent to a single level decision tree with a very sensibly preferred multivariate split condition. Weka uses a specific efficient optimization algorithm inside Sequential Minimal Optimization (SMO) for SVM. The goal is to increase the margin of the separating hyper plane:

$$\text{Objective function} = \frac{||\overline{W}||^2}{2} + C \cdot \sum_{i=1}^{n} \xi_i \tag{4}$$

ξ_i is a Slack parameter whose purpose is to incorporate soft margins and C adjusts the importance of margin and slack necessities. Nonlinear SVM are focus of the present era of research which are learned using kernel methods. Here the pair wise dot product between different training instances and between different test instances are used as similarity values, which in turn open the gates for transformations of data into multidimensional space. Kernel function (dot product) is:

$$K(\overline{X}, \overline{Y}) = \phi(\overline{X}) \cdot \phi(\overline{Y}) \tag{5}$$

Performance evaluation of classifiers are evaluated from two different perspectives, i.e., Visualization Techniques (ROC analysis and Reject Curves) and Statistical techniques (Confusion Matrix, Precision, Recall, Sensitivity, specificity and F-Measure).

3.2 Classification Metrics

Consists a list of parameter normally used for finding out the classification accuracy of the classifier.

Sensitivity (**also called *Recall sensitivity, recall, hit rate or true positive rate* (*TPR*)**)]: Sensitivity is the share of genuine positives that are properly acknowledged as positives by the classifier.

$$\text{Sensitivity} = \text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \tag{6}$$

Specificity (**also called *True Negative Rate***): Specificity is capability of classifier to isolate negative results.

$$\text{Specificity} = \text{TNR} = \text{TN}/(\text{TN} + \text{FP}) \tag{7}$$

Precision [**positive predictive value** (**PPV**)]: Measure of relevant retrieved instances.

$$\text{Precision} = \text{PPV} = \text{TP}/(\text{TP} + \text{FP}) \tag{8}$$

Accuracy: Gives the share of correctly classified instances.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{9}$$

***F*$_1$ score** (**also *F*-score or *F*-measure or balanced *F*-score**): It is the harmonic mean of recall and precision:

$$F_1 = 2 * \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

3.3 Regression Metrics

Consists of statistical parameters like MAE, MSE and RMSE, etc.

4 Results and Discussion

In this section the outputs from experiments are calculated in the mathematical terms and the same are plotted in both tabular and graphical form.

In this paper, four most widely used data mining algorithms available from the literature are applied on the two benchmark cancer datasets using all time WEKA toolkit [14]. The algorithms used are Naive Bayes, J48, k-NN, and SVM in Weka tool and corresponding classification parameters are noted down. Weka offers a special type of Wrapper feature selection facility known as WrapperSubsetEval [15] which outputs the optimal feature subset from feature space. Two types of feature selection are widely used in the literature [16–18]:

Filter Models: Independent of the specific algorithm being used, a hard principle on a feature or set of features is the trademark of this model that has the usability in evaluating the suitability of classification [19–21].

Wrapper Models: Here algorithm and feature selection process is packed together which makes the feature selection process algorithmic specific. This technique believes on the fact that different algorithms may exertion better with different feature vectors [22–24].

The same four algorithms are applied on the modified datasets of the original datasets and corresponding classification parameters are noted down again. The results when compared showed an increase in the classification accuracy of all four algorithms. An increase from 92.6186 to 97.0123 on Wisconsin dataset and from 87.4126 to 89.5105 on BCDR-D01 dataset in case of Naive Bayes occurred. In the same way an increase from 92.9701 to 96.8366 in case of Wisconsin dataset and from 80.4196 to 90.9091 on BCDR-D01 dataset in case of J48 occurred. Similarly an increase from 96.1336 to 97.3638 on Wisconsin dataset and from 93.7063 to 97.9021 on BCDR-D01 dataset in case of k-NN occurred. Finally an increase from 97.891 to 97.9123 on Wisconsin dataset and from 91.6084 to 95.1049 on BCDR-D01 dataset in case of SVM occurred as visible from the above Table 1 and Fig. 2. The corresponding misclassification of all the four algorithms also decreased accordingly.

Table 1 Accuracy improvement on Wisconsin and BCDR-D01 datasets on four algorithms

Accuracy				
	Initial Wisconsin dataset (%)	Modified Wisconsin dataset (%)	Initial BCDR-D01_Dataset (%)	Modified BCDR-D01_Dataset (%)
Naive Bayes	92.6186	97.0123	87.4126	89.5105
J48	92.9701	96.8366	80.4196	90.9091
k-NN	96.1336	97.3638	93.7063	97.9021
SVM	97.891	97.9123	91.6084	95.1049

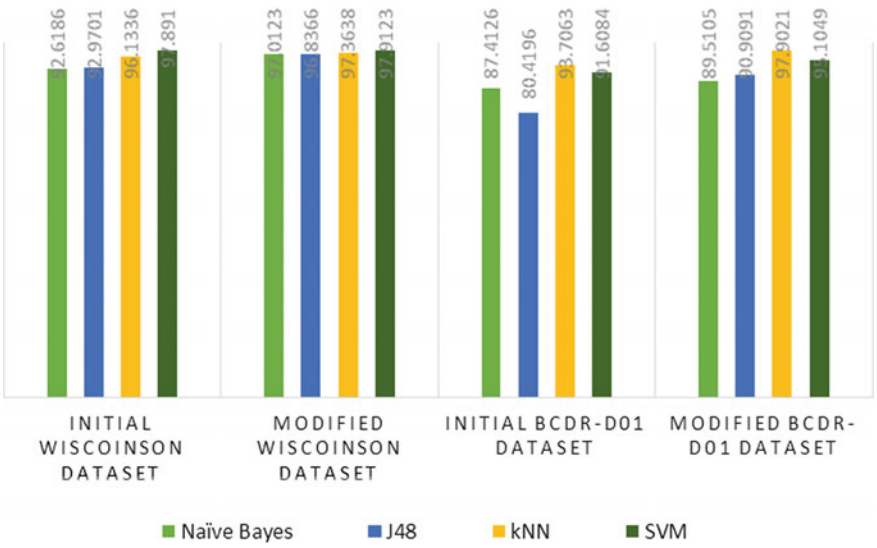

Fig. 2 Graphical representation of accuracy improvement on the modified datasets

5 Discussion

Mining of Data and Big data are the hot research topics of the last few decades and Machine Learning is the backbone for framing up a practical shape to all the concepts related to these. After the industrial revolution was witnessed by human history where the muscular energy of living beings was transformed into the moving engines, these mechanical systems now do the tons of amount of work in a very short amount of time, for which manually it was taking both a lot of effort and time. Carrying on these foundations, human brain was always in search of bringing about a new revolution where they were thirsty not only to produce motion in machines but also impart them with a sense to make judgment, take reasonable decisions and solve complex computational problems by using the prior experiences. This all gave birth to Artificial Intelligence where data is the biggest asset and to mine it for drawing conclusive results is done by different Classifiers.

In this work, we used the four most used data mining classifiers on the two medical science cancer datasets and corresponding results are noted down. The experiments are carried out in famous Weka tool. A Wrapper kind of dimensional reduction technique of WrapperSubsetEval is applied on both the datasets and again the same parameters are noted down. Initially Naive Bayes got a classification accuracy of 92.6186%, which after dimensional reduction reached up to 97.01235 on Wisconsin dataset. Same way Naive Bayes got classification accuracy of 87.4126%, which after dimensional reduction reached up to 89.5105%, on BCDR-D01_ Dataset. Carrying the same calculation on same datasets using J48, the results showed an increase from 92.9701 to 96.8366% in case of Wisconsin dataset and increase from 80.4196

to 90.9091% in case of BCDR-D01_ Dataset. Same way k-NN got 96.1336% as initial classification accuracy which got improved to 96.8366% in case of Wisconsin dataset and same way got initial accuracy of 93.7063% which upon modifying dataset reached up to 97.9021%, % in case of BCDR-D01_Dataset. Finally the SVM got 97.891% as initial classification accuracy which got improved to 97.9123% on Wisconsin dataset. Lastly, the SVM got 91.6084% as initial classification accuracy which got improved to 95.1049% on BCDR-D01_ Dataset.

6 Conclusions

The techniques of dimensional reduction has been of widely use in the literature. It has been proving as the promising result oriented treasure making the field of Machine Learning more mature. Lot of techniques of reducing the number of dimensions does exist and each one is showing good results on different type of Classifiers. Ranging from Linear Discriminant Analysis (LDA), Idempotent Component Analysis (ICA), Principal Component analysis (PCA), Generalized discriminant analysis (GDA), Backward Feature Elimination (BFE), Forward Feature Construction (FFC), etc. it has got a wide range in the IT world nowadays. In future, more dimensional reduction techniques using soft computing concepts are coming as these are the bridges for making the mining of data smooth.

Acknowledgements This work is partly supported by Research Fellowship of the "Visvesvaraya Ph.D. Scheme for Electronics & IT", Ministry of Electronics & Information Technology (MeitY), Government of India (GoI), Vide Grant no. PHD-MLA/4(39)/2015-16.

References

1. B.R. Matheus, H. Schiabel, Online mammographic images database for development and comparison of CAD schemes. J. Digital Imaging **24**(3), 500–506 (2011)
2. I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardos, J.S. Cardoso, INbreast: toward a full-field digital mammographic database. Acad. Radiol. **19**(2), 236–248 (2012)
3. A. Nandakumar, N. Anantha, T.C. Venugopal, R. Sankaranarayanan, K. Thimmasetty, M. Dhar, Survival in breast cancer: a population-based study in Bangalore, India. Int. J. Cancer **60**(5), 1–5 (2006)
4. S. Swaminathan, *Consensus Document for Management of Breast Cancer* (Indian Council of Medical Research, New Delhi, 2016), pp. 12–20
5. D.M. Parkin, Global cancer statistics in the year 2000. Lancet Oncol. **2**, 533–543 (2001)
6. H.D. Nelso, K. Tyne, A. Naik, C. Bougatsos, B.K. Chan, L. Humphrey, Screening for breast cancer: systematic evidence review update for the US Preventive Services Task Force. Ann. Intern. Med. **151**(10), 727, 1–22 (2009)
7. L. Tabar, L. Vita, T.H.H. Chen, A.M.F. Yen, A. Cohen, T. Tot, S.Y.H. Chiu, S.L.I.S. Chen, J.C.Y. Fann, J. Rosell, H. Fohlin, R.A. Smith, S.W. Duffy, Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology **260**(3), 658–663 (2011)

8. R.R. Pollán, G.M. López, C.S. Ortega, D.G. Herrero, F.J. Valiente, R.M. Solar, P.N. González, M. Vaz, J. Loureiro, I. Ramos, Discovering mammography-based machine learning classifiers for breast cancer. J. Med. Syst. **36**(4), 2259–2269 (2012)
9. J. Diz, G. Marreiros, A. Freitas, Using data mining techniques to support breast cancer diagnosis. *New Contributions in Information Systems and Technologies*, vol 1 (Springer, Berlin, 2015), pp. 689–700
10. K. Rodenacker, A feature set for cytometry on digitized microscopic images. Anal. Cell. Pathol. **25**(1), 1–36 (2001)
11. J.S. Suri, D.L. Wilson, S. Laxminarayan, *Handbook of Biomedical Image Analysis*, vol 2 (Springer Science & Business Media, Germany, 2005)
12. Data repository for machine learning. http://archive.ics.uci.edu/ml/datasets.html. Last visited 28-10-2017
13. M.A.G. López, N. Posada, D.C. Moura, R.R. Pollán, J.M.F. Valiente, C.S. Ortega, M. Solar, D.G. Herrero, I. Ramos, J. Loureiro, T.C. Fernandes, B.M.F. Araújo, BCDR: a breast cancer digital repository, in *15th International Conference on Experimental Mechanics, FEUP-EURASEMAPAET*, Porto/Portugal, ISBN: 978-972-8826-26-02, 22–27 July (2012)
14. M. Hall, The WEKA data mining software: an update. SIGKDD Explor. **11**(1) (2009)
15. R. Kohavi, G.H. John, Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
16. S. Jeyasingh, M. Veluchamy, Modified bat algorithm for feature selection with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset. Asian Pac. J. Cancer Prev. **18**, 1257–1264 (2017)
17. S.A. Josephine, K. Shannon, *Application of Data Mining Techniques in Improving Breast Cancer Diagnosis*, vol 9420 (2016), pp. 1–10
18. S. Sasikala, S.A. Balamurugan, S. Geetha, Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set. Appl. Comput. Inform. **12**, 117–127 (2016)
19. S. Sasikala, S.A. Balamurugan, S. Geetha, A novel feature selection technique for improved survivability diagnosis of breast cancer, in *2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)*, vol 50. Procedia Computer Science, Elsevier (VIT University Chennai, India, 2015), pp. 16–23
20. L.T. Vinh, S. Lee, Y. Park, B.J. Auriol, A novel feature selection method based on normalized mutual information. Int. J. Appl. Intell. **37**(1), 100–120 (2011)
21. S. Lin, S. Chen, Parameter determination and feature selection for C4.5 algorithm using scatter search approach. Int. J. Soft Comput. **16**(1), 63–75 (2011)
22. X. Lu, X. Peng, P. Liu, Y. Deng, B. Feng, B. Liao, A novel feature selection method based on CFS in cancer recognition, in *IEEE 6th International Conference on Systems Biology (ISB)* (IEEE Computer Society, China, 2012), pp. 226–231
23. M.D. MonirulKabi, M.D. Shahjahan, M. Kazuyuki, A new local search based hybrid genetic algorithm for feature selection. Int. J. Neuro Comput. **74**(17), 2914–2928 (2011)
24. T. Ruckstieb, C. Osendorfer, P.V.D. Smagt (2012) Minimizing data consumption with sequential online feature selection. Int. J. Mach. Learn. Cybern. **4**(3), 235–243 (2012)