Breast cancer diagnosis using SVM classifier with PSO

Ashima Kalra , Bhawna Tandon

Chandigarh Engineering college,Landran,

uppal.ashi@gmail.com, bhawna.ece@cgc.edu.in

Abstract.

Since the past four decades, India's investment in cancer research has led to a greatly improved understanding of the numerous diseases known as cancer, taking into account some of the biological pathways that contribute to the formation, progression, and spread of cancer. This knowledge has improved the therapy for some tumors and allowed patients suffering from conditions that were typically deadly to live longer. Numerous risk factors arerecognized by the National Cancer Act for the classification of cancer based on type. Problems arise when cancer is divided into different types. In our proposed work, we classify dataset1 in 5 type of cancer and dataset2 in 6 type of cancer. The classification is performed when these datasets are compatible with our classification method. So we perform pre-processing step and then Particle Swarm Optimization is used for feature selection of each gene. Now on basis of features, Support Vector Machine performs well to classify data in predicted classes. Accuracy of classification method iscompared with proposed method.

***Keywords*—Support Vector Machine, Particle Swarm Optimization**

1. INTRODUCTION TO BREAST CANCER

The scientific review led to recommendations based on evidence on breast cancer screening for women of average risk, intermediate and high. From the beginning, breast cancer has been fought with therallying cry of early detection. There is no known exact cause of breast cancer, but evidence suggests that it must be caused by a combination of inherited and environmental factors. Our current scientific knowledge of the disease hampers efforts to prevent breast cancer. Many established breast cancer risk factors, such as being a woman, age, breast density, family background, heredity, or a past breast cancer diagnosis, cannot be changed.

Positively, research shows that maintaining a healthy body weight, engaging in physical activity, eatingsensibly, and limiting alcohol intake can all lower therisk of developing breast cancer. However, early detection has been the foundation of hope and action for breast cancer because personal risk reduction is still not a guarantee. Early detection will continue to be our top strategy for lowering breast cancer-related death and sickness until we can successfully prevent the disease.

Breast cancer detection is the process of diagnosing the disease earlier than would otherwise be the case. Breast cancer screening, also known as secondaryprevention, is the routine testing of people who do not have any symptoms with the goal of finding breast cancer as early as possible so that appropriate treatment can be provided. Breast cancer mortality in Canada has decreased by an estimated 25–30% since the late 1980s due to the introduction of breast cancerscreening.

The recommendations suggest specific technologies for each of the three risk groups. Digital or film mammography is specified for average-risk women, digital supplemented by ultrasound in intermediate- risk women, and digital supplemented by MRI in high-risk women. The differences in technologies are related to differences in breast density and tumor characteristics in higher-risk women. Some imaging systems are better than others for detecting cancer or dismissing it given certain characteristics of an individual woman’s breasts.

2. RISK FACTORS OF BREAST CANCER CLASSIFICATION

Since the recommendations use three risk groups to sort out differences in screening routines, it may be wondering how to fit the risk profile. Some people would consider just being a woman to be the most common risk factor for breast cancer. While men do get breast cancer, women are about 100 times more likely to get it. Thus the vast majority of women (at least 80%) are considered “average” risk [1]. The prevalence of breast cancer increases with age, which makes advancing age the most commonly recognized risk factor. The median (middle) age for a breastcancer diagnosis is 61.

After age, the risk of breast cancer increases significantly with family history, certain benign lesions, breast density, a history of previous breast cancer and hormonal factors. According to Dr. Eisen [2], only about 20% of women who have had breast cancer have had a first or second degree relative who also had it. About 5% of women with breast cancer have a very strong family history of the disease, and about a quarter of high-risk cases are due to known genetic mutations such as BRCA1 and BRCA2. That leaves about three quarters of familial risk unexplained, possibly due to environmental factors [3], but more likely genetic factors yet unidentified.An array of other risk factors has been explored. Some lifestyle factors like obesity, particularly in postmenopausal women, can increase risk. A specific diet to reduce breast cancer risk has been hard to pin down [4], but women who exercise regularly and vigorously may decrease their risk. Alcohol is a known risk factor. Low Vitamin D is also a possible risk factor.Reproductive factors also affect breast cancer risk. Both early menarche (the age of first menstruation) and late menopause are associated with increased risk. Breast cancer risk also increases with null parity (having no full-term births) or having a first birth at alater age [5]. According to Dr. Eisen [2], it was common over the past century for women to start families in their twenties and to continue giving birth well into their 30s. Today, women are more likely to have a first birth in their 30s [6]. So it is possible that the social trend of having families later may be increasing breast cancer risk.

Exposure to estrogen is an important risk factor for breast cancer because these hormones stimulate tumor growth. Women who are menopausal and obese have higher circulating estrogen. Excess estrogen can also be ingested with birth control pills or hormonal replacement therapy after menopause [7]. After 2001, breast cancer incidence appeared to decrease with a substantial drop in prescriptions after hormone replacement therapy was recognized as a possible breast cancer risk.

3. PROBLEM STATEMENT

New methods have been used in the framework of the human genome model to make it easier to implement experiments in parallel on a large number of genes at once. A notable example are DNA microarrays, also referred to as DNA chips. This method focuses on simultaneously measuring the mRNA levels for numerous genes in specific cells or tissues. A tumour biopsy's array is hybridised after the mRNA has been extracted, tagged, and processed. The intensity value that results from measuring the quantity of label on each spot should be connected to the abundance of the matching RNA transcript in that sample.

Let Y= {y_1,y_2, y_n} be the random variables for gene g1, g2----- gn respectively. Let C be the random variable for the class labels such that C={1,2,3------k) where k is the no of classes .Let t={t.y_1,t.y_2, ------ t.y_n}be the expression values corresponding to n genes. We aregiven be with the training set of m tuples

T ={(t_1,c_1),(t_2,c_2),(t_3,c_3)(t_m,c_m)}.where ci is the class label of the tuple ti..Let X=(t_1,t_2,-----t_s) be the test set.

A classifier is the function with two arguments T and X where T is the training set and X is the test set. We have to identify the class label of all the unknown samples in the test class by using the knowledge or information available from the training sample and then check the accuracy the classifier. Classification accuracy is defined as the number of samples which is predicted accurately by using the classifier trained on the training samples.

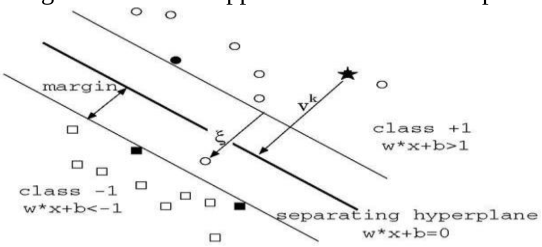
4. HYBRID METHODS

There are two methods used for implementation .first is support vector machines and second is particle swarm optimization.

4.4.1 Support vector machines

Support vector machines have the finest theoretical and outstanding empirical level of start-up. They have been used for projects like database classification, object detection, and handwritten digit detection. Support-vector machines are first presumpted for numerical data. The basic idea underlying SVM is to look at separators in the search space that can most effectively separate the various cancer kinds. With SVM, we can anticipate the lowest true error if we are able to identify a pattern for the concept of very low risk. Its likelihood for h is that randomly chosen test objects will develop a problem. .The concept is shown well as shown in fig 1 .

Figure 1: How a Support Vector Machine operates



The main advantage of the SVM technique is that since it tries to discover the optimum distinction in the feature space by defining the appropriate hybrid[8] of features, it is quite robust to high dimensionality.

4.4.2 *Particle Swarm Optimization*

PSO is a technique that uses numbers. Swarm intelligence was the fundamental theory used to develop this method [9]. PSO was created based on the social behaviour of flock-survival birds. There are a lot of responses in the beginning of this procedure, or rather, a counting of replies in which any random pick is termed a particle. This method operates in an iterative manner and approaches the ideal outcome. Every particle makes an effort to move across space at a dynamic velocity vi that directs the particle toward the best possible outcome at each iteration. In the search space, n particles are initially distributed at random [10].

5. PROPOSED METHODOLOGY

Flowchart of proposed methodology is shown in figure . Microarray experiments allow us to quantify the expression level of thousands of genes instantly. These either monitor each gene several times under distinct conditions or each gene into a single state, but in the type of different tissues. First type of experiment identify the genes which are related to each other due to common expression while the latter type of experiment identify the genes whose expression are good diagnostic indicator. In order to obtain the significant information from the gene expression profiles various methods have been used. Firstly we need to prepare data which compatible for our proposed work. The step performed before implementation is known as pre-processing. Following certain steps are required to perform:

1. Select the dataset1 and dataset2 from a hugdatabase.
2. Perform pre-processing step to make datasetcompatible for our proposed method.
3. Define classes of cancer type and applySVM algorithm for classification of dataset on basis of classes.
4. Evaluate the results.
5. Define the classes of cancer type and use a feature selection technique PSO. The selected feature categories according to classes type using SVM method.
6. Evaluate the results.
7. Compare the results of both techniques.

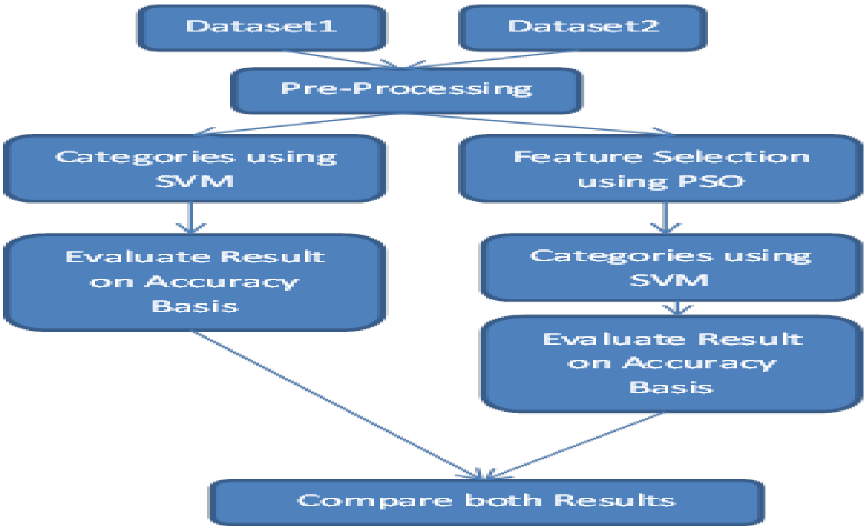

Dataset1
Dataset2
Pre-Processing
Categories using SVM
Feature Selection using PSO
Evaluate Result on Accuracy Basis
Categories using SVM
Evaluate Result on Accuracy Basis
Compare both Results

Figure 2: Flowchart of proposed work

6. DATABASE

1. Dataset 1:GSE22226-GPL1708:

It consists of 129 columns and 44270 rows. Column represents the no of patients or samples while rows represent genes corresponding to each patient. It consist of 5 classes of cancer named as luminal A, luminal B,basal-like,her2 enriched and normal .The no of samples corresponding to each type of cancer is shown in table I.

2. Dataset 2: GSE10866-GPL1390:

It consists of 199 columns and 22575 rows. Column represents the no of patients or samples while rows represent genes corresponding to each patient. It consist of 6 classes of cancer named as luminal A, luminal B,basal-like,her2 enriched , normal and claudin. .The no of samples corresponding to each type of cancer is shown in Table II.

Table I: Patients corresponding to cancer type

Cancer Type	No of patients
Luminal A	32
Luminal B	25
Basal-like	43
Her 2 enriched	21
Normal	8

Table II: Patients corresponding to cancer type

Cancer Type	No of patients
Luminal A	32
Luminal B	25
Basal-like	43
Her 2 enriched	21
Normal	8
Claudin	21

7. RESULT ANALYSIS

Many types of database are available on which different type of existing and proposed classification methods implemented. Performance of each method provides us the merits and demerits regarding that method. We assign two different datasets to the proposed method. Each dataset have 129 and 199 samples respectively with multiple classes. These datasets are assigned as input of SVM classifier and evaluate the performance. Some issue generates in front of us one of them is feature selection. This issue resolved using PSO technique with SVM classifier. The same both datasets are assigned to this hybrid approach and evaluate the performance. Now we compare both results and get a conclusion about classification method. Different reading has been taken that based on the different proportion of training and testing set.

1. Results Evaluation for Accuracy on Dataset 1:GSE22226-GPL1708

Accuracy of SVM without PSO

Accuracy of SVM with PSO

Table III: Results of SVM classifier on dataset 1

Table IV: Results of PSO with SVM classification on dataset 1

SVM	Training Testing Ratio
63.18%	80:20
61.49%	75:25
52.79%	70:30
43.83%	65:35
59.14%	60:40

SVM	Training Testing Ratio
65.41%	80:20
61.90%	75:25
53.05%	70:30
56.69%	65:35
58.47%	60:40

Table III represents the results of SVM classification which consist of 129 samples. Each sample consist 44270 genes. There are no genes available which know its parent sample. We can say there is no technique used for feature selection of each gene. There are 5 cancer classes exists in this dataset. The classification performed on these types classes of cancer. Different type of cancer classes are named as luminal A, luminal B, basal-like, her2 enriched and normal cancer.Table IV represents the accuracy result of SVM classification with PSO feature selection technique which has been implemented on dataset 1 in a way to increase the classification accuracy by selecting certain number of genes.

2. Results for Dataset 2: GSE10866-GPL1390

Accuracy of SVM without PSO.

Table V: Results of SVM classification on dataset 2

SVM	Training Testing Ratio
78.83%	80:20
81.33%	75:25
80.19%	70:30
78.26%	65:35
74.16%	60:40

Accuracy of SVM with PSO

Table VI: Results of PSO with SVM classification on dataset 2

SVM	Training Testing Ratio
79.88%	80:20
79.67%	75:25
79.60%	70:30
78.93%	65:35
72.38%	60:40

Table V represents the results of SVM classification which consist of 199 samples. Each sample consist 22575 genes. There are no genes available which know its parent sample. We can say there is no technique used for feature selection of each gene. There are 6 cancer classes exists in this dataset. The classification performed on these types classes of cancer. Different type of cancer classes are named as luminal A, luminal B, basal-like, her2 enriched and normal cancer and claudin. Table VI represents the result of KNN classification with PSO feature selection technique on dataset 2 which consist of 199 samples. Each sample consist 44270 genes. This technique has been implemented in a way to increase the classification technique because large no of features can make the classification process complex.

8. CONCLUSION & FUTURE SCOPE

Basic research is always required for further new proposal. It is key of transformative to discoveries about classification of cancer type. A mechanism is used for classification of data. Current research work delivery emphasizes overwhelmingly the classification of acute cancer disease rather than protection and preservation of overall health. The instability performance of SVM classifier helps in categories dataset in different type of cancer. The top level accuracy is 61.49% for dataset 1 and 80.18% for dataset 2 by SVM classifier without using feature selection technique. So it is transparent to us that accuracy of SVM classifier is high when dataset has more samples. Now we deploy a feature selection technique PSO with SVM classifier. This deployment is implemented on both same datasets and evaluates results. The top level of accuracy of SVM with PSO during execution is 65.41% for dataset 1 and 79.88% for dataset 2. This is concluding that SVM with PSO perform very well fewer samples in dataset.

REFERENCES

1. Reuben SH. Promoting healthy lifestyles: policy, program, and personal recommendations for reducing cancer risk. President's Cancer Panel 2006-2007 annual report. Bethesda (MD): National Cancer Institute; 2007.
2. Dr. Eisen, "Earlier Detection and Diagnosis of Breast Cancer: A Report from its About Time! A Consensus Conference" Canadian Breast Cancer Foundation - Ontario Region, Amended October 13, 2010.
3. Reuben SH. Reducing environmental cancer risk: what we can do now. President's Cancer Panel 2008-2009 annual report. Bethesda (MD): National Cancer Institute; 2010.
4. World Cancer Research Fund International/American Institute for Cancer Research Continuous Update Project Report: Diet, Nutrition, Physical Activity, and Breast Cancer Survivors. Breast Cancer Survivors Report 2014.
5. Zikmund-Fisher BJ, Fagerlin A, Ubel PA. Risky feelings: why a 6% risk of cancer does not always feel like 6%. Patient Educ Couns. 2010.
6. Lindsay M. Monte and Renee R. Ellis, "Fertility of Women in the United States: 2012", U.S. Department of Commerce Economics and Statistics Administration,U.S. Census Bureau, Issued July 2014.
7. Brar, K.K., Kalra, A., Samant, P. (2020). Computer-Aided Textural Features-Based Comparison of Segmentation Methods for Melanoma Diagnosis. In: Jain, S., Sood, M., Paul, S. (eds) Advances in Computational Intelligence Techniques. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-2620- 6_6.
8. Kalra, A. and Pahwa N., Heart Disease AnalysisUsing Crossover with Multi-Class Support Vector Machine Method, INTERNATIONAL JOURNAL OF RESEARCH INELECTRONICS AND COMPUTERENGINEERING, IJRECE VOL. 7 ISSUE 1 (JANUARY- MARCH 2019) pp494-501.
9. Ajith Abraham, Crina Grosan Vitorino Ramos (Eds.) E- book "Swarm intelligence in Data mining", (April,2006).
10. Tanya Taneja, Balraj Sharma, "Text Classification Using PSO & Other Technique", International Journal of Recent Development in Engineering and Technology, Volume 3, Issue 1, July 2014.

Biographies

Currently working as an Assistant professor in Electronics and Communication Department in Chandigarh group of colleges, Landran, (Mohali)Punjab,India. Gold Medalist in B Tech in Electronics from Kurukshetra University, Kurukshetra in 2003 .Received M tech degree from Punjab Technical University, Jalandhar in 2008 and pursuing Ph.D in the field of soft Computing. She has teaching experience of 18+ years at Post Graduate and Under Graduate Level. Her research activities include designing model identification using neural networks, fuzzy systems, supervised learning, machine learning. Published more than 40 research papers in reputed journals .Published4 book chapters in springer series and 5 text books .Published 6 Indian patents.

Dr. Bhawna Tandon received her BTech in Electronics and Instrumentation in 2001, ME in Instrumentation and Control in 2009 and PhD in Control System from Kurukshetra University, Kurukshetra, Panjab University, Chandigarh and PEC (Deemed to be University) Chandigarh respectively. She is having teaching experience of 20 Years and currently working as an Associate Professor at Chandigarh Engineering College, Landran, Mohali. She is having more than 20 research publications in reputed International journals. Her current research interests include robust control, non-linear control and optimisation techniques.