Optimizing Breast Cancer Gene Data Analysis: A Review on Feature Selection and Classification Techniques

Priya Deshmukh, Monika Patil, Ujwal Joshi, Rasika Kulkarni,
Shweta Pawar & Pournima Naik*ÿSavitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT

Cancer classification is very important in the field of bioinformatics for cancer diagnosis and drug discovery. There are two problems that have been examined in bioinformatics for cancer classification i.e. class discovery and class prediction. Accurate prediction of different type of cancer is very important for providing better treatment and for avoiding the additional cost associated with wrong therapy. Large numbers of methods have been proposed in recent years for classifying the cancer but still lot of problems exist which need to be addressed. In this survey paper, we present various cancer classification methods such as SVM, KNN, Naïvebayes as well role of feature extraction method for classifying gene expression data.

KEYWORDS: Support Vector Machine, Particle Swarm Optimization, Principal Component Analysis, K nearest Neighbor, relief, gene data, Breast cancer.

INTRODUCTION

Research for cancer is one of the central research in medicinal field [8]. Breast cancer is the most common type of cancer among the women with high mortality rate. Different subtype of cancer respond differently to treatment and therefore it is essential to classify the cancer so that accurate or better treatment can be given to patient and it also reduces the medical cost associated with the unnecessary treatment [7]. Various approaches for diagnosis (to detect the particular type of cancer) and prognosis (to predict how it will behave in future) exist that are based on the microarray gene expression data but the problem with using this expression data is its higher dimensionality; few samples for too many genes but this problem can be overcome by using data preprocessing techniques such as feature extraction or feature selection which help to reduce the dataset size also improve the performance. Some of the key issues in using the microarray expression dataset are: 1) curse of dimensionality. 2) High level of noise in data. 3) Large number of irrelevant features which does not have significance. There are various clinical covariates associated with the breast cancer such as age, size, grade, histological grade, ER, PR, HER and subtype etc [1]. These covariates help in the diagnosis and prognosis of breast cancer and to provide the better treatment. The various subtypes of breast cancer are Luminal A, luminal B, basal-like, normal breast cancer and HER2 [1]. Histological or tumor grade of breast cancer provides clinical prognostic information. Prognostic factor in breast cancer can be determined by tumor grade. About half of the patient suffering from breast cancer has histological grade 1 or 3 status (with low or high risk of recurrence) and others have histological grade of 2. The emerge of post-genomic technology has provided an opportunity to decipher the genome origin of human diseases. Thus the gene expression analysis using microarray of DNA allowed improved classification and prognosis of breast cancer or cancer of other types also. Several studies have produced different signature for the diagnosis and prognosis of breast cancer. Some of signatures are 1) Mamma print (70 gene signature) which classify patients into good or poor prognosis group. 2) Veridex (76 gene signature) which identify the patient with high risk of metasis. 3) Oncotype DX (21 gene signature) which predicts likelihood of breast cancer recurrence. It has been believed that the cancer is derived from the driver genes that change the large amplitude representation of the genes that interact with the driver genes. In many patients, Microscope or clinical evident metastasis have already occurred by the time the primary cancer was diagnosed. Certain treatments like chemotherapy and hormonal therapy reduce this risk of metastases. Therefore it is challenging task to predict accurate outcome that will help the physician to give the accurate treatment to the patients. The key challenge in the microarray expression data is to reduce the dimension of the data because only the small amount of genes are required to diagnose a particular type of cancer which can be achieve with the help of feature selection. The large number of features led to poor generalization and high execution time [1]. The main idea behind gene selection is to eliminate the genes which do not have any significance. Several methods for feature selection exist such as PSO (Particle swarm optimization), I-relief, PCA (Principle Component Analysis). There are also some wrapper based and filter based methods which refined the features based on some fitness criteria. After feature selection our next aim is cancer classification. This classification is done in order to predict the type of cancer person is having so that accurate treatment can be given to the person on time. Different methods of classification exist each having its own advantages and disadvantages. Some of the methods of classification are SVM (support Vector Machine), KNN (k nearest neighbors), and Naïve Bayes.

Steps of classification

One of the main challenges with gene expression data is classification of different type of cancer into correct types. The problem with the gene expression data i.e. the curse of dimensionality, the small number of samples

with large features, makes the classification task more difficult and challenging. In past several decades, many dimension reduction as well as classification methods have been proposed. Each classification method involves learning phase in which known samples are used to train and test samples are used to predict the accurate class of unknown samples .some classification methods that have been applied to cancer classification are SVM (support vector machine), Naive bayes, K nearest neighbor.

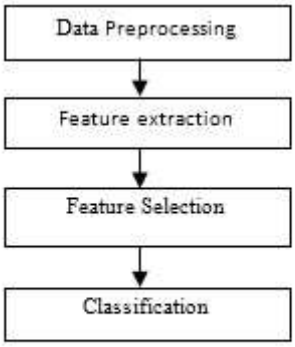

Figure 1: Steps Of Classification

Feature Selection

The issue with the microarray expression data is its large dimensionality which leads to poor generalization and high execution time so there is great need for feature selection. The key concept of feature selection is to eliminate the irrelevant feature. Irrelevant features can be removed by feature selection .Several methods for feature selection exist such as PSO (Particle swarm optimization), I-relief, PCA (Principle Component Analysis). There are also some wrapper based and filter based methods which refined the features.

Particle Swarm Optimization(PSO)

PSO was discovered by Dr.Eberhart and Dr.Kennedy which is an optimization technique to optimize a problem by repetitively trying to enhance the candidate solution [14]. The main idea behind this technique was derived by the social behavior of bird flocking and fish spooling. It is very simple concept which requires some mathematical operators and also inexpensive in term of storage as well as speed [16] [15]. It resolves the problem by initializing the system with the population of random solution and keep on moving the particle for discovering the optimal solution**.** The movement of the particle from one place to another place depends on 2 things namely the motion of the particle to its optimal position and secondly the motion of the particle to the most optimal position with respect to its neighbor particles. It basically requires the two vectors .one is the position vector and other is the position vector. Position vector depends upon the velocity vector[15][16]. There is repetition of updating the position and velocity of particle for large no of generations and the process is continued until we reach the objective or if the maximum no of iterations is reached. The movement of each particle is affected by its local best position and is focused toward its best known position. It is very much similar to genetic algorithm .The main advantage of using PSO is it is simple to understand also require very few parameters for its implementation.

Binary PSO

Initially PSO was developed only for continuous search space and later on it has been extended to discrete valued search space In binary PSO ,each particle is represented by two vectors i.e. velocity vector or position vector. The possible values for the position vector are 0 or 1 which decides whether to select that feature or reject. If the value of the position vector is 1 it is selected otherwise rejected .The particle position is defined by the vector $X_i=(X_{i1},X_{i2},X_{i3}\ldots\ldots\ldots X_{iM})$ where X_{iM} represent the M^{th} dimension of the i^{th} particle position and the velocity vector is given by $V_i=(V_{i1},V_{i2},V_{i3}\ldots\ldots\ldots V_{iM})$.local best position is given by the $PB_i=(PB_{i1},PB_{i2},PB_{i3}\ldots\ldots\ldots PB_{iM}$ and the global best position is given by the $GB=(g_1,g_2,g_3,\ldots\ldots\ldots g_M)$.The velocity vector is given by the equation (a)

$$v_{id}^{k+1} = \left(wv_{id}^{k}\right) + r_1C1(p_{id} - x) + r_1C2(g_d - x_{id}) \qquad \text{(a)}$$

Where I = (1,2,3…..n) where n is the number of the particle and m is the dimension of the multidimensional space. R1 and R2 are uniform random nos. C1 and C2 are two constants. The position of the particle is given by the equation (b)

$$x_{id}^{k+1} = x_{id} + v_{id}^{k} \qquad \text{(b)}$$

Recursive PSO

In this approach the PSO is applied on each step .Initially the whole search space is randomly explored and then at each successive step the search space is refined .The method is applied recursively until there is no reduction of feature set.

RPSO is applied on different levels. At each level the number of features are reduced and the method stop when there is no further reduction and returns the set of selected feature .The figure 1 below illustrates the process of recursive PSO .Initially we filter out set of feature using linear SVM. At first level we apply the PSO it extract only 10 features then again the PSO is applied on these extracted features and extract only the features which have value 1 and rest all are eliminated so it extract only 6 features and again apply the PSO .The process is repeated until we left only with the feature which cannot be refined further so at last we left with only 3 features.

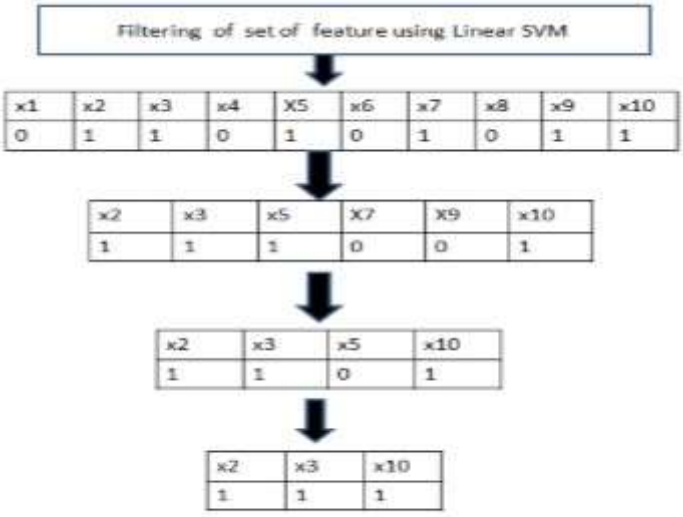

Figure 2.

Principal component Analysis(PCA)

Principal Component Analysis (PCA) is a statistical method used to limit high dimensional data while retaining the number of variations. It was discovered by pearson [1901] and hoteling [1933]. It was put into ecology in by Goodall [1954] under the name 'factor analysis first'. PCA uses orthogonal transformation so that correlated variables can be translated into linearly uncorrelated variables called principal components. This translation is done in such a way that number of principal component after translation is less than or equal to the number of original variables. The resulting vectors are uncorrelated orthogonal set. PCA provides lower dimensional view since the visualization of high dimensional dataset is complex.

Working of PCA:

Step1*:* Compute the mean of the data matrix.

Mean is calculated by the sum of observation divided by total no of observation.

Step 2: Subtract the mean from each feature of sample.

Step 3: Compute the covariance matrix.

For this step variance of each dimension and covariance between dimensions is needed. The diagonal terms should be variances in covariance matrix and other terms will be covariance.

Covariance and variance is calculated by equation given below:

$\mathrm{Cov}_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)} = \frac{\sum(xy)-n\bar{x}\bar{y})}{(n-1)}$

$\mathrm{Var}_{x} = \frac{\sum_{i=1}^{n}(x_i-x)^2}{n}$

Step 4: Compute the eigen values.

The next step is to get eigen values by solving a determinant function.

The equation of determinant function is given by

$(A - \lambda I) = 0$

The calculated eigen values will be the sum of variance. Sum of eigen values will always equal to the sum of variance.

Step 5: Compute the eigen vectors according to eigen values.

After getting eigen values there is a need to calculate the eigen vector by solving a matrix X in such that :
$[A - \lambda I] * [X] = [0]$.

Step 6: Arrangement of eigen vectors.

Once eigenvectors are detected, the very next step is to arrange them in decreasing order i.e. from highest to lowest, the eigenvector with most eigen value will be the first principal component and the Components with lower eigen values have lesser importance so they can be ignored. This will provide final dataset with reduced dimension.

Step 7: To acquire the coordinates

The next step is to acquire coordinates of data point in the direction of eigen vector. We can acquire this by multiplying centered data matrix to the eigen vector matrix. Variance of the projection on the line of principal components is to be obtained which is equal to the eigen values of the principal components. First eigen vector is able to explain about 99% of the variance.

Step 8: Compute the feature matrix

From the choosen components feature vector matrix is to be formed by taking the eigenvectors that are choosen.

Data Classification Techniques

Different classification techniques exist for classification exists and have been applied for different purposes. Some of common methods are

Support Vector Machine(SVM)

SVM was invented by Vapnik and co-workers and then extended by other researchers [19]. It most widely used algorithm for classification. It is supervised learning algorithm in which some known sets of data is used to train the machine known as training set and the test data is to predict the type of cancer for an unknown sample.SVM constructs hyper plane or set of hyper plane in high dimensional space and the hyper plane which has maximum distance to the nearest data point in training set is the best choice [19] [18]. On both sides of hyper plane a margin is defined as the width that the boundary could be increased by before hitting a data point.

Linear SVM

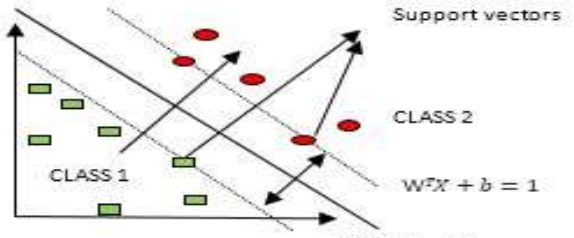

Figure 3

SVM mainly focuses on the training vector that lie exactly on the margins and is known as the support vectors. The decision boundary of a linear classifier can be written as follows

$$W.X + b = 0 \ .$$

The goal of the linear classifier is to classify all the training data into correct classes.

$$WX_i + b <= 1 \quad If\ y_i = +1$$
$$WX_i + b >= 1 \quad If\ y_i = -1$$
$$Y_i(WX_i + b >= 1) \quad \text{For all i}$$

And the second goal is to maximize the margin M=2/|w and is same as minimizing 1/2w^{t}w.
Now we have Find **w** and b such that

$$\phi(w) = \frac{1}{2}W^TW \text{ is minimized;}$$
$$\text{And for all } \{(X_i, y_i)\}: y_i(W^TX_i + b) \geq 1$$

This optimization problem can be solved using Lagrange multiplier.

Non-Linear SVM

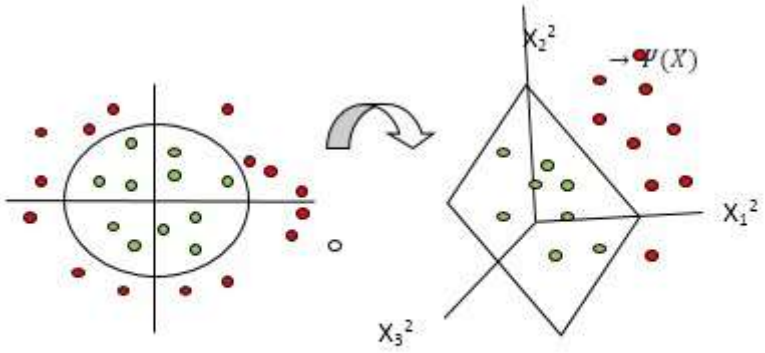

Figure 4

Sometime there may be a situation when it is not possible to construct a linear hyper planes between the training data points because the data points are very closely related to each other, in such a situation we need to map the data in low dimensional space to high dimensional space via some transformation **x → φ(x),** which could be possible only with the use of kernel functions. A kernel function specifies the inner product of data points in expanded feature space. Some of the kernel functions are:

- Linear kernel:
$$K(x_i, xj\text{--}) = x_i^T xj$$
- Polynomial kernel:
$$K(x_i, xj\text{--}) = (1 + x_i^T xj)^p$$
- Radial-basis kernel:
$$K(x_i, xj\text{--}) = exp(-\|x_i - xj\|^2 \ / 2\sigma^2)$$
- Sigmoid kernel:
$$K(x_i, xj\text{--}) = tanh(\beta_\circ x_i^T xj + \beta_1)$$

MULTI-SVM

SVM is well known method for classification in machine learning for binary classification problem with 2 classes' i.e. positive class vs. negative class. However sometimes there may be a situation when we have more than 2 classes as that the case in breast cancer classification .The solution to this is to implement the SVM with multiple classes and is generally known as multi-class SVM's. Multi-class SVM's are implemented by combining several binary SVM's [3]. Multi-class SVM is difficult to implement as compared to binary SVM because output could be more than one class and must be classified into mutually exclusive classes .There are various ways to implement multi-class SVM such as one against one and one against all classifier.

One against All SVM

For N class problem (N>2) N binary classifiers are constructed [3].The i^{th} SVM is trained with all sample in i^{th} class labeled as positive and all other samples as negative.

One Against One SVM

In this we construct N(N-1)/2 classifiers [3]. Each classifier is qualified with samples of first class as positive and the samples of other class as negative and then the max –win strategy decide to which class sample belong.

Applications

SVM gives good performance for binary classification and has been used successfully for variety of problems such as in bioinformatics it is used for cancer classification and protein classification, for image classification, text and hypertext classification and hand-written character recognition

Bayesian classification

Bayesian classification is supervised learning model for classification which is probabilistic in nature and is based on bayes theorem and hence known as naive bayes classifier [5]. Naive bayes is simple technique for constructing the classifier which is based on the assumption that value of a particular feature is independent of the value of the feature. It uses the knowledge of prior event in order to predict the new events and is based on method of maximum likelihood. The major advantage of this learning model is that it requires only small amount of training data in

order to determine the accurate parameter which is necessary for classification. It is conditional probability model in which the test sample x is represented by vector $x=\{x_1,x_2,x_3,- - - - x_n\}$ with n distinct features is classified according to the maximum posteriori probability

i.e. $\text{class}(X) = \text{argmax}(\log p\left(\frac{M_i}{X}\right))$

Where $p(M_i/x)$ is posteriori probability that M_i is true given the test sample x.
By Bayes rule,

$$P\left(\frac{M_i}{X}\right)\ p(X) = p\left(\frac{X}{M_i}\right)\ p(M_i)$$

By assuming equal prior probabilities, we obtain

$$\text{Class}(X) = \text{argmax}(\log p\left(\frac{X}{M_i}\right))$$

I.e. the test sample is being classified into the class for which samples have greatest likelihood.

K Nearest neighbor(KNN)

KNN is the one of the simplest supervised learning algorithm in machine learning used for the classification which is based on the similarity measure. It is a non-parametric lazy learning technique in which new samples are classified based on the distance function. As it is lazy learning algorithm it makes decision based on entire training data. KNN is based on certain assumption that data is in feature space and can be scalar or multidimensional vector. Given the dataset, each of the training samples is represented by set of vectors and a class associated with each vector. A test sample is classified by a majority vote of its neighbors i.e. test sample belongs to the class which is most common among its k nearest neighbor measured by distance function. We are given with the number k which decides how much neighbor influence the classification. When the value of k=1 then sample is assigned to class of nearest neighbor

KNN method can either work for 2 class problem one of which is positive and other is negative and can also perform well with arbitrary number of classes.

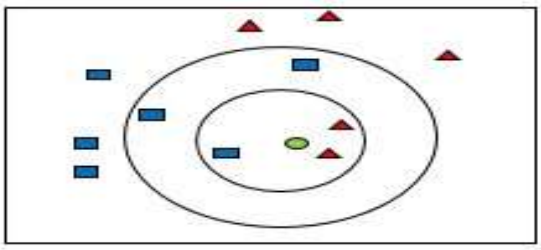

Figure 5

As an example of KNN classification we want to classify the sample which is in green to either of one blue square or red triangle. When k=3 then the test sample is classified to class of red triangle because of its majority as represented in inner circle. When k=5 then it is classified to class of blue square because of the majority of blue square as represented in outer circle.

CONCLUSION

Different classification techniques have different domains .We present our conclusion for different technique used in this paper. SVM gives better performance for binary classification and generally require large data sample and can also be used for multiclass problems. However KNN classification requires large amount of storage space and also very sensitive for feature selection but can also perform when sample size is small. Naive bayes algorithm requires little storage space for training and classification and is also robust to missing values by simply ignoring these values while computing the probabilities.

Classifier	Advantages	Disadvantages
SVM	1) It provides the ability to handle the large feature space. 2) It can control problem of over fitting. 3) Only support vectors are used to specify the separating hyper plane.	1) It does not give accurate results for classifying the multiclass problem 2) It does not identify the attributes which are most useful for classification

Naive Bayes	1) It is scalable and robust to noise. 2) Gives good performance even if input size is small.	1) It is based on assumption that class attributes values are independent. 2) It assumes equal prior probabilities.
KNN	1) Simplest method for classification.	1) Its error rate is at most twice when compared to Bayesian.

REFERENCES

[1] Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., & Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, *104*(4), 311-325.

[2] Desper, R., Khan, J., & Schäffer, A. A. (2004). Tumor classification using phylogenetic methods on expression data. *Journal of theoretical biology*,*228*(4), 477-496.

[3] Chamasemani, Fereshteh Falah, and Yashwant Prasad Singh. "Multi-class support vector machine (SVM) classifiers--an application in hypothyroid detection and classification." In *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on*, pp. 351-356. IEEE, 2011.

[4] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1-3), 389-422.

[5] Keller, A. D., Schummer, M., Hood, L., & Ruzzo, W. L. (2000). Bayesian classification of DNA array expression data. *Technical Report UW-CSE-2000-08-01*.

[6] Li, H., Zhang, K., & Jiang, T. (2005, August). Robust and accurate cancer classification with gene expression profiling. In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE* (pp. 310-321). IEEE.

[7] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., ... & Desmedt, C. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.*Journal of the National Cancer Institute*, *98*(4), 262-272.

[8] Lu, Y., & Han, J. (2003). Cancer classification using gene expression data.*Information Systems*, *28*(4), 243-268.

[9] Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A. M., ... & Daidone, M. G. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen.*BMC genomics*, *9*(1), 1.

[10] Saini, A., Hou, J., & Zhou, W. (2014). RRHGE: A Novel Approach to Classify the Estrogen Receptor Based Breast Cancer Subtypes. *The Scientific World Journal*, *2014*.

[11] Sun, Y., Goodison, S., Li, J., Liu, L., & Farmerie, W. (2007). Improved breast cancer prognosis through the combination of clinical and genetic markers.*Bioinformatics*, *23*(1), 30-37.

[12] Prasad, Y., & Biswas, K. K. (2015, March). Gene Selection in Microarray Datasets Using Progressively Refined PSO Scheme. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[13] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[14] Jensen, Richard. "Performing feature selection with ACO." In *Swarm Intelligence in Data Mining*, pp. 45-73. Springer Berlin Heidelberg, 2006.

[15] Bai, Qinghai. "Analysis of particle swarm optimization algorithm." *Computer and information science* 3, no. 1 (2010).

[16] Nandpuru, Hari Babu, S. S. Salankar, and V. R. Bora. "MRI brain cancer classification using support vector machine." In *Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on*, pp. 1-6. IEEE, 2014.

[17] Brown, Michael PS, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Manuel Ares, and David Haussler. "Support vector machine classification of microarray gene expression data." *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09* (1999).

[18] Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines."*Machine learning* 46, no. 1-3 (2002): 389-422.

[19] Furey, Terrence S., Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, and David Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data."*Bioinformatics* 16, no. 10 (2000): 906-914.