ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Various Techniques for Predicting Cervical Cancer

M.D.Krithiga[1], V.P.Sumathi[2] , G.Prema Arokia Mary[3]

PG Scholar, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India [1,]

Assistant Professor (SRG), Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India [2]

Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India [3]

ABSTRACT: The healthcare analytics in big data are developing towards digitization of medical records, as pharmaceutical companies and other organization has been researched towards the development of data in electronic database. In healthcare cervical cancer is one among the leading cancer worldwide and also the most typical in developing countries, if it is detected in early stages it is easy to determine which stage it belongs and correct treatment has given in time. The cervical cancer data are in the form of "big data," and the big data is not only for its volume but for its variety, velocity. In this paper, various techniques for detecting cervical cancer such as penalized matrix decomposition (PMD), nonnegative matrix factorization (NMF), meta sample based SR classification (MSRC), tumor classification based on correlation filters and gene co-expression network have been discussed.

KEYWORDS: Cervical Cancer, PMD, NMF, MSRC, Correlation Filters, Co-expression network

I. INTRODUCTION

The healthcare analytics in big data are developing towards digitization of medical records, as pharmaceutical companies and other organizations has been researched towards development of data in electronic database [6]. Development of new technologies such as capturing devices, sensors, and mobile applications. Due to the development of new technologies collection of genomic information became cheaper, patient social communications in digital forms are increasing and more medical knowledge/discoveries are being accumulated. In healthcare cervical cancer is one among the leading common cancers moving ladies worldwide and also the most typical in developing countries, if it is detected in early stages it is easy to determine which stage it belongs and correct treatment has given in time. The cervical cancer data are in the form of "big data," and the big data is not only for its volume but for its variety, velocity. Pharmaceutical-industry experts, buyers, and providers are now beginning to analyze big data to obtain insights. Gene expression profiling has been widely used for predicting cancer at three different stages. Gene expression patterns have been used in many types of cancer along with the statistical techniques. Individual genes are usually studied in various cancer cells to draw a general conclusion about their behavior in more than one type of cancer. However, only a few studies have attempted such an approach on a genomic scale. With recent interest in biological networks, a gene co-expression network has emerged as a novel holistic approach for microarray analysis. For detecting cervical cancer various other techniques can be utilized such as screening techniques for detecting cervical cancer include penalized matrix decomposition, nonnegative matrix factorization, meta sample based SR classification, tumor classification based on correlation filters and gene co-expression network. The screening strategies mentioned above though applicable to the developed world may not be cost effective enough for widespread application in the underdeveloped countries.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

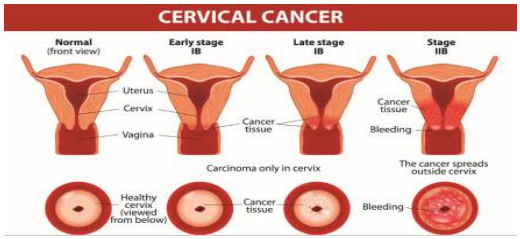

Fig.1. Cervical cancer stages

In this paper, various techniques for detecting cervical cancer has been discussed.

II. METHODS

A. *Penalized Matrix Decomposition(PMD)*

C. H. Zheng, 2011 described that the PMD is for predicting cancer in which meta samples are extracted from the gene expression data. A meta sample is a linear combination of original samples. By using PMD to extract a small number of meta samples, each meta sample can capture the inherent structures of the samples belonging to the same class. At the same time, the samples can be clustered by mapping themselves to the extracted meta samples. Moreover, the number of meta samples, i.e., the number of clusters, could be determined according to the changing trend of factor extracted by PMD.

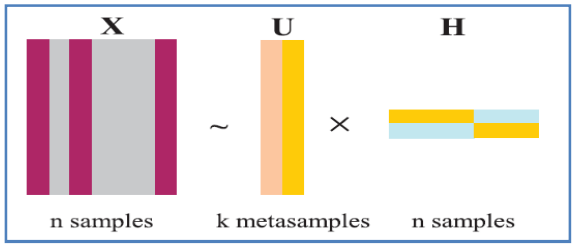

Fig.2. Penalized matrix decomposition

This method is applied it to cancer subtypes and cell differentiation. Three cancer data sets, i.e., the acute leukemia data set, the central nervous system tumor data set, and the cervical cancer data set.

Disadvantage: In addition, how to introduce the biological interpretation into the meta sample calculation process is another problem of using PMD. It overcomes the limitations of them is that each sample can only be clustered into one class, which may not be identical to the facts in some instance, e.g., borderline tumors and compound tumors.

B. *Nonnegative Matrix Factorization(NMF)*

D. S. Huang, 2009 described that gene selection and explicitly enforcing sparseness are introduced into the factorization process. Particularly, independent component analysis is employed to select a subset of genes so that the effect of irrelevant or noisy genes can be reduced. The NMF and its extensions, sparse NMF and NMF with sparseness constraint, are then used for tumor clustering on the selected genes. A series of elaborate experiments are performed by varying the number of clusters and the number of selected genes to evaluate the cooperation between different gene selection settings and NMF-based clustering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

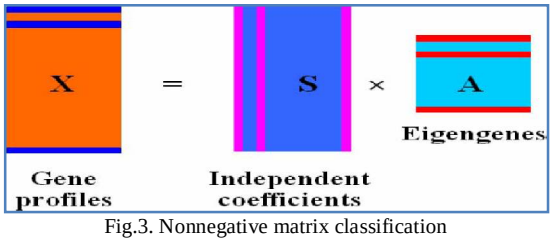

Fig.3. Nonnegative matrix classification

The selected gene expression data are represented as a matrix Y of size $m \times n$, whose rows contain the expression levels of the m selected genes in the n cell samples, and each column represents the expression level of all genes in one sample. All the entries in the gene expression matrix are nonnegative. The NMF methods resort to factor the gene expression matrix Y into the product of two matrices of nonnegative entries $Y \approx VH$ where matrix V is of size $m \times k$ with each of the k columns defining a meta gene, matrix H is of size $k \times n$ with each of the n columns representing the meta gene expression pattern of the corresponding sample, and k is a desired rank. Based on the rank the gene data are separated and formed as a cluster.

Disadvantage: One disadvantage of them is that they cluster the microarray dataset from the thousands of genes directly, in which the clustering results are not satisfied. This method was validated on the leukemia dataset, embryonal tumors dataset from the central nervous system, and the cervical cancer dataset. It can be found that improved clustering results were achieved by selecting the key genes using ICA. From the experimental results, it can see that the ICA-based gene selection is useful to detect the subsets of relevant genes for tumor clustering, especially when coupled with the NMF clustering method. It should be noted that although the three datasets used in this experiments have similar number of genes, i.e., about 5000, our method has no constraints on the number of genes contained in the data. To overcome this gene selection has to be performed before clustering to reduce noise and to achieve the better clustering results.

C. *Meta sample-based SR Classification (MSRC)*

C. H. Zheng, 2011 says a set of meta-samples are extracted from the training samples, and then an input testing sample is represented as the linear combination of these meta-samples by regularized least square method. Classification is achieved by using a discriminating function defined on the representation coefficients. Since minimization leads to a sparse solution, the method is called meta-sample-based SR classification (MSRC). Extensive experiments on publicly available gene expression data sets show that MSRC is efficient for tumor classification, achieving higher accuracy than many existing representative scheme. Classification is achieved by using a discriminating function of the representation coefficients on the meta-samples obtained by regularized least square. Since minimization could lead to sparse solution, our approach is then named as meta-sample based sparse representation classification (MSRC).

A meta sample is a linear combination of the gene expression profiles of samples, which can capture the alternative structures inherent to the data. The samples are analyzed by summarizing their gene expression patterns in terms of expression patterns over the meta samples.

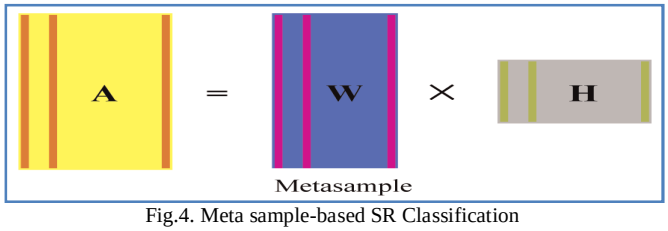

Fig.4. Meta sample-based SR Classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

The experimental results also show that, compared with SRC, MSRC is a better choice if there are 10 or more than 10 training samples. The reason may be that when there are 10 or more than 10 training samples, meta samples can capture the intrinsic structural information of the data of each class, and thus MSRC shows superior classification performance to SRC.

Disadvantage: On the other hand, if the number of training samples is less than 10, the trained meta samples may not be able to capture sufficient intrinsic structural information of each class, and the performance of MSRC is similar to or slightly worse than SRC. This is one weakness of the proposed method, i.e., the training samples for meta sample training cannot be too limited.

D. *Tumor classification method based on correlation filters:*

S. L. Wang, 2012 says tumor classification method based on correlation filters to identify the overall pattern of tumor subtype hidden in differentially expressed gen0es. Concretely, two correlation filters, i.e., Minimum Average Correlation Energy (MACE) and optimal tradeoff synthetic discriminant function (OTSDF), are introduced to determine whether a test sample matches the templates synthesized for each subclass. The experiments on six publicly available data sets indicate that it is robust to noise, and can more effectively avoid the effects of dimensionality curse. Compared with many model-based methods, the correlation filter-based method can achieve better performance when balanced training sets are exploited to synthesize the templates. Particularly, correlation filters can detect the similarity of overall pattern while ignoring small mismatches between test sample and the synthesized template. And it performs well even if only a few training samples are available.

MACE and OTSDF. Both of them can produce a sharp correlation peak at its origin while keeping the rest of output energy plane as low as possible when the test sample is similar to the synthesized template.

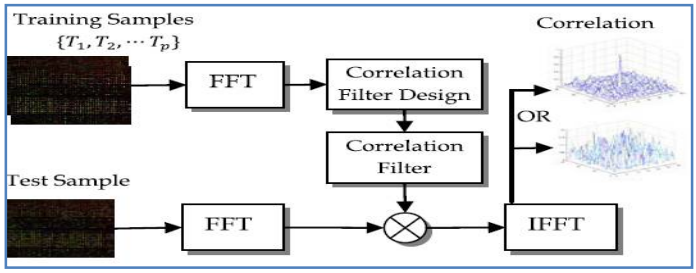

Fig.5. Block diagram for correlation process and how correlation process works

Disadvantage: The similarity of overall pattern emerging from the differentially expressed genes can be detected while ignoring small mismatches between test sample and the synthesized template because correlation filters are based on integration operation.

E. *Gene co-expression network*

Su-Ping Deng, 2016 co-expression was measured by Pearson product–moment correlation. The top 10 absolute correlation values were kept. P-values were calculated using R 3.0.2.26 due to the high degrees of freedom, the P-value after correction for multiple testing in each correlation measurement returned a significance of 6.53E-13. Cancer gene and function were retrieved from the NCBI Gene database. Cancer gene disease associations were provided in the COSMIC Cancer Gene Census List.

A gene co-expression network (GCN) is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them. Having gene expression profiles of a number of genes for several samples or experimental conditions, a gene co-expression network can be constructed by looking for pairs of genes which show a similar expression pattern across samples, since the transcript levels of two co-expressed genes rise and fall together across samples. Gene co-expression networks are of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex.

The direction and type of co-expression relationships are not determined in gene co-expression networks; whereas in a gene regulatory network (GRN) a directed edge connects two genes, representing a biochemical process such as a reaction, transformation, interaction, activation or inhibition co-expression network is very effective in discovering the modular structures in microarray data, both for genes and for samples. As the method is essentially parameter-free, it may be applied to large data sets where the number of clusters is difficult to estimate.

III. CONCLUSION

In this paper, various techniques for detecting cervical cancer has been discussed such as PMD in which the datasets can be clustered by mapping themselves to the extracted meta samples. It overcomes the limitations of them is that each sample can only be clustered into one class, which may not be identical to the facts in some instance, e.g., borderline tumors and compound tumors. NMF is for tumor clustering on the selected genes and the major disadvantage of them is that they cluster the microarray dataset from the thousands of genes directly, in which the clustering results are not satisfied and gene selection has to performed to overcome this problem Meta sample-based SR Classification also has a disadvantage that the trained meta samples may not be able to capture sufficient intrinsic structural information of each class. Among them co-expression network is very effective in discovering the modular structures in microarray data, both for genes and for samples. As the method is essentially parameter-free, it may be applied to large data sets where the number of clusters is difficult to estimate.

Table .1. Comparison of different Methods of problem and advantages

s.no	Problem identified	Methodology used	advantage	inference
1	The biological interpretation into the metasample calculation process is problem.	Penalized matrix decomposition	powerful method in cancer class discovery.	Identification of tumor
2	The clustering results may be different.	Non negative matrix factorization	Combining sequence and expression data for improving functional gene annotation	Clustering tumor for the identification of cancer using NMF
3	The training samples are too limited.	Metasample-based SR classification (MSRC).	Provides a simple, scalable frame-work	Tumors are classified in terms of matrix to identify the affected genes
4	False matching is larger	correlation filter-based method	detect the similarity of overall test sample	Identifies tumor from large scale tumor data sets
5	Relationship	Co expression network	Discover modular structure	Identifies cancer at early stage

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

REFERENCES

1. Su-Ping Deng, Lin Zhu, and De-Shuang Huang "Predicting Hub Genes Associated with Cervical Cancer through Gene Co-Expression Networks" IEEE/ACM transactions on computational biology and bioinformatics, vol. 13, no. 1, january/february 2016.
2. S. L. Wang, Y. Zhu, W. Jia, and D. S. Huang, "Robust classification method of tumor subtype by using correlation filters," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 9, no. 2, pp. 580–591, Mar./ Apr. 2012.
3. C. H. Zheng, L. Zhang, V. T. Y. Ng, S. C. K. Shiu, and D. S. Huang, "Molecular pattern discovery based on penalized matrix decomposition," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 8, no. 6, pp. 1592–1603, Nov./Dec. 2011.
4. C. H. Zheng, L. Zhang, V. T. Y. Ng, S. C. K. Shiu, and D. S. Huang, "Metasample-based sparse representation for tumor classification," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 8, no. 5, pp. 1273–1282, Sep./Oct. 2011.
5. C. H. Zheng, D. S. Huang, L. Zhang, and X. Z. Kong, "Tumor clustering using non-negative matrix factorization with gene selection," IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 4, pp 599– 607, Jul. 2009.
6. World Health Organization, World Cancer Report 2014. pp. Chapter 5.12, 2014.