International Journal of Computer Science Engineering & Technology
APPROVED BY
NATIONAL SCIENCE LIBRARY (NSL)
NATIONAL INSTITUTE OF SCIENCE COMMUNICATION AND INFORMATION RESOURCES (NISCAIR)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR)- NEW DELHI INDIA .
ISSN :2455-9091

INVESTIGATION ON CERVICAL CANCER USING MULTIKERNAL SUPPORT VECTOR MACHINE

[1]Dr.N.UmaDevi, [2] R.Jenitha Mary
[1]Associate Professor & Head, [2] M.Phil Research Scholar,
[1,2]Dept of CS & IT
[1, 2] Sri Jeyendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore.

Abstract- Selection of relevant genes for sample classification is a common task in most gene expression studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance (for instance, for future use with diagnostic purposes in clinical practice). Many gene selection approaches use univariate (gene-by-gene) rankings of gene relevance and arbitrary thresholds to select the number of genes, can only be applied to two-class problems, and use gene selection ranking criteria unrelated to the classification algorithm. In contrast, Multikernal Support Vector Machine (MK-SVM) is a classification algorithm well suited for microarray data: it shows excellent performance even when most predictive variables are noise, can be used when the number of variables is much larger than the number of observations and in problems involving more than two classes, and returns measures of variable importance. Thus, it is important to understand the performance of random forest with microarray data and its possible use for gene selection.

Keywords: [Probability, Accuracy, Precision, Recall, SVM, Cervical Cancer]

1.INTRODUCTION

Cervical cancer is, potentially, one of the most preventable cancers. Unlike many other cancers there is an easily detectable and normally prolonged premalignant phase. The incidence and mortality rates for cervical cancer have leveled off during the past 40 years. With early detection through Papanicolaou (Pap) test screening, cervical cancer can be prevented. Minority populations and persons of low socioeconomic status, however, still have high incidence and mortality rates. The world-wide occurrence of the cancer cervix cases show that only 20% of these cases occur in the developed nations while 80% of the cases are found in the developing countries. The mortality rate of this cancer can be affected by the alteration of earlier diagnosis and improved treatment in the natural history of the disease, and recent clinical research has identified that HPV is one of the main triggers for cervical cancer. The patterns discovered when mining cervical cancer screening data may support these observations, or suggest additional triggers.

One of the emerging issues in medical research is data mining. The widespread use of computers makes it easy to gather and manage large amounts of data from many different sources. A well-organized system can make available clinical, biological, genetic

data, and all other information collected about patients. This data is often complex, meaning that it contains many elements related in non-obvious ways or characterized by explicit or implicit relationships and structures. Such integration is increasingly considered necessary in order to produce more accurate diagnoses. Cervical cancer remains one of the leading causes of cancer-related death among women globally. Even though the morbidity and the mortality have been decreasing in recent years, the morbidity rates of cervical cancer are the second leading type in women and the mortality rates are the sixth of the top ten cancers in Taiwan. There are few researches on its relationship between recurrent events and the mortality and incidence rate. Indeed, recurrent cervical cancer is a devastating disease for those women unfortunate enough to suffer such an event. Patients with recurrent disease or pelvic metastases have a poor prognosis with a 1-year survival rate between 15 and 20%. Cancer is a group of abnormal cells that are formed from cells that grow continuously, not limited, not coordinated with the surrounding tissue and does not function physiologically. The tissue is destructive and can spread to other body parts that generally would be fatal if left unchecked. The growth of cancer cells will cause the tissue to be big which is called a tumour. Cancer cells spread through the blood vessels and lymph vessels. Cancer has different characteristics, those which can grow rapidly or slowly. Cervical cancer is a common type of cancer that found in the cervix. Cervical cancer often has no symptoms in the early stages, but the common symptoms that occur are unusual vaginal bleeding, which occurs after having sex between periods or after menopause.

1.1 Types of Cervical Cancer

There are 2 main types of cervical cancer:

1. Squamous Cell Cancer
2. Adenocarcinoma

They are named after the type of cell that becomes cancerous. Each one is distinguished by the appearance of cells under a microscope.

Squamous cell cancer

These cell cancer are the flat, skin-like cells that cover the outer surface of the cervix (the ectocervix).Between 70 and 80 out of every 100 cervical cancers (70 to 80%) are squamous cell cancers.

Adenocarcinoma

These types of cells are developed in the glandular cells that line the upper portion of the cervix. These cancers make up 10 to 20 percent of cervical cancers. Adenocarcinoma is a cancer that starts in the gland cells that produce mucus.

2. RESEARCH METHODOLOGY

2.1 Existing System

Employing functions of several bioinformatics tools, we selected 143 differentially express edgenes (DEGs) associated with the cervical cancer. The results of bioinformatics analysis show that these DEGs play important roles in the development of cervical cancer. Through comparing two differential co-expression networks (DCNs) at two different states, we found a common sub-network and two differential sub-networks as well as some hub genes in three sub-networks. Moreover, some of the hub genes have been reported to be related to the cervical cancer. Those hub genes were analyzed from Gene Ontology function enrichment, pathway enrichment and protein binding three aspects.In the transcript me analysis, differential co-expression analysis (DCA) emerged as a unique complement to traditional differential expression analysis. DCA investigates differences in gene interconnection by calculating the expression correlation changes of gene pairs between two conditions. The rationale behind DCA is that changes in gene co-expression patterns between two contrasting phenotypes (e.g., healthy and disease) provide hints regarding the disrupted regulatory relationships or affected regulatory sub-networks specific to the phenotype of interest. Therefore, among the many growing directions of DCA, there is

the so-called differential regulation analysis (DRA), which integrates the transcription factor (TF)-to target information to probe upstream regulatory events that account for the observed co-expression changes.

Recently, researchers have integrated the concepts of differential co-expression and differential expression concepts to propose a novel Regulatory Impact Factor (RIF) that could be used to prioritize disease-causative TFs. Additionally, some researchers have begun to perform DCA of micro RNAs. And some tools have been developed for differential expression analysis based on microarray, such as R packages named as LIMMA, SAMR, WGCNA, and so on. Constructed and analyzed two differential co expression networks (DCNs) under different conditions, then found some hub genes associated with the cervical cancer.

2.1.1 Disadvantages

Cannot be used when there are many more variables than observations.

Cannot be used both for two-class and multi-class problems of more than two classes.

Has bad predictive performance even when most predictive variables are noise, and therefore it require a pre-selection of genes.

2.2 Proposed Work

For finding out the genes associated with Cervical Cancer, the MKSVM is proposed. Kernels are employed in Support Vector Machines (SVM) to map the nonlinear model into a higher dimensional feature space where the linear learning is adopted. Every kernel has its advantages and disadvantages. Preferably, the 'good' characteristics of two or more kernels should be combined. Through the implementation for average molecular weight in polyacrylonitrile productive process, it demonstrates the good performance of the proposed method compared to single kernel. Selection of relevant genes for sample classification (e.g., to differentiate between patients with and without cancer) is a common task in most gene expression studies. When facing gene selection problems, biomedical researchers often show interest in one of the following objectives:

To identify relevant genes for subsequent research; this involves obtaining a (probably large) set of genes that are related to the outcome of interest, and this set should include genes even if they perform similar functions and are highly correlated.To identify small sets of genes that could be used for diagnostic purposes in clinical practice; this involves obtaining the smallest possible set of genes that can still achieve good predictive performance (thus, "redundant" genes should not be selected).

2.2.2 Advantages

Can handle a mixture of categorical and continuous predictors.

Incorporates interactions among predictor variables.

Can be used when there are many more variables than observations.

Can be used both for two-class and multi-class problems of more than two classes.

Has good predictive performance even when most predictive variables are noise, and therefore it does not require a pre-selection of genes (i.e., "shows strong robustness with respect to large feature sets").

3. Experimental Results

The most often metric used to determine the performance of classifier is accuracy. Since the accuracy is inappropriate when data is imbalanced, we used another metrics to compare the performance. The standard technique for evaluating classifier on imbalanced class is Receiver Operating Characteristic. It shows SVM has constant accuracy even though the data has been randomized 30 times. Random Forest Tree can classify the result better than other classifiers. Recall measures how often a positive class instance in the dataset was predicted as a positive class instance by the classifier. Precision measure how often an instance that was predicted as positive that is actually positive.

Survival Probability:

Evaluate memory usage for each algorithm with the same datasets as the runtime tests. Our algorithm, it guarantees

Survival Probability as good as that of the state-of-the-art algorithm. Moreover, our algorithm presents the most outstanding results in many cases.

No of Web Documents	DCA	DCN	Multi Kernel SVM
100	0.682	0.73	0.80
200	0.693	0.74	0.82
300	0.71	0.76	0.83
400	0.73	0.78	0.85

Table 1: Survival Probability Results

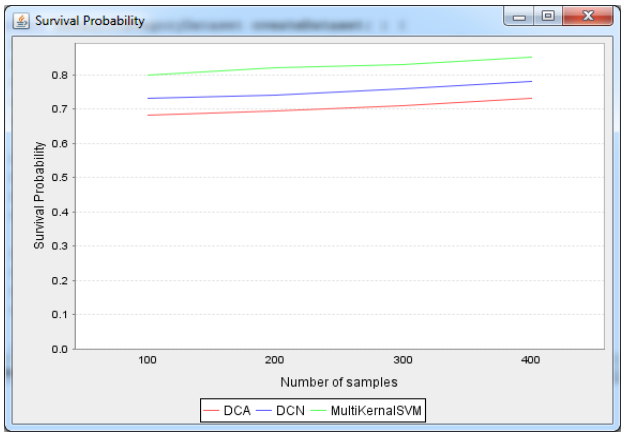

Figure 1: Survival Probability

Classification Accuracy (%):

Proposed linear structure to its trees instead of the previous tree form in order to minimize access times to search nodes. As a result, its advantages have a positive effect on reducing runtime in whole experiments. Especially as the minimum support threshold becomes lower, the difference of runtime between our algorithm and the others is bigger.

No of Web Documents	DCA	DCN	Multi Kernel SVM
100	69.5	73.6	83.6
200	69.9	75.6	85.6
300	69.5	77.6	87.6
400	70.8	78.6	89.1

Table 2: Classification Accuracy Results

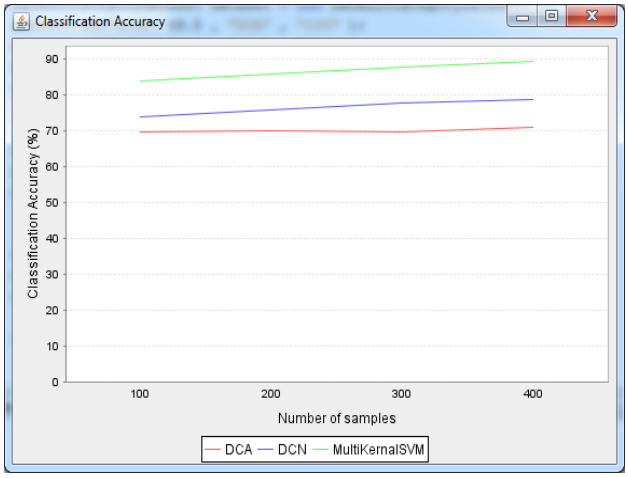

Figure 2: Classification Accuracy

Precision (%):

Proposed algorithm shows the best Precision while the others have relatively poor performance, which indicates that our scheme can store these increasing attributes more efficiently than the other structures of the competitor algorithms. Through the above experimental results, we know that the proposed algorithm, outperforms the others with respect to increasing transactions and items in terms of scalability as well as runtime and memory usage for the real datasets.

No of Web Documents	DCA	DCN	Multi Kernel SVM
100	71.7	74.7	86.7
200	72.8	76.1	88.6
300	73.1	77.7	89.8
400	74.2	79.1	90.0

Table 3: Precision Results

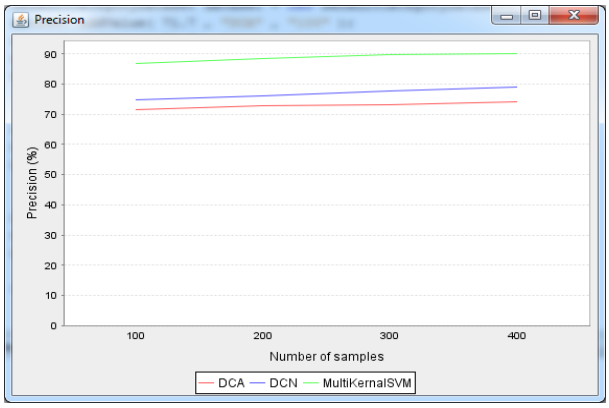

Figure 3: Precision

Recall (%):

Through the above experimental results, we know that the proposed algorithm, outperforms the others with respect to increasing transactions and items in terms of scalability as well as runtime and memory usage for the real datasets.

No of Web Documents	DCA	DCN	Multi Kernel SVM
100	76.0	80.6	90.0
200	77.4	81.4	90.4
300	79.1	82.5	91.9
400	79.5	83.8	92.5

Table 4: Recall Results

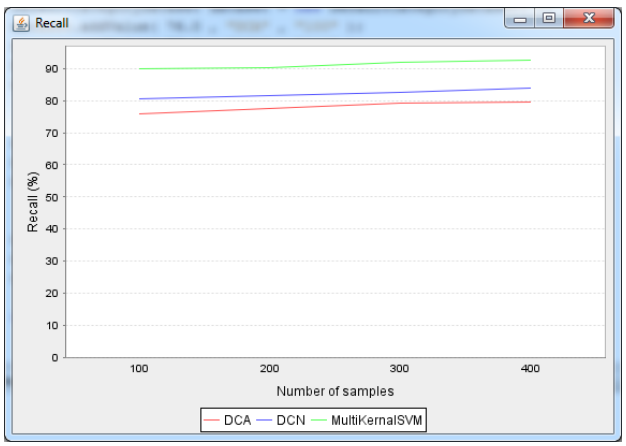

Figure 4: Recall

Experimental results showed that SVM outstanding performance in terms of accuracy, precision, recall, memory usage, and scalability. We could also observe that our algorithm outperformed the previous algorithms especially in the runtime experiments due to the reduced pointer accesses. The techniques and strategies described in this paper can be applied to not only general frequent cancer mining but also a variety of data mining fields such as closed/maximal pattern mining, and graph mining.

CONCLUSION

Multikernal Support Vector Machine (MKSVM) is proposed. Kernels are employed in Support Vector Machines (SVM) to map the nonlinear model into a higher dimensional feature space where the linear learning is adopted. Every kernel has its advantages and disadvantages. Preferably, the 'good' characteristics of two or more kernels should be combined. Through the implementation for average molecular weight in polyacrylonitrile productive process, it demonstrates the good performance of the proposed method compared to single kernel.Selection of relevant genes for sample classification (e.g., to differentiate between patients with and without cancer) is a common task in most gene expression studies. When facing gene selection problems, biomedical researchers often show interest in one of the following objectives: To identify relevant genes for subsequent research; this involves obtaining a (probably large) set of genes that are related to the outcome of interest, and this set should include genes even if they perform similar functions and are highly correlated. To identify small sets of genes that could be used for diagnostic purposes in clinical practice; this involves obtaining the smallest possible set of genes that can still achieve good predictive performance

REFERENCES

[1] World Health Organization, World Cancer Report 2014. pp. Chapter 5.12, 2014.

[2] (2014, Feb.). World Health Organization. Fact sheet No. 297:Cancer.

[3] A. Gadducci, C. Barsotti, S. Cosio, L. Domenici, and A. G.Riccardo, "Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: A review of the literature,"Gynecological Endocrinol.,vol. 27, no, 8, pp. 597–604, 2011.

[4] C. Stuart and M. Ash,Gynaecology by Ten Teachers (18 ed.). London, U.K.: Hodder Education, 2006.

[5] C. M. Croce, "Oncogenes and cancer," N. Engl. J. Med., vol. 358, no. 5, pp. 502–11, 2008.

[6] A. G. Knudson, "Two genetic hits (more or less) to cancer,"NatureRev. Cancer, vol. 1, no. 2, pp. 157–62, 2001.

[7] D. S. Huang and H. J. Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based onthe numbers of adjacent amino acids," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 10, no. 2, pp. 457–467, Mar./Apr. 2013.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] D. S. Huang and C. H. Zheng, "Independent component analysis based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol. 22, no. 15, pp. 1855–1862, 2006.

[13] S. Ramaswamy et al., "Multiclass cancer diagnosis using tumorgene expression signatures,"Proc. Nat. Acad. Sci. USA, vol. 98, pp. 15149–15154, 2001.

[14] D. R. Rhodes et al., "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," Proc. Nat. Acad. Sci. USA, vol. 101, pp. 9309–9314, 2004.

[15] E. Segal et al., "A module map showing conditional activity of expression modules in cancer,"Nat. Genet., vol. 36, pp. 1090–1098, 2004