International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Cancer Detection using Support Vector Machines Trained with Linear Kernels

Gerardo Alfonso

University Autonoma Barcelona

Abstract: ***In this article support vector machines are used for determining is cancer is present in lung, liver and cervix tissue using multiple kernels. The results indicate that linear kernel in this regard seems to be a better approach than using polynomial or Gaussian kernels. It was also found that using support vector machines trained with a linear kernel seems to also produce more accurate results than using a backpropagation neural network with 10 neurons. The accuracy of classification decreases when methylation in blood samples is analyzed, rather than direct tissue samples, to determining the presence of cancer.***

Keywords: Methylation, cancer, kernel, support vector machine

1. Introduction

DNA methylation remains a very active area of research due to its suspected effect in areas as diverse as development [1], aging [2] and cancer [3]. DNA methylation remains an area not well understood, likely due to its very high level of complexity, and it seems intertwined with many biological processes. As a biological marker DNA methylation has proven a very useful technique and it is likely to generate a large amount of research in years to come. Technological advancements have made available an increasing amount of methylation data from patients that undergo procedures or that volunteer for research. This increase in data availability has put pressure to developed better and more efficient statistical models to try to understand these processes. This increase in data availability is almost certain to continue in the future. This paper attempts to utilize a well know statistical tool called Support Vector Machines ("SVM") to the task of differentiating healthy tissue from tissue with cancer using methylation data. SVMs are a general statistical tool that can, and has, been applied to a multitude of different problems. It is likely that in the near future SVMs will continue finding new areas of application as the amount of data created in many scientific and engineering disciplines increases and simultaneously computing power, which allows such enormous amount of data to be processed, also continues to increase. SVM use the concept of separating data into the different sides of a hyperplane in order to categorize such data. It is a remarkably flexible technique and of relatively simple use. A SVM needs, in the context of this paper, the methylation levels for each CpGs, which is a number ranging from 0 to 1 and a binary identification, defining if the sample comes from a tissue with cancer or from a healthy tissue. Currently is possible to obtain thousands of CpGs methylation data quickly from a patient sample using relatively affordable techniques. This creates a mismatch between the number of samples in studies, typically from a few dozens to a few hundreds, and the thousands of data points available for each patient. In this context SVM attempt to categorize the methylation for each patient into two categories: 1) caner and 2) no cancer. CpGs are just a bond between two bases, a Cytosine and a Guanine and they have proven rather important in several biological processes receiving a considerable amount of interest by researchers. Having a quantifiable indicator of cancer could be useful for the doctors making diagnosis as well as a potential tool for confirmation of such diagnosis. It will be shown that training the SVM with a linear kernel for the three tissues analyzed (liver, lung and cervix) produced more accurate results than using other kernels, such as polynomial or Gaussian. These results were rather consistent among the three data sets with direct tissue data (no blood samples). The approach of using an SVM trained with a linear kernel seems also to produce results more accurate than using a simple backpropagation neural network trained with 10 neurons. It will also be shown that the results are less accurate when the analysis is performed on blood samples, rather than using directly methylation data from lung, liver or cervix. The results regarding what type of kernel to use in this case are less conclusive. This last point is likely a good area for further research.

2. Literature Review

In this article only a brief description of SVM is presented, for the reader interested in a more mathematically detailed explanation of SVM we point to [4], [5] or [6]. These are all very good articles and they go into details into formal mathematical issues. The mathematical formalism for support vector machines is not particularly simple and getting into its details is outside of the scope of this article, which focuses on applying such techniques to the specific case of detecting cancer though SVM using methylation data as an input. Plainly speaking a SVM tries to create a boundary (hyperplane) between the two sets of data which is trying to classify. This boundary should be as far away from the data as possible while containing all of them. This clearly leads to a Lagrange multiplier type of situation in which a function needs to be maximized while certain constraints must be kept [7]. There has been a lot of interest both theoretically [8] as well as regarding practical applications of SVM [9], [10].

There are some articles in the literature applying this technique for imaging processing (radiology). For instance, [11] applied this technique to breast cancer data and [12] applied it to lung cancer data. Imaging processing is clearly a natural candidate for application of SVM as it removes, at

least to some degree, the subjectivity of the radiologist when examining MRI images to determine the presence of cancer. This process clearly depends heavily on the experience of the radiologist with some degree of subjectivity when analyzing unclear images or cancer in early stages. This is an area in which a great deal of automation could be applied and in fact it is currently a vibrant area of research. Perhaps less attention has received the application of support vector machines using methylation data as inputs. One interesting article in this regard is [13], which successfully applied the technique to breast cancer. In this article the input data used were not only methylation levels but also gene expression data. In another interesting article [14] used neural networks as a classification for differentiation between healthy tissue and lung cancer. The literature in this regard is expanding rapidly due to the clear practical applications of these techniques and the ever increasing amount of data available.

3. Methodology

All the data used in this article is publically available in the GEO database and come from other research reports. There are the dataset containing methylation sample from cancer and control cases. The first data set contains cases with liver cancer (GSE57956) and comes from [15] article. There are 120 samples. The second dataset is from a lung cancer study [16] contains 88 cases and has the GEO database code (GSE49996). Half of the samples (44) are from lung tissue with cancer and the other half from healthy lung tissue. This dataset is from a cervical cancer article [17] and has the accession code (GSE30759) in the Geo Database. These are the three datasets containing methylation information from organs. A fourth dataset was used, in this case, rather than having sample from organs the methylation data was extracted from blood samples. This information was obtained from bladder cancer research published by [18] with the GEO database code (GSE50409), 120 samples. All the dataset contain DNA methylation information of patients obtained with the Illumina 27K. There are in excess of 27,000 CpGs methylation data points for each patient present in the dataset as well as an indicator representing if the data comes from a cancer sample or otherwise. All the data used in this article is publically available and obtained from the Geo Database [19]. The algorithm used to detect cancer was a support vector machine, trained with three different kernels: liner, polynomial or Gaussian. The objective is to obtain the smallest, out of sample, classification error possible. The three previously mentioned kernels can be defines as follows: 1) Linear kernel $= x * y$, 2) polynomial kernel $= (x * y + 1)^a$, where a is the degree of the expression and 3) Gaussian $= e^{-|x-y|^2}$. Deciding which kernel to use is of clear importance and can potentially have a substantial impact on the accuracy of the data classification. This decision, of what type of kernel to use, depends on the specific application. It is not easy, in principle, to decide a priori without actually comparing the results of different kernels which one to use.

As an additional step and comparison purposes the results from the SVM were also compared with the results from a simple neural network with one hidden layer, 10 neurons and trained using back propagation. The same process was applied for all the four data sets, regardless if the methylation data came from organs of from blood samples. 100 simulations were performed on each case to obtain a probability distribution. Then a Wilcoxon test was performed comparing the results obtained using SVM, with linear, polynomial and Gaussian kernels, as well as with neural networks. All the calculations were performed using the commercially available software package Matlab.

4. Results

Liver cancer

The lowest median error obtained using support vector machines for detection of cancer in liver tissue (out of sample data) in the 120 sample studied was obtained with a linear kernel (Table 1). According to a Wilcoxon test the result was statistically significant at a 5% significance level (Table 2). The approach of using support vector machines with a linear kernel appeared to produce better results than using polynomial or Gaussian kernels. The linear SVM approach also produced a more accurate result than using a neural network with back propagation and 10 neurons in the hidden layer. This NN approach generated a median error of 0.0556 with a standard deviation of 0.0329. All the simulations (for both SVM and NN) were repeated 100 times each. The error using SVM was statistically significantly smaller (Table 3) for linear and polynomial kernels when compared to the NN approach but that was not the case when using a Gaussian kernel.

Table 1: Error rates for SVM using three different kernels.

	Linear	Polynomial	Gaussian
Median	0.0250	0.0333	0.0833
Mean	0.0233	0.0347	0.0851
σ	0.0033	0.0032	0.0051

Table 2: Results of Wilcoxon test for SVM using different kernels

	p	h
Liner – polynomial	1.7e-39	1
Linear – Gaussian	3.6e-38	1
Polynomial – Gaussian	1.8e-38	1

Table 3: Comparison of NN results with NN (Wilcoxon)

	Linear	Polynomial	Gaussian
P	1.29e-5	1.19e-5	1.99e-26
H	1	1	1

Lung cancer

Similarly to the case of liver, the median error obtained using an SVM with a linear kernel is smaller (Table 4) that the one obtained using either a polynomial or a Gaussian kernel. This hypothesis was tested with a Wilcoxon test (Table 5). The median error obtained using backpropagation in a NN with 10 neurons was 0.1538 with and standard deviation of 0.0971. In this case, the error was statistically smaller using any of the three kernels and SVM when compared to neural networks (Table 6). The confusion matrix and NN accuracy

information can be seen in (Figure 1) and (Figure 2). All the compared errors were obtained using untrained data. In other words, data not used for training purpose by the algorithm,

Table 4: Error rates for SVM using three different kernels.

	Linear	Polynomial	Gaussian
Median	0.1023	0.1364	0.1136
Mean	0.0972	0.1356	0.1198
Standard deviation	0.0067	0.0122	0.0085

Table 5: Results of Wilcoxon test for SVM using different kernels

	p	h
Liner – polynomial	4.2e-35	1
Linear – Gaussian	2.0e-33	1
Polynomial – Gaussian	9.1e-19	1

Table 6: Comparison of NN results with NN (Wilcoxon)

	Linear	Polynomial	Gaussian
P	1.20e-3	7.80e-3	1.30e-3
H	1	1	1

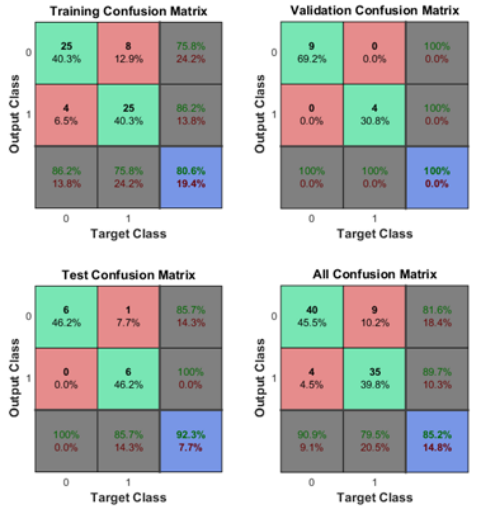

Figure 1: Confusion matric sample obtained for a single lung cancer NN simulation

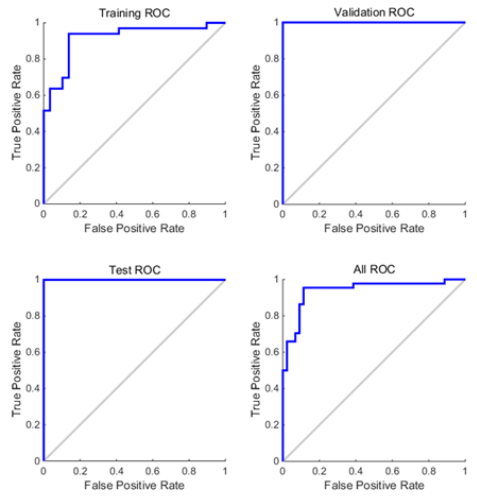

Figure 2: Lung cancer NN simulation

Cervical cancer

The results using tissue samples from the cervix (63 patients in total) are consistent with the ones obtained from lung and liver samples (Table 7). The approach of using SVM with linear kernel seems to produce the smallest error and to be statistically significantly smaller than the median errors obtained using either polynomial of Gaussian kernels (Table 8). The median result, after 100 simulations, obtained using backpropagation and a NN was 0.1111 with a 0.1008 standard deviation. Using, once more (Table 9) a Wilcoxon test the values obtained using SVMs and NNs were compared. SVMs using linear and polynomial kernels had statistically significantly smaller errors than the NNs. The major difference with the previous cases is that for the cervical cancer data set the hypothesis that the medians for the error obtained using SMVs with Gaussian kernel and the NN being equal cannot be rejected.

Table 7: Error rates for SVM using three different kernels

	Linear	Polynomial	Gaussian
Median	0.0010	0.0159	0.0317
Mean	0.0006	0.0092	0.0263
Standard deviation	0.0005	0.0079	0.0120

Table 8: Results of Wilcoxon test for SVM using different kernels

	p	h
Liner – polynomial	3.2e-35	1
Linear – Gaussian	3.5e-33	1
Polynomial – Gaussian	1.5e-19	1

Table 9: Comparison of NN results with NN (Wilcoxon)

	Linear	Polynomial	Gaussian
P	0.0019	0.0008	0.0691
H	1	1	0

Bladder cancer

The approach used in the bladder cancer section was different from the previous three cases as the methylation data come from blood samples from the patients rather than from tissue samples from the area potentially affected by cancer. The idea was to see if the results can be extrapolated to analyzing the methylation of blood, which can be obtained with much less invasive techniques than organ tissue samples. The obtained median errors are substantially higher than in the previous cases (when using sample directly from the organs). In this case, the SVM with the smallest error (120 patients) is the one using a polynomial kernel (Table 10), which is in clear contrast with the previous cases. There is also no statistically appreciable difference between the results using a linear or a Gaussian kernel (Table 11). There appears also not to be a statistically significant difference when using neural networks compared to both a linear and a Gaussian kernel in a SVM.

Table 10: Error rates for SVM using three different kernels.

	Linear	Polynomial	Gaussian
Median	0.4167	0.3667	0.4166
Mean	0.4166	0.3682	0.4140
Standard deviation	0.0229	0.0217	0.0204

Table 11: Results of Wilcoxon test for SVM using different kernels

	p	h
Liner – polynomial	2.9e-26	1
Linear – Gaussian	3.2e-1	0
Polynomial – Gaussian	2.1e-26	1

Table 12: Comparison of NN results with NN (Wilcoxon)

	Linear	Polynomial	Gaussian
P	0.0732	0.0214	0.0617
H	0	1	0

5. Conclusion

For the data sets analyzed, the results indicate that when using DNA methylation data from the liver, lung or cervix, to determine the presence of cancer using a support vector machine a linear kernel training generates results that are more accurate, than using other training kernels such as polynomial or a Gaussian kernels. The difference was statistically significant (tested with a Wilcoxon test). The results were also more accurate than the ones obtained using a simple backpropagation NN with 10 neurons. These results were also statistically significant. The dynamics seems to be rather different when the methylation analysis is performed on blood samples, rather than tissue from the previously mentioned organs. In this case the accuracy of the method seems to be substantially smaller and there appears to be less statistically significance differences between using SVM and NN.

6. Future Work

While the results seem to indicate that training a SVM with linear kernel is more accurate that a polynomial or a Gaussian kernel for the three sample tissues analyzed the results for blood are less conclusive an open a further area of investigation for future work. Further work is necessary to determine the best type of kernel to use with SVM when using methylation data for cancer detection purposes.

References

[1] Smith Baxter, Lagrance multipliers tutorial in the context of support vector machines. http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf. 2012.
[2] Horvath, Steve. DNA methylation age of human tissues and cell types. Genome biology. 2013.
[3] Massie Charles. The importance of DNA methylation in prostate cancer development. The journal of steroid biochemistry and molecular biology. Volume 166. 2017.
[4] Zhou Yan. Murat Kantarcioglu, Thuraisingham Bhavani. Adversarial support vector machine learning. 2012.
[5] Shawe-Taylor John. Support vector machine. Cambridge University Press. 2000.
[6] Ahmed Ashfaq, Sultan Aljahdali. Comparative prediction performance with support vector machine and random forest classification techniques. International journal of computer applications. 2013.
[7] Smith Baxter, Lagrance multipliers tutorial in the context of support vector machines. http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf. 2012.
[8] Astorino Annabella, Fuduli Antonio. Support vector machine polyhedral separability in semisupervised learning. Journal of optimization theory and applications. Volume 164. 201.5
[9] An Sung-Hoon, U-Yeol Park, Kang Kyung-In Kang, Moon-Young Cho. Application of support vector machines in assessing conceptual cost estimates. Journal of computing in civil engineering. Volume 21. 2007.
[10] Du Shicang, Liu Changping, Xi Lifeng. A selective multiclass support vector machine ensemble classifier for engineering surface classification using high definition metrology. Journal of Manufacturing Science and Engineering. Volume 137. 2015.
[11] Ebrahim Edriss, Feng Wu Zhi. International Journal of Science and Research (IJFR). Breast cancer classification using support vector machine and neural network. 2012.
[12] Kohad Rashmee, Ahire Vijaya. Diagnosis of lung cancer using support vector machine with ant colony optimization technique. International journal of advances in computer science and technology (IJACST). Vol 3. No. 11. 2014.
[13] List Markus. Hauschild Anne-Christin, Tan Qihan, Batra Richa. Classification of breast cancer subtypes by combining gene expression and DNA methylation data. Journal of integrative bioinformatics. 2014.
[14] Hosseinzadeh F, Ebrahimi Goliaei, Shamabi N. Classification of lung cancer tumors based on structural and physiochemical properties of proteins by bioinformatics models. PLOS ONE 7(12). 2012.
[15] Mah WC, Thurnherr T, Chow PK, Chung AY. Methylation profile reveals distinct subgroup of hepatocellular carcinoma patient with poor prognosis. PLoS One 9(8). 2014. (GSE57956)
[16] Lenka G, Tsai MH, Lin HC, Hsiao JH et al. Identification of Methylation-Driven, Differentially Expressed STXBP6 as a Novel Biomarker in Lung Adenocarcinoma. *Sci Rep* 2017 Feb 15;7:42573. (GSE49996)
[17] Zhuang J, Jones A, Lee SH, Ng E. The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. PLos Genet 8(2). 2012. (GSE30759)
[18] Langevin Scott. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. University of Cincinnati College of Medicine. Department of environmental health. 2015. (GSE50409)
[19] Geo database. www.ncbi.nim.nih.gov.