A REPRESENTATION OF CHANGES OF IMAGES AND ITS APPLICATION FOR DEVELOPMENTAL BIOLOLOGY

Abstract. In this paper, we consider a series of events observed at spaced time intervals and present a method of representation of the observations. The procedures are sketched with comments and details in hopes of explaining the ideas that deal with a set of gene expression data, which are obtained from developmental biology. We describe representation as a process of choosing a function that highly correlates with the observed data. In other words, we are systematizing and quantifying the data by representing it as a mathematical function. By using this representation and incorporating a machine-learning method, we can classify the expression data, as well as any other data that was obtained over a certain period.

§1. Introduction

This work was motivated by questions of several medical/biological scientists of whom I have had chances to work with. Here are some of the projects that had been discussed:

1. Changes of gene expression rates in the brains of mice with/without electric shock for equally spaced time intervals
2. Association study between changes of gene expression rates and radiation sensitivity

3. Changes of clinical tests including international prostate symptom score (IPSS) while a series of dose of some medicine are applied

The projects listed above are typical everlasting problems in the biological/medical world. Today, in order to analyze them, scientists use recent techniques, such as DNAchip, gene expression, and micro-array analysis. However, despite all the innovative technological advances, the statistical analysis and interpretations have not been as successful as the technological advances. This could be a result of a lack of understanding the problems in depth.

Though these three problems look unrelated, the intrinsic properties are the conceptually equivalent. Here we introduce a universal method, which may put those three problems into a single setting, which is ready for application of machine-learning methods, such as SVMs.

Our approach could lead to a new paradigm of not only genetic research (including developmental research), but also of diagnosis, such as progresses of diseases. Throughout this paper, for convenience, we will consider the case of a single gene expression data, keeping in mind that we can apply the concept to other problems as well as multiple genes.

DNAchip and micro-array techniques convert the expression rates into densities of stained images, which may be recorded as a series of numbers. All the experiments and phenomenon infer that the numbers fluctuate over the time and if we plot them as a graph, it is similar to a portion of the following shape:

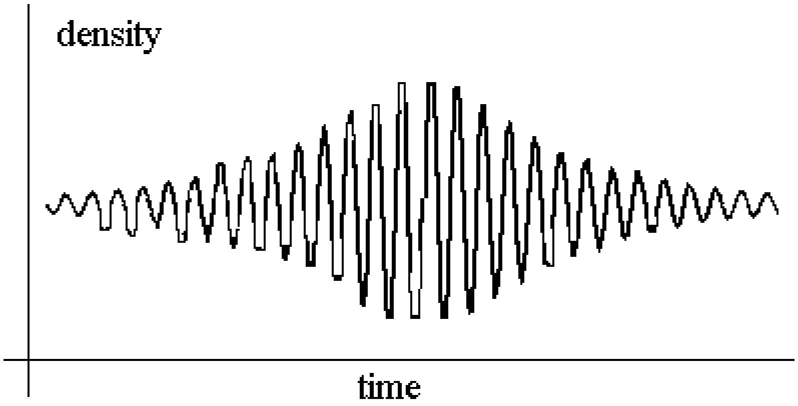

Fig. 1

So to speak, the expression rates will start to increase slowly, reach a peak, and decrease to a certain level. To proceed with the description of this phenomenon, we need to find a function for the data.

§2. Representation

In this section, we will describe the procedures of representing a set of gene expression data as a vector. Each step will be explained thoroughly.

Step 1. Fixing a candidate function

Since we need a periodic-looking, fluctuating functions, it will be wise to start with $sin(t)$ or $\cos(t)$, while, for increasing and decreasing effects, the exponential function would be the feasible choice. Consequently, a possible function for representing the changes of the expression rate would be of the form $exp(kt)sin(mt)$ or $exp(kt)cos(mt)$, where k and m are real constants. The exponential functions have their shares in science, especially, in modeling problems and theories, so it is not surprising that exponential functions make their appearances here. Although we are comfortable and familiar with the function, once in a while, one might ask the following question: why do the exponential functions appear so often? Although the answer to this question is not obvious, I would like to justify my choice of the exponential function here. The first clue is in the profound experimental fact, i.e., that the radioactive decay is measured in terms of *half-life* – the number of years required for half of the atoms in a sample of radioactive material to decay. Mathematically this is expressed as

$$y^{'} = ky$$

Here y represents the mass and k is the rate constant. Then the general type of a solution looks like $y = Cexp(kt)$, where t represents time and C is a constant. Second, every object is made of atoms, bulbs, engines, living cells, enzymes, DNA, RNA so on, and, as we deal in this paper, in most of cases, it might be used to observe some sort of *life expectancy* of a certain phenomenon or behavior.

Here are a few examples of functions possible for fitting, depending on the nature and behavior of observed data.

Examples

1. $f(t) = C_1\ exp(k_1t)\text{sin}(m_1t)$ for $-a\leq t\leq 0$

   $C_2\ exp(k_2t)sin(m_2t)$ for $0\leq t\leq a$

2. $f(t) = C_1\ exp(k_1t)sin(m_1t+\varphi_1)$ for $-a\leq t\leq 0$

   $C_2\ exp(k_2t)sin(m_2t+\varphi_2)$ for $0\leq t\leq a$

   where $C_2sin\varphi_1 = C_2\ sin\varphi_2$

3. $f(t) = -m_1t+b_1,\ 0 \le \mathrm{t} \le \frac{\mathrm{b}_1}{\mathrm{m}_1}$

$$m_2t+b_2,\ \frac{b_2}{m_2} \le t \le \frac{b_3 - \mathrm{b}_2}{m_2 + \mathrm{m}_3}$$

$$-m_3t+b_3,\ \frac{b_3 - \mathrm{b}_2}{m_2 + \mathrm{m}_3} \le t \le \frac{b_3}{m_3}$$

$$\vdots \qquad \vdots$$

$$m_{2k}t+b_{2k},\ \frac{b_{2k}}{m_{2k}} \le t \le \frac{b_{2k+1} - \mathrm{b}_{2\mathrm{k}}}{m_{2k} + \mathrm{m}_{2\mathrm{k}+1}}$$

$$-m_{2k+1}t+b_{2k+1},\ \frac{b_{2k+1} - \mathrm{b}_{2\mathrm{k}}}{m_{2k} + \mathrm{m}_{2\mathrm{k}+1}} \le t \le \frac{b_{2k+1}}{m_{2k+1}}$$

,where $\frac{\mathrm{b}_{2\mathrm{i-1}}}{\mathrm{m}_{2\mathrm{i-1}}} = \frac{\mathrm{b}_{2\mathrm{i}}}{\mathrm{m}_{2\mathrm{i}}}$, for each *i=1,.., k*.

where *a, k's, m's* are positive real numbers.

Step 2. Determining coefficients, *C's* and *k's*

Once we fix a candidate function, it remains to determine the coefficients *C's* and *k's* etc., for each set of data. This may be achieved by using the least square sum principle with high accuracy set to our own standard. Commercial software, such SAS and SPSS, are available for such calculation, namely, R-squared. The least square sum method, as the most popular one for fitting a curve/function with experimental data, finds the coefficients of a function of given type, by minimizing the sum of square of errors, or deviations. More precisely, given a set of data points, (x_{1}, y_{1}), (x_{2}, y_{2})... (x_{n}, y_{n}) and a candidate function f with undetermined coefficients, the unknowns in f would be determined so that the summation of errors, given by

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2 ,$$

be minimized. As in the case of choosing a candidate, why do we have to use the power two? Why don't we use $\sum_{i=1}^{n} | f(x_i) - y_i |$, $\sum_{i=1}^{n}(f(x_i) - y_i)^3$ or $\sum_{i=1}^{n}(f(x_i) - y_i)^4$ for measuring the degree of errors? For odd number like three, the sum cancels each other and does not reflect our purpose, i.e., the degree of errors. When the exponent is two, it is the smallest and good for further manipulation, i.e., we could use many tools, calculus, involving differentiation unlike the absolute value function, | |.

Step 3. Representing each set of data as a vector

From the first two steps, we have obtained a functional representation for observed data, i.e., a function fitting with the data. Consequently, with respect to the fixed type, each function is represented as a set of coefficients calculated in the step two. Suppose we observe a single gene expression rate of an object, *A*, for a certain fixed period and, for simplicity, we assume for a moment that the data fits well with model,

$$y = Cexp(kt)sin(mt),$$

where y is the expression rate. Then we can say that the set of numbers (C, k, m) represent the object, *A*. In other words, the object may be identified with the triple

(*C, k, m*), analogous to students and their corresponding ID numbers. For the general case, i.e., the gene expression rates of *n* multiple genes, we will get (C_{1}, k_{1}, m_{1}), (C_{2}, k_{2}, m_{2}),.., (C_{n}, k_{n}, m_{n}), which form a vector, made by enumerating the n triples, in the 3*n* dimensional Euclidean space.

Once again, we are in a familiar position, ready to apply the support vector machine (or a machine learning tools such as neural network and decision tree etc.) to find a criterion for separating one from another. (For more details, see [1], [2], [3] and [4]).

§3. Discussions

The method in the previous section can be applied for observed events during a certain period. Reactions of drugs, any changes of substance in changing environments, and many others can be classified and used for a criterion of diagnosis of a disease or analysis of statuses.

On the matter of choosing a function of a proper type, besides the exponential function, in quantitative analysis of science including genetics, we often encounter linear functions, $ax+b$ where x is the variable, and a and b real constants. In many cases, the linear functions or, sometimes, quadratic functions are simply assumed without any explanations. For readers who are in a hurry to grasp the rest of the story, it is hardly expected to spend time in getting a convincing answer for that. I believe that we could find an answer to this on one hypothesis and a famous mathematical theorem of calculus.

1. Hypothesis

   Any quantitative measurement can be expressed as a function of some variables.

2. Taylor theorem

   Roughly speaking for a function of a variable, any smooth function f may be expressed as follows;

   $$f(x)=f(0)+f^{'}(0)x+f^{(2)}(0)/2!x^{2}+f^{(3)}(0)/3!x^{3}+\ldots\ldots+f^{(n)}(0)/n!\,x^{n}+R_{n}(x)$$

   where $R_n(x)=f^{(n+1)}(z)/(n+1)!\,x^{n+1}$

The variables in the hypothesis are not observable with modern technology or maybe impossible for human, but theoretically, we can assume that there exist such variables. For the last several decades, we have witnessed the rapid development of computers, super computing powers. This luxury enables us to raise the power of exponents, in other words, we do not need to restrict ourselves to function of degree one or two. This might be a bit wild, but what if we allow singularities in the quantitative function, then we might have to replace Taylor by the Maclaurin theorem.(See for [5])

References

[1] Chul Ahn and MyungHo Kim, “Support Vector Machines for the Estimation of Diagnostic Accuracy”, submitted

[2] Seung-chan Ahn, Gene Kim and MyungHo Kim, “A Note on Applications of Support Vector Machine”, (http://xxx.lanl.gov/abs/math.OC/0106166)

[3] Gene Kim and MyungHo Kim, “Application of Support Vector Machine to detect an association between multiple SNP variations and a disease or trait”, DIMACS workshop of Rutgers University, On the Integration of Diverse Biological Data, 2001 (http://dimacs.rutgers.edu/Workshops/Integration/program.html), submitted for patent (Application No. : 10/128,377)

[4] Gene Kim and MyungHo Kim, “Pattern Recognition of Protein 2D Gel Image and its Application for Diagnosis of a Disease”, DIMACS workshop of Rutgers University, On Complexity in Biosystems: Innovative Approaches at the Interface of Experimental and Computational Modeling, 2002 (http://dimacs.rutgers.edu/Workshops/Complexity/program.html), submitted for patent (US Application No: 10/336,334)

[5] Thomas and Finnley, “Calculus and Analytic Geometry” 9th edition, Addison-Wesley, 1996