|  |
| --- |
|  |
| Analyzing a Microarray Classification Algorithm |
| A Digital Image Processing Approach |
| CSC 570 – Digital Image Processing |
| Justin D. Orr |
| UIN**:** 656335315 |

12/2/2017

|  |
| --- |
|  |

INTRODUCTION

In the field of medical research, a large portion of particularly valuable time is lost to the process of analyzing data, classifying lab and research results. This is especially true for “on the edge” cancer research which requires meticulous analysis of tens of thousands of genes extracted from a living patient; many of which need results quickly so that treatments can begin. One form of technology used to study these genes is called Microarray technology. [1] Khalilabad explains in this journal article, “Microarray technology involves [a] gene chip, collections of gene sequences in known locations on a solid surface [where] each gene is call a spot. Images obtained from this chip consist of a matrix with several blocks and each block contains a number of rows and columns of spots.” Khalilabad goes on to introduce the proposed algorithm which takes as input a microarray image and outputs the diagnosis of the disease, if one is determined present. This algorithm will drastically reduce the need for human presence during the analysis phase and save massive amounts of time, better utilizing prepping microarrays and handling the results returned.

CURRENT ALGORITHMS

Prior to the proposal of this paper, there were a few programs for analyzing microarray images. [1] “These [programs] quantify microarray images by using different algorithms in the gridding, segmentation and extraction of light intensity proportional to the gray level of spots.” However, there are a few large issues with different phases of the analysis these programs run.

For example, in the gridding phase; the phase in which the spots are recognized, some of the programs used have difficulty discovering the spots if noise is a major factor. Others have a difficult time if the spots to be discovered are not uniform in their displacement across the microarray. Either of these two issues can cause for a loss of significant data, and can lead to improperly classified results. In the segmentation phase; the phase in which the foreground spots are retrieved from the background image, most of these programs use the best-known algorithms. However, the issue lies in the fact that user input is required to tune the methods based on the extracted images.

[1] Khalilabad explains it best when the paper states “Most image analysis software require the user to set several parameters and subsequently manually adjust individual spots to obtain satisfactory gridding results. Moreover, when images have problems such as noise and rotation, the spots are not properly recognized.”

IMAGE DATASET

This algorithm is based on microarray images which are images of a scientific tool known as a DNA microarray. [2] The National Human Genome Research Institute states that “the DNA microarray is a tool used to determine whether the DNA from a particular individual contains a mutation in genes and consists of a small glass plate encased in plastic. On the surface, each chip contains thousands of short, synthetic, single-stranded DNA sequences, which together add up to the normal gene in question, and to variants (mutations) of that gene that have been found in the human population.” In other words, a clinical trial retrieves a DNA microarray from a candidate and a high resolution image of the microarray is created. This allows for the clinical trial to retain the living glass plate and have only the image needed in repeated testing.

As stated previously, the layout of the microarray is a combination of “spots” spread evenly across a rectangular plate. There are multiple blocks per plate which are broken down into equal rows and columns. These rows and columns correlate to different gene sequences ready for analysis.

PROPOSED ALGORITHM

ROTATION DETECTION

This algorithm begins where the currently accepted algorithms do not and this is by checking for a skewed microarray plate during image creation. There are times when the technician loads the plate with a non-zero rotation angle and starting with the correct image is critical as it will affect all steps of the algorithm going forward. This non-zero rotation angel is corrected using the Radon transform method which is obtained by the integral transform consisting of the integral of a function over straight lines and is defined as [1]:

IMAGE BLOCKING

After the image has been corrected for any non-zero angles, the images blocks must be discovered. These blocks are defined in rows and columns and indicate the different sequences to be tested in the image. In order to more accurately see the blocks, the images contrast ratio is increased using a method called histogram equalization. It is possible for small white spots to appear in the background of the image, which are then removed using the Wiener filter to reduce noise. Finally, the pixel values in each row are summed together and each block is discovered by identifying the local minima value.

BLOCK GRIDDING

Despite having dealt with the noise introduced by the contrast ratio, there are actually two other types of noise that can be present in a microarray image: dust particles present on the slide, and lens reflection or light anomalies created on the slide plate during image creation. To remove the dust particles, the average brightness of all pixels is taken and compared to every pixel individually. If its value is less than the average, the value is set to zero. The second type of noise, light interference, is controlled during the process of gridding each spot in a block. This is handled using the morphological operator’s erosion and closing, which removes the interference and highlights the individual spots. Lastly, pixel brightness is calculated using the following equation:

The local minima discovered in this summation of the images rows are the location in which spots exist.

SEGMENTATION

The segmentation process is the most complicated step in the analysis of the image. It involves a process by which the image is split into two parts: the foreground (the spots) and the background (everything else). In order to do this, [1] the paper utilizes the association of a total variation regularization method with an L1-norm precision term (TV + L1) and applies the two to a complex formula for isolating the spots:

The result of this formula is a gray level matrix related to the spots in which the background of the image has been removed.

GENE EXPRESSION CALCULATION

After segmenting the spots from the background, the next step is to compare the spots to one another via gene expression. On a standard microarray, there are samples of both the suspected unhealthy tissue, as well as known healthy tissue. Gene expression is a series of mathematical operations which compares the dissimilarities and similarities between the individual genes. This process takes two equations, the first of which is used to calculate the brightness intensity:

Next, this intensity is used to calculate the expression level of each of the spots. In this equation, ‘cy5’ is the unhealthy gene sample and ‘cy3’ is the healthy gene sample:

If the value calculated here is close to zero, it means that the cy5 and cy3 values are equal, whereas if the values are greater or less than zero in a significant way it means that there is a strong deviation in the genes.

NORMALIZATION

After calculating the gene expression values for each of the spots, it then necessary to normalize the values in order to see the ‘biological differences’ between them. This is performed using a normalization function where [1] “the first term is related to the gene expression, the second term shows the fitted curve f(x). The Loess approach estimates f(x) of gene expression versus spot brightness.”

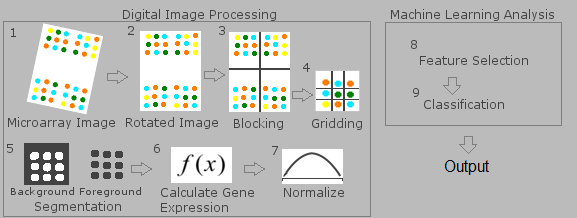
GENE SELECTION

Gene selection is the first step in the two step machine learning process. The data created up to this point is all a product of digital image processing methods. With that in mind, since so much data has been calculated, an information gain technique must be utilized to calculate how ‘relative’ the data is to the values necessary for accurate classification. This technique calculates the entropy of a sample set of data for any one feature that is being analyzed and highlights values that should be used in the classification step.

DATA CLASSIFICATION

In the final step, the values calculated during gene selection are used to classify the data and indicate the type of tumor discovered and the stage that the disease could be in. A classification algorithm known as the J48 decision tree algorithm is used in that instance. The J48 decision tree is an altered version of the standard C4.5 classification algorithm which uses entropy and information gain to run its calculations. The results from this classification stage is the type of tumor and the stage in which it is in, both of which are then used to determine the treatment options for the patient going forward.

FLOWCHART



RESULTS

The research discussed is based on three major diseases: breast cancer, lymphoma cancer, and myeloid leukemia. There are various parts that need to be tested in order to prove the accuracy of the results. The first of which is its ability to calculate the rotational angle of the inserted microarray plate. Out of multiple tests run, the algorithm was able to calculate the degree of rotation with anywhere from 98% to 100% accuracy. Just as well, the blocking and gridding accuracy was calculated with a range of 98.8% to 100%. The segmentation, gene expression, and normalization methods are mostly mathematical operations and therefore don’t require a substantial accuracy analysis. The remaining steps are feature selection and classification. With any form of machine learning or data mining, the performance of the classification is based on the strength of the training data provided. This system was tested using the top one hundred genes from each of the three major cancers. The accuracy was calculated by whether or not the provided genes were analyzed to be the appropriate type of cancer as well as the stage discovered. With this in mind, the accuracy for breast cancer, lymphoma cancer, and myeloid leukemia was 95.45%, 100%, and 94.11%, respectively.

ADVANTAGES

There are two clear advantages to this algorithm, the first of which is all of the steps from the slide being photographed to the resulting classification, is in one process. There is no need for external alterations or interference which could cause compounded issues from human error. Human error is a significant factor in the need to automate processes as it can have an effect on any stage of analysis, invalidating the results.

The second advantage to this algorithm is the ease of upgrading and testing it due to new research discoveries. The entire algorithm, in its essence, is mathematical. If any new method is discovered for image rotation or gene expression calculation, it is a simple process to update the segment of the algorithm. This creates a modular and “future proof” design that is essential in a field based on constant research and new revelations.

DISADVANTAGES

However, this algorithm does not come without some significant disadvantages. A major disadvantage of machine learning and classification is a particular set of issues. A data set can be over or under trained and can cause classification to return a large number of false positives or true negatives. It is a delicate art to find the right amount of varied data to train your model correctly. This issue is compounded by the number of diseases the algorithm needs to be able to analyze. Each new disease calls for a long training process on a wide range of already known data.

IMPROVEMENTS

Increasing precision and accuracy in an algorithm such as this is the most important issue. Reducing false positives and true negatives and increasing true positives and false negatives is significant. Though the resolution of the images is not discussed, the first place an improvement could occur is with the quality of the images being analyzed. As image resolution continues to grow, the machine running the analysis needs to be at the forefront to guarantee results are always growing in accuracy.

Another improvement that could be made is in the training dataset used in the data mining process. This algorithm is currently being trained on only one hundred genes; increasing this number would increase the accuracy of the classification phase. Though this does run the risk of overtraining the data, it also has the promise of improving the algorithm itself.

Overall, this algorithm is a great advance in the manner in which certain types of cancer are analyzed and diagnosed. The more accurate this algorithm becomes, the faster patients can learn that they may have a disease and treatment options can be discussed. This is a necessary advance in the right direction for cancer research and disease detection.

REFERENCES

[1] Nastaran Dehghan Khalilabad and Hamid Hassanpour. 2017. Employing image processing techniques for cancer detection using microarray images. Computers in Biology and Medicine 81 (2017), 139–147. DOI:<http://dx.doi.org/10.1016/j.compbiomed.2016.12.012>

[2] Anon. DNA Microarray Technology. Retrieved December 2, 2017 from https://www.genome.gov/10000533/dna-microarray-technology/