Student Research Contest

Extended Abstracts

1. Information Technology to Support Student Enrollment for an Innovative College Sustainability
   Robert Brewer, Davis Simpson, Georgia Gwinnett College.
2. Investigating Face Recognition using Hyperspectral Images and Principal Components
   Marco Chang, Montclair State University.
3. Load Balancing: Viewing Breast Cancer Tissue X-Rays using Graphic Processing Units
   Nicole Dawson, Spelman College.
4. Accelerating Filter Back Projection on a Graphic Processing Unit
   Stevie-Marie Hawkins, Spelman College.
5. The Robot Snack Food Coach for Reducing Childhood Obesity : Developing a Vision Algorithm
   Amelia Henderson, Spelman College.
6. Computer Analysis of Plant Immune Response
   Whitney Sanders, Wofford College.
7. Semi Automatic Parallelization of Multiple GPUs
   Rashidah Slater, Spelman College.
8. Information Technology to Support the Analysis of the College Growth and its Influence on the Traffic Dynamics
   Curtis St. John, Georgia Gwinnett College.
9. Improving the Accuracy of Lung Cancer Diagnosis by using Datamining Techniques
   Nam-Phuong Tran, The University of Texas at Dallas.
10. Development of a Web Tool for Teaching Principles of Chromatography with Mass Spectrometric Detection
    Thresa Wells, Spelman College.

Information Technology to Support Student Enrollment for an Innovative College Sustainability

Robert Brewer, Undergraduate Information Technology Student
Davis Simson, Undergraduate Information Technology Student

Advisor: Dr. Anatoly Kurkovsky
Georgia Gwinnett College, a member of the University System of Georgia

Abstract

Within this project we demonstrate how to implement the United Nations initiative on achieving sustainability in educational areas for the newest innovative higher education institution: Georgia Gwinnett College that was established several years ago as a unit of the University System of Georgia. We analyze a relatively small portion of the Georgia Gwinnett College sustainability related to system analysis of the enrollment dynamics at the college. Various Information Technologies play a pivotal role here because our systems analyze is intended to produce various data and statistics for a future computer simulation that can best achieve the purpose of estimating faculty needs. The goals of the project are to discuss and to formulate the problem of the college's student body growth and how it is affects faculty needs from systems dynamic point of view. We have looked at the conceptual model our system, its components as well as the particular relationships that exist between the identified components.

There exists several simulation packages in our area of interest; these include professional packages such as Virtual U. This software simulates entire university systems with very detailed settings and outputs, including specific enrollment numbers, faculty requirements, and resource needs The project includes a study from the ground up to understand the basics and essentials of analyzing university growth for a specific future model. We have identified boundaries within our system to set a clear scope to our analyses and have a clear focus on data that needs to be formulated. It prevents over-analyzing of components that are tied into our specific area of focus. Various information sources and approaches have been used for the identified subject domain: from traditional concepts on

system analysis to particular data on university enrollment. They have been used to collect data, establish information flows, and to analyze that data to properly document the enrollment process at the college.

We identified a set of specific system variables that can be applied to a system model. The variables were divided into a set of

a) input: I1 – Students currently enrolled (Deterministic); I2 – Transfer students – in and out (Stochastic); l3 – Students who graduate from college (Deterministic); l4 – Students who drop out of college (Stochastic/Deterministic); I5 - Faculty to student ratio (Constant – Max number of students for a Professor),
b) output: O1 – Number of expected students for future semesters; O2 - Number of faculty needed for Future Semester
c) control: C1 - Maximum Number of Faculty Hired

subsets variables based on their relationship to the system. These variables are further analyzed to be either stochastic deterministic or both. After looking at enrollment data and breaking down the various numbers we decided that in addition to knowing the current number of students enrolled we would also need the number of students who transferred, students who graduated, and students who dropped out of the college. We have established the information flow of these variables in order to create the simulation model.

The obtained results of our analyses make it is possible to create and run a special simulation model. This model will give as a possibility to predict the enrollment growth of the college and to support the college sustainability. After looking at the broken down enrollment data, it was decided that in order to make the simulation of enrollment as precise as possible meant creating equations for estimating each of the individual variables. By using these various variables and equations to model enrollment and subsequently estimating faculty needs through simulation, we will try to greatly increase the sustainability of the college.

Keywords: Higher Education, University System, System Dynamics, Sustainability, Simulation.

Investigating Face Recognition using Hyperspectral Images and Principal Components

Author:
Marco Chang
Montclair State University, Montclair, NJ, changreynam1@mail.montclair.edu

Faculty Sponsor:
Dr. Stefan Robila
Department of Computer Science, Montclair State University, Montclair, NJ 07043 robilas@mail.montclair.edu

Abstract:
The human eye is able to perceive only a limited range of the electromagnetic spectrum (400 to 800 nanometers) and combines the sensed information into a single color composite image. Considerably richer in expressivity, hyperspectral data are collections of tens or hundreds images of the same scene with each image corresponding to the reflected energy within a narrow spectrum interval and the overall range of the spectrum often far exceeding the human sensing ability (e.g. including the infrared and far infrared ranges). Due to this, hyperspectral sensors have been employed on aerial and satellite platforms for a variety of applications ranging from crop assessment and monitoring, mineral identification, and pollution monitoring, to camouflage detection, and land cover classification. More recently, as the technology costs decreased, hyperspectral sensing has been introduced in new fields such as medical imaging, industrial quality control, or law enforcement.

Face recognition, i.e. the ability to identify or verify a person based on facial traits is a task easily solved by humans but still difficult to tackle by an automated computational system. While a tremendous body of work has been done in bridging the gap between the human and the computer based recognition, most of the research and applications have focused on grayscale or color images that are similar to the ones seen by the eye. It is worthwhile though to investigate what contributions would hyperspectral images make to improve recognition, given that the data provide information beyond the normal human perception. Previous studies have identified parts of the infrared spectrum (that reveal blood vessels in the face) as promising to improve face recognition.

Our work has focused on a set of hyperspectral face images collected over two years. It included 17 subjects, each with five facial expressions and viewing angles (see figure below for an example). Each hyperspectral image is 640x640 pixels, 120 bands/images (5nm wide) covering the 400 to 1000 nanometers interval. The images were cropped during the data collection to 200x200 pixels. The images were collected using a Surface Optics (SOC) 700 instrument available at the host institution (Montclair State University).

Three distinct steps were employed: *data fusion, face detection, and face recognition*. In the *data fusion*, various approaches can be used to reduce the 120 bands to a single grayscale image while preserving information from across the covered spectrum. This was done using a weighted average of each pixel across the bands where the coefficients were computed based on Bilateral Filtering. For each pixel in an image, Bilateral Filtering computes the difference between pixel and the weighted average of intensity values from nearby pixels. The normalized differences for the same pixel across the bands are being used as weights for the fusion. In *face detection*, a simple skin texture detection algorithm was employed to delimitate the face area and the enclosing rectangle was created. To ensure uniformity across the images, the largest rectangular shape among the images was selected as pattern rectangle. For each image, the original rectangle was expanded to match the pattern rectangle, while trying to ensure the center of the shape remains the same.

Finally, for face recognition, we employed eigenfaces. Given a collection of face images, the eigenfaces are generated by performing Principal Component Analysis (PCA). Introduced in the last decade or so, eigenfaces are often considered to be the set of base vectors from which any original

image in the set can be reconstructed (using different sets of coefficients). Projecting a mean subtracted image on the eigenfaces we obtain set of coefficients that indicate how much the new image differs from the mean face. Images with coefficient vectors close to each other (according to the Euclidean distance) can be considered similar, whereas images with coefficients significantly separated from each other can be considered different (i.e. different subjects).

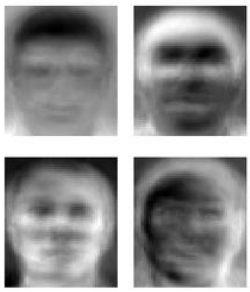


Fig 2. Examples of Eigenfaces

Our work was implemented in Matlab, allowing for the use of the many vector and matrix manipulation operators and functions available. To test the approach we have randomly selected one face for each subject as test and considered the other four faces as training data.

To test the impact of Bilateral Filtering we compared the results with ones obtained from regular band averaging or and averaging of selected visible blue and green bands and an infrared bands. Our results show that Bilateral Filtering based averaging slightly improves regular averaging. We expect that such improvement to be increased when the filtering is done of spectra segments and not on all the bands.

Next, we investigated the impact of face detection on the recognition accuracy. In this case, the accuracy decreased. We believe that this is due to the fact that, as the images are now focused on the face itself, the ability of outside features (such as background shape or texture) to influence the results is being eliminated. As such, if the images available would have had a more varied background, the results on full images are expected to have been significantly reduced. As future work we plan to expand the dataset library.

We note that as part of the same project we have also investigated face recognition based on spectra, i.e. on the matching of pixel vectors extracted across the bands. The work, performed by another undergraduate student (Nisha D'Amico) is presented separately but includes as common component the face detection step. In that case, the automatic cropping has resulted in increased accuracy and provided an automated way for spectra selection. A general future direction of the project is the fusion of spectral based and image based recognition is a single automated system.

Overall, our work shows promise on the use of "beyond the human eye" vision and warrants additional efforts.

Acknowledgement:
This work was supported by the National Science Foundation's REU Program under Grant Number IIS-1004447. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Load Balancing: Viewing Breast Cancer Tissue X-Rays using Graphic Processing Units

Nicole Dawson
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
ndawson1@scmail.spelman.edu

James Hale
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
jhale@spelman.edu

Abstract

The Load Balancing: Viewing Breast Cancer Tissue x-rays by using Graphic Processing Units (GPU) project focuses on gaining higher quality breast cancer images, improving breast cancer diagnosis and the overall improvement of traditional mammography. This project seeks to improve the many load balancing algorithms with a concentration in getting more breast cancer images from traditional mammography and analyzing those images with the help of graphic processing units.

This Load Balancing algorithm will take the images given from the mammogram and, instead of analyzing the images one at a time or with a Central processing unit, give each image to a Graphic processing unit to be analyzed. This method would be a better method because Graphic processing units specialize in viewing images and image reconstruction (which is what we need in order to look at Breast Cancer tissue x-rays) and because the image analysis would be faster because of the GPU's specialization.

Materials and Restrictions

In order to understand the Load Balancing Project a series of tutorials and information was referenced. **Load Balancing, GPUs, Breast Cancer and how it is diagnosed and detected, the programming language OpenCL, and Pthreads** were all items that needed to be understood,

Load Balancing is normally used in computer networking, it is described as a technique to distribute workload evenly amongst two or more computers, network links, CPUs, hard drives and or other resources (**http://en.wikipedia.org/wiki/Load_balancing_(computing)**). The project at hand will focus on using a typical Load Balancing algorithm to divide images amongst Graphic Processing Units.

Graphic processing units, also known, as GPU's are the computationally intensive part of an application. Graphic Processing units specialize in rendering 2-D and 3-D images. GPUs are very efficient at manipulating computer graphics, and their highly parallel structure makes them more effective than general-purpose CPUs for a range of complex algorithms (http://en.wikipedia.org/wiki/Graphics_processing_unit).

Breast Cancer is an uncontrolled growth of breast cells. There are 3 common ways to detect Breast Cancer; Clinical Breast Examination where a physician feels the breasts for lumps or irregularities, the Breast Self-Exam in which a patient checks breasts for lumps or irregularities and finally, the focus of the Load Balancing Project, the mammogram, this process uses low-dose amplitude x-rays to examine the human breast and is used as a diagnostic and a screening tool(http://www.breastcancer.org).

The programming language OpenCL and Posix threads were used for implementation purposes. OpenCl is the language that is easily recognized by GPU's. Posix threads or Pthreads were used to divide up the images and pass them to the GPU's.

The platform used to implement the project was NVIDIA's GPU platform. NVIDIA is a multinational corporation which specializes in the development of graphics processing units and chipset technologies for workstations,personal computers, and mobile devices (http://www.google.com/search?hl=en&defl=en&q=define:Nvidia&sa=X&ei=y_SHTM6HJ4L7lwfL6On0Dg&ved=0CBIQkAE). Throughout the research NVIDIA's programming guide was used to understand the structure of their GPU's and programming in OpenCL.

Research Problems & **Future Research**

I studied under Dr.David Kaeli at Northeastern University and one of his graduate students Dana Schaa.

1.One problem was finding an equation that divided 15 images amongst 8 graphic processing Units. The problem at hand was how do we divide the images amongst each GPU. To solve this problem several equations where considered but because 8 does not divide into 15 evenly we ran the risk of having a GPU with half an image or a GPU with too many images.

2.Also the algorithm could not be specific to the needs of 15 images and 8 GPU's. The algorithm would need to be very broad so that any amount of GPU's and images could be used and implemented.

Because these 2 problems were not solved due to time constraints, some future research would be finding n equation that is both broad and able to handle odd amounts of images and/or GPUs in order to be analyzed.

Accelerating Filter Back Projection on a Graphic Processing Unit

Stevie-Marie Hawkins
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
shawkin7@scmail.spelman.edu

James Hale
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
jhale@spelman.edu

Abstract

The goal of this project was to port a serial CT (computed tomographic) image reconstruction algorithm on a GPU. The image reconstruction used was called Filtered Back Projection, also known an FBP, which is the most commonly used algorithm in tomographic construction. The process of FBP includes running projections back through an image, hence the name back projection, to obtain a rough approximation to the original image. Then a high pass filter is used to eliminate any blurring that might happen in the rough approximation. Before the algorithm was run to collect data, there was a lot of background work that had to be done first. The initial algorithm was written in a language called CUDA, Compute Unified Device Architecture, which is a parallel computing architecture developed by NVIDIA.

Since the algorithm was written in CUDA it needed to be written in OpenCL, which stands for Open Computing Language. OpenCL was initially developed by Apple in 2008, but is now managed by a non-profit associated call the Khronos Group. It is a framework for developing applications that execute on a range of heterogeneous platforms, including GPUs and CPUs, and is the first open, royalty-free standard for cross-platform, parallel programming frameworks. OpenCL has many purposes, but two major ones. The first is to harness the parallel computing power of GPUs to accelerate data-parallel workloads. The second being to allow programmers to easily target multiple platforms, such as multi-core CPUs and the latest GPUs. The difference between CUDA and OpenCL is that CUDA can only be ran on an NVIDIA platform and OpenCL can ran on both AMD and NVIDIA platforms.

The Robot Snack Food Coach for Reducing Childhood Obesity : Developing a Vision Algorithm

Amelia Henderson
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
ahende18@scmail.spelman.edu

Andrea Lawrence
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
lawrence@spelman.edu

The Robot Snack Food Coach for Reducing Childhood Obesity : Developing a Vision Algorithm

Abstract

The growing increase in childhood obesity is a serious public health concern facing the United States today. Obesity has been attributed to a combination of genetic, behavioral, and environmental factors. The numerous health problems resulting from childhood obesity include high blood pressure, high cholesterol, Type 2 diabetes, asthma, liver disease, and sleep apnea. Also, obese children often suffer from low self-esteem, anxiety, and depression.

One means of addressing childhood obesity is through changing behavior by educating children on proper nutrition and healthy food choices. An effective and interactive way to provide this education will be through programming a robot to act as a snack food coach. This research explores the development and use of a vision algorithm to detect barcodes on snack foods.

Computer Analysis of Plant Immune Response

Whitney Sanders
Wofford College
Spartanburg, SC
sanderswe@email.wofford.edu
Faculty Sponsor: Dr. Natalie Spivey, Biology

Just like animals, plants regularly encounter bacterial pathogens. Upon infection, the plant induces a defense response by altering its gene expression. The mechanisms that regulate this transcriptional reprogramming are largely unknown. Additionally, some pathogens likely have the capability to alter the plant's defense response, meaning that some of the changes in gene expression may be initiated by the pathogen.

I sought to better understand the plant immune response by studying gene expression data taken from an experiment in which Arabidopsis plants were injected with two different strains of the bacterial pathogen Pseudomonas syringae. The experiment recorded expression levels of 30,000 genes at five time points, for six treatments. This massive amount of data is typical in the field of bioinformatics, which involves problems in biology that must be solved using computing.

Plotting expression levels of nearly all the Arabidopsis genes over five time points allowed comparison of the responses to the control, avirulent, and virulent treatments. Genes were clustered together based on similar behavior during infection. I found prominent patterns of expression seen in the 48 hours after injection. In those prominent clusters of genes, I looked for enriched functional characterization. I grouped commonly expressed genes (genes which are clustered together in the three treatments) and looked for enriched functional characteristics and enriched transcription factor binding sites in these groups.

The previous analysis suggested several transcription factor binding sites of interest. I looked for the clusters most heavily populated by the genes whose expression is likely regulated by these binding sites. Finally, I analyzed those subsets of genes for functional characterization.

The results of my analysis provide the basis for experimentation by a plant geneticist to further our understanding of the interaction between plants and plant pathogens.

Semi Automatic Parallelization of Multiple GPUs

Rashidah Slater

Rashidah Slater
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
rslater@scmail.spelman.edu

Alfred Watkins
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
alwatkins@spelman.edu

The overall goal of this research effort is to create a tool that utilizes multiple graphics processing units (GPUs) semi-automatically. The use of semi-automatic parallel GPUs can be used to solve problems in a litany of fields such as architecture, gaming, and biomedicine in which rendering several high detailed images is frequently necessary. Today, imaging systems used in biomedical applications typically rely on one or more central processing units (CPUs) with at most one GPU to process images. These systems tend to take a significant amount of time to render images that are often of insufficient quality. When patients go to the hospital to get screened for things like cancer, it is important that the images that are taken are as accurate and detailed as possible so that doctors will be able to more easily detect abnormalities in a more timely fashion. This tool will allow for faster rendering times and better quality images to help enhance the biomedical imaging process for doctors and patients.

This project utilizes a new industry standard programming language called OpenCL to program GPUs. OpenCL can be used to program GPUs made by NVIDIA and AMD, two major companies that produce graphics processing units. Therefore, the choice to use OpenCL provides the hope of code portability and should make performance analysis and comparisons across different GPUs much easier.

Graphics processing units are designed to offload 2D and 3D graphics rendering tasks from the CPU(s) in computationally intensive applications. When multiple graphics processing units are used, more computations can be done in parallel, consequently reducing the computation time. A wrapper file was created to encase all the necessary functions needed for writing a multi-GPU program in an effort to make it general and effective enough to be used with several different programs. The algorithm will include POSIX threads to help distribute data more efficiently, in turn using less system overhead. For the larger images the algorithm will use Message Passing Interface

(MPI), which will allow the program to access needed hardware from other machines. Initial testing of the multi-GPU algorithm is being conducted on NVIDIA GPUs exclusively to prove the concept and to accurately measure the impacts on performance as changes are made to the algorithm's parameters. In the future the algorithm will be run on AMD GPUs as well to compare their run times to that of NVIDIA's.

The use of semi-automatic parallel graphics processing units is predicted to become wide spread as costs come down and better algorithms for effective utilization are developed. Ultimately, these algorithms should positively impact the quality of care doctors can give their patients.

Information Technology to Support the Analysis of the College Growth and its Influence on the Traffic Dynamics

Author: Curtis St. John, Undergraduate Information Technology Student
Georgia Gwinnett College
Lawrenceville, Georgia
Cstjohn1@ggc.edu
Faculty Sponsor: Dr. Anatoly Kurkovsky, School of Science and Technology

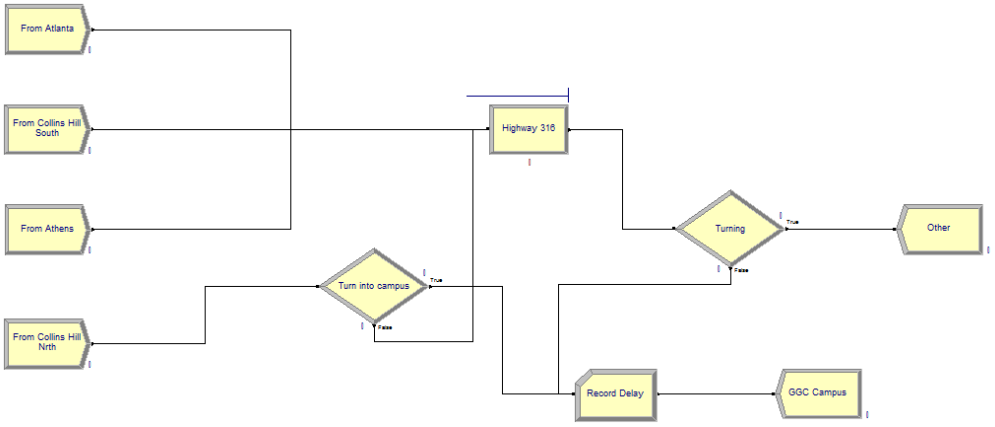
Abstract

Georgia Gwinnett College's (GGC) sustainable growth depends on many elements such as student and faculty, the economy, its facilities, and technology. It needs to have a good balance of all of these things in order for it to expand. A great number of students and faculty drive to the campus, and have to either drive down or cross over Highway 316. These people, along with other motorists driving to and from the Metro Atlanta area, are increasing the traffic on Highway 316 nearby the college. This has become so bad that at peak hours, during the morning and in the afternoon and takes an alarming amount of time to get to one's destination. The expansion of the college's student body every semester continues to put pressure on traffic, increasing the amount of time it takes to arrive at the college campus from one's home.

Our goals within the project are to discuss and to formulate the problem of the college's growth and how it is affected by highway 316's traffic from systems dynamic point of view. Our system consists of the intersection of Highway 316 and Collins Hill Rd. The conceptual model for our system is made up of factors from both inside and outside the system. Some of the inside factors include the current traffic made up of the college population, the growth of this traffic, current travel time, and future travel time that are relative to the growth of GGC. Factors that affect our system from the outside are peak traffic times, traffic on GA 316, and delays on GA 316.

During this study, we have found that there are many simulation models that pertain to our particular system of traffic. For example, Transmodeler and miniTraff are models that simulate traffic flow and with various analysis such as accidents, number of lanes, traffic lights, and natural disasters. The Statistical Visual Computing Lab provides a new way of gathering data about traffic dynamics by using traffic videos. By doing this they can determine, car speeds, individual lane speeds, and lane occupancy. The Gwinnett County Department of Transportation shared with us real traffic information that we have used as a primary source of data for our simulation. This data source includes particularly the amount vehicles that go in each direction for the GA 316/Collins Hill Rd intersection per hour.

The software we have chosen to use for our model is ARENA simulation software. It has proven a very useful and efficient tool. ARENA is graphical professional simulation environment that can be used to simulate many scenarios, and systems, including our own. Below is the base simulation model, made in ARENA, which we used for the project.



We identified three types of system variables: input, output, and control. Some of our input variables are the college's population per semester, and the college traffic, which is the amount of the college vehicles that contribute to traffic. Our output variables include the average travel time to the college, and the change in travel times during peak hours, which is our delay. Our control variable is the amount of roads that allow access to the college campus, the amount of the road lanes available, and the traffic light rules.

Keywords: Higher Education, System Dynamics, Sustainability, Traffic, Simulation

Improving the Accuracy of Lung Cancer Diagnosis by using Datamining Techniques (Extended Abstract)

Nam-Phuong Tran
The University of Texas at Dallas,
Richardson, Texas
kelly.v.tran@gmail.com
Faculty Advisor: Dr. Quoc-Nam Tran, Lamar University CS Department*

September 10, 2010

Approximately 1.5 million new cancer cases are expected to be diagnosed in 2010 [6]. The past decade has seen an explosive growth in biomedical research including the identification and study of the human genome by discovering large-scale sequencing patterns and gene functions [5]. As a result, new techniques are needed to analyze, manage, and discover sequence, structure, and functional patterns or models from these large sequence and structural databases [8]. Somatic alterations in cellular DNA underlie almost all human cancers [7]. In recent years, microarrays have opened the possibility of creating the data sets of molecular information to represent many systems of biological or clinical interest. However, finding the responsible gene for a cancer is not an easy task because a typical microarray dataset has only a small number of records while having up to thirty thousand of attributes. This kind of dataset creates a high likelihood of finding false predictions that are due to chance because irrelevant and redundant attributes have negative impacts on the accuracy of classification algorithms. Finding the most relevant genes is often the key phase in building an accurate classification model. Since many interesting sequential pattern analysis and similarity search techniques have been developed in data mining, data mining has become a powerful tool and contributes substantially to DNA analysis. Data mining refers to extracting or "mining" knowledge from large amounts of data [5]. One form of data analysis is data classification. Data classification is the process of finding a set of models that describe and distinguish data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Classification is a two-step process. In the first step, a model is built. In the second step, the model is used for classification. The predictive accuracy of the model is estimated. The accuracy of a model is the percentage of samples that are correctly classified by the model. If the accuracy of the considered acceptable, the model can be used to classify future data tuples for which the class label is not known.

In this Research Experience for Undergraduate (REU), I experiment many different datamining techniques for diagnosing the lung cancer disease using a dataset of lung cancers from the Broad Institute [2]. I used Weka, a data mining software, to implement data mining techniques [3, 1]. Weka is a collection of machine learning algorithms and data preprocessing tools. It provides implementations of learning algorithms that you can easily apply to your dataset. First, I used

*Supported in part by an NSF Research Experience for Undergraduate grant at Lamar University. Reference: Dr. Kami Makki (KAMI.MAKKI@lamar.edu)

attribute evaluator algorithms to rank the attributes. I took different number of best attributes given by each algorithm to create new datasets. Classification may need to be preceded by relevance analysis, which attempts to identify irrelevant attributes that do not contribute to the classification process. These attributes can then be excluded. Many of the attributes in the data may be irrelevant or redundant to the classification task. As a result, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. Including such attributes may otherwise slow down, and possible mislead, the learning step [5]. I then applied classification methods to the new datasets. Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size. As a result, data preprocessing is a very important issue for data mining. Data preprocessing is an important step in data mining, since quality decisions must be based on quality data. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the classification process [5]. One technique for data preprocessing is data reduction. Complex data analysis and mining of huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data reduction can reduce the data size. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results. One method of data reduction is discretization, where raw data for attributes are replaced by ranges. Discretization techniques can be used to reduce the number of values for a given continuous attribute by replacing actual data values with intervals[4].

We compares the accuracy of 17 different classifiers (figures in the Appendix give details of the diagnosing accuracy for lung cancer for each of these classifiers) on datasets that have been filtered. Three different methods have been used to select small sets of relevant genes. We also compare the accuracy of 17 different classifiers on discretized attributes.

In conclusion, many of the attributes in the data may be irrelevant or redundant to the classification task. Relevance analysis, with the aim of removing any irrelevant or redundant attributes, helps improves classification accuracy and efficiency. Data preprocessing techniques can improve the accuracy and efficiency of the classification process by improving the quality of the data.

References

[1] *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2008.

[2] http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, 2009.

[3] http://www.cs.waikato.ac.nz/ml/weka, 2009.

[4] U. Fayyad and K. Irani. The attribute selection problem in decision tree generation. In *Proc. of AAAI-92*, pages 104–110, 1992.

[5] J. Han and K. Micheline. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2 edition, 2006.

[6] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun. Cancer statistics, 2009. *CA Cancer J Clin*, 59:225–249, 2009.

[7] B. Weir, X. Zhao, and M. Meyerson. Somatic alterations in the human cancer genome. *Cancer Cell*, pages 433–43, 2004.

[8] N. Ye, editor. *The Handbook of Data Mining.* Lawrence Erlbaum Associates, 2003.

DEVELOPMENT OF A WEB TOOL FOR TEACHING PRINCIPLES OF CHROMATOGRAPHY WITH MASS SPECTROMETRIC DETECTION

Thresa Wells
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
twells8@scmail.spelman.edu

Alfred Watkins
Spelman College
350 Spelman Ln.
Atlanta, GA 30314
alwatkins@spelman.edu

ABSTRACT

A major challenge in interdisciplinary collaborations is a basic understanding of the fundamental principles that are essential in the adjoining disciplines. As computer science becomes an integral part of the applied sciences (biology, chemistry, physics, etc.), it is increasingly important to strengthen the common understanding between all participating scientists to enhance the efficiency of the data processing software. In this presentation we focused on the development of a web tool that presents the basics of the technology of gas chromatography with mass spectrometric detection (GC/MS). GC/MS instruments are used in a number of important industrial, environmental and forensic applications in which complex samples of organic compounds are separated and identified for substance characterization. We have been developing a novel technique, comprehensive two-dimensional gas chromatography (GCxGC) that provides separation improvements over GC/MS, but that require advanced data processing techniques for the interpretation of the information contained in the complex multi-dimensional images. In order to provide an interactive platform for our chemistry and computer science students to learn the intricacies of signal processing in GC/MS and GCxGC, we have developed an anthropomorphic model of chromatography to generate signals that are analogous to those that are commonly seen in the aforementioned technologies.

In chromatography, a mixture of chemical compounds are separated by their physical properties to produce a signal in which the x-axis denotes the retention time (or the amount of time it takes for a given compound to exit the system) and the y-axis is a measure of the amount of the substance in question. In our anthropomorphic model, we transpose chemical compounds to human beings in order to explain the separation process to non-chemists in a more perceptive manner. Human beings, like chemicals, can also be distinguished in terms of a number of characteristics, such as their names, heights, ages, etc. We use an "instrument" called a namograph to separate mixtures of names, and the resulting namograms are thus displays of the separation of these names in which the x-axis is a measure of the name length and the y-axis is a measure of the person's height. In addition to the namograph, an alphabetograph (analogous to a mass spectrometer) is also used to further elucidate the identity of the compounds in the mixture. Using this model, we can explain more subtle aspects of chromatographic separations. Concepts such as stationary phase selection and co-elution of analytes, as well as the challenges of one-dimensional chromatography to fully separate complex mixtures. As we develop new chemometric algorithms for multi-dimensional gas chromatography, this web tool will serve as a useful primer to our field as well as a good training tool for our students and collaborators to assess the impact of a variety of parameter alterations.