CIS 691 Capstone Presentations

April 21, 2017, 1-4:30 pm, KEN 330

Data Mining Techniques Applied to the Hydrogen Lactose Breath Test

Vanaja Danda

For the assessment of gut microbiome functional activity, the hydrogen breath test is one of the most common tests in use nowadays. The main purpose of this test is to learn about gut microbial function activity by identifying the hydrogen patterns. These hydrogen patterns can be identified by statistical methods like heat map, principal component analysis and clustering. My approach is based on a research paper. The authors used the Matlab software, but my approach is to use the R language. My goal is to compare both software packages. In addition, I performed interactive visualization using 3D scatter plot by considering all the variables present in the dataset using plotly tool. This paper is the first one to apply the data mining techniques to hydrogen breath tests, so the conclusions in this paper are quite conservative and research work is going on further to link these hydrogen patterns to different set of symptoms occur in metabolic activity of gut flora. This is considered as the initial step that needs further research.

Diabetes Mellitus Prediction in Women

Rama Prathyusha Musti

Motivation: Diabetes Mellitus is a chronic, lifelong condition that effects the body, its diagnosis will help to improve the health of the individuals. Many of the researchers started using the bioinformatics and knowledge discovery to help in better diagnosis of this disease. The goal of the paper is to predict the occurrence of diabetes taking various factors into consideration.

Materials and Methods: The dataset was taken from the UCI Machine learning repository (Pima Indian Diabetes dataset). The statistical or machine learning models used are Logistic regression, Random forest, Support Vector Machines(SVM). To construct these models various steps such as checking the variability of the data, feature selection methods were performed. The analysis was done using the R software.

Results: Three models were constructed to predict the occurrence of diabetes. The first one was the logistic regression model with an accuracy of 80.51%, second being the random forest model with accuracy of 81.82% and for the SVM model, the accuracy was 80.51%.

Visualizing Smoking Trends in the US

Kavya Vital Naram

Introduction: The main aim of my project is to create a visualization of behavioral use of tobacco in the United States of America in the past 20 years and how it is trending in the current population in the US. This was done by taking

the data from past years on four levels of use of tobacco in the US, that is, who smoked regularly, who smoked occasionally, who never smoked and who quit smoking after a period of time.

Methods: My approach is to use data from each state of the United States from 1995-2010 and use two different visualization approaches: Tableau software and programming in the Python language to have an explicit view of the use of tobacco among the population of the United States. The data is extracted from data.cdc.gov, and the data that focus on males, females and youth usage of tobacco is extracted from http://www.americashealthrankings.org.

Results: Tobacco use can be in any form of use like use of cigarettes, cigars, small cigars or pipe smoking. This use varies in different race/ethnicity, education, age, gender, education etc.as well as the United States, hence it varies depending on sociodemographic factors too. The data results show that there is increased prevalence and trends in the tobacco use in the US. The results show that currently the woman have a high percentage of smoking in the US, followed by the youth in most of the states when compared to the males. The results show that there is a necessity of increasing the evidence-based prevention methods to ultimately decrease the increasing effects of tobacco and its related diseases on the population of the US.

Conclusion: The survey has showed that there is reduction of life expectancy by 10 years in people who smoke tobacco compared to people who never smoked in the United states. The survey results also show that people who quit smoking at an average of 35-40 years of age, reduces the risk of being affected with the fatal diseases by 90%. It also showed that rural population use tobacco higher than the urban populations.

Furthermore, we investigated how effective these three visualization tools were for specific variables. Geospatial visualization had most interactive features and was most user friendly in Tableau when compared to Python and R visualizations. Python plotting for different levels of smoking is interactive along with geospatial visualization and the user needs to create his own code for the visualization. R visualizations are static and the script needs to be documented. I conclude that for specific and fewer variables the best interactive visualization can be done using Tableau.

Health Information Exchange in OPENMRS using HL7 & FHIR

Shivanshu Gupta

The aim of this project is to implement, and analyze the healthcare interoperability using two different interoperability technologies, HL7 and FHIR, and determine, which one is better and for what reasons. The EHR used was OPENMRS, an open-source EHR available with a multitude of modules implemented. OPENMRS uses the MySQL database, backend coded in JAVA, using Apache Tomcat as server and JavaScript as user interface. The interface engine, MIRTH was used for health information exchange between the EHRs. HAPI-FHIR is an open source framework available in JAVA which was being used by the FHIR interface. According to Meaningful Stage 2 Consolidated Clinical Document Architecture (C-CDA) comprises of 70 documents sections, this project was also to know that how many of them are being currently supported by OPENMRS. A sample HL7 and FHIR message was generated within the OPENMRS and was being transferred to a different one, where data being saved in MySQL.

Prediction of Parkinson's Disease Using Machine Learning Approaches

Gnana Velagapalli

Parkinson's disease(PD) is the second most common neurodegenerative disease after Alzheimer's disease. Prediction of PD is most difficult and challenging issue for physicians. It's hard for the physicians in most of the cases to take a decision on patient whether he/she can develop the disease in future or not. In such cases machine learning techniques, can be most helpful to the physicians for better clinical decision support and prediction of the disease. In this paper, I had performed 4 machine learning models were used to predict the Parkinson's disease. It was observed that Random Forest and SVM has given the overall accuracy of 87% and 88% when compared to other approaches that were used in this paper. These techniques can help the physicians to make better decisions without reviewing of similar or prior cases.

A Data-Driven Approach To Predict Cerebral Aneurysm Ruptures Using Discretized Data

Bhanu Yandrapragada

Cerebral aneurysm are deformations of the cerebral vessels characterized by a bulge of the vessel wall. It poses a major clinical threat and upon diagnosis, it is complex, lengthy and costly. There are two methodologies for evaluating and treating patients with cerebral aneurysms. They are computed tomography angiography (CTA) and digital subtraction angiography (DSA) for evaluation of cerebral aneurysms. These methods evaluate the rupture status, and other imaging characteristics that guide surgeon to make appropriate and timely treatment recommendations. The main goal of this capstone current research is the creation of a data mining classifier that supports the use of Computed Tomography Angiography (CTA) to identify a cerebral aneurysm and predict Subarachnoid Hemorrhage (SAH). This study attempted to use 101 patient data with 40 features each. The features include basic demographics, clinical information, and morphological characteristics. These data were collected from the Aneurisk team at Emory University. They have shared the collected data to improve the understanding and the therapy of cerebral aneurysms. Therefore, to discover the data mining classifier, the data is first preprocessed and is then discretized using the R's discretize function. In this study, decision tree algorithm is used as it is a good fit for this scenario; it selects the best features and provides clear rules. To access the suitability of this technique, the first unprocessed and original continuous data was fed into R's rpart function from the caret library. To overcome the continuous barrier of the decision tree, a novel approach of using multiple decision trees for feature selection along with Apriori algorithm is used for the best association rules. To access this R's arules library is used.

Comparing Classification Algorithms to Predict Vertebral Column Disorder

Sowjanya Ratho

Spinal disorders are extremely common in two third of adults. In this project, we are analyzing a dataset on biomechanical features of the Vertebral Column to classify patients as normal or abnormal (disk hernia or

spondilolysthesis). The dataset has 12 physical parameter measurments. After normalizing the data, we performed t-tests to determine the significance of all variables. To improve the accuracy of prediction, it is also important to determine the correlation between the variables and their impact on classification on spinal disorders. Principle component analysis, calculation of the correlation matrix and feature selection libraries are used for feature selection. The selection of the key features resulted in 6 variables left for analysis. These 6 variables explain about 95% of the data. Furthermore, we fitted a logistic regression model, we used Random forest and SVM algorithms for classification on new data and we compare their performance based on the AUC (area under curve) and the confusion matrix. The analysis was performed using R Studio and some built-in libraries that are helpful for us to automate procedures.

The results showed that the logistic regression model attained the best fit. It classified data with 0.9513 accuracy, but this model is not recommended because it may overfit the data. Hence, we recommend the random forest model, it predicted the data with 0.903486 accuracy. This model also shows the importance of the variables for model selection and is especially helpful for forward/backward stepwise selection.

Comparative Study of Classification Algorithms in Breast Cancer Prediction

Saikrishna Kotha

Objective(s): This study addresses the comparison of classification models for diagnosing breast cancer. The dataset has been analyzed previously with various classification techniques (naïve Bayes, decision tree, random forest, support vector machine and adaboost) with the goal to classify breast cancer as benign or malignant. Our goal is to determine how the above mentioned classification models are comparable with regard to classification of breast cancer.

Materials and Methods: The dataset used for the study was the Wisconsin Breast Cancer Diagnostic dataset obtained from the UCI Machine learning repository. Metrics used to evaluate the effectiveness of the classification models were confusion matrix statistics, the accuracy of the results and the kappa statistic. Feature selection methods such as Pearson correlation, recursive feature elimination or feature selection based on the random forest algorithm were also employed to identify the best possible features to improve the accuracy of the respective models. For the analysis R software was used.

Results: Among all the classification models the support vector machine model could achieve the best accuracy in classifying the data. The accuracies of the other models were also comparable to the support vector machine model with the adaboost and random forest models being closest.

Comparing the Performance of Scalpel to GATK-HaplotypeCaller Using Simulated Reads

Matthew Lueder

The ability to accurately call variants from next-generation sequencing data (NGS) is a necessity for the success of NGS in clinical genomics. Therefore, there is a need for continuous in-depth reporting on the accuracy of state-of-the-art variant calling algorithms. In this paper, the performance of two local de novo reassembly-based variant calling

tools are benchmarked using a simulated dataset. Genome Analysis Tool Kit HaplotypeCaller (GATK-HC) is consistently reported to be one of the best performing variant callers. Scalpel is a newer tool which has recently been reported to outperform GATK-HC in calling insertion/deletion elements (INDELs). The goal of this study is to provide an up to date and in-depth comparison of these two variant callers using a realistic simulated dataset. Simulated reads were generated using the tools VarSim and ART, then aligned to a reference genome using BWA-MEM. Precision, recall, and F1-scores were calculated by comparing variants called by GATK-HC and Scalpel to a truth-set of variants using PrecisionFDA's comparison tool. GATK-HC was observed to have higher precision and recall for single nucleotide polymorphisms (SNPs) and INDELs.