120th ASEE Annual Conference & Exposition
FRANKLY, WE DO GIVE A D*MN
June 23-26, 2013

Paper ID #6905

A Study of Application of Data Mining Algorithms In Healthcare Industry

Dr. Reza Sanati-Mehrizy, Utah Valley University

REZA SANATI MEHRIZY is a professor of Computing Sciences Department at Utah Valley University, Orem, Utah. He received his MS and PhD in Computer Science from University of Oklahoma, Norman, Oklahoma. His research focuses on diverse areas such as: Database Design, Data Structures, Articial Intelligence, Robotics, Computer Integrated Manufacturing, Data Mining, Data Warehousing and Machine Learning.

Jeffrey H Wright

Jeff is a senior at Utah Valley University pursuing a BS in Computer Science with an emphasis in Database Engineering. He is currently employed at Lucid Software Inc.

Dr. Afsaneh Minaie, Utah Valley University

Afsaneh Minaie is a professor of Computer Engineering at Utah Valley University. Her research interests include gender issues in the academic sciences and engineering elds, Embedded Systems Design, Mobile Computing, Wireless Sensor Networks, and Databases.

Dr. Ali Sanati-Mehrizy

Dr. Ali Sanati-Mehrizy is a graduate of the Milton S. Hershey Pennsylvania State University College of Medicine. He completed his undergraduate studies in Biology from the University of Utah. In July 2013, he will begin a Pediatrics residency at the UMDNJ-Newark University Hospital. His research interests involve pediatric hematology and oncology as well as higher education curricula, both with universities and medical schools.

Paymon Sanati-Mehrizy, Icahn School of Medicine at Mount Sinai

Paymon is currently a medical student at the Icahn School of Medicine at Mont Sinai. Paymon completed his Bachelor of Arts in Biology in May 2012. Currently, his research interests consist of higher education curricula, particularly in fields that incorporate science with medicine.

A Study of Application of Data Mining Algorithms In Healthcare Industry

Abstract

Data mining is a relatively new area of computer science that brings the concept of artificial intelligence, data structures, statistics, and database together. It is a high demand area because many organizations and businesses can benefit from it. There is no doubt that it is a great idea to use data mining tools in the healthcare systems to improve decisions made by the experts in this field. It has been seen that certain groups of people living in certain areas of the world with certain food diets carry certain diseases more than some other groups of people. It will be a great contribution to the healthcare system to know what the root of this type of problems is. The most critical challenge in this contribution is the issue of data management and utilization. Data mining algorithms can be used to find useful patterns in the patients' statistical data to find the associations among certain disease and possible causes of them. Thus, health care organizations are finding value as well as strategic applications to mining patient data, in general, and community data, in particular. While significant gains can be obtained and have been noted at the organizational level of analysis, much attention has been given to the individual, where the focal points have centered on privacy and security of patient data. While the privacy debate is a salient issue, data mining (DM) offers broader community-based gains that enable and improve healthcare forecasting, analyses, and visualization. This paper will study a few cases where data mining has been used to mine patient data to make decisions about patients and investigates how data mining can be beneficial in the context of healthcare system.

Introduction

The purpose of this research paper is to study the effect that data mining has had in the field of medicine. The field of medicine produces an extremely large amount of data, and like most agencies that produce a lot of data, they need to have a way of analyzing that data effectively and efficiently. The benefits of understanding, centralizing and making the data easily readable and accessible to medical professionals are potentially earth shattering. In order to show what they are doing, we have taken the opportunity to look at three different studies which cover different areas of the medical field. In the first study it views the effect of data mining has been able to have on the ability of doctors to prescribe hearing aids, the second study views the benefits of data mining in researching cancer genes, and the last one looks at the potential gains and also the obstacles that must be overcome in order to effectively analyze DNA data received from sequencing machines. As time goes on, it appears that data mining will continue to have an incredible influence on the progression of medicine and the ability of nurses and doctors to treat their patients. It will forever change the way we know and understand medicine.

Data mining is becoming more and more popular in all realms of business. It only makes sense then that data mining would be able to have a huge impact in the healthcare industry as well as searching for better ways to correctly diagnose different diseases and give proper treatment based not only on the experience of the doctor and his staff, but of the past experience of a large part of the medical community. Srinivas et al[4] have mentioned that the healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. But, there is a lack of effective analysis tools to discover hidden relationships and trends in data. However, knowledge discovery and data mining have found numerous applications in business and scientific domain.

Some of the benefit that has been found already was laid out by Koh & Tan[2]. They have mentioned that data mining is becoming increasingly popular in the healthcare system, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services.

In order to understand the value that data mining is having and will have when the processes and algorithms have been refined and optimized for healthcare, we will look at a few examples of what data mining has discovered and what they are hoping to accomplish. First we will look at a study done on hearing aids and how we can use that data to help doctors to correctly treat their patients suffering from hearing loss. Second, we will see research done to study and identify candidate genomes for cancer and see the benefits and possibilities that provides. Last, we will view a paper written on DNA sequencing and how data mining will shed new light as researchers develop the needed tools and solutions to delve into the deepest level of human biology.

Hearing Aid Data Set

Anwar & Oakes[1] conducted research based on audiology records to help audiology technicians when the decision between an ITE, hearing aid worn ‘in the ear’, and a BTE, hearing aid worn ‘behind the ear’. Audiology technicians generally will make the decision based on the audiogram results and a consultation with the patient and the decision to be made at that point is generally pretty clear. However, there are occasions where the data leaves them in a spot where they can go either way. In these cases, they might “benefit from a second opinion given by an automatic system with an explanation of how that second opinion was arrived at”[1]. The data that Anwar & Oakes collected consisted of the audiograms, which graphs the auditory thresholds of the patient at 6 different frequencies, some categorical data including gender, diagnosis, and hearing aid type, and some brief notes from the technicians.

Then, they took that data and performed a PCA, Principal Component Analysis, on the audiogram data retrieved to group them by audiogram types. Once the audiogram types were established, they assigned each patient to one of the main audiogram groups and used chi-squared to look for relationships between the audiogram type and the hearing aid type. The chi-squared test showed that the audiogram is very clearly related to the type of hearing aid that that patient will or should in the end receive. On top of that data processing, they also used a k-means clustering algorithm and were able to essentially reproduce the same results as they had already achieved with the ITE’s being associated with the lighter to moderate hearing loss and the BTE going for the more sever hearing loss cases.

Once having determined the strong correlation between the audiogram type and the hearing aid type, they continued to use the chi-squared quest to also test how the categorical data and tech notes played into the probabilities. Using chi-squared, they were able to determine that males ended toward the ITE hearing aid while the majority of females ended with the BTE hearing aid. They also checked on patients that were diagnosed with tinnitus, which is a ringing in the ear. Those results showed that patients with tinnitus generally ended up getting the ITE. They continued through the other attributes determining that using a tinnitus masker, patient age, mould type, and the text gathered from the technician notes were also significant to the end classification, showing that all attributes in the data set were significant in the classification of the patients hearing aid type.

The next step in their data processing was to create a "single logistic regression model." They set up the data for training and testing by separating out 80 percent into one subset and the remaining 20 percent into another subset. The subset containing 80 percent of the data was to be used for training purposes, leaving the last 20 percent for the testing of the newly created model. They limited their datasets to records with no missing values for the attributes in relation to the right ear only. Then they discretized the data and created the model and processed it to get the 'best model', as they say in statistics[1].

Having made all of the adjustments and refining their statistical model, they discovered the following:

> Testing of their logistic regression models showed that overall there was 82.21% agreement between the predictions of their model and the actual hearing aid chosen by the audiologist. The agreement rate was higher for patients fitted with ITE aids (88%) than for those fitted with BTE aids (75%). The results were analyzed according to precision, recall and F-measures.

This modeling allows for the audiologist to easily get a second opinion from an unbiased source and includes in the classification a description of the attributes that weighed most heavily in the prediction, thus allowing for the audiologist to make an informed decision about whether to continue with his original diagnoses or reconsider. At the end they mentioned that "The overall agreement is much better than random (50%)," but that they really needed to compare it to the simplest possible algorithm which happens to come in at 54.6% (simply assigning all people the ITE). There are quite a few more considerations they might take into account as they continue to process and refine more data, but it seems that this will certainly add to the confidence of a clinic that they are making the correct choice when diagnosing a hearing aid type to their patients.

Mining Cancer Genes

Next, we delve deeper into the inner workings of the human body as we analyze human Genes for signs of cancer. Inho Park, Kwang Lee and Doheon Lee[3] published an article discussing "a new method to detect candidate cancer genes for developing molecular biomarkers or therapeutic targets from cancer microarray datasets." Previously, most research on this topic focused on "identifying genes that have altered expression during cancer development." That method is limited, however, to only find the expression patterns across the entire group of cancer patients. Further research was done to improve upon this standard that was originally followed. The improvements that have been made in this field are based on outlier profiling, improved statistical analysis of those outlier and improvements in defining how an outlier is classified.

Even with these improvements, they still lack the ability to be effective in cases where the normal expression levels are similar to that of the outliers, and when there are a small number of very large outliers.

The authors of this article believe that comparisons should be made not only with the cancer group as a whole, but “if expression levels of a gene are significantly altered in a subset of cancer samples, the expression levels of the samples in the subset would be located close to each other at the end of the ordered list of its expression values. Thus, we propose the RS score based on a weighted running sum statistics.” Their approach would be what they determined to be the “first approach in the context of detecting individual cancer genes.”

The data used for this study appears to have come from 4 older studies, 2 of which were used for testing and 2 for training. It appears they only used data that had less than 20% of its data missing. They then performed a quantile normalization, which is used to normalize the subsets of a particular data set. In order to work well with the outliers, they performed a median-MAD normalization. They then performed the weighted running sum algorithm they created and were able to discover “102 over- and 88 down-expressed genes whose adjusted-p value is less than 0.005 after the Bonferroni corrections for multiple comparisons.” Contained within the genes identified were several known prostate cancer genes. The authors explain the remarkable results gained by this research[3] and conclude by saying that they were able to successfully identify many different candidate genes associated with prostate cancer. The results of their classification analysis showed that their study performed more accurately while taking into consideration the outliers than the other studies done on the outlier profile analysis. They also mention ideas for improvement on their chosen method as they closed their research.

Here is how data mining can be extremely beneficial in healthcare system. Imagine if we were able to identify these genes for any cancer types and do so well in advance of any adverse health conditions. If the candidate genes were found early enough, it may even be possible to isolate these candidate genes and prevent the cancer from ever really gaining a foothold in the human body. Obviously, that would require some kind of regular interaction with the doctor on the part of the individual, but if they could revise the process well enough that you could easily run the needed tests as part of a more ‘in-depth’ check-up or something that you could do once every six month or a year. It might not be enough to catch all of the instances of cancer, but it would be incredible to see the number of people whose lives likely were able to be saved by this form of early detection. The implications of being able to classify these genes are astounding to say the least.

DNA Speculation

Justin Zobel[5] explores where data mining is headed. He explains the potential of DNA sequencing and what challenges are presented in trying to extract useful information from the data gathered by the DNA sequencing process.

The Human Genome Project was a joint effort to catalog all the different genes found in human DNA. Since its completion in 2000, “the cost of sequencing DNA fell by a factor of around a million, and continues to fall.” This is significant because “Applications of sequencing in health include precise diagnosis of infection and disease, lifestyle management, and development of highly targeted treatments" and because the "knowledge of DNA and its function is key to the understanding of living organisms." He also mentions his belief that reading a person's genome may soon be less than $1000[5].

Sequencing is used throughout the medical field for many different purposes, and the machines used to do this sequencing produce an incredible amount of data. That data, though, must go through a substantial amount of pre-processing in order to be useful to anyone that wanted to research it. The DNA is read in fragments that can be anywhere between 35 and 500 bases (the different combinations of these bases are the make-up of all the different genes in the human body). Once read in, the DNA must then be replicated in order to increase the chances of getting enough useable data. Since the fragmented reading of these machines is non-uniform, it causes many errors in the data from missing and misread bases.

There are a significant number of obstacles that come in at this point. First the author mentions that the sizes of the actual genomes are incredibly variant. Anywhere from "a few kilobases for a virus to over 100 gigabases for some plants; the human genome is 3 gigabases." The next issue that he describes is dealing with the amount of data produced. He writes: "On the order of a hundred gigabases of sequencing data is required to construct the genome of a human individual or to quantify activity within a cell. The computational cost of analysis may greatly exceed that of the sequencing that produced the data, and the cost of reliable storage of the raw data for a single year could soon exceed that of production." Finally there are many computational challenges to overcome, including simply modifying existing methods to work on a scale as large as this data which the author believes will soon be the same scale as the web[5].

Presenting the data in a useable and concise fashion will be the final hurdle to jump, but once accomplished will present the medical world with tools it has only begun to appreciate and understand. The ability to work on not even just a molecular level, but the ability to examine the DNA itself, the building blocks of life, is an incredible accomplishment that will be met and achieved by using advanced methods of processing the insane amount of data that will be created. The World Wide Web is a very complex thing and the amount of data traveling around there is mind numbingly vast, yet it pales in comparison to the amount of data contained in each human body. Some potential uses for this data are "to examine how bacterial communities respond to stresses such as new drugs; to examine how lifestyle and genetics interact; and to cheaply assemble and annotate of the complete genomic sequence of an individual."

Conclusion

The purpose of writing this paper is to see how data mining is helping us to progress in our knowledge of medicine. We, the authors, feel that these brief studies have been done go a long way toward showing the progress that is being made in the medical field. The authors chose these particular papers to review for two reasons. First, we started with the ear research. This research is being used at the current time. In the sense of accomplishments, it is passed. In the sense of the body, it is the outermost layer. Second, we reviewed the paper on identifying candidate genomes for cancer. The research in this area is ongoing with many new ideas on how to improve what has been done and many anxious to continue what has already been discovered.

It is the ‘present’. It’s where we are today. In relation to the body, it is one of the most inner layers. Finally, we observed the short paper on DNA analysis. It is the future of the applications of data mining in biomedicine. It is also the ‘core’ so to speak of the body. It is what makes the whole thing come together into what we see and experience every day.

The longer data mining exists in the medical world, the more we are able to explore and quantify what previously would have been incomprehensible. That is what I see as ‘how data mining advances our medical knowledge’. I see it as the bridge, which was once missing, for us to cross into a whole new, unexplored territory that will have far reaching consequences. Our understanding of medicine will and should be way different as we come into this ‘new world’ of knowledge. No one will argue against the statement that our current medical practices are flawed. This is not to say that they will reach perfection by any means, but, as with anything else in this world, if you know and understand what it is that it’s built out of, if you understand how it works together, then, you can understand how to influence it. In the medical field, this can be accomplished by generating distinct medical models, in order to foresee a patient physical condition or recommend medical remedy[6].

Bibliography

1. Anwar, M. N., & Oakes, M. P., Data Mining of Audiology Patient Records: Factors Influencing the Choice of Hearing Aid Type. *CIKM '11 International Conference on Information and Knowledge Management*, pp. 11-17, Association for Computing Machinery, New York, 2011.

2. Koh, H. C., & Tan, G., Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, pp. 64-72, http://www.himss.org, 2005.

3. Park, I., Lee, K. H., & Lee, D. Mining Cancer Genes with Running Sum Statistics. *CIKM '09 Conference on Information and Knowledge Management* , pp. 35-42, Association for Computing Machinery, New York, 2009.

4. Srinivas, K., Kavihta Rani, B., & Govrdhan, A., Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*, pp 250-255, http://www.enggjournals.com/, 2010.

5. Zobel, J. Data, Health, and Algorithmics: Computational Challenges for Biomedicine. *CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management*, p. 3, Association for Computing Machinery, New York 2011.

6. Sethukkarasi, R., Keerthika, U. and Kannan, A., A self Learning Rough Fuzzy Neural Network Classifier for Mining Temporal Patterns, *Proceeding of International Conference on Advances in Computing, Communications and Informatics, 2012.*