DATA MINING APPLICATIONS

For the Healthcare Industry

Abstract

The role of big data in medicine is one where we can build better health profiles and better predictive models around individual patients so that we can better diagnose and treat disease[1]. The unique challenge we have at hand is generating actionable insight for individual patients. There are several different data mining techniques that can be applied to the healthcare datasets for medical value creation. These applications will be further discussed in this portfolio.

Ian Ustanik

IST 707

Table of Contents

Data Exploration

The healthcare field presents unique challenges in data exploration. There are massive amounts of healthcare and medical data for every individual who has ever visited a doctor or been treated at a hospital. The role of big data in medicine is one where we can build better health profiles and better predictive models around individual patients so that we can better diagnose and treat disease[1].

As we know, data can be broken up into structured and unstructured data. Most structured medical data are inputted into a health profile or Electronic Medical Record (EMR) that is being constantly updated every time you visit a medical professional. Each one of these visits is known as an encounter. The EMR includes features such as age, gender, and medical history. Another large proportion of structured data is collected from medical instrumentation and sensors. Examples of this type of structured data include features such as blood pressure, body temperature, and weight. Unstructured medical data may include clinical notes, imaging, genomic sequence data, and even biopsy samples. There are perhaps more features in the healthcare industry than any other.

The unique challenge we have at hand is generating actionable insight for groups of patients in a population as well as individual patients from all of this raw data. The data must have meaning, quality, and it may be necessary to make some transformations. Data quality is imperative for this application. Outliers should be detected and analyzed carefully. Anomaly detection is used in disease and cancer screenings. Summary statistics can be used in building better health profiles. Data privacy concerns add a whole other level of complexity as well. Many individuals are concerned with disclosing too much personal medical information to third parties like insurance companies and employers. There may be a need to generate synthetic medical data in order to perform data mining algorithms like Associated Rules Mining, clustering, and classification.

There are several different data mining techniques that can be applied to the healthcare datasets that have been collected. The these include the aforementioned Associated Rules Mining, clustering, and classification techniques (Decision Tree Analysis, Naïve Bayes, Support Vector Machines, k-Nearest Neighbor, and Ensemble Learning) along with other techniques such as Deep Learning and Natural Language Processing. Each technique is best used in confronting particular issues within the healthcare industry. All of these particular issues underline the ultimate goal of creating better predictive models around individual patients so that we can better diagnose and treat disease[1].

Model Evaluation

The best predictive models are created using a randomly extracted stratified sample of the best data and most effective features within that data. These models must then be evaluated based on certain metrics such as Accuracy, Precision and Recall, and techniques like Cross Validation and Hold-Out tests to ensure that they are in fact the best models possible. This is an important aspect of any predictive model built for the healthcare industry.

The first task in building a predictive model is feature selection. Attributes selected in model training have perhaps the highest influence on model performance. It has been stated that feature selection and data cleaning should take up the majority of the model-building process. There are different methods to perform feature selection. The most popular technique is correlation between features and the decision class. How much does a change in a particular feature explain a change in the decision-making process of the model? Stronger correlated values will have a correlation closer to positive or negative one. Weaker correlated values will have a correlation closer to zero. Information gain-based feature selection is another popular technique. Information gain can be defined as the change in information entropy from a prior state to a state that takes some information. Those attributes that contribute more information will have a higher information gain and should be preferred over attributes that have a lower information gain.

Cross-validation and the Hold-Out test are two evaluation methods that have been used to measure model fitness before using it for predictions on outside data. The Hold-Out test is performed by splitting the training data into two subsets for training and testing. A train-test split is the ratio used for creating the split. This is a quick way of evaluating model accuracy, but the test result may be subject to high variability based on changes in the train-test split. Cross-validation is the average accuracy of n-validations for a particular training set. N is the number of equal-sized splits that are made in the training set. Every split acts as a test set against the other splits for a single validation of model accuracy. Once all splits have served as the test set, n-validations are averaged for accuracy. This model evaluation method tends to be less variable, but it may take much longer to compute depending on the n-number of folds chosen.

There are several metrics that have been historically used in model selection. Accuracy (Correctly Classified Instances/Incorrectly Classified Instances) is the most common measure, but it has limitations on skewed data sets. A confusion matrix can be used to visualize Precision and Recall. Precision is a metric that determines among all positive predictions, how many are correct? (True Positive Instances/True Positive Instances + False Positive Instances). Recall is a metric that determines among all positive examples, how many are correctly predicted? (True Positive Instances/True Positive Instances + False Negative Instances). Precision and Recall measures can differentiate between False Negative and False Positive errors. This is important when the weights of these errors differentiate. For example, in a cancer detection classification model you would want a False Positive more than a False Negative error. One could mean incredible temporary distress, but the other could mean a possible preventable death.

Data Mining Techniques

Association Rules Mining

An important advantage of Association Rules Mining is the clarity of the results that come from the rule discovery process. Additionally, the frequent itemset and rule generation processes are easy to understand and audit. This sort of clarity is paramount in healthcare applications and the medical field.

Association Rules Mining can be used as a computational intelligence approach to flag those who have a higher potential for different cancers or disease. Heart Disease is a good candidate for this technology to be used. Currently, three different rule generation algorithms - Apriori, Predictive Apriori, and Tertius are being used with biological databases in rule discovery. These biological databases house anonymous patient data with features like ECG, weight, and even general profession. These technologies are used to find new combinations of health factors to flag patients who were previously unknown to be predisposed to Heart Disease.

This would be a cost-effective first step and proactive medical application. Of course, it would be most effectively used in combination with existing early warning systems. Inversely, new rule discoveries may be used to identify healthy individuals who are living a lifestyle that leads to prolonged longevity. These factors could be used to make future health recommendations and other new discoveries.

Clustering Analysis

Automatic cluster detection is a tool for undirected data mining, because the automatic cluster detection techniques find patterns in the data with no target variable. Clustering has the ability to deal with many different types of attributes. This is important in the healthcare field when there are so many different types of structures and unstructured data features that I mentioned earlier on in the data exploration process.

Different clustering algorithms like K-means can tackle applications where a reliable label output is not provided. This can include identifying diabetic and non-diabetic fluid group structure and further identifying similar patients based on their health risk score. This further step may even occur after preliminary classification methods to make data-driven decisions about how patients may adjust their lifestyles in order to lower their health risk score and live longer, healthier lives.

This process works by collecting patient medical data and any classification algorithm results that better define the individual’s health risk score. This data can then be input through the K-means algorithm where parameter adjustments can minimize overall model SSE. An accurate model would then be able to create K-clusters that can separate patients into groups based on minimizing intra-cluster distances and maximized inter-cluster differences. These groups can be assigned health risk scores, where personalized solutions may lead to a reverse in unhealthy trends before it is too late. This would help save resources before the patient becomes a drag on the health system due to doctor visits, prescription costs, and expensive operations.

Classification Methods

The Decision Tree Analysis is one of the more popular classification techniques used today. This algorithm generates a decision tree model to make some sort of supervised learning prediction. One of the algorithm's strengths is that it is very effective in interpreting several different data types when building the tree. Additionally, Decision trees allow for transparency in rules decision-making. You are always able to refer to your model's build in order to answer any questions regarding how it produced the final classification. Decision trees have been used in healthcare applications for over twenty years due to the algorithm's accuracy and robustness.

One of the most common applications of Decision Tree Analysis is in patient diagnosis. The data from an individual's health profile can be inputted into a decision tree model created from training data that contains thousands of other health profiles with similar attributes. Decisions made at the leaf nodes would be certain conditions or diagnosis. A very practical first step that doctors may use in the diagnosis process. The J48 and C4.5 algorithms are examples of tools that can be used in building these predictive models.

Another similar application for the use of this algorithm would be treatment decision-making. Medical treatments can be very expensive, so it is paramount that the most effective and economic strategy is taken. When the correct decision is made earlier on, success rates skyrocket. These decision tree models would be built in a similar manner to the patient diagnosis model, only this time the decisions made at the leaf nodes would be treatment options. Whenever new and accurate decision aids can be used in economic healthcare decision-making, the entire industry becomes more sustainable and powerful for individual patients.

Similar to the Decision Tree Analysis method, Naïve Bayes is a classification algorithm used as a supervised learning technique. The Naïve Bayes Algorithm is based on the Bayes theorem which states that "The presence of any particular feature in a class is completely independent of the presence of any other feature". Furthermore, the algorithm works by calculating the posterior probability of each class at each level of the model for the dataset. The classes with the highest posterior probabilities will be the chosen result.

The Decision Tree Algorithm performs better for small amounts of classes while The Naïve Bayes Algorithm is highly accurate for big data applications. This is fitting for the healthcare field where there are perhaps more features in the industry than any other. Naive Bayes classification has been demonstrated to be superior to several other classification methods when applied specifically to medical data[2].

The best application for using the Naïve Bayes classification algorithm in healthcare would again be in health prediction and patient diagnosis. An effective model like this would be an invaluable tool to assist doctors in the diagnosis process fast and accurately. The process for building this model would be very similar to that of the Decision Tree Analysis, except the change in parameters need to be noted. Instead of binarySplits, pruning, and confidence factor, parameters such as SupervisedDiscretization must be considered.

There are so many classification methods that can be used in healthcare machine learning applications. We have already discussed Decision Tree Analysis and Naïve Bayes Classification and their possible contributions to the industry. Now we can discuss Instance-Based Learning (k-Nearest Neighbor), Support Vector Machines, and Ensemble Learning methods. These algorithms are based on different mathematical concepts and therefor have different strengths and weaknesses in the field.

The k-Nearest Neighbor algorithm is an example of Instance-Based Learning. This method stores training examples without performing calculations or even building a model. Classification is delayed until new examples are provided in the test data. This procedure is also known as “lazy learning” because the predictions are delayed. Classifications are made by analyzing the distance between the test sample and training data. The k-value is the parameter that refers to the number of nearest neighbors to include in the majority of the voting process. This algorithm may be very flexible because its decision boundary has no pre-defined shape and it makes no assumptions like Naïve Bayes. Unfortunately, this means that one must consider the possibility of model overfitting. The k-NN algorithm is the most popular lazy learning algorithm used for pattern recognition in the healthcare field. This would be perfect for identifying at-risk patients for Heart Disease and predicting heart attacks before they happen. It would be paramount though for the sake of model accuracy that any noise in the data be removed or limited before the Instance-Based Learning began.

An important topic that must be discussed before introducing Support Vector Machines is that of "linear inseparability". Decision Tree Analysis partitions linearly on each dimension for its decision boundary. Naïve Bayes Classification partitions a single line for its decision boundary. Many times however, data is not neatly partitioned linearly in any of these manners. This is a common issue for medical data. This is where Support Vector Machines prove their worth. Support Vector Machines (SVMs) find a hyperplane that maximizes the margin between the support vectors that make up the decision boundary. These support vectors are the training examples that are located on the margins. If an appropriate linear boundary is not found in the two-dimensional plane, a kernel parameter can be used to map the data to higher dimension space. Data sets that are linearly inseparable can now have accurate decision boundaries using this kernel polynomial. I foresee SVMs accurately building classification models for Breast Cancer diagnosis. This is because of how effective the model performs in high dimensional spaces.

The accuracy of any one individual classification method can be enhanced through the use of Ensemble Learning Methods. Ensemble learning models essentially aggregate the predictions used by multiple classifiers and then make predictions based on majority vote. This way over half of the models would have to be incorrect in order to make a false classification. An example of a supervised learning ensemble learning algorithm is Random Forests. Random Forests create random vectors from training data, use these vectors to build multiple decision trees, and then combines the decision trees for the final classification. Other methods such as Boosting or Bagging can improve the accuracy of any classification model as well. These enhancements can apply to most classification models in the healthcare industry.

Deep Learning

Artificial Neural Networks excel in creating machine learned features that are easy to adapt and fast to learn. This is ideal for the medical industry that produces very large amounts of unstructured data that would be difficult to manually feature engineer. Like Support Vector Machines, a multilayer network can represent convex regions to separate data that is linearly inseparable. There could be several applications of this technology in the field, but I would like to focus on cancer detection using image recognition algorithms.

These Neural Networks mimic the human brain in how it solves problems. Individual neurons have one or more weighted input connections, its own activation function, and then a different output connection. Many neurons may make up a layer, and at least multiple hidden layers must be used to represent the convex regions that were mentioned before. Back-propagation is then used to fine tune the weights of the neural net.

According to medical professionals, detection of subclinical Breast Cancer on screening mammography is challenging as an image classification task[3]. Convolutional Neural Networks can quickly feature engineer and back-propagate their models in order to assist in this image classification. This technology can assist in cancer detection through the detection of anomalies in mammograms and other medical scans. The same is true for other medical imaging tasks like Melanoma diagnosis. These algorithms are now in many ways more accurate than human radiologists. Though I do not foresee humans being replaced by the algorithms in this application, Artificial Neural Networks would be an excellent supplement in the diagnosis process and serve to great value in regions that may have a shortage of medical professionals in the field.

Natural Language Processing

Natural Language Processing and Text Mining are two specialized branches of Artificial Intelligence that include computational techniques for the interpretation and manipulation of human-like language. The healthcare industry can use applications of this technology for medical value creation. This includes information extraction for diagnoses, procedures, and symptoms and even human-to-machine natural language instructions.

Natural Language Processing and Text Mining processes require large quantities of data to produce accurate results. Think about the millions of medical texts across the globe that would take impossible lengths of time to mine. Information extraction and automatic summarization algorithms can use tokenizers and then convert documents into vectors in order to quickly skim texts. Python NLTK is an example of a toolkit that can be used as a platform for processing these sorts of tasks. Information retrieval engines can be used for diagnoses and utilize the lowest levels of Natural Language Processing to stem words and help medical professionals find the phrase or words they were looking for.

This can be taken a step further in the development of human-computer interfaces. Human-to-machine natural language instruction systems can be developed in the future for applications like robot-assisted procedures guided by doctors and surgeons. This nascent natural language algorithm would probably need to be developed using some sort of neural network in order to train and create an accurate model.

Moving Forward

A combination of various data mining techniques such as Association Rules Mining, Clustering Analysis, Classification Methods, Deep Learning, and Natural Language Processing may provide a solution in generating medical value for groups of patients in a population as well as individual patients in the healthcare industry. These methods have the proven capability to augment medical professional abilities in diagnostics and treating disease. These tasks can be done more accurately and economically with the assistance of applied Big Data. In an industry that could deeply use the advantages of greater computing power we can look forward to results that mean real good to real people. Instead of maximizing profits we can maximize a patient’s years on Earth and lengthen the amount of time they spend healthy, happy, and with meaningful purpose.

Works Cited

[1] Chilukuri, Sastry, and Dr. Eric Schadt. “The Role of Big Data in Medicine.” McKinsey & Company, 2015, www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-role-of-big-data-in-medicine. Accessed 5 Apr. 2020.

[2] Hickey, Stephanie J. "Naive Bayes classification of public health data with greedy feature selection." *The Free Library*, 2013 www.thefreelibrary.com/Naive Bayes classification of public health data with greedy feature...-a0351818405. Accessed 5 Apr. 2020.

[3] Shen, Li. Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography. p. 2, 2018, Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography.