Applying Machine Learning Improvements Derived From Diabetes Prediction To Macro Healthcare Systems

ROLAND BABAIEV

Colorado Technical University 4435 North Chestnut Street, Colorado Springs, CO 80907 USA
Corresponding Author Email:- rolandhb37@gmail.com
http://dx.doi.org/10.22147/jucit/110101

Acceptance Date 2nd March, 2020, Online Publication Date 6th March, 2020

Abstract

Gunapati Venkata Krishna (GVK) Emergency Management and Research Institute (EMRI) is based in India and is an entity consisting of a partnership between public and private sectors. The entity responds to 30 million emergency calls and saves a million lives annually by deploying 9000 ambulances and managing 20,000 emergencies daily. Although a large scale healthcare system such as EMRI already leverages business intelligence to help enhance its capabilities, further improvements may be achieved by taking advantage of what is learned from smaller, micro-level research studies associated with machine learning prediction of diseases such as diabetes. Machine learning based diabetes prediction research has leveraged feature reduction methods, ensemble machine learning models, hybrid combinations of supervised and unsupervised learning, ‘white box’ modelling techniques, and more, to help enhance prediction accuracy. By integrating such specific improvements, macro-level BI systems such as EMRI can lower operational and treatment costs, more effectively allocate its human and physical resources, improve patient outcomes through personalization, and more effectively predict disease and complication risks.

Key words: machine learning, diabetes, business intelligence, healthcare systems.

Introduction

There are various examples of large scale business intelligence applications to healthcare. The area of Public Health Informatics (PHI), for example, is a systematic approach in the application of technology and information science for the benefit of population health. PHI involves data collection, analysis, and action. Its goals include addressing problems such as disease prevention, disease surveillance, bioterrorism, epidemics, and more. Coye (2016) conducted a case study of Gunapati Venkata Krishna (GVK) Emergency Management and Research Institute (EMRI) which is based in India. This entity is a partnership of public and private sectors in India which provides free emergency medical services to 750 million people (Coye, 2016). Coye (2016) noted that the entity responds to 30 million emergency calls and

This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-sa/4.0)
saves a million lives annually by deploying 9000 ambulances and managing 20,000 emergencies daily. By analyzing millions of emergency requests which occur in a real-time manner and by being fed data about the outcomes of emergency dispatches and hospital interventions, the decision support analytics capabilities of the entity are improved in a real-time manner (Coye, 2016). EMRI therefore illustrates a larger scale application of Business Intelligence to public health.

**Problem:**

Although a large scale healthcare system such as EMRI already leverages business intelligence to help enhance its capabilities, further improvements may be achieved by taking advantage of what is learned from smaller, micro-level research studies associated with machine learning and prediction in healthcare. Such smaller studies include those seeking to improve the capabilities of disease prediction through machine learning. Specifically, this paper will explore how large scale BI based healthcare systems such as EMRI may be improved using what is learned from smaller scale, machine learning based research conducted for diabetes risk prediction.

Type 2 diabetes (T2D) is a global epidemic (International Diabetes Foundation [IDF], 2019). The International Diabetes Foundation estimates that in 2017, there were 425 million people globally living with diabetes. By 2045, the organization estimates the number to rise to 629 million. Many studies have been conducted related to type 2 diabetes (T2D) prediction using machine learning models, with many novel approaches and techniques being developed.

**Problem Statement:**

The problem to be addressed in this paper is the potential absence of applying particular machine learning improvements seen in smaller studies focused on specific disease prediction in healthcare, namely type 2 diabetes, to larger scale healthcare BI systems.

**Purpose:**

This study will assess how the particular machine learning improvements, which were developed within the specific context of T2D prediction, can be applied to larger scale, BI healthcare systems, such as EMRI (Coye, 2016).

**Research Justification:**

This study will help to summarize, define, and assess what has been achieved thus far within the realm of type 2 diabetes (T2D) prediction and its applicability to larger scale healthcare systems. By presenting this analysis, public health systems may be improved. Furthermore, by understanding the current state and the proposals of this research, future research efforts can be better targeted.

**Machine Learning for Diabetes Prediction and Complication Management:**

Machine Learning is a branch of Artificial Intelligence which is focused on enabling systems to automatically learn on their own without specific programming. Within the domain of data mining, machine learning has been used for purposes of uncovering hidden patterns and knowledge in data. Machine learning is also being used to create models for disease prediction through the processes of supervised and unsupervised learning methods. Machine learning models which have been researched to help predict diabetes include support vector machines, decision trees, artificial neural networks, and more.

As an example of a machine learning application to diabetes prediction, Srivastava, Sharma, Sharma, and Kumar used an artificial neural network to predict diabetes. The study used a Pima Indians data set of female diabetes patients as its data source to train the neural network. The data possessed various predictor attributes which were used to train an artificial neural network to predict whether a record was that of a diabetic patient or not. The attributes included the below, where the class variable indicates whether a patient is diabetic or not:

- Number of times participant was pregnant
- Plasma glucose concentration at 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skinfold thickness (mm)
- 2-Hour serum insulin (μU/ml)
- Body mass index (weight in kg/(height in m²)
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)

An artificial neural network is inspired by biological systems, namely the human brain. In a neural network, a layer of nodes represents neurons which accept continuous data inputs from a source such as a data set. The initial layer transmits the inputs to yet another interconnected layer where each neuron in the layer aggregates its inputs using a mathematical activation function. The neuronal outputs of the neuronal layer are fed as weighted inputs to the next layer where the same process occurs, until the final layer makes a prediction of diabetic or not. The weights are associated with the interconnections of the neurons and are adjusted using a backpropagation method during the process of learning the data to construct a predictive model. Using an artificial neural network, the researchers were able to achieve an accuracy rate of 92% in predicting diabetes using the Pima Indians data.

Figure 1 below depicts a single neuron in a neural network. Multiple data inputs to the node, or neuron, are weighted and netted together using a mathematical function. The netted value is then passed to an activation function which converts the weighted sum to a final output value, or activation. The weights determine the final outputs, or predictions, of the network, and they are what the model adjusts through a process of backpropagation as it makes multiple passes through the training data.

In yet another study by Contreras and Vehi, it was noted that the patient variability associated with diabetes has not been accounted for in diabetes management processes and protocols which tend to use generalized models of the disease. They noted that patient variability is of consequence when attempting to manage the disease. As such, a number of patient stratification, or grouping, mechanisms were discussed in their review of prior research. They described studies which leveraged machine learning models such as random forest and regression algorithms to classify patients into complication risk groups such as retinopathy, neuropathy, and nephropathy which can all occur as a result of diabetes. Other studies sought to stratify patients with the disease by drug usage, HbA1C profiles, or biomechanical foot profiles. Figure 2 is a portion of a table constructed by Contreras and Vehi to summarize the stratification methods used by different historical studies. The method ‘ANN’ is an Artificial
Neural Network, ‘RF’ is Random Forest, ‘RA’ is a Regression Algorithm, ‘SVM’ is a Support Vector Machine, ‘NB’ is a Naïve Bayes Algorithm, and ‘DT’ is a Decision Tree.

A study conducted by Brigham and Women’s Hospital used machine learning to uncover the most important characteristics that can predict heart failure in type 2 diabetes patients. The study found that 10 factors, including BMI, HDL-C, age, hypertension, creatinine, QRS duration, diabetes control (fasting plasma glucose), coronary artery bypass grafting, and myocardial infarction 2 were significant predictors. These factors were uncovered by using data from over 8756 type 2 diabetes patients, of which, 319 suffered from heart failure. The factors were used to calculate a risk score for patients. Furthermore, the risk score calculation mechanism was made available as an online tool for patients and soon will be made available on electronic health records of patients. This is an example of how machine learning is being used to help manage and prevent complications associated with the disease.

Improvements in diabetes prediction have been made through combinations of different machine learning models. Wu, Yang, Huang, He, and Wang conducted a study where K-Means clustering and logistic regression were combined to improve type 2 diabetes prediction. Zhu, Idemudia, and Feng improved K-Means Clustering through Principal Components Analysis which was then applied to medical health records data followed by Logistic Regression to enhance diabetes prediction results. These studies exemplify how supervised and unsupervised learning methods can be combined to improve the overall prediction accuracy for T2D.
Within the context of social network and text data mining, Marir, Said, and Al-Obeidat (2019) applied data mining on social network based text in order to uncover insights into diabetes risk factors, symptoms, and treatments. Another study mined social network text data in India to understand the relationship between sentiments towards foods, physical activity, and diabetes risk factors. Questionnaires completed by diabetic and pre-diabetic patients in China have also been analyzed using data mining techniques to extract significant predictor attributes for use in models for diabetes prediction.

Ensemble and feature reduction techniques have also been used to improve T2D prediction accuracy. Ensemble based methods including Adaboost and bagging using J48 decision trees were successfully used to enhance the accuracy of diabetes prediction by Perveen, Shahbaz, Guergachi, and Keshavjee (2016). In Ensemble based learning, multiple models are trained and used to make predictions. The different models each make a prediction decision and an algorithm is used, such as voting or averaging, to make a final decision. In particular, Adaboost uses multiple weak model learners and combines them into a single model with greater predictive power than any one alone.

A visual depiction of Adaboost is presented in Figure 3. The weak learner represented by model B1 predicts the ‘+’ objects on the left side of the data set well but does a poor job classifying the right side. Similar constructs occur with models B2 and B3. However, the combination of models represented by B4 classifies the data quite well.

Principal components analysis was used to isolate attributes in data that were most influential in predicting diabetes. Principal Components Analysis seeks to reduce the feature space, or attribute set, in a prediction problem by deriving a smaller number of independent and uncorrelated predictive features from the original feature space. Reducing the number of features used for modelling and deriving uncorrelated features in this manner can help improve model accuracy and generate a more understandable model explained by fewer predictor attributes.

As can be seen in Figure 4, the PCA process transformed a multi-dimensional, or multi-attribute, set of data described by 3 genes, into one which is two dimensional, with only two newly derived variables, or ‘principal components’, that describe the data. The newly derived principal components are generated from the original attribute set of the data and can be used as predictors for a machine learning model in place of the original, larger, attribute set.

Machine learning algorithms and their generated predictive models can seem like black boxes at times. A new approach to diabetes prediction was
proposed by Hayashi and Yukita which used an improved rule extraction method generating data rules and patterns in the data that were human readable. The study employed a ‘transparent’ box approach to data mining rather than the traditional black box approaches.

Applications to Larger Healthcare BI Systems:

Actionable intelligence only has value if it is delivered promptly and therefore can result in a change in business outcomes. The Gunapati Venkata Krishna (GVK) Emergency Management and Research Institute (EMRI) exemplifies this concept through their use of business intelligence. This entity is a partnership of public and private sectors in India which provides free emergency medical services to 750 million people. Coye (2016) noted that the entity responds to 30 million emergency calls and saves a million lives annually by deploying 9000 ambulances and managing 20,000 emergencies daily. As described earlier, by analyzing the millions of emergency requests which occur in a real-time manner and by being fed data about the outcomes of emergency dispatches and hospital interventions, the decision support analytics capabilities of the entity are improved in a real-time manner (Coye, 2016). Such an entity, however, can improve its operations even further by applying the results of the machine learning based diabetes studies described earlier.

Social Network and Questionnaire Based Improvements:

Social network data can be used by EMRI in a manner similar to Ramsingh and Bhuvaneswari in which diabetes sentiment analysis was applied to unstructured data. EMRI serves millions of patients and can seek to tap into the network of unstructured data created by their patients in social media regarding the services it has provided. The mined data can uncover feedback from patients that can be leveraged to further improve healthcare delivery.

Questionnaires can be used by EMRI, as was done in study, to uncover the most salient factors and variables associated with different aspects of the EMRI healthcare delivery system. Analyzing questionnaires completed by the millions of patients served using machine learning techniques can uncover the most significant factors associated with ambulance dispatchments, emergency room visits, emergency number calls, and more. Attempting to analyze the large amounts of data such questionnaires would...
produce for patterns on an ‘adhoc’, manual basis, would be complex and burdensome exercise.

**Machines Learning Model Based Improvements:**

If not already in use, ensemble approaches such as Adaboost and bagging can be used to improve the accuracy of machine learning models used by EMRI. As described earlier, the Adaboost algorithm leverages a suite of weak learners in order to generate a single model which is a strong classifier. Using Adaboost is an example of leveraging multiple instances of a single classifier type, namely a decision tree by default. However, the hybrid model techniques described earlier, namely using the unsupervised method of K-Means clustering followed by the supervised method of logistic regression, can be used to improve model predictions as well. Principal components analysis (PCA) can be used by EMRI as a feature reduction method to help improve model performance by reducing the predictive feature space. PCA derives new orthogonal, or uncorrelated, features from existing ones, resulting in a reduced feature space, helping to improve model accuracy. The improvements noted above may be applied to help predict and recommend the best medical treatments and drug therapies, to estimate the recovery time for medical procedures, and to predict potential disease and medical disorder onset.

**Risk factor Derivation:**

The heart failure risk factor derivation approach described in the Brigham and Women’s Hospital (2019) study can be applied by EMRI as well. The entity can isolate through machine learning algorithms which factors are most predictive of different medical disorders and medical procedures as well as the factors which contribute most to emergency room visits and emergency hotline calls. Additionally, as was described in the study by Contreras and Vehi (2018), patient stratification methods can be employed by EMRI to group patients into different risk factor categories based on their attributes. As exemplified earlier, patients can be segmented by EMRI as being more at risk for retinopathy, neuropathy, heart failure, etc. Once the risk factors are determined, a risk weightage, or score, can be applied by EMRI to individual patients, regions of interest, or districts, with a higher risk or probability for making emergency calls or for developing certain conditions.

**Rule Extraction Based Improvements:**

As noted earlier, the models developed by machine learning processes and algorithms can be black boxes. EMRI can use a Recursive Rule extraction algorithm employing J48 graft to derive rules and relationships in its data. The identified readable rules and relationships can be used by EMRI to identify what antecedent conditions lead to certain consequences. Examples include:

A. What months of the year are associated with the highest rates of ambulance dispatches?
B. Which patient factors and conditions lead to longer hospital stays?
C. Which treatments lead to the most successful patient outcomes measured by lower patient readmission rates?

**Conclusion**

Machine learning research findings focused on solving specific problems such as diabetes prediction can be leveraged to further improve large scale business intelligence applications in the healthcare sector. Diabetes prediction and risk factor analytics through the use of machine learning has led to a number of novel approaches to address the problem of disease prediction accuracy. For example, social network data analysis focused on uncovering public sentiment related to diabetes, physical activity, and food has been conducted. Furthermore, a white box machine learning approach to diabetes prediction was developed using a Recursive Rule extraction algorithm employing J48 graft. Also, prediction performance improvements were made through hybrid applications which combined K-Means Clustering, Logistic Regression, and/or PCA. Ensemble approaches which leveraged bagging and Adaboost techniques have also been leveraged to improve prediction accuracy of diabetes.

These improvements can be used by healthcare entities with large scale BI systems such as EMRI in different ways. They can be used to:
A. Improve the BI system’s ability to recommend short term and long term treatments/procedures.
B. Enhance the BI system’s ability to isolate patterns and trends related to ambulance dispatches, emergency room visits, emergency number calls, and more.
C. Enhance the BI system’s ability to estimate the duration of surgical recovery and predict the probability of success of medical procedures before and after they are applied.

Furthermore, by applying these improvements, an entity such as EMRI can more efficiently:
A. Allocate its physical and human resources based on predicted and estimated need, thereby reducing costs.
B. Identify its most problematic operational issues as perceived by its patients and staff.
C. Reduce costs by leveraging machine learning based treatment recommendations that have higher probabilities of success over others.
D. Develop personalized disease treatment plans based on model predictions and methods for its patients.
E. More effectively isolate disease risk factors and predict complication risks.
F. Stratify patients into different risk and complication groups during and after treatments/procedures are applied for more targeted methods of preventing possible complications.

Future research should focus on analyzing focused and specific studies associated with machine learning applications to disease beyond diabetes prediction. Such diseases include, but are not limited to, heart disease, chronic obstructive pulmonary disease (COPD), lung cancer, breast cancer, and dementia. As was demonstrated in this analysis and review, these studies can present new applications and opportunities for improving large scale healthcare institutions within the context of their business intelligence systems. Furthermore, the reverse is also true. The analysis and review of large scale institutions and their applications of business intelligence may lead to new insights, applications, and opportunities for specific machine learning research focused on disease risk prediction. By studying macro level healthcare business intelligence systems in tandem with more focused studies on machine learning based disease prediction, synergies can be uncovered along with adoption opportunities which may otherwise remain unnoticed.

References
1. Aziz, H. A., A review of the role of public health informatics in healthcare. *Journal of Taibah University Medical Sciences*, 12(1), 78-81. doi:https://doi.org/10.1016/j.jtumed.2016.08.011 (2017).
2. Brigham and Women’s Hospital., Predicting risk of heart failure for diabetes patients with help from machine learning. Retrieved from https://www.sciencedaily.com/releases/2019/09/190913191451.htm (2019).
3. Carter, K. B., *Actionable Intelligence: A guide to delivering business results with Big Data fast!* Hoboken, NJ: Wiley (2014).
4. Contreras, I., & Vehi, J., Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. *Journal of medical Internet research*, 20(5), e10775. doi:10.2196/10775 (2018).
5. Geeks for Geeks. *Boosting in machine learning, Boosting and Adaboost.* Retrieved from https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/
6. Hayashi, Y., & Yukita, S., Rule extraction using Recursive-Rule extraction algorithm with J48 graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. *Informatics in Medicine Unlocked*, 2, 92-104. doi:https://doi.org/10.1016/j.imu.2016.02.001 (2016).
7. International Diabetes Foundation. (IDF; 2019). *Diabetes facts and figures*. Retrieved from https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
8. Joel Coye, M. D. M. P. H. M., Informatics: The Frontier of Innovation in Health and Healthcare. *Engineering*, 2, 37-39. doi:10.1016/J.ENG.2016.01.009 (2016).
9. Larose, D. T. & Larose, C. D., *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: Wiley (2015).
10. Mahboob Alam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., . . . Abbas, Z., A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204. doi:https://doi.org/10.1016/j.imu.2019.100204 (2019).

11. Marir, F., Said, H., & Al-Obeidat, F., Mining the Web and Literature to Discover New Knowledge about Diabetes. *Procedia Computer Science*, 83, 1256-1261. doi:https://doi.org/10.1016/j.procs.2016.04.261 (2016).

12. Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q., Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences*, 29(2), 93-99. doi:https://doi.org/10.1016/j.kjms.2012.08.016 (2013).

13. Mishra, P., *A layman's introduction to Principal Components*. Retrieved from https://hackernoon.com/a-laymans-introduction-to-principal-components-2fca55c19fa0 (2018).

14. Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K., Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*, 82, 115-121. doi:https://doi.org/10.1016/j.procs.2016.04.016 (2016).

15. Ramsingh, J., & Bhuvaneswari, V., An efficient Map Reduce-Based Hybrid NBC-TFIDF algorithm to mine the public sentiment on diabetes mellitus – A big data approach. *Journal of King Saud University - Computer and Information Sciences*. doi:https://doi.org/10.1016/j.jsuci.2018.06.011 (2018).

16. Rigla, M., García-Sáez, G, Pons, B., & Hernando, M. E., Artificial Intelligence Methodologies and Their Application to Diabetes. *Journal of Diabetes Science and Technology*, 12(2), 303-310. doi:10.1177/1932296817710475 (2018).

17. Srivastava, S., Sharma, L., Sharma, V., Kumar, A. & Darbari, H., Prediction of Diabetes Using Artificial Neural Network Approach: ICoEVCI 2018, India. 10.1007/978-981-13-1642-5_59. (2019).

18. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X., Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, 10, 100-107. doi:https://doi.org/10.1016/j.imu.2017.12.006 (2018).

19. Zhu, C., Idemudia, C. U., & Feng, W., Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 100179. doi:https://doi.org/10.1016/j.imu.2019.100179 (2019).