Portfolio Report

Data Mining on Artificial Pancreas system

CSE 572 Data Mining

Islam Hamed
Ira A. Fulton School of Engineering
Arziona State University
ID: 1221217088

I. INTRODUCTION

Humans are creating 2.5 quintillion bytes every day [1], this volume of data were very difficult to process or use in last decades. As Edwards said "You can't manage what you don't measure." [2]. Fortunately, nowadays there are the computerize capabilities that can measure and process these amount of data, and produce patterns in data, where we can get knowledge. Accordingly, we can define Data Mining as the process of finding patterns in big set of data using machine learning, artificial intelligent and database systems [3]. Data Mining is being used in different fields, such as business, science and medical. In this project we are focusing in data mining in medical field. The medical field is a very sensitive area, because it is connected directly to human's health.

The Data Sets

Before describing every problem, let us taking an overview of project data sets that are generated from sensors.

These data sets were generated from a Medtronic 670G system, which is an Artificial Pancreas medical control system. This system contains continuous glucose monitor (CGM) and the Guardian Sensor (12), which is used to read blood glucose measurements every 5 minutes. The sensor should be replaced after every single use; this single use can be 7 days continuously. User is required to get blood glucose measurements using a Contour NextLink 2.4 glucosemeter, in case of sensor replacement and recalibration procedure. Also, this procedure needs manual intervention.

Datasets: there are two different datasets:

1) from the Continuous Glucose Sensor (CGMData.csv): columns of interested are:

a. data time stamp (Columns B and C combined), and

b. the 5 minute filtered CGM reading in mg/dL, (Column AE).

2) from the insulin pump (InsulinData.csv): columns of interested are:

a. data time stamp (Columns B and C combined), and

b. auto mode exits events and unique codes representing reasons (Column Q).

The time stamp in CGM data and time stamp in insulin data are not the same, as these two devices work asynchronously.

The Project is combination of three problems. First problem is Time Series Properties of Glucose Levels Extraction in Artificial Pancreas. Second problem is Machine Model Training. And third problem is Cluster Validation.

For the first problem, the requirement is 18 metrics for two cases Manual and Auto modes. The 18 metrics describes the percentage time in GCM reading measures, for 3 different time interval: daytime, overnight, and whole day.

For the second problem, the requirement is training machine model to classify the data set time series between meal and no-meal.

For the third problem, the requirement is clustering meal and no-meal data and validate this clustering.

II. SOLUTIONS

A. The First Problem

In the first problem, first step is determining date and time threshold to distinguish between auto and manual mode. So, the first AUTO MODE ACTIVE PLGM OFF in the column "Q" of the InsulinData.csv message should be found – which indicate the start of Auto Mode–, with considering of the data is in reverse order of time. Typically, we need to search in data from the end to the beginning to get the message. Once reaching the message, this time and date is our threshold. So, any date and time after this threshold is considered as Auto Mode, and any time and date before that will be Manual Mode. As mentioned before, the two files' time stamp are not the same, so we need to use the nearest (or later than) time stamp to threshold in CGMData.csv file.

Then, data pre-processing by eliminating dates that having missing value more than 20% (data cleaning). According to sensor data, it gives reading every 5 minutes, so every day has to have 288 CGM values. So, any day has CGM values less than 230 values, will be saved in a list. This list will be checked when looping in data, if the day date is in the eliminated list ignore this day and continue. Also, to keep a high data quality, outliers should be eliminated. Outliers here are the CMG reading's values that have a high difference value from the pervious and forward reading's values. In such a case, this reading's value should be eliminated.

After that, using Python Pandas [4] to loop over all data. While looping, we check the CGM values according to interval time, mode, and GCM reading range. So, we can increase the count of each stretch in particular GCM reading range, particular mode and particular time interval.

Finally, we get the average of each stretch by dividing each stretch count over (288 * number of days).

B. The Second Problem

In the second problem, first step is pre-processing data by eliminating dates that having missing value more than 20% as we did in the first problem.

Then, we categorize each time stretch and CGM data into meal time, and no-meal time. For the meal time (tm), the stretch should be 2.5 hours (2 hours after taking the meal, and half before it). If there is meal token time (tp) inside the stretch (tm: tm+ 2hrs), we ignore tm stretch, and we consider the meal token time (tp) as new (tm). For the no-meal time, the stretch time is 2 hours when there is not any token meal inside this time period.

Then, we need to extract and select features for the two classes meal and no-meal (feature extraction). I select a feature after many trials and comparisons, to have the most accurate results and optimize the time and memory of data mining process. The Selected feature is the raw data itself. Examples of trials features are:

a) Feature Creation:
   1. stretch data average over its variance, and
   2. maximum CGM value minus minimum CGM value over its average.

b) Feature Extraction:
   1. Fourier transform.

Then using Python Scikit-learn [5], we divide the data into 70% training data and 30 % test training data, and train the machine learning model using Decision Tree classification in scikit-learn. Many classification techniques (Support Vector Machine SVM – Neural Network) were used in trials process, to get best accuracy with best computation cost. Also, Entropy method is used to select the split, instead of GINI. Which gives higher accuracy.

Finally, the model evaluation. In order to evaluate the performance of the classification model, I used the Confusion matrix, as shows in the table [6]:

	Predicted Class		
Actual Class		+	-
	+	TP	FN
	-	FP	TN

I used this matrix to calculate the accuracy and performance of output model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ [6]}$$

.

C. The Third Problem

In the third problem, we do the same pre-processing and feature extraction process as the second problem.
The features selected here are (1) stretch data average over its variance, (2) maximum CGM value minus minimum CGM value over its average, (3) and, the bin value. As if we use the raw data as the second problem, it will cost a lot of unnecessary computations. This decision after many trials and comparisons.
In order to get every stretch bin value, a simple equation has been developed:

$$bin = \frac{carb\ value - \min(carb\ value)}{20} + 1$$

These bins are considered as ground truth for the clustering, after using K-means, and DBSCAN to cluster. We can compare these clustering with the ground truth. By calculating Purity, Entropy and SSE for each cluster technique (K-means, and DBSCAN).

$$Entropy = -\sum_{i=0}^{n-1} P(x_i) * \log_2(P(x_i)) \text{ [6]}$$

$$Purity = \sum_{i=1}^{k} \frac{\max(m_i)}{m} \text{ [6]}$$

$$SSE(C_1, C_2, \dots, C_k) = \sum_{j=1}^{k} \sum_{i=1}^{m} w_{ij}^{p}\, dist(x_i, c_j)^2 \text{ [6]}$$

III. Results

For the first problem, the output is 2x18 matrix printed in csv file. Two rows for two modes, manual mode and auto mode. Eighteen columns for 6 GCM reading ranges as follows:

a) Percentage time in hyperglycemia (CGM > 180 mg/dL),

b) percentage of time in hyperglycemia critical (CGM > 250 mg/dL),

c) percentage time in range (CGM >= 70 mg/dL and CGM <= 180 mg/dL),

d) percentage time in range secondary (CGM >= 70 mg/dL and CGM <= 150 mg/dL),

e) percentage time in hypoglycemia level 1 (CGM < 70 mg/dL), and

f) percentage time in hypoglycemia level 2 (CGM < 54 mg/dL).

Each range has 3 time intervals daytime (6 am to midnight), overnight (midnight to 6 am) and whole day (12 am to 12 am).

For the second problem, the output is a trained machine learning model (model.sav) to classify the GCM readings to meal and no-meal using Decision Tree classification and csv file with matrix of size Nx24, where 24 is the size of CGM time series and N the size of unlabeled data. This model has accuracy 94.14%, when test it with the 30% training test data.

For the third problem, the output is 1x6 vector that contains SSE, Entropy, Purity for each clustering technique used to cluster the dataset (K-means, and DBSCAN). In order to compare between these two technique by the ground truth we already have. According the output, the K-means SSE is very high, however the K-means Entropy is better than DBSCAN, and for the Purity, it is the approximately the same for both.

IV. LESSONS LEARNED

I explored to new technology and techniques to process on data, moreover new methods to solve problems.

In real world problems, it is very rare to find a perfect dataset. Because of may be human errors, sensors' failures, and collecting data errors. So, we need to implement data quality (data cleaning), and data preprocessing techniques and strategies to tackle these kind of limitations in datasets. [6].

Data cleaning is a process that tackles data noise and artifacts, data outliers, and missing data. On the other hand, data preprocessing is a technique that applied to datasets, in order to make them more suitable for data mining with respect to speed, memory, and accuracy [6].

I explored to many data cleaning techniques, as follows:

a) Missing data: in this problem domain, the missing data would be happened because of replacing sensors after 7 days continuously usage, or human error in manual mode. This issue can be tackled by eliminate entire day, or interpolation. Indeed, I used a recommended threshold to eliminate days that has more than 20% missing data. If the day's segment less than 230 readings, it should be eliminated.

b) Outliers: usually they are objects with dissimilar characteristics to other data objects in the same dataset [6]. In this problem domain, I searched and deleted any CGM reading values with high difference between the two other values, the value before and the value after. The difference may be suddenly increase or decrease.

Also, I studied various data preprocessing techniques in order to select the most suitable techniques to this problem domain, as follows:

a) Dimensionality reduction: data's high dimensionality means, the data has many features that could affect the time and memory of data processing operation. One of the most popular linear technique to tackle this problem is Principal Components Analysis (PCA). It is an algebra method to find new feature from another continues features. [6]

b) Feature subset selection: this is another way to reduce dimensionality and optimize data techniques time and memory. In this method, we ignore some duplicates or irrelative features, and use the other features. [6]

c) Feature creation: also called Feature Extraction. This is a process of creating or extract new feature from the original raw data. The new created features should capture the important information in the original raw data.

Problem two and three asked for machine learning models, do I get to know in practical what is machine learning, supervised and unsupervised learning, classification, and clustering.

Machine learning is an artificial intelligence (AI) application, which make the system model able to learn independently from human using previous experience without explicitly programmed [7].

Machine learning has many techniques:

a) Supervised learning technique: we can summarize it as learning from datasets that are labeled or already classed. The learned model can predict the class label of new unlabeled data, such as classification [7].

   I. Classification model: it is a function that represents the relationship between features set and class label. Some classification techniques examples: Decision Tree, Nearest Neighbor, Naïve Bayes, Artificial Neural Network, and Support Vector Machine. [6]

b) Unsupervised learning technique: it is the process of learning from unclassified or labeled datasets. The learned model can categorize the new data with pervious data, such as clustering.

   I. Clustering: is the process of dividing the data into meaningful and useful groups. Some clusters techniques examples: K-means, and DBScan [6].

c) Semi-supervised learning: It is somewhere between supervised (labeled) and unsupervised (unlabeled) learning. So, it combination of both labeled and unlabeled data. Typically, the unlabeled data more than the labeled data [7].

In order to evaluate the model in the second problem, I used the Confusion Matrix, which is provide four attributes to calculate the accuracy, also we can calculate the precision, recall, and F1. Good model should maximize the both values, precision and recall, so we can use F1 measure. [6]

$$Recall(r) = \frac{TP}{TP+FN} [6]$$

$$Precision(p) = \frac{TP}{TP+FP} [6]$$

$$F1 = \frac{2rp}{r+p} [6]$$

I explored to python pandas [4], it is an open source for data analysis. I used it to process and manipulate data from CSV files. It has many helpful functions such as (Max, Min, Average, and Locate).

Also, I used Python Scikit-learn [5]. It is a machine leaning tool for python. Where there are many classification modules such as Decision Tree, Support Vector Machine (SVM), and Neural Network. Also, there are many clustering methods, such as K-means, DBSCAN. Moreover, there are Regression, Dimensionality reduction, Model selection, and Preprocessing techniques.

In addition to Python Pickle module, which is used to save the learning machine model. In order to be able to use it again in different script and to classify unlabeled data.

Using Trial method to select best technique or method, when there are many, in respect of accuracy and computation cost. In data mining, many techniques available to tackle the same problem. So, to select the best one, we need to consider the data domain. Although, there may be many techniques also for the same domain. Here, the best way to determine the best techniques is Trialing.

V. REFERENCES

[1] J. Bulao, "Techjury," 22 January 2021. [Online]. Available: https://techjury.net/blog/how-much-data-is-created-every-day/#gref.

[2] A. McAfee and E. Brynjolfsson, "Harvard Business Review," October 2012. [Online]. Available: https://hbr.org/2012/10/big-data-the-management-revolution.

[3] ACM ;SIGKDD, "Data Mining Curriculum," 30 04 2006. [Online]. Available: https://www.kdd.org/curriculum/index.html. [Accessed 27 01 2014].

[4] W. McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," 02 March 2021. [Online]. Available: https://pandas.pydata.org/docs/.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, "Scikit-learn: Machine Learning in Python," *JMLR 12*, pp. 2825-2830, 2011.

[6] P. N. Tan, Introduction to Data Mining, University of Minnesota, 2005.

[7] E. Team, "Expert.ai," 6 May 2020. [Online]. Available: https://www.expert.ai/blog/machine-learning-definition/.