Proceedings of the 7th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2012)
H. Yang, D. Zeng, O. E. Kundakcioglu, eds.

Discovering Meaningful Cut-points to Predict High HbA1c Variation

Si-Chi Chin, University of Iowa [1]
W. Nick Street, University of Iowa
Ankur Teredesai, University of Washington - Tacoma

Abstract

Monitoring HbA1c, the measurement for the average blood glucose level, is important to diabetic patients and may help improve their treatment. This study aims to train learning algorithms that predict patients with high variation in their HbA1c readings. Attributes in clinical data are often continuous, such as age, blood pressure, and lab tests. However, many machine learning algorithms work better – or work only – with categorical attributes. We propose using discretization processes to identify meaningful cut points for continuous attributes. For example, the study could help understand questions like, What is the range of age and BMI for diabetic patients that have high variation in their HbA1c readings? The discretization process finds the number of intervals and the boundaries for the intervals. The process may reveal meaningful cut points in continuous attributes and contribute knowledge to the medical domain. Discretization process also allows us to discover suitable intervals and boundaries for a particular attribute depending on variety of parameters that can be controlled by the experimenter.

Keywords: HbA1c Monitoring, Discretization algorithm, Decision Tree

1 Introduction

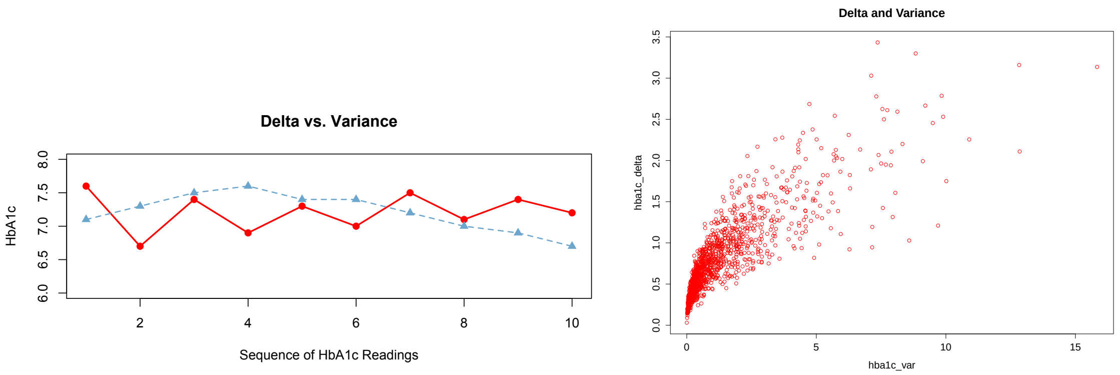
HbA1c is a measurement for the average blood glucose level and is commonly used to monitor diabetes patients. In this paper, we explore methods to analyze patients who exhibit high variation for their HbA1c monitoring. We adopt Time Abstraction (TA) based analysis [1] to derive useful summaries of time series. In addition, we employed discretization algorithms to identify meaningful cut-points among continuous features to enhance the analysis. The goal of the proposed methods is to improve predictive generalization and translate the discretization results into interpretable clinical knowledge.

In clinical data analysis, discrete features are easier to interpret for both data scientists and clinicians. Discrete features are closer to a knowledge-level representation than continuous ones. Prior research also indicates that discretization makes learning faster and more accurate [2]. Discretization methods determine “cut-points” (or split-points) for continuous features, dividing a range of continuous values into intervals of various lengths. The discovery of meaningful cut-points would enhance knowledge to the question: what is the range of age and BMI for diabetic patients that are more inclined to have high variance in their HbA1c readings?

Variation for HbA1c can be observed from two aspects: the variance and the delta. Variance indicates the usual difference between a measurement and the mean of those measurements, while the delta is the mean of the differences between consecutive readings. Figure 1a exemplifies ten HbA1c readings for two patients, represented separately by a solid red line (Patient A) and a dashed blue line (Patient B). Both patients have the same variance (0.081) for their HbA1c time series. However, Patient A has higher delta (0.46) than Patient B (0.15). Although Figure 1b shows that the variance and the delta are linearly correlated, they represent different concepts.

In this paper, we aim to learn methods that explain and predict patients with high variance and high delta for their HbA1c. We define that a patient has high variance and delta if the value is higher than one standard deviation away from the mean of the data.

[1] This work is done during the author’s visit to the Center for Web and Data Science at the University of Washington - Tacoma.



(a) Illustration of delta and variance

(b) Correlation between delta and variance

Figure 1: Variance vs. Delta

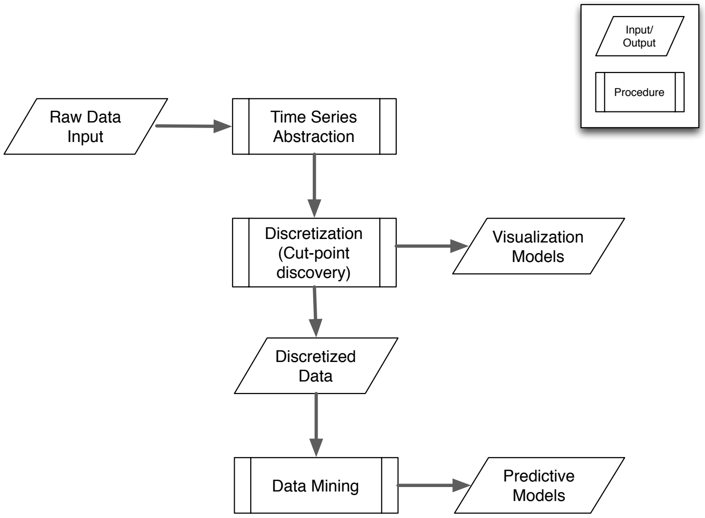


Figure 2: Analysis framework

The remainder of the paper is organized as follows. In the next section, we propose an analysis framework that incorporates the TA-based analysis and the discretization methods. In Section 3, we summarize the data and detail our experiments. In Section 4, we discuss the experimental results. Section 5 concludes the work and outlines directions for future research.

2 Methods

We incorporated TA based analysis and two discretization methods. Figure 2 shows the proposed analysis. The preprocessed input data contains the sequence of measurements for diastolic blood pressure, systolic blood pressure, BMI, and patient age, ordered by the sequence of visits. Each sequence of measurements is a time series feature. As shown in Figure 2, we first summarize the preprocessed time series to obtain abstract descriptions of the data. To provide this higher level description of the time series, we computed the mean, the variance, and delta (i.e. the difference between two consecutive data), and the variance of delta for each time series feature.

As described in the proposed framework (see Figure 2), we apply discretization methods on these summaries of time series to discover meaningful cut-points. We then use visualization techniques to examine the cut-points derived from different discretization algorithms. The visualizations of different cut-points support the interpretation of the data. The discretized features is then used for data mining tasks to learn predictive models. The proposed framework aims to investigate how to discover meaningful cut-points through visualizations and whether these summaries could be used to learn some characteristics of the patients.

We adopted Minimum Description Length Principle (MDLP) and ChiMerge to discretize continuous features. MDLP is a supervised, top-down (splitting) discretization method. It selects cut-points that maximizes the information entropy. ChiMerge is a supervised, bottom-up (merging) discretization method. It uses the chi-square statistic to determine if the class frequencies of the two intervals are significantly different.

Our predictive models include logistic regression model and decisions trees. We compare the performance of logistic regression model on the original continuous features and the discretized features. The logistic regression model over continuous attributes is used as a baseline to compare with a discretized logistic regression. The performance is measured in AUC and F1 scores, averaging from 10-fold cross validation.

3 Data Description

We used real patient data extracted from the patient record system at the University of Iowa Hospitals and Clinics. The data contains the engineered summary of time series clinical data. The data includes 1,208 patients who have at least 10 recorded HbA1c readings. All patients have a diagnosis of Type II diabetes.

Table 1 describes the variables used in the data. The processed data include time abstractions (i.e. mean, variance, delta, and variance of delta) for diastolic blood pressure, systolic blood pressure, BMI, and the patient age. Since the class label is determined by the feature *hba1c_var, and hba1c_delta*, we excluded the two features from the data for the classification tasks.

Table 2 shows the correlation results of the 14 explanatory variables (x-variables) and 2 predicted variables (y-variables, i.e. hba1c_var and hba1c_delta). The results indicate that the four time abstractions provide different information for the predicted variables. As shown in Table 2, although bp_sys_mean does not have significant correlation to hba1c_var and hba1c_delta, bp_sys_var and bp_sys_delta exhibit correlations. In addition, although bmi_mean is significantly correlated to hba1c_var and hba1c_delta, bmi_var is not.

In this paper, we aim to learn methods that explain and predict patients with high variance and high delta for their HbA1c. We define that a patient has high variance and delta if the value is higher than one standard deviation above the mean of the data. The final processed data for classification tasks contains 131 patients are labeled positive for having high variance for HbA1C and 165 patients are labeled positive for having high delta for HbA1c.

Table 1: Variable description. HbA1c related metrics are excluded from predictive models.

Index	Variable	Description
1	bp_dia_mean	Mean of diastolic blood pressure
2	bp_dia_var	Variance of diastolic blood pressure
3	bp_dia_delta	Difference between two consecutive diastolic blood pressure
4	bp_dia_delta_var	Variance of the difference between two consecutive diastolic blood pressure
5	bp_sys_mean	Mean of systolic blood pressure
6	bp_sys_var	Variance of systolic blood pressure
7	bp_sys_delta	Difference between two consecutive systolic blood pressure
8	bp_sys_delta_var	Variance of the difference between two consecutive systolic blood pressure
9	bmi_mean	Mean of BMI
10	bmi_var	Variance of BMI
11	bmi_delta	Difference between two consecutive BMI
12	bmi_delta_var	Variance of the difference between two consecutive BMI
13	pat_min_age	Lowest recorded age for a patient
14	pat_max_age	Highest recorded age for a patient
15	hba1c_var	Variance of HbA1c
16	hba1c_delta	Difference between two consecutive HbA1c readings
17	class_var	Patients with HbA1c variance larger than one standard deviation
18	class_delta	Patients with HbA1c delta larger than one standard deviation

Table 2: Correlations between variable. * indicates the significance level <0.05.

index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
hba1c_var	0.20*	0.24*	0.22*	0.21*	0.02	0.15*	0.11*	0.13*	0.09*	0.007	0.06*	0.06*	-0.17*	-0.19*
hba1c_delta	0.22*	0.23*	0.22*	0.18*	0.04	0.16*	0.13*	0.15*	0.11*	0.01	0.08*	0.07*	-0.19*	-0.21*

Table 3: Performance Comparison for AUC and F1

		AUC			**F1**		
Learner	**Y**	**Raw**	**MDLP**	**ChiMerge**	**Raw**	**MDLP**	**ChiMerge**
Logistic Regression	Variance	0.657	0.735	0.728	0.07	0	0.535
	Delta	0.694	0.752	0.811	0.045	0.109	0.579
Decision Tree	Delta	0.615	0.527	0.646	0.052	0.066	0.11

4 Predictive Models

We constructed logistic regression models and decision trees using both the numeric features and the discretized features. Table 3 compares the performances of three models – a logistic regression model to predict high variance, a logistic regression model to predict high delta, and a decision tree model to predict high delta. We also compare the results between the original continuous features, the features discretized by MDLP method, and the features discretized by the ChiMerge method. We use the models that constructed with the original continuous features as the baseline to evaluate the performance.

The results show that using features discretized by ChiMerge methods consistently enhance the performance of AUC and F1 score for the three predictive models. Although using features discretized by MDLP method introduces mixed results to the models, it outperforms the baseline on both AUC and F1 scores for the logistic regression model that predicts high delta. The MDLP method also achieves higher performance on F1 score for the decision tree model and has higher AUC for the logistic regression model that predicts the high variance. The experimental results show promise using discretized features to enhance predictive models.

5 Cut-points Visualization

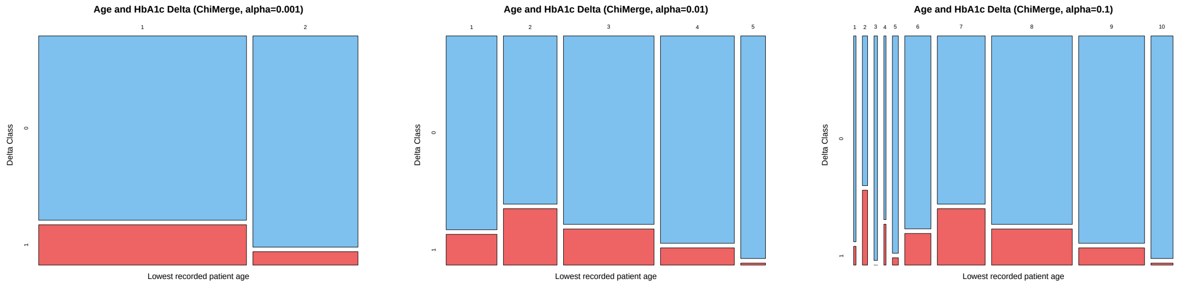
The choice of cut-points could affect the perceptions and the understanding of the data. Figure 5 shows how the choices of alpha parameters changes the number of selected cut-points. Experts in medical domain may provide input on how the parameters could be tuned. For example, Figure 3c is simpler to interpret than Figure 3b and Figure 3c. However, Figure 3b provides a more granulated view of the class distribution, indicating that the negative correlation observed in Table 2 may mostly exhibit among patients older than 39 years old. This pattern does not exhibit in a simpler discretization results. It requires experiments and collaborations with domain experts to select the most appropriate and useful number of cut-points.

Figure 5 shows six examples of cut-points visualization. Each sub-figure in Figure 5 shows the distribution of class label for each discretized interval. The width of each vertical columns reflects the size of intervals – the number of patients within the interval. The goal of cut-point visualizations is to provide graphical representations of the data to reveal patterns that are concealed in correlation analysis.

In general, MDLP method provides a discretization result that is easier to interpret. However, with different choices of alpha parameter, ChiMerge methods has the flexibility to represent the data with more granularity. For example, Figure 4b suggests that the group of patients between age and 32 has lower probability for high variance in HbA1c despite the general linear trend that the variance decrease as the age increases.

6 Conclusion

Data mining techniques are increasingly used in clinical domains such as treatment planning. However, simply building a better model – traditionally the goal of the data mining community – is not enough in these problems. Health care



(a) (b) (c)

Figure 3: Comparisons between choices of alpha parameter. Figure (a) has the cut-point at 56.5, Figure (b) has the cut-points at 38.5, 46.5, 56.5, 68.5, and Figure (c) has the cut-points at 12.5 19.5, 22.5, 23.5, 27.5, 38.5, 46.5, 56.5, 68.5.

Table 4: Cut-points for Figure 5

X	Y	MDLP	ChiMerge
pat_min_age	var	{51.5} (Fig. 4a)	{19.5, 32.5, 51.5} (Fig. 4b)
pat_min_age	delta	{56.5} (Fig. 4c)	{38.5, 46.5, 56.5, 68.5} (Fig. 4d)
bmi_delta	delta	{0.94} (Fig.4e)	{0.58} (Fig. 4f)

professionals are unlikely to trust models which are not understandable. Variable discretization can be an important tool in building models that are both accurate and interpretable.

In this paper we explored the use discretization to aid in the prediction and understanding of which diabetic patients will have trouble controlling their blood sugar. Our conclusions are as follows. First, we note that HbA1c variation is better measured using the delta measurement, or change in consecutive readings, as opposed to the more traditional variance measure. Clinically, we are looking for patients whose readings show large variations in the short term, while a large variance may indicate a continuous but undramatic trend. Applying discretization methods to continuous features such as age, blood pressure, and BMI provide insight into the profiles of patients with different variation properties, especially when used in conjunction with interpretable predictive models such as decision trees. For example, patients in their 40s show a much greater probability of high variation than those in their 60s and 70s. Finally, the discretization process often leads to improved generalization, as was seen with all of the models we tested.

In the future we will expand our list of predictive variables and explore the use of multi-dimensional discretization methods. The goal is to build easily recognizable profiles of patients who may have difficulty controlling their blood sugar in order to help guide treatment planning.

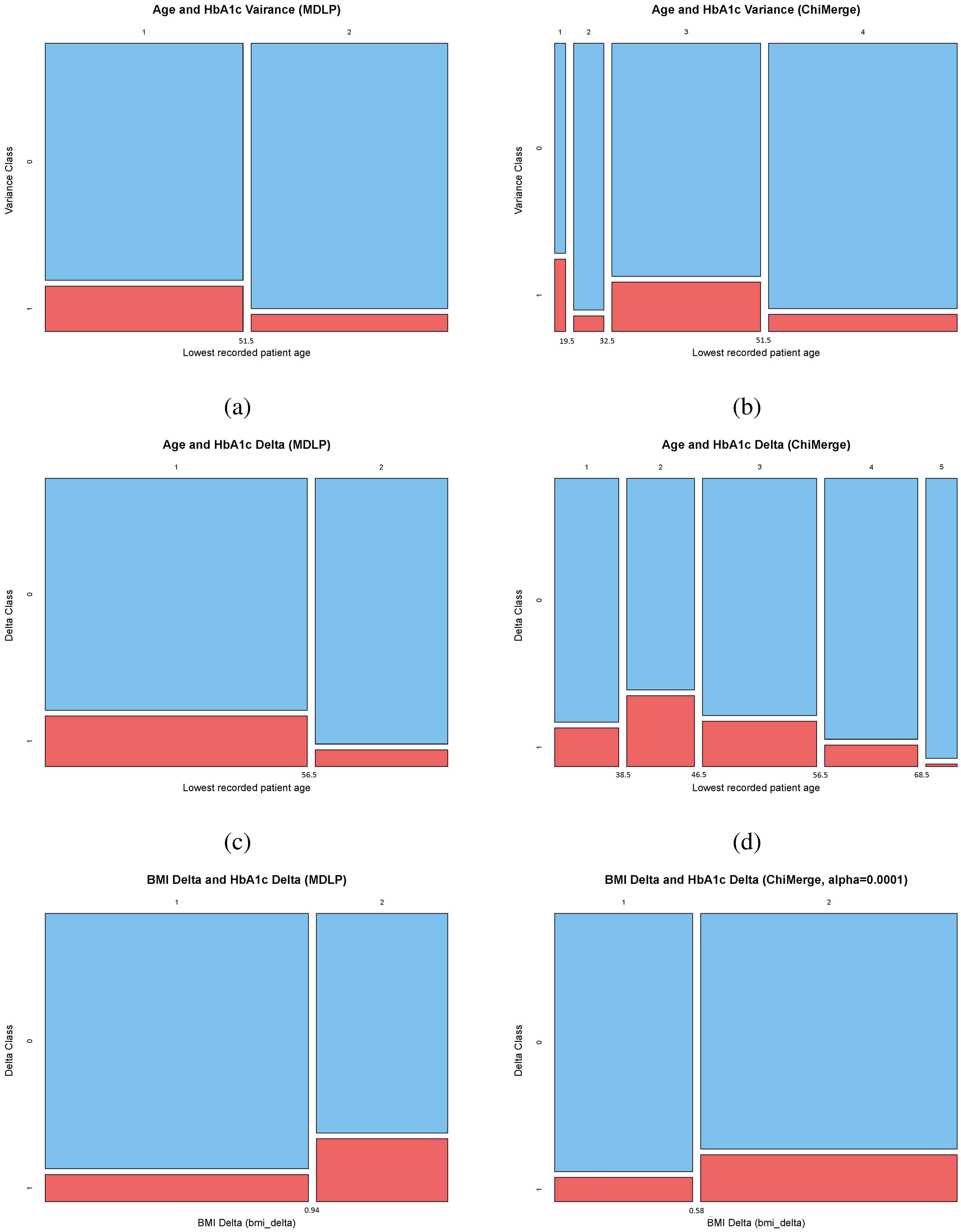
Acknowledgement

This work was made possible by Grant Number UL1RR024979 from the National Center for Research Resources (NCRR), part of the National Institutes of Health (NIH).

References

[1] Riccardo Bellazzi and Ameen Abu-Hanna. Data mining technologies for blood glucose and diabetes management. *Journal of diabetes science and technology*, 3(3):603–612, May 2009. PMID: 20144300.

[2] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.



Age and HbA1c Vairance (MDLP)

(a)

Age and HbA1c Variance (ChiMerge)

(b)

Age and HbA1c Delta (MDLP)

(c)

Age and HbA1c Delta (ChiMerge)

(d)

BMI Delta and HbA1c Delta (MDLP)

(e)

BMI Delta and HbA1c Delta (ChiMerge, alpha=0.0001)

(f)

Figure 4: Visualization examples for cut-points. See Table 4 for the values of cut-points.