35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Discretized Data Pattern in Endoscopic Gastritis Images using Dynamic Window and Pairwise Gini Criterion

Yasmin M. Yacob, Harsa Amylia Mat Sakim, Nor Ashidi Mat Isa and Zuriani Sobri, *Member, IEEE*

Abstract— **Curent standard clinical procedure for gastritis is via endoscopy by performing an invasive procedure. The procedure takes tissue samples from patient's antrum and diagnoses based on pathological evaluation. Several non-invasive computer-aided visualization studies have been conducted to perform feature extraction from the endoscopic gastritis images. Based on an extensive literature search, studies to extract data patterns from the images has never been conducted. Discretization or data pattern extraction is one of the data pre-processing technique that promotes classification. However, data pre-processing is often overlooked by many researchers because it takes up time from the overall classification process. Thus, data pre-processing studies offer faster pre-processing time and compromise with the error rate. Trade-off has been a prolonged issue in discretization studies. Often discretization time is reduced, and the error rate is compromised. However, the proposed discretization algorithm implemented on extracted features from gastritis images has reduced not only the discretization time but also the error rate. As a result of discretization process, it generates good generalization of the data patterns to the endoscopic gastritis extracted features. Thus, determining discretized data patterns from the extracted endoscopic gastritis images may improve the overall classification process in terms of accuracy and learning time.**

I. Introduction

The standard clinical diagnosis for gastritis is via endoscopy procedure. This procedure performs a biopsy from the patient's antrum to be examined by the pathologist. However, this procedure is invasive thus, may cause discomfort for the patient. Due to rapid development in medical image processing, several non-invasive computer-aided visualization studies have been conducted to perform feature extraction from the endoscopic gastritis images. The studies implement feature extraction techniques such as texture-based [1], color-based [2], texture-color-based [3], color-wavelet-based [4][5], and hybrid color-texture-wavelet-based [6]. The feature extraction method employed in this study is texture-color-based and is discussed briefly in this paper. Detail explanation of the feature extraction method employed is elaborated in [7].

The work was funded by the Universiti Sains Malaysia (USM) Research University Grant #814082.

Yasmin M. Yacob, and Zuriani Sobri are postgraduate students from the School of Electric and Electronic Engineering, Universiti Sains Malaysia, 14300, Penang, Malaysia (E-mail: ymy.ld09; zs10_eee058@student.usm.my).

Harsa Amylia Mat Sakim and Nor Ashidi Mat Isa are lecturers with the School of Electric and Electronic Engineering, Universiti Sains Malaysia, 14300, Penang, Malaysia (Phone: 604-559-5821; Fax: 604-559-5555; E-mail: amylia; ashidi@eng.usm.my).

Many researchers regard the data pre-processing as depletion of time. Thus, the data pre-processing study provides faster pre-processing time and compromises with the error rate. Discretization or data pattern extraction is one of the data pre-processing strategy that promotes classification. However, discretization often encounters the trade-off issue. Current researchers have to tolerate that whenever the discretization time is reduced, the error rate is compromised and vice-versa. Discretization is anticipated by many researchers to improve not only the discretization time but also the error rate in order to promote the overall classification process.

The study proposes extraction of data pattern that reduces not only the discretization time but also the error rate to the endoscopic gastritis images. As far as my literature search, studies to extract data patterns from extracted features of endoscopic gastritis images has never been conducted. The proposed study is developed based on benchmark data sets from the UCI machine learning repository [8]. Although the proposed discretization method is applied to the digitized endoscopic gastritis images as a case study, it has great implication to the medical research, especially critical diseases. This is because extraction of essential data patterns with faster data pre-processing time may improve the classification process of the diseases, including critical diseases such as breast cancer and brain tumor. The study proposes a gain-based boundary cut-point of changing class's technique. This technique skips inessential cut-points, thus generate better boundary cut points. The succeeding section provides brief discretization background, methodology of feature extraction and discretization, experimental result and concluding remarks of the study.

II. DISCRETIZATION

Discretization partitions ascending ordered continuous features and transform it into interval form. Discretization causes an unpatterned raw data to be simplified and generalized for easier usage and comprehend. Good discretize point generates good representation of the continuous data in interval form whereby it is written as a range of values. Since the proposed method is class-based, the discretize point that generates the interval is chosen so that it best describes the class. The transformation of data during discretization process is described in Figure 1. As illustrated in Figure 1, discretization generates interval as a model during training. Then, the continuous data is transformed into interval form. During testing, the test data is transformed based on the model and subsequently channeled for evaluation via classifier.

978-1-4577-0216-7/13/$26.00 ©2013 European Union

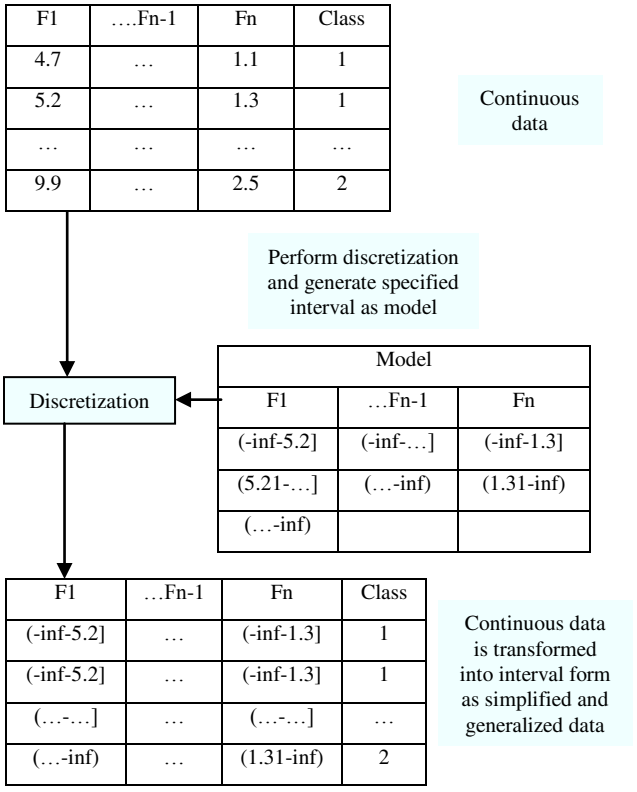


Figure 1. Data transformation during discretization process

It is mentioned earlier, the study proposed gain-based boundary cut point of changing class's technique. It is an improvement to the current technique [9][10][11]. The improvement has potential to address the prolonged trade off issue based on background study and technical literature review.

III. METHODOLOGY

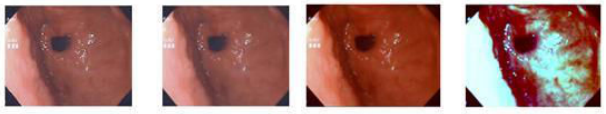
The computer-aided endoscopic gastritis diagnosis is divided into two parts. The first part involves image pre-processing and feature extraction from the acquired image. The second part involves data pre-processing of the output generated from the feature extraction part. It results in a discretized data pattern in a regularized form. The discretized data patterns then undergone the classification process. The overall flow of the computerized endoscopic gastritis diagnosis is shown in Figure 2.

A. *Image Pre-Processing and Feature Extraction*

The digitized endoscopic gastritis images are captured via Evis Exera II video imaging system provided by Olympus. Image pre-processing involves image enhancement, cropping and color conversion. In order to enhance the image, Gaussian filter is employed to smooth the image and eliminates minor noise. The samples of enhanced gastritis images are shown in Figure 3. Then, the images go through manual cropping of 100 x 100 size whereby the redness must be positioned in the middle of the region of interest (ROI). After image cropping, the image requires the color conversion process into intensity images. This is because the feature extraction process employs Gray-Level Co-Occurrence Matrix (GLCM) which is devoted to gray-level images. GLCM features are extracted from Discrete Wavelet Transform (DWT), and colored features are extracted from enhanced images.

B. *Dynamic Window and Pairwise Gini Criterion (DWPG)*

The proposed discretization algorithm is known as Dynamic Window and Pairwise Gini Criterion (DWPG). It proposed entropy-gain-based boundary cut-points of changing class's technique. The technique selects boundary points of changing classes as cut points, which skip inessential cut points. DWPG searches better boundary



(a) (b) (c) (d)

Figure 3: (a) Original image of endoscopic; (b) effect of Gaussian filter; Contrast enhancement using (c) w_{min}=1.0 and w_{max}=0.009, (d) w_{min}=0.25 and w_{max}=0.75.

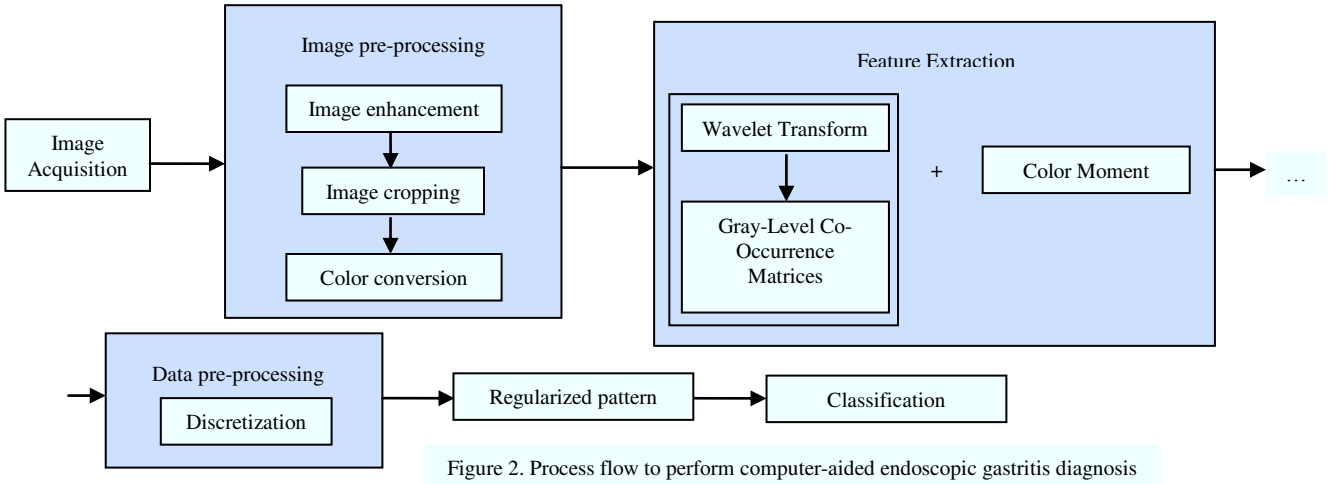


Figure 2. Process flow to perform computer-aided endoscopic gastritis diagnosis

points. The basic concept of DWPG algorithm is a good cut point exists when there is a drop in a pair of consecutive Gini Gain reading from boundary cut points of changing classes. The good cut point is at the position where the Gini Gain value is higher than the other one. DWPG implements the algorithm via possible and optimal cut points stages. In order to find possible cut points, there are two layers of screening to determine good cut points. Both layers adopt the concept of a drop in a pair of consecutive Gini Gain reading. Later, the optimal cut point is determined from set of possible cut points via threshold measure. The search to determine the possible and optimal cut points is via sequential traversal of the changing class's boundary cut points. The DWPG algorithm is implemented as follows:

1. Compute the Gini parent entropy of the data set S. The Gini parent entropy is described in (1). Let $j = 1, 2, ..., k$, where k is the number of classes in the data set.

$$\text{Gini(S)} = 1 - \sum_{j=1}^{k} p_j^2 \quad (1)$$

2. For each attribute A
   (a) Map each attribute with the class and sort them in ascending order.
   (b) If the point is a boundary cut point
      i. Compute Gini class entropy induced by the boundary cut point T given by (2)

$$\text{Gini(A; T,S)} = \frac{|S_1|}{S} * \text{Gini}(S_1) + \frac{|S_2|}{S} * \text{Gini}(S_2) \quad (2)$$

      ii. Compute Gini Gain induced by the boundary cut point T

$$\text{Gini Gain(A;T,S)} = \text{Gini(S)} - \text{Gini(A;T,S)} \quad (3)$$

   (c) Repeat step (b) for a pair of consecutive boundary cut points
   (d) Determine a higher Gini Gain value
   (e) Store the higher Gini Gain value and its position in array
   (f) Repeat step (b) to (e) until end of boundary cut point
   (g) If there is a drop in a pair of consecutive Gini Gain reading from the stored array
      i. Choose the cut point with the higher Gini Gain value from the stored array
      ii. Compute Normalized Gini Gain of the chosen cut point from the stored array given by (4)

$$\text{Normalized Gini Gain(A; T,S)} = \frac{\text{Gini Gain(A; T,S)}}{\log_2 n} \quad (4)$$

      iii. Compute Normalized Gini Threshold Measure of chosen cut-point from the stored array given by (5)

$$\text{Normalized Gini Threshold Measure(A;T,S)} = \frac{\frac{\log_2(N-1)}{N} + \frac{\Delta(A;T,S)}{N}}{\log_2 n} \quad (5)$$

for both (4) and (5), where n is the number of partition, T is the chosen cut-off point, N is the number of instances and

$$\Delta(A;T,S) = \log_2(3^k-2) - ((k*\text{Gini}(S) - k_1*\text{Gini}(S_1) - k^2*\text{Gini}(S_2)) \quad (6)$$

where k_i is the number of class labels represented in set S_i.

      iv. The optimal cut point point is determined if and only if
      Normalized Gini Gain > Normalized Gini Threshold Measure (7)
      The Normalized Gini Gain (NGG) and Normalized Gini Threshold Measure (NGTM) act as threshold criteria to determine set of optimal cut-off points.
   (h) Repeat step (g) until end of pairwise boundary cut point in the stored array

C. Experiment Set Up

Out of 500 regions of interest (ROI) cropped from the digitized endoscopic gastritis images, there are 250 normal and 250 abnormal cases. The abnormal cases are the most common type of gastritis, which is atropic gastritis. The features are extracted up to 60 features with 36 colored-features and 24 textured-features. Textured-features collected are significant GLCM features, which are contrast, correlation, dissimilarity, angular second moment, entropy and inverse different moment. Each type of the selected GLCM feature is extracted for 0^0, 45^0, 90^0, and 135^0 directions. The GLCM features are extracted from Discrete Wavelet Transform(DWT). On the other hand, mean, standard deviation, entropy and skewness are colored-features whereby each feature is extracted from nine color channel (R, G, B, H, S, V, Y, Cb, Cr). The feature extraction process is performed via the Weka machine learner.

The output from the feature extraction process is then channelled into data pre-processing stage. In the data pre-processing stage, the study employed a 90% training and 10% testing proportion, thus there are 450 cases training and 50 cases testing. The training process consists of 225 normal and 225 abnormal data. Whereas, the testing process consists of 50 cases with 25 normal and 25 abnormal cases. The discretization method runs on 10-fold cross validation of C5.0 classifier. During discretization, the training process created a data pattern model in interval data form. Testing involves transformation of continuous features into interval form based on the model created during training.

The performance of discretized versus non-discretized features is evaluated based on error rate. Discretization time is not recorded because non-discretized features do not perform discretization. The analysis is performed among non-discretized method, MDLP [9], EDA-DB [11] and the proposed DWPG technique. The error rate result for each compared method is taken for both the training and testing process.

IV. EXPERIMENTAL RESULTS

The experimental results of discretized data pattern method from the endoscopic gastritis data set is compared among non-discretized or continuous scheme and popular peer methods in the boundary cut point of changing class's technique. The two most popular discretization methods in the boundary cut point technique are MDLP and EDA-DB. Discretization time for non-discretized scheme is not recorded since it is not applicable. This paper addresses the impact of improved boundary cut point of changing class's technique to the prolonged trade-off issue of compromise between error rate and discretization time. The result is presented in Table I and Table II. Whereas Table I and Table III focus on the impact of discretized versus non-discretized scheme to the error rate and classification time.

Referring to Table I, the proposed method shows the best error rate among compared methods not only during training but also testing. The discretization time of DWPG is the best among analyzed methods as shown in Table II. The proposed method succeeded to address the trade-off issue whereby it improves not only the error rate but also the discretization time of the tested data set.

Another essential issue is the impact of performing discretization versus non-discretized scheme. Results in Table I show DWPG's error rate seems comparable with the continuous scheme, but the other compared methods; MDLP and EDA-DB, did not show betterment. The discretized scheme seems to generate less significant impact to the error rate compared to non-discretized scheme but the proposed method improves both the training and testing error rate. On top of that, although classification time of the compared methods during testing is insignificant, discretized scheme improves the classification time. The classification time of the analyzed methods are shown in Table III.

TABLE I. ERROR RATE FROM TRAINING AND TESTING

Error Rate	Training	Testing
Continuous	28.00±1.80	32.00
MDLP	35.30±1.20	38.00
EDA-DB	29.60±1.90	36.00
DWPG	27.30±2.50	26.00

TABLE II. DISCRETIZATION TIME AMONG POPULAR METHODS OF BOUNDARY CUT POINT TECHNIQUE

	Discretization Time (sec)
MDLP	13.853
EDA-DB	16.714
DWPG	11.311

TABLE III. CLASSIFICATION TIME FROM TRAINING AND TESTING

Classification Time	Training	Testing
Continuous	0.4	0.1
MDLP	0.3	0.1
EDA-DB	0.3	0.1
DWPG	0.3	0.1

V. CONCLUSION

The proposed discretized data pattern method not only improves the discretization time but also error rate. Thus, discretization may generate impact to the overall classification process. Faster and essential discretized data patterns can be extracted and benefit not only to the endoscopic gastritis studies but also to the critical diseases such as brain tumor and breast cancer.

ACKNOWLEDGMENT

Authors of this paper would like to acknowledge the assistance given by Dr. Nor Aizal Che Hamzah, Dr. Nazri Mustafa, Dr. Lee Yeong Yeh of the Endoscopy Unit, Hospital Universiti Sains Malaysia (HUSM), Kubang Kerian, Kelantan, Malaysia.

REFERENCES

[1] P. Wang, S. M. Krishnan, C. Kugean, M. P. Tjoa, "Classification of Endoscopic Images Based on Texture and Neural Network", in *Proceedings of the 23rd Annual EMBS International Conference*, pp. 3691-3695, 2001.

[2] Quan Zhang, Xiao-ying Tai, "Endoscope Image Retrieval Based on Color Feature Fusion", *Congress on Image and Signal Processing*, pp. 247-251, 2008.

[3] Marta P. Tjoa, Shankar M. Krishnan, "Feature Extraction for the Analysis of Colon Status from the Endoscopic Images," *Biomedical Engineering Online*, pp. 1-5, 2003.

[4] Stavros A. Karkanis, Dimitris K. Iakovidis, Dimitris E. Maroulis, Dimitris A. Karras, M. Tzivras, "Computer-Aided Tumor Detection in Endoscopic Video Using Color Wavelet Features", *IEEE Trans. on Information Technology in Biomedicine*, vol. 7, no. 3, pp. 141-152, 2003.

[5] C.S. Lima, D. Barbosa, J. Ramos, A. Tavares, L. Monteiro, L. Carvalho, "Classification of Endoscopic Capsule Images by Using Color Wavelet Features, Higher Order Statistics and Radial Basis Functions", in *30th Annual International IEEE EMBS Conference*, pp. 1242-1245, 2008.

[6] C.-R. Huang, P.-C. Chung, B.-S. Sheu, "*Helicobacter Pylori*-Related Gastric Histology Classification Using Support-Vesctor-Machine-Based Feature Selection", *IEEE Trans. On Information Technology in Biomedicine*, vol. 12, no. 4, pp. 523-531, 2008.

[7] Zuriani Sobri and Harsa Amylia Mat Sakim, "Texture Color Fusion Based Features Extraction for Endoscopic Gastritis Images Classification", *International Journal of Computer and Electrical Engineering (IJCEE)*, vol. 4 no.5, pp. 674-678, 2012.

[8] A. Frank, A. Asuncion, A., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[9] U. Fayyad, K.Irani, "Multi-interval discretization of continuous attributes for classification learning",in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp 1022–1027, 1993.

[10] P. Perner and S. Trautzsch, A. Amin, D. Dori, P. Pudil, and H. Freeman, "Multi-interval discretization methods for decision tree learning", *Advances in Pattern Recognition*, vol. 1451, pp. 475-482, 1998.

[11] A. An and N. Cercone, "Discretization of Continuous Attribute for Learning Classification Rule," *Lecture Notes in Artificial Intelligence (PAKDD 1999),* vol. 1574, pp. 509-514, 1999.