Using Data Mining to Help Detect Dysplasia

Extended Abstract

Avi Rosenfeld[1], Vinay Sehgal[2], David G. Graham[2], Matthew R. Banks[2], Rehan J. Haidry[2], Laurence B. Lovat[2]
[1]Department of Industrial Engineering Jerusalem College of Technology (JCT), Jerusalem, Israel
[2]Department of Gastroenterology, University College London Hospital (UCLH), London, United Kingdom
rosenfa@jct.ac.il, v.sehgal@ucl.ac.uk

I. Introduction

Over the past 10 years, data mining has been increasingly used within the medical field. To date, models have typically focused on using a set of patient-specific information to predict a medical outcome or to help support doctors make a clinical diagnosis [1]. Several industrial partners have begun to help develop this field. One example is IBM's WatsonPath project in conjunction with the Cleveland Clinic (http://www.research.ibm.com/cognitive-computing/watson/). However, as opposed to other medical applications, we consider a challenging problem in which the patient specific information is unclear and open to interpretation, making previous methods unsuitable.

In this paper we explore how data mining can be applied to gastroenterology, and specifically to aid in the diagnosis of patients with high-risk lesions within Barrett's oesophagus (BE). BE is the only identifiable premalignant lesion for oesophageal adenocarcinoma (OA), a tumor whose incidence has been rising rapidly in the Western World [2]. Finding abnormalities in the oesophagus, known as dysplasia or early-stage cancer, thorough endoscopic surveillance is critical as OA can more easily be cured without invasive treatment at this stage. Current surveillance strategies typically rely on white-light endoscopy (WLE) to obtain four-quadrant biopsies through every 2cm of the Barrett's segment [5]. Unfortunately this approach samples less than 5% of the Barretts epithelium and is therefore likely to miss the dysplasia representing early-stage cancer [4]. A novel endoscopic image enhancement technology, i-Scan (PENTAX), has been developed to help overcome these shortcomings. i-Scan utilizes post processing light filter technology to help provide real-time analysis and enhancement of different elements of the mucosa to help improve dysplasia detection. We focus on creating decision trees from data mining patients' i-Scan videos and multiple experts' interpretation of these videos.

This paper makes two key contributions. First, as patient information is open to interpretation, we demonstrate that composite rules learned from multiple experts can be more accurate than that of one expert alone. Even expert doctors interpret endoscopy scans differently, potentially making it important to aggregate multiple opinions. Second, we demonstrate that decision trees can generate simple rules for dysplasia diagnosis. These rules can either be used to encapsulate the rules of the most accurate expert for training purposes or to help identify diagnostic errors.

II. Methods

Forty-seven High Definition (HD) videos recordings were collected from patients with non-dysplastic (ND-BE) and dysplastic (D-BE) BE undergoing endoscopy at University College London Hospital. A strict protocol was used to record areas of interest (lesions) after which a corresponding biopsy was taken to confirm the histological diagnosis. In a blinded manner, videos were shown to 3 expert endoscopists who were asked to interpret them based on their mucosal (M) and vascular patterns (V), presence of nodularity (NOD) and ulceration (ULC) as well as overall suspected diagnosis. The M, V, NOD attributes were rated as either being normal or abnormal (including cases of suspected abnormality), and the experts rated the video quality as either being excellent, satisfactory or poor, and self-rated their certainty on a scale of 1 to 6. Acetic acid (ACA) chromoendoscopy was added in thirteen of the 47 lesions, creating a total of 60 videos within the dataset. Of the 47 videos, 23 were later found to be dysplastic and 24 were non-dysplastic. Within the cases in which ACA was used, 7 had ND-BE and 6 D-BE.

The goal of the study was to determine if standard data mining techniques could provide insight into the medical diagnosis of dysplasia. To do so, we used the Weka data mining package to construct C4.5 decision trees [3], [6] based on all data inputs. The advantage of using decision trees did not necessarily lie within the accuracy of this algorithm, as we found that other algorithms did yield similar results. Instead, the advantage in using decision trees lies in their ability to output the exact if-then rules behind the model. This in turn allowed us to analyze the outputted rules; something we found was useful in the diagnosis of these lesions.

III. Results

Table 1 summarizes the study's results. Columns 1–3 present the experts' recall of no dysplasia (row 1), dysplasia (row 2) and overall accuracy (row 3). Note that all three experts easily found most cases of non-dysplasia, but had more difficult in finding the important pre-cancerous

dysplasia cases. In fact, in 6 dysplasia cases (26%) all 3 experts failed to detect D-BE, thus signifying the inherent complexity of the task. The addition of ACA had no net effect in diagnosing dysphasia. For two experts ACA did not change any of their diagnoses, and in a third expert ACA increased the bias to diagnose dysplasia in two lesions. However, even within these two cases the results were mixed with the expert reversing a correct diagnosis of no dysplasia once and only once correctly finding a new case of actual dysplasia. As we found no significant difference between using ACA or not, we then considered how data mining could be used on the dataset without ACA.

	E-1	E-2	E-3	Best Rule	Best E-3
No Dysplasia	0.83	0.8	0.91	1	0.96
Dysplasia	0.7	0.7	0.39	0.65	0.52
Overall Accuracy	77%	75%	66%	83%	74%

As experts interpreted the videos differently, we posited that composite models in which the different experts' opinions are aggregated may be helpful. Within this model we noted the total number of experts whom reported that a lesion had abnormal mucosal line, vascular structure etc. We then considered two different models, one that explicitly included the expert opinions (dysplasia non-dysplasia) and one that does not. The model including the expert opinion was the most accurate model (79%) with the following rule:

Abnormal-Vascular $<= 1$
| DYSPLASIA $<= 1$: NO DYSPLASIA
| DYSPLASIA > 1: DYSPLASIA
Abnormal-Vascular > 1: DYSPLASIA

According to this rule the most important attribute (the root of the tree) is the lesion's vascular pattern. If the vascular pattern is found to be abnormal by more than one doctor (root of the tree), then the lesion is dysplastic. Otherwise, if the doctors still think it is dysplastic it is otherwise it is not. Somewhat surprisingly, taking the experts' input out of the tree produces the following simplified rule:

Abnormal-Vascular $<= 1$: NO DYSPLASIA
Abnormal-Vascular > 1: DYSPLASIA

which performs only slightly worse (77% accurate) and still equal to the accuracy of the most accurate expert. However, differences still exist in the recall of the dysplastic cases. While the decision tree finds only 13 of the 23 cases, the most accurate expert was able to find 17 of the 23 cases. Thus, while this rule is mathematically is equally accurate, it does leave significant room for understanding what led the expert to find the dysplasia in the other cases.

These above models assume that all experts' opinions should be considered equally. Note from Table 1 that at times experts differ in their opinion and accuracy, and thus it may be useful to consider what rules the most accurate expert used so that other people can learn from their example. Using decision trees allows us to consider such rules without bias. The C4.5 algorithm found that the first expert yielded the most accurate decision tree with the following rule (column 4 of Table 1):

NOD $=$ Definite: DYSPLASIA
NOD $=$ Probable
| Normal-Vascular $<= 0$: DYSPLASIA
| Normal-Vascular > 0: NO DYSPLASIA
NOD $=$ None: NO DYSPLASIA

Last, we found that decision trees could help correct a person's decision process. Note that Expert 3 was accurate 66% of the time. While an outside observer likely does not know the internal thought process of that expert, decision trees were able to construct the simple rule presented below that was much more accurate (74%) than expert's diagnosis (column 5 of Table 1). Thus, we posit that using decision trees in this fashion could potentially aid this doctor to better diagnose dysplasia.

NOD = None: NO DYSPLASIA
NOD = Probable/Definite: DYSPLASIA

IV. Conclusion and Future Work

Experts are able to endoscopically diagnose D-BE correctly in up to three-quarters of cases using i-Scan enhanced imaging. We demonstrate the decision trees can be useful in helping doctors understand the rules by which dysplasia can be found. These rules, which focus on irregular vascularity and nodules, predict dysplasia with a similar level of accuracy and are easier to learn than conventional classification systems. They could be used to train non-expert endoscopists in dysplasia detection, or to help even experts better their diagnosis. For future work, we hope to study how to a-priori identify which doctors are the most accurate overall and / or for which specific types of dysplasia. We believe this direction holds great promise in this and other domains where diagnoses are dependent on subjectively observed patient information.

References

[1] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: Current issues and guidelines. *I. J. Medical Informatics*, 77(2):81–97, 2008.

[2] Elfriede Bollschweiler, Eva Wolfgarten, Christian Gutschow, and Arnulf H. Hölscher. Demographic variations in the rising incidence of esophageal adenocarcinoma in white males. *Cancer*, 3(92):549–555, 2001.

[3] J. Ross Quinlan. *C4.5: Programs for Machine Learning – Morgan Kaufmann Series in Machine Learning*. January 1993.

[4] Prateek Sharma and Kennesh McQuaid et. al. A critical review of the diagnosis and management of barrett's esophagus: the aga chicago workshop. *Gastroenterology*, 1(127):310–330, 2004.

[5] Kenneth K. Wang and Richard E. Sampliner. Updated guidelines 2008 for the diagnosis, surveillance and therapy of barretts esophagus. *Am J Gastroenterol*, (103):788–797, 2008.

[6] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June 2005.