RESEARCH ARTICLE

Colorectal Cancer Staging Using Three Clustering Methods Based on Preoperative Clinical Findings

Saeedeh Pourahmad[1], Soudabeh Pourhashemi[2], Mohammad Mohammadianpanah[3]*

Abstract

Determination of the colorectal cancer stage is possible only after surgery based on pathology results. However, sometimes this may prove impossible. The aim of the present study was to determine colorectal cancer stage using three clustering methods based on preoperative clinical findings. All patients referred to the Colorectal Research Center of Shiraz University of Medical Sciences for colorectal cancer surgery during 2006 to 2014 were enrolled in the study. Accordingly, 117 cases participated. Three clustering algorithms were utilized including k-means, hierarchical and fuzzy c-means clustering methods. External validity measures such as sensitivity, specificity and accuracy were used for evaluation of the methods. The results revealed maximum accuracy and sensitivity values for the hierarchical and a maximum specificity value for the fuzzy c-means clustering methods. Furthermore, according to the internal validity measures for the present data set, the optimal number of clusters was two (silhouette coefficient) and the fuzzy c-means algorithm was more appropriate than the k-means clustering approach by increasing the number of clusters.

Keywords: Colorectal cancer - clinical tumor staging - cluster analysis

Asian Pac J Cancer Prev, **17 (2)**, 823-827

Introduction

Colorectal cancer is one of the most prevalent and leading causes of cancer death worldwide. This neoplasm is the third most common cancer in men and the second in women worldwide (Jemal et al., 2010). In Iran, colorectal cancer is the fifth common cancer in men and the third in women (Hoseini et al., 2014; Maeda et al., 2015). This malignancy tends to present at late stages and has poor outcome. Stage at presentation is the most important prognostic factor in patients with colorectal cancer (Jee et al., 2015; Roder et al., 2015). Currently, the American Joint Committee on Cancer (AJCC), the tumor node metastasis (TNM) staging system is commonly used for pathologic staging of colorectal cancer (Hari et al., 2013). After establishing the pathologic diagnosis, the locoregional and distant extent of the disease should be determined to provide a baseline for defining preferred therapy and prognosis. Preoperative clinical staging is usually based on the findings of physical examination and imaging studies particularly Computed Tomography (CT) scan of the abdomen and pelvis, and chest imaging (Kijima et al., 2014). Further imaging studies such as MRI, transrectal ultrasonography and PET scan may improve the accuracy of preoperative clinical staging; however, they impose an additional cost on patients (Petersen et al., 2014; Li et al., 2015).

Generally, data classification or clustering is one way to control and manage the information. Clustering is a type of classification methods in which data has been separated based on their similarities in some common characteristics. Usually, these similarities are calculated based on the distance formulas (Xu and Wunsch, 2005; Bataineh et al., 2011). No statistical assumptions are needed for data distribution in most of the clustering algorithms. Hence, they are very useful classification methods when there is no prior knowledge about the data.

The widespread use of different clustering algorithms in medical researches includes clustering of the disease risk factors (Lee, 2014), clustering the symptoms of a disease (Shahrbanian et al., 2015), gene expression data (Sarkar and Maulik, 2015), image processing (Nguyen et al.; Ryali et al., 2015), pattern recognition (Yang et al., 2015) and so on. As the fuzzy logic comes into the statistical analysis, the clustering approaches have more interesting applications in clinical studies (Belhassen and Zaidi, 2010; Bataineh et al., 2011; Bunyak et al., 2011; Clifford et al., 2011; Ekong et al., 2011; Fallahi et al., 2011; Hirsch et al., 2011; Keller et al., 2011; Pang et al., 2012; Xu et al., 2013). In fuzzy clustering algorithms, each case can belong to more than one cluster simultaneously with different possibilities. Therefore, more comprehensive

*[1]Colorectal Research Center, Faghihi Hospital, [2]Biostatistics Department, Medical School, [3]Colorectal Research Center, Shiraz University of Medical Sciences, Shiraz, Iran *For correspondence: mohpanah@gmail.com*

operations may be done by considering these different possibilities.

This study aimed to investigate a clustering system for prediction of preoperative clinical staging of colorectal cancer. Fuzzy c-means clustering and two classical algorithms including k-means and hierarchical clustering methods were applied.

Materials and Methods

Data set

This retrospective study was carried out at Colorectal Research Center of Shiraz University of Medical Sciences. One hundred and seventeen patients with histologically proved colorectal adenocarcinoma were treated and followed-up between January 2006 and January 2014 at our department. Patients with other epithelial pathologies such as squamous cell carcinoma, or non-epithelial tumors, and recurrent disease were excluded. Moreover, we excluded the patients with missing or incomplete medical records or those who lacked complete pathological reports or had received neoadjuvant therapy. Tumors were restaged according to the 7th edition of the AJCC TNM staging system (2010). All patients with non-metastatic disease were initially treated with standard curative surgery. Also, patients with resectable metastatic disease were initially treated with surgery. However, those patients with disseminated disease were staged clinically and initially treated with systemic therapy. Preliminary evaluation included comprehensive history and physical examination, colonoscopy, chest, abdominal and pelvic computed tomography (CT) scans for all primary sites and pelvic MRI and/or transrectal ultrasonography for rectal tumors. PET/CT scan was performed in selected cases (Omidvari et al., 2015).

The pretreatment information was obtained from the patients' records. We collected 25 clinical and pathological variables including the patients' characteristics (age, sex, weight and BMI), and presentations (symptoms duration, anemia, abdominal pain, colicky abdominal pain, constipation, weight loss, rectal bleeding, jaundice, nausea and vomiting), Tumor characteristics (differentiation, location, growth pattern (base colonoscopic findings, and obstruction), and the results of liver function test (Alkaline phosphatase and bilirubin level). In this study, imaging findings of CT scan, MRI, transrectal ultrasonography and PET/CT scan were not included as variable data set.

Accurate staging was defined according to postoperative pathological findings (in locoregional tumors) and imaging findings (in metastatic disease). Furthermore, metastatic disease was confirmed by biopsy in those with suspected or limited metastatic foci.

Statistical analysis

Based on data separation method, clustering techniques consist of hierarchical and non-hierarchical approaches (Xu and Wunsch, 2005). K-means clustering algorithm is one famous non-hierarchical method in which the number of clusters is known (k) at the beginning. The center of k clusters is selected randomly among the data and updated during an iterative process. Indeed, an objective function is constructed based on the Euclidian distances of all data from k centers. Then, data assignment to the clusters is performed based on minimization of the objective function. Afterwards, by an iterative process, the clusters' centers are updated and the mean value of the data in each cluster is treated as the new center. These steps will continue until no change happens in the clusters' members (Xu and Wunsch, 2005). This method requires getting the number of clusters as the input parameter at the beginning. Therefore, hierarchical clustering is recommended when there is no prior information about the data, even the number of the clusters. This method includes two techniques called agglomerative and divisive methods (Xu and Wunsch, 2005). In the agglomerative method, each case is allocated to a separate cluster. Cluster integration process continues until a certain criterion is met. In contrast, all cases are allocated to one cluster in divisive method at first. Then, separation process starts till the stop criterion occurs. In the present study, the agglomerative technique was performed with three different clustering techniques (Nearest-neighbor or Single-Linkage method; based on the smallest distance between the members of two clusters, Farthest-neighbor or Complete-Linkage method; according to the largest distance between the members of two clusters and Average-Linkage method; in terms of the mean distance among all the members of two clusters).

The other clustering approach utilized in the present study was fuzzy c-means clustering method. It is the extension of k-means clustering technique in which an extra real-valued parameter in [0, 1] is added to the objective function representing the membership degrees of each datum to k clusters. These membership degrees should be summed to one for each datum over the k clusters (Bataineh et al., 2011).

10 fold cross-validation method was used for system validation. To evaluate the methods, external validity (accuracy rate, sensitivity and specificity values) and internal validity measures (Silhouette and Dunn's partition coefficients) were utilized (Belhassen and Zaidi, 2010). For external validity measures, the values were calculated for each stage and the mean values were reported.

Results

The preoperative information of 117 patients with colorectal cancer was used in the present study. The age range of the patients was 58.3±12.9 year. More than half of them were men (54.7%). According to the pathology results of the patients after the surgery, there were 17.1 % in stage 1, 33.3 % in stage 2, 44.4% in stage 3 and 5.2% in stage 4. Table 1 shows the patients' characteristics according to their hospital records.

To start the clustering algorithms, data set was divided into two testing and training subsets (80% and 20%, respectively) and four clusters were considered (k=4). By the training set, the clusters were made and by the testing set, the methods were compared. Table 2 displays a summary of the results. For the initial parameters in the analysis, the assumed values in MATLAB software were considered as follows: 2 for the fuzzification parameter,

Table 1. Descriptive Statistics of 117 Patients with Colorectal Cancer

Quantitative variables	No. (%)	Qualitative variables	Mean ± SD
Gender		Age (year)	58.3 ± 12.9
Man	64 (54.7)		
Woman	53 (45.3)		
Anemia		BMI	24.8 ± 6.2
Yes	35 (29.9)		
No	82 (70.1)		
Abdominal pain		Weight (kg)	68.6 ± 18.5
Yes	58 (49.6)		
No	59 (50.4)		
Constipation		Height (cm)	163.4 ± 23.4
Yes	59 (50.4)		
No	58 (49.6)		
Weight loss		ALP level	229.4 ±150.6
Yes	63 (53.8)		
No	54 (46.2)		
Rectal bleeding		Total-Bilirubin level	0.8 ± 0.4
Yes	85 (72.6)		
No	32 (27.4)		
Obstruction		Presentation duration (day)	176.6 ±142.6
Yes	29 (24.8)		
No	88 (75.2)		
Napkin ring			
Yes	6 (5.1)		
No	111 (94.9)		
Tumor growth pattern			
Ulcerative	83 (70.9)		
Fungative	37 (31.6)		
Diffuse infiltrative	47 (40.2)		
The cell differentiation			
Well differentiation	83 (49.8)		
Moderately differentiation	27 (22.1)		
Poorly differentiation	9 (28.1)		
Jaundice			
Yes	10 (8.5)		
No	107 (91.5)		
Vomiting			
Yes	12 (10.3)		
No	105 (89.7)		
Abdominal cramp			
Yes	21 (17.9)		
No	96 (82.1)		
Pain with bowel movement			
Yes	18 (15.4)		
No	99 (84.6)		

Abbreviations: BMI, Body Mass Index; ALP, Alkaline phosphatase

Table 2. The results of three clustering algorithms by 10-fold cross validation method

Clustering algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
K-means	52.3	51	52
Hierarchical, (nearest-neighbor)	85.9	85	5
Fuzzy c-means	43.9	44	65

100 iterations for convergence and e-5 for stop criterion. For hierarchical clustering algorithm, the nearest-neighbor method was considered since it had more accuracy value than the others (not shown here). In addition, for fuzzy c-means clustering algorithm, the maximum membership degree of each patient was considered as the assignment criterion to the clusters. Moreover, among 5-fold, 7- fold and 10-fold cross-validation methods, 10-fold method had more accuracy than the others (not shown here). To select the training and testing subsets, the division of 80% vs. 20% led to more accurate results than 70% vs. 30%.

Discussion

Providing a system for cancer staging before the surgery is a valuable prediction in disease's therapy and inhibition (Ludwig and Weinstein, 2005). Indeed, the decision to perform urgent surgery, chemotherapy or radiation before the surgery and the type of the surgery depends on the cancer stage (Cirocco, 2000; Omidvari et al., 2013). This topic has been less attended in previous researches. Therefore, the present study is important in two aspects: prediction of cancer staging before the

surgery and comparison among three different clustering algorithms (k-means, hierarchal and fuzzy clustering algorithms). The use of fuzzy clustering in cancer staging may be interesting in the sense that the border between the stages of a cancer cannot be considered as a crisp border. In other words, there is not an exact definition for passing a patient from one to the next stage clinically. Fuzzy clustering technique calculates the possibility of belonging to all the cancer's stages for an individual. Although the maximum membership degree is used for final decision, the next maximum one can be considered for further attempts in therapy (Karemore et al., 2010).

To the best of our knowledge, there are a few studies on cancer's staging by clustering algorithms (Nguyen et al.) and no study for fuzzy clustering application in this issue. However, fuzzy clustering approach is utilized and compared with other techniques for disease diagnosis (Bunyak et al., 2011; Ekong et al., 2011; Keller et al., 2011), image classification (Belhassen and Zaidi, 2010; Fallahi et al., 2011; Pang et al., 2012; Xu et al., 2013), pattern recognition (Hirsch et al., 2011) and genome information (Clifford et al., 2011). Of these, some studies evaluated fuzzy clustering technique better than the classical clustering approaches (Bunyak et al., 2011; Fallahi et al., 2011) and some others represented more accurate results for classical methods such as hierarchal approach, similar to our results (for instance (Clifford et al., 2011)).

These three algorithms were utilized on 117 patients with colorectal cancer who underwent surgery in the present study. The results revealed that hierarchal clustering method had more accuracy in prediction. In addition, fuzzy c-means with maximum specificity and hierarchal clustering method with maximum sensitivity were specific and sensitive methods for cancer staging respectively. Furthermore, the results of Dunn's partitioning coefficient showed that fuzzy c-mean was proper than k-means clustering algorithms for more clusters (0.78, 0.63, 0.62, 0.55 and 0.49 for one to six clusters, respectively).

As a result, hierarchal clustering algorithm was a proper technique in colorectal cancer staging according to this dataset. However, there were some problems concerning the data. Due to the small number of patients in the first and forth stages (17.1%, 33.3%, 44.4% and 5.2% in the first to fourth stage respectively), the internal validity measures such as Silhouette coefficient (0.43, 0.34 and 0.33 for two to four clusters, respectively) suggested two clusters for the optimal number of clusters. In addition, in this study, some important laboratory factors such as carcinoembryonic antigen (CEA) were not available in the majority of patients' hospital records. These clinical findings may improve the cancer staging process before the surgery. Therefore, complete dataset based on the study's objects should be applied for representing a prediction system or evaluation of the methods.

Acknowledgements

This work was supported by the grant number 93-7219 from Shiraz University of Medical Sciences Research Council. This article was extracted from Soudabeh Pourhashemi's Master of Science thesis. The authors are thankful to the Research Improve¬ment Center of Shiraz University of Medical Sciences and Colorectal Research Center of Faghihi hospital for their help in data gathering. The authors would like to thank the Center for Development of Clinical Research of Nemazee Hospital and Dr. Nasrin Shokrpour for editorial assistance.

References

Bataineh K, Naji M, Saqer M (2011). A comparison study between various fuzzy clustering algorithms. *Editorial Board*, **5**, 335.

Belhassen S, Zaidi H (2010). A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET. *Medical physics*, **37**, 1309-24.

Bunyak F, Hafiane A, Palaniappan K (2011). Histopathology tissue segmentation by combining fuzzy clustering with multiphase vector level sets. In 'Software tools and algorithms for biological systems', Eds Springer, 413-24

Cirocco WC (2000). Complex decisions after local (endoscopic) resection of early rectal cancer: opening Pandora's box of staging and treatment options. *Gastrointest Endosc*, **52**, 309-10.

Clifford H, Wessely F, Pendurthi S, et al (2011). Comparison of clustering methods for investigation of genome-wide methylation array data. *Frontiers in genetics*, **2**.

Colon and Rectum (2010). In 'American joint committee on cancer, AJCC cancer staging manual', Eds Springier, New York, **143**.

Ekong V, Onibere E, Imianvan A (2011). Fuzzy cluster means system for the diagnosis of liver diseases. *Int J Computer Science Technol*, **2**, 205-9.

Fallahi A, Pooyan M, Ghanaati H, et al (2011). Uterine segmentation and volume measurement in uterine fibroid patients' MRI using fuzzy C-mean algorithm and morphological operations. *Iranian J Radiol*, **8**, 150.

Hari DM, Leung AM, Lee JH, et al (2013). AJCC Cancer Staging Manual 7th edition criteria for colon cancer: do the complex modifications improve prognostic assessment? *J Am Coll Surg*, **217**, 181-90.

Hirsch O, Bösner S, Hüllermeier E, et al (2011). Multivariate modeling to identify patterns in clinical data: the example of chest pain. *BMC Med Res Methodol*, **11**, 155.

Hoseini S, Moaddabshoar L, Hemati S, et al (2014). An Overview of Clinical and Pathological Characteristics and Survival Rate of Colorectal Cancer in Iran. *Ann Colorectal Res*, **2**, 17264.

Jee Y, Oh CM, Shin A (2015). recent decrease in colorectal cancer mortality rate is affected by birth cohort in Korea. *Asian Pac J Cancer Prev*, **16**, 3951-5.

Jemal A, Center MM, DeSantis C, et al (2010). Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol Biomarkers Prev*, **19**, 1893-907.

Karemore G, Mullick JB, Sujatha R, et al (2010). Classification of protein profiles using fuzzy clustering techniques: an application in early diagnosis of oral, cervical and ovarian cancer. *Conf Proc IEEE Eng Med Biol Soc*, **2010**, 6361-4.

Keller B, Nathan D, Wang Y, et al (2011). Adaptive multi-cluster fuzzy C-means segmentation of breast parenchymal tissue in digital mammography. In 'Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011', Eds Springer, 562-9

Kijima S, Sasaki T, Nagata K, et al (2014). Preoperative evaluation of colorectal cancer using CT colonography,

MRI, and PET/CT. World J Gastroenterol, 20, 16964-75.
Lee PH (2014). Association between adolescents' physical activity and sedentary behaviors with change in BMI and risk of type 2 diabetes. *PLoS One*, **9**, 110732.
Li L, Chen S, Wang K, et al (2015). Diagnostic Value of Endorectal Ultrasound in Preoperative Assessment of Lymph Node Involvement in Colorectal Cancer: a Meta-analysis. *Asian Pac J Cancer Prev*, **16**, 3485-91.
Ludwig JA, Weinstein JN (2005). Biomarkers in cancer staging, prognosis and treatment selection. 5, 845-56.
Maeda Y, Sadahiro S, Suzuki T, et al (2015). Significance of the mucinous component in the histopathological classification of colon cancer. Surg Today.
Nguyen HT, Jia G, Pohar KS, et al (2014). Improving Bladder Cancer Staging by using quantitative DCE-MRI with k-means clustering.
Omidvari S, Hamedi SH, Mohammadianpanah M, et al (2013). Comparison of abdominoperineal resection and low anterior resection in lower and middle rectal cancer. *J Egypt Natl Canc Inst*, **25**, 151-60.
Omidvari S, Zohourinia S, Ansari M, et al (2015). Efficacy and safety of low-dose-rate endorectal brachytherapy as a boost to neoadjuvant chemoradiation in the treatment of locally advanced distal rectal cancer: A Phase-II Clinical Trial. *Ann Coloproctol*, **31**, 123-30.
Pang Y, Li L, Hu W, et al (2012). Computerized segmentation and characterization of breast lesions in dynamic contrast-enhanced MR images using fuzzy c-means clustering and snake algorithm. *Comput Math Methods Med*, **2012**.
Petersen RK, Hess S, Alavi A, et al (2014). Clinical impact of FDG-PET/CT on colorectal cancer staging and treatment strategy. *Am J Nucl Med Mol Imaging*, **4**, 471-82.
Roder D, Karapetis CS, Wattchow D, et al (2015). Colorectal cancer treatment and survival: the experience of major public hospitals in south Australia over three decades. *Asian Pac J Cancer Prev*, **16**, 2431-40.
Ryali S, Chen T, Padmanabhan A, et al (2015). Development and validation of consensus clustering-based framework for brain segmentation using resting fMRI. *J Neurosci Methods*, **240**, 128-40.
Sarkar A, Maulik U (2015). Gene microarray data analysis using parallel point-symmetry-based clustering. *Int J Data Min Bioinform*, **11**, 277-300.
Shahrbanian S, Duquette P, Kuspinar A, et al (2015). Contribution of symptom clusters to multiple sclerosis consequences. *Quality Life Res*, **24**, 617-29.
Xu R, Wunsch D (2005). Survey of clustering algorithms. Neural Networks, *IEEE Transactions on*, **16**, 645-78.
Xu Z, Allen WM, Baucom RB, et al (2013). Texture analysis improves level set segmentation of the anterior abdominal wall. *Medical physics*, **40**, 121901.
Yang G, Raschke F, Barrick TR, et al (2015). Manifold Learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering. *Magn Reson Med*, **74**, 868-78.