ORIGINAL ARTICLE

Study on specificity of colon carcinoma-associated serum markers and establishment of SVM prediction model

Lu Li, Xuhui Ma *

Department of Oncology, Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou 450008, China

Received 27 October 2016; revised 28 December 2016; accepted 7 January 2017
Available online 26 January 2017

KEYWORDS
Colon carcinoma; Tumor marker; Logistic regression; Specificity; Bhattacharyya distance; Support vector machine

Abstract
We aimed to evaluate the specificity of 12 tumor markers related to colon carcinoma and identify the most sensitive index. Logistic regression and Bhattacharyya distance were used to evaluate the index. Then, different index combinations were used to establish a support vector machine (SVM) diagnosis model of malignant colon carcinoma. The accuracy of the model was checked. High accuracy was assumed to indicate the high specificity of the index. Through Logistic regression, three indexes, CEA, HSP60 and CA199, were screened out. Using Bhattacharyya distance, four indexes with the largest Bhattacharyya distance were screened out, including CEA, NSE, AFP, and CA724. The specificity of the combination of the above six indexes was higher than that of other combinations, so did the accuracy of the established SVM identification model. Using Logistic regression and Bhattacharyya distance for detection and establishing an SVM model based on different serum marker combinations can increase diagnostic accuracy, providing a theoretical basis for application of mathematical models in cancer diagnosis.

© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Colon carcinoma is a common type of malignant tumor of the alimentary system. In recent years, as the daily diet of many individuals has changed, the incidence and mortality associated with colon carcinoma have increased worldwide. In America, the incidence of colon carcinoma has increased dramatically, making this cancer type the third highest among common malignant tumors (Levin et al., 2008). Since the onset of colon carcinoma is insidious with ambiguous symptoms, the opportunity for early treatment is often missed, and most patients have been in middle or late stages when they are diagnosed (Onouchi et al., 2008). Therefore, the early diagnosis of colon carcinoma is particularly important for the management of this disease.

There are several diagnostic methods for detection of colorectal cancer. Firstly, fecal occult blood tests (FOBTs), a common method used to detect colon carcinoma, can be used to identify fecal occult blood which is one of the symptoms of early-stage colon carcinoma. Secondly, endoscopy performed generally by a fibrocolonoscope or electronic colonoscope, is
the most effective means to identify and diagnose colon carcinoma and is also the primary method for early-stage diagnosis with high accuracy. This method can be used to directly observe colonic lesions and perform qualitative biopsy (Young and Cole, 2007). Thirdly, tumor marker test is common method to diagnose all tumors. To date, carcinoembryonic antigen (CEA) is broadly used as a cancer marker in the clinical setting. Additionally, some carbohydrate antigens are evaluated as indexes of early-stage colon carcinoma; these antigens include CA199, CA242, and CA50. These indexes alone or their combinations are helpful for the early diagnosis of colon carcinoma in the clinical setting. Fourthly, gene-based diagnosis is also adopted for detection of colon carcinoma which is a multigenic disease involving several carcinogenic steps. The occurrence and development of colon carcinoma involve changes in multiple cancer-associated genes. Mutations in genes such as APC, KRAS, p53, and DCC can occur during the process of carcinogenesis and metastasis (Owing and Clevers, 2002). Finally, enzymes such as telomerase (TLMA) (Hauguel and Bunz, 2003) and cyclooxygenase 2 can be used as markers of colon carcinoma.

Serum tumor markers usually refer to the substances in blood produced and released by tumor tissues. Analysis of tumor markers has been broadly applied in the clinical setting; however, this method has several limitations (Kawamura, 1996). The optimal serum tumor markers are sensitive and specific. However, among dozens of serum tumor markers currently used for detection of colon carcinoma, most are not sensitive or specific (Ocin et al., 1997). Thus, analysis of the sensitivity and specificity of tumor markers is clinically meaningful. In this report, we present a combined mathematical and bioinformatics analysis of the specificity of a few common tumor markers.

2. Materials and methods

2.1. General data

Group with Colon Carcinoma: A total of 100 patients who visited Affiliated Cancer Hospital of Zhengzhou University from January 2013 to December 2013 and underwent surgery for colon carcinoma, were enrolled in this study, including 56 men and 44 women (average age: 59.0 years, range: 25–82 years). According to the World Health Organization (WHO) standards on pathological types and degrees of differentiation, there were 72 cases of colorectal tubular adenocarcinoma, 17 cases of mucinous adenocarcinoma, and 11 cases of papillary-tubular adenocarcinoma; and there were 12 cases exhibiting poorly differentiated tumors, 79 cases exhibiting moderately differentiated tumors, and nine cases exhibiting well-differentiated tumors. All patients were confirmed by operation and pathology, and imaging and surgical exploration demonstrated that no patients showed metastasis to other tissues or organs.

Control Group with Benign Tumors: Fifty patients who were admitted to Affiliated Cancer Hospital of Zhengzhou University within the same time period, including 21 men and 29 women (average age: 52.5 years, range: 32–80 years) were also enrolled in this study. There were 20 cases of colitis, 16 cases of polyposis coli, nine cases of colorectal tubular adenoma, and five cases of rectal-villous-papillary epithelioma. All diagnoses were proven by clinical analysis, endoscopy, and pathological examination.

All patients agreed to participate in the study and provided written informed consent.

2.2. Clinical examination of tumor markers

All patients were phlebotomized after fasting, and 2 mL isolated serum was cryopreserved at −80 °C, being prepared for centralized serum examination. The levels of CEA, neuron-specific enolase (NSE), heat-shock protein 60 (HSP60), CYFRA21-1, tissue plasminogen activator (TPA), alpha-feto protein (AFP), CA199, CA242, CA724, CA125, CA153, and UGT1A8 in the serum were measured by enzyme-linked immunosorbent assays and a COBAS 6000 automatic electrochemiluminescence immunoassay analyzer (Roche, Switzerland).

2.3. Index screening by Logistic regression

Taking 12 indexes as covariate and pathological diagnosis results as dependent variable, Logistic regression analysis was used to screen indexes for benign and malignant tumor differentiation.

2.4. Screening indexes using Bhattacharyya distance

Bhattacharyya distance was used to sequence and screen the indexes. Bhattacharyya distances show the upper bounds of the minimum error rate of Bayes in sample normal distributions. This method is linked to error rate, and it can theoretically gain the advantageous features of classifications but hardly obtain analytic solutions. For selection of features, multidimensional and low-dimensional data were both feasible. The definition of the Bhattacharyya distance of each index between colon carcinoma samples and normal samples is shown in Eq. (1) (Xuan et al., 2006). Larger Bhattacharyya distances were associated with better classified effects.

In this formula, \( \mu_{+} \) and \( \sigma_{+} \) are the mean and variance of colon carcinoma samples, respectively, and \( \mu_{-} \) and \( \sigma_{-} \) are the mean and variance of the sample in the control group, respectively. In this study, the calculations for the Bhattacharyya distances were carried out using MATLAB.

2.5. Accuracy validation by SVM

The specificity of indexes screened by Bhattacharyya distances was validated using SVMs, and the establishment, training, and validation of SVM models were all implemented based on the MATLAB tools program (Chang and Lin, 2011).

First, 150 patients were normalized. The malignant regions of samples were marked as 1, and the benign regions were marked as 0. Eighty out of 100 patients with malignant tumors and 40 out of 50 patients with benign tumors were chosen, yielding a matrix of \( 120 \times 12 \). The samples were input into the SVM for training. During the training, penalty parameter C and nuclear parameter \( \gamma \) were gradually optimized to achieve better results. The remaining 20 patients with malignant tumors and 10 patients with benign tumors were evaluated as the testing samples and input into the SVM network after training; the corresponding results (1 or 0) were obtained.
The accuracy could be determined by comparison with the objective.

3. Results

3.1. Results of serum content analysis

The results from tumor marker analyses for the 150 samples in the two groups are listed in Table 1. The 12 indexes were CEA, NSE, HSP60, CYFRA21-I, TPA, AFP, CA199, CA242, CA125, CA153, and UGT1A8.

3.2. Logistic regression analysis of each index

Through Logistic regression analysis, CEA, CA199 and HSP60 were finally screened out, and the corresponding P values were 0.000, 0.000 and 0.008 respectively (Table 2).

3.3. Bhattacharyya distance of each index

The Bhattacharyya distance of each index was calculated according to Eq. (1). The results are shown in Table 3. The Bhattacharyya distances of NSE, CEA, CA199, and AFP were larger, followed by those of CYFRA21-I and CA125.

3.4. Establishment of different diagnosis models by SVM

Based on the 12 indexes and the six indexes screened out by Logistic regression and Bhattacharyya distance respectively, two SVM models were established. 30 test samples were input into SVM models to perform stimulation, and results are shown in Figs. 1 and 2. Results show that the accuracy, sensitivity and specificity of the SVM model based on 12 indexes were 73.3%, 75.0% and 70.0%, and those of the SVM model based on six screened-out indexes were 90.0%, 85.0%, and 100.0%.

---

**Table 1** Analysis of 12 serum markers in the two groups (means ± standard deviations).

| Indexes groups | Colon cancer group | Control group |
|----------------|-------------------|-----------|
| CEA            | 29.31 ± 8.31 (ng/mL) | 4.28 ± 1.39 (ng/mL) |
| NSE            | 11.76 ± 2.33 (ng/mL) | 2.45 ± 1.01 (ng/mL) |
| HSP60          | 587.29 ± 477.44 (pg/mL) | 201.45 ± 120.97 (pg/mL) |
| CYFRA21-I      | 8.75 ± 2.22 (ng/mL) | 1.98 ± 1.04 (ng/mL) |
| TPA            | 0.87 ± 1.25 (U/mL) | 0.081 ± 0.54 (U/mL) |
| AFP            | 17.68 ± 5.15 (ng/mL) | 2.78 ± 0.98 (ng/mL) |
| CA199          | 52.03 ± 38.34 (U/mL) | 24.03 ± 12.22 (U/mL) |
| CA242          | 18.55 ± 10.09 (U/mL) | 5.06 ± 1.47 (U/mL) |
| CA724          | 5.87 ± 1.25 (U/mL) | 1.06 ± 0.77 (U/mL) |
| CA125          | 43.05 ± 9.73 (U/mL) | 10.31 ± 7.65 (U/mL) |
| CA153          | 21.40 ± 3.95 (U/mL) | 15.14 ± 2.83 (U/mL) |
| UGT1A8         | 8.52 ± 2.03 (ng/mL) | 34.6 ± 12.16 (ng/mL) |

---

**Table 2** Variables in Logistic regression equation.

| Step | Variable(s) entered | Coefficient | Standard Error | Wald | df | Sig. | Exp (B) |
|------|---------------------|-------------|----------------|------|----|------|---------|
| 1    | CA199               | 1.839       | 0.420          | 19.158 | 1  | 0.000 | 6.291   |
|      | Constant            | 0.024       | 0.220          | 0.012 | 1  | 0.913 | 1.024   |
| 2    | CEA                 | 1.806       | 0.450          | 16.138 | 1  | 0.000 | 6.086   |
|      | CA199               | 1.922       | 0.447          | 18.508 | 1  | 0.000 | 6.834   |
|      | Constant            | -0.640      | 0.283          | 5.102  | 1  | 0.024 | 0.527   |
| 3    | CEA, HSP60, CA199   | 1.721       | 0.462          | 13.911 | 1  | 0.000 | 5.592   |
|      | Constant            | 0.096       | 0.325          | 0.9357 | 1  | 0.332 | 0.369   |

* Variable(s) entered on step 1: CA199.

* Variable(s) entered on step 2: CEA.

* Variable(s) entered on step 3: HSP60.

**Table 3** Bhattacharyya distances of tumor markers in the two groups.

| Index      | CEA | NSE | HSP60 | CYFRA21-I | TPA | AFP | CA199 | CA242 | CA724 | CA125 | CA153 | UGT1A8 |
|------------|-----|-----|-------|-----------|-----|-----|-------|-------|-------|-------|-------|-------|
| Bhattacharyya Distance | 3.4608 | 4.2157 | 1.2176 | 2.7314 | 0.9357 | 3.2135 | 1.0877 | 1.7578 | 3.4332 | 2.4567 | 1.0739 | 2.3742 |
4. Discussion

There are many methods for analyzing the specificity of certain features. The most common method is the measurement of distance. Statistical pattern recognition states that as the distance between two categories becomes larger, the classification becomes easier, and the error rate becomes lower. Distance measure is also called class separability criteria or scatter criteria. The study of class separability in statistical pattern recognition is relatively deep. Distance is an important concept in statistical pattern recognition and is often analyzed using Euclidean distances, Mahalanobis distances, and Bhattacharyya distances (Li, 2009). Euclidean distances and Mahalanobis distances are defined in terms of space, whereas Bhattacharyya distances are defined in terms of probability. With regard to selection of features, subsets that can result in the largest classifying distance and the lowest error rate should be selected. The Bhattacharyya distance is usually applied to the feature analysis of gene expression profiles and is applicable to both multidimensional and low-dimensional data.

Serum tumor markers were screened by Logistic regression and Bhattacharyya distance, and six indexes were screened out (CEA, CA199, HSP60, NSE, CA724, AFP), which was consistent with a few previous studies. Several reports have demonstrated that CEA and CA724 are of high value in the diagnosis of colon cancer (Gebauer and Muller, 1997). A study by Wong (2006) found that the serum CEA content in patients with colon cancer increased markedly, with a positive rate of 32.26%. Some studies have also reported that when CEA is used to diagnose alimentary canal neoplasms, colorectal cancer exhibits the highest positive rate (Kim et al., 2003). Chen et al. (2008) demonstrated that the sensitivity of AFP, NSE, CEA, and CA125 is 55.8% when these indexes are combined to detect gastric cancer and colon cancer. In addition, AFP is now recognized as the tumor marker with the highest specificity in primary liver cancer. However, 30–40% of samples are negative for APF (Jia et al., 2012). A study by Dai (2008) showed that AFP is statistically meaningless in the diagnosis of colon cancer. Notably, however, nonspecific serum tumor markers are of certain clinical value for diagnosis, assessment of lesion range and degree, evaluation of surgical outcomes, and examination of metastasis and postoperative recurrence in patients with colon carcinoma (Yamamoto et al., 2001).

From the results of our study, we concluded that when 12 indexes were combined to establish an SVM model, the accuracy was 73.33%, which was not optimal. However, when six indexes were combined to establish the SVM model, the accuracy was 90%. These data indicated that if too many indexes were used, the effective indexes could be influenced by the redundant indexes, thus decreasing the accuracy. A study by Fu et al. (2012) also found that when a single index was used to detect cancer, the difference between the indexes was not statistically significant. However, when five indexes, including CEA, CA199, CA724, and others were combined to detect cancer, the specificity and sensitivity were dramatically improved. Thus, these findings demonstrated that fewer indexes do not necessarily indicate better results but may lead to instability and unreliability of the results.

\[
B_i = \frac{1}{4} \left( \frac{\mu_x - \mu_y}{\sigma_x + \sigma_y} \right)^2 + \frac{1}{2} \ln \left( \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \right)
\]

5. Conclusion

High accuracy can’t be achieved using too many tumor marker indexes. The application of Bhattacharyya distance can effectively screen out indexes with high specificity, and the combination of specific indexes can be used to establish an SVM diagnosis model with high accuracy. However, it is not necessarily good to use fewer indexes. The number of indexes should be controlled properly to avoid occasionality of the results.

Acknowledgments

The authors thank all individuals who contributed to this study by providing advice and comments. And thanks Affiliated Cancer Hospital of Zhengzhou University for providing research materials and experimental base. And this study is supported by the Open cooperation project of Henan Province, China (Grant No. 132106000064) and by the Program of research in base and cutting-edge technologies of Henan Province, China (Grant No.152300410151).

References

Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM. T. Intel. Syst. Technol. 2, 389–396.

Chen, T., Su, X.X., Quan, S., 2008. Contrastive study on serum SGF, CEA, AFP, NSE, CA125 in clinical diagnosis of malignant tumors. In: The Seventh National Conference on Laboratory Medicine of Chinese Medical Association, Chongqing, 333.

Dai, P., 2008. The significance of serum tumor marker detection in colorectal cancer (Master’s degree). Shanxi Medical University, Xi’an, Ed. 6, 5087–5090.

Gebauer, G., Muller, R.W., 1997. Tumor marker concentrations in normal and malignant tissues of colorectal cancer patients and their prognostic relevance. Anticancer Res. 17, 2939–2942.

Hauguel, T., Bunz, F., 2003. Haploinsufficiency of hTERT leads to telomere dysfunction and radiosensitivity in human cancer cells. Cancer Biol. Ther. 2, 679–684.
Jia, B.C., Luo, X.L., Liang, R., Yue, H.F., Ge, L.Y., Yuan, W.P., Shen, X.Y., 2012. Diagnostic value of serum GP73 and AFP detection in primary hepatic carcinoma. Chin. J. Cancer Prev. Treat. 19, 832–835.

Kawamura, T., 1996. Current advancement of assay of tumor markers and the perspective in future. Nihon Rinsho. 54, 1642–1648.

Kim, S.B., Fernandes, L.C., Saad, S.S., Matos, D., 2003. Assessment of the value of preoperative serum levels of CA242 and CEA in the staging and postoperative survival of colorectal adenocarcinoma patients. Int. J. Biol. Markers 18, 182–187.

Levin, B., Liebemnn, D.A., McFarland, B., Smith, R.A., Brooks, D., Andrews, K.S., Dash, C., Giardiello, F.M., Glick, S., Levin, T.R., Pickhardt, Perry., Rex, D.K., Thorson, A., Winawer, S.J. American Cancer Society Colorectal Cancer Advisory Group; US Multi-Society Task Force; American College of Radiology Colon Cancer Committee, 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Cancer J. Clin. 58, 130–160.

Li, P., 2009. Study on Features Gene Selection of Gastric Cancer Based on Gene Expression Data (Master’s thesis). Beijing University of Technology.

Ocin, Y., Okabe, H., Inui, T., Yamashiro, K., 1997. Tumor markers present and future. Rinsho Byori 45, 875–883.

Onouchi, S., Matsushita, H., Moriya, Y., Akasu, T., Fujita, S., Yamamoto, S., Hasegawa, H., Kitagawa, Y., Matsumura, Y., 2008. New method for colorectal cancer diagnosis based on SSCP analysis of DNA from exfoliated colonocytes in naturally evacuated feces. Anticancer Res. 28, 145–150.

Oving, I.M., Clevers, H.C., 2002. Molecular causes of colon cancer. Eur. J. Clin. Investig. 32, 448–457.

Wong, Z.Y., 2006. The Clinical Significance of CEA, CA19-9 and CA242 Detection of Colorectal Cancer (Master’s thesis). Wenzhou Medical University.

Xuan, G.R., Zhu, X.M., Chai, P.Q., Zhang, Z.P., Fu, D.D., Shi, Y.Q., 2006. Feature Selection Based on the Bhattacharyya Distance. In: 18th International Conference on Pattern Recognition (ICPR’06), Hongkong. 3, 1232–1235.

Yamamoto, H., Miyake, Y., Noura, S., Ogawa, M., Yasui, M., Ikenaga, M., Sekimoto, M., Monden, M., 2001. Tumor markers for colorectal Cancer. Gan To Kagaku Ryoho. 28, 1299–1305.

Young, G.P., Cole, S., 2007. New stool screening tests for colorectal cancer. Digestion 76, 26–33.