Bayesian Multiclass Classification of Gene Expression Colorectal Cancer Stages

Monika Simjanoska, Ana Madevska Bogdanova and Zaneta Popeska

Ss. Cyril and Methodious University, Faculty of Computer Sciences and Engineering,
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
m.simjanoska@gmail.com, ana.madevska.bogdanova@finki.ukim.mk,
zaneta.popeska@finki.ukim.mk

Abstract. Recent researches of Colorectal Cancer (CRC) aim to look for the answers for its occurrence in the disrupted gene expressions by examining colorectal carcinogenic and healthy tissues with different microarray technologies. In this paper, we propose a novel generative modelling of the Bayes' classification for the CRC problem in order to differentiate between colorectal cancer stages. The main contribution of this paper is the solution of the distinguishing problem between the critical CRC stages that remained unsolved in the published materials - distinguishing the stage I with stage IV, and stage II with stage III. The Bayesian classifier enabled application of the 'smoothing procedure' over the data from the third stage, which succeeded to distinguish the probabilities of the mentioned stages. This results are obtained as a continuation of our previous work, where we proposed methodologies for statistical analysis of colorectal gene expression data obtained from the two widely used platforms, Affymetrix and Illumina. Furthermore, the unveiled biomarkers from the two platforms were used in our generative approach for modelling the gene expression probability distribution and were used in the Bayes' classification system, where we performed binary classifications. This novel approach will help in producing an accurate diagnostics system and precising the actual stage of the cancer. It is of great advantage for early prognosis of the disease and appropriate treatment.

Keywords: Microarray Analysis, Machine Learning, Bayes' Theorem, Colorectal Cancer, Classification

1 Introduction

Colorectal cancer is the fourth most common cause of death from cancer worldwide. The incidence, mortality and prevalence research showed that it mostly occurs in the developed regions with a total incidence of 1,234,000 cases in 2008 [1]. Recently, the colorectal cancer (CRC) occurrence is considered to be tightly connected to the gene expression phenomena. The whole genome gene expression has been observed with different types of microarray technologies in order to detect increased or decreased gene expression levels of particular genes. Gene

expression profiling by microarrays is expected to advance the progress of personalized cancer treatment based on the molecular classification of subtypes [2].

In our previous research, we analysed colorectal gene expression from the two commonly used microarray chips, Affymetrix and Illumina. In our previous work we concluded that even though some scientists claim the two platforms produce equal outputs when examining same tissue; when considering a particular cancer, the analysis showed that both of them require different statistical approach. Therefore, we proposed methodologies for distinguishing significant genes, i.e., the biomarkers, for tissues probed with both microarrays, respectively [3]. The two biomarker sets were appropriately preprocessed for the prior distribution modelling and therefore applied in the Bayes' theorem to compute the posterior probabilities for each of the carcinogenic and the healthy class. The procedure reliability has been confirmed with the classification of new and unknown patients for the classifier, who are already diagnosed with CRC.

However, once we obtained very accurate Bayesian binary classificator, we confronted the challenge of producing Bayesian multiclass classificator capable of predicting the patient's current CRC stage. Current staging tests as: CAT scan, Magnetic Resonance Imaging, PET scan, Surgery, Complete Blood Count, etc. [4], are based either on imaging, or, blood tests and the analysis may last longer and may evoke additional stress to the patients. We believe that this type of classification is essential since the results are obtained immediately and it does not require additional microarray analysis.

The rest of the paper is organized as follows. In Section 2 we present the latest work related to the multiclass classification of the CRC stages. In Section 3 we present the methodology used to extract information from the biomarkers for the different CRC stages and the classification process itself. The experiments and the results are given in Section 4, and eventually, we present our conclusions and plans for future work in the final Section 5.

2 Related Work

In this section we briefly present some work related to the problem of gene signature revealing and the use of appropriate classifier to diagnose CRC.

Eschrich et al. [5] state that even though the Dukes' staging system, A to D, is the gold standard for predicting CRC prognosis; however, accurate classification of intermediate-stage cases, C and B, is problematic. Therefore, they propose molecular staging neural network classifier based on a core set of 43 genes that seem to have biologic significance for human CRC progression in order to discriminate good from poor prognosis patients. Another prove that stage II and III, according to the American Joint Committee on Cancer TNM staging system, are problematic for prognosis prediction is presented by Salazar et al. [6]. They present the development and validation of a gene expression signature of 18 genes that is associated with the risk of relapse in patients with stage II or III CRC. Their classifier identifies two thirds of patients with stage II colon cancer who are at sufficiently low risk of recurrence who may be safely managed without

adjuvant chemotherapy. Similarly, Donada et al. [7] examined 120 stage II colon cancer patients in order to investigate the combined role of clinical, pathological and molecular parameters to identify those stage II patients who better benefit from adjuvant therapy. Farid in his research [8], compared the unsupervised artificial neural networks (ANNs) to the histopathological TNM staging system and proved that ANNs were significantly more accurate for diagnosis and survival prediction than the TNM staging system. Frederiksen at al. [9] used a nearest neighbour classifier to classify normal, and Dukes' B and C samples with less than 20% error, whereas Dukes' A and D could not be classified correctly.

The microarray experiments from patients diagnosed with different cancer stages that are used in this paper, are also applied in different researches. Here we present part of the literature related to those sets.

Laibe et al. [10] profiled both stage II and stage III carcinomas. They realized that expression profile of stage II colon carcinomas distinguishes two patterns, one pattern very similar to that of stage III tumors, based on a 7-gene signature. The function of the discriminating genes suggests that tumors have been classified according to their putative response to adjuvant targeted or classic therapies. Tsukamoto et al. [11] performed gene expression profiling and found that the overexpression of OPG gene may be a predictive biomarker of CRC recurrence and a target for treatment of this disease. Hong et al. [12] aimed to find a metastasis-prone signature for early stage mismatch-repair proficient sporadic CRC patients for better prognosis. Their best classification model yielded a 54 gene-set with an estimated prediction accuracy of 71%. Another problem of limited discrimination for Dukes stage B and C disease is presented by Jorissen et al. [13]. They conclude that metastasis-associated gene expression changes can be used to refine traditional outcome prediction, providing a rational approach for tailoring treatments to subsets of patients. Finally, three of the five microarray data sets used in this paper, have also been used by Schlicker et al. [14]. They model the heterogeneity of CRC by defining subtypes of patients with homogeneous biological and clinical characteristics and match these subtypes to cell lines for which extensive pharmacological data is available, thus linking targeted therapies to patients most likely to respond to treatment.

3 The Methodology

In this section we present the methodology used for finding significant gene signature and its application in the Bayesian multiclass classification.

3.1 Microarray Experiments

Colorectal stages systems are designed to enable physicians to stratify patients in terms of expected predicted survival, to help select the most effective treatments, to determine prognoses, and to evaluate cancer control measures [15]. The microarray experiments we used in this paper are retrieved from Gene Expression Omnibus database [16] using the following GEO accession IDs: GSE37892,

GSE21510, GSE9348, GSE14333 and GSE35896. The experiments have been performed using the Affymetrix Human Genome U133 Plus 2.0 Array which contains 54675 probes, but the unique genes observed are 21050. All data is organized into four CRC stages [17]:

- *Stage I* - In this stage cancer has grown through the superficial lining, i.e., mucosa of the colon or rectum, but has not spread beyond the colon wall or rectum. This set contains gene expression from 137 patients.
- *Stage II* - In this stage cancer has grown into or through the wall of the colon or rectum, but has not spread to nearby lymph nodes. The set contains gene expression from 257 patients.
- *Stage III* - In this stage cancer has invaded nearby lymph nodes, but is not affecting other parts of the body yet. The set contains gene expression from 182 patients.
- *Stage IV* - In this stage cancer has spread to distant organs. This set contains gene expression from 81 patients.

In order to unveil the biomarker genes in Section 3.2, we used the microarray experiment with GEO accession ID GSE8671, where 32 carcinogenic and 32 adjacent normal tissues were probed with the same Affymetrix platform.

3.2 Biomarkers Selection

The biomarkers selection methodology consists of few steps necessary for producing reliable results. Once we have retrieved both CRC and healthy tissues data, we use the following procedure which reduces the number of genes in every step:

- **Normalization**. As a suitable normalization method we use Quantile normalization, since it makes the distribution of the gene expressions as similar as possible across all samples [18] and we are interested in the genes that show significant changes in their expression.
- **Filtering methods**. In order to remove the genes with almost ordered expression levels, we used an entropy filter which measures the amount of information, i.e., disorder about the variable.
- **Paired-sample t-test**. Considering both the carcinogenic and healthy tissues are taken from the same patients and that the whole-genome gene expression follows normal distribution [19], we used a paired-sample t-test.
- **False Discovery Rate**. This method solves the problem of false positives, i.e., the genes which are considered statistically significant when in reality there is not any difference in their expression levels.
- **Volcano Plot.** Previous methods identify different expressions in accordance with statistical significance values and do not consider biological significance. In order to display both statistically and biologically significant genes we used the volcano plot visual tool.

3.3 Modelling The Prior Distributions

The biomarkers revealed in Section 3.2 showed very high precision while diagnosing both carcinogenic and healthy patients [3]. This intrigued us to test their ability to correctly classify patients into the different cancer stages we defined in Section 3.1. In order to apply the biomarkers in the Bayes' theorem, at first we must model the prior distributions for each CRC stage distinctively. Considering the little variation in the biomarkers probability distributions among the CRC stages, we used the following preprocessing procedure:

- **Round-up threshold method**. Some genes, due to noise, are negatively expressed. One way to remove the genes with negative expression is to transform all gene expression values below some threshold cut-off value to that threshold value [20]. This method is known as Round-up threshold method. In order to avoid eventual gene accumulation at one point, and thus, sustain the prior distribution shape, we chose a whole interval instead of particular value. Therefore, we map any expression value below the threshold value of 2 into the interval [0,2].
- **Normalization**. Even though the noisy gene expression values from the experiments have been previously normalized using the Quantile Normalization, we additionally used the normalization in (1) so that the overrepresented genes will be leading factor in the histogram distribution shape. Let $S_i(j,k)$ be the current stage i, for a particular gene j and a given patient k, where $i \in \{1,2,3,4\}$, $j \in \{1,...,m\}$ and $k \in \{1,...,n\}$. The number of biomarkers is m and n is the number of patients. Then the normalized gene expression is calculated as:

$$N_i(j,k) = |\frac{S_i(j,k) - \mu}{\sigma}|, \tag{1}$$

  where μ and σ are the mathematical expectation and the standard deviation of $S_i(j, 1:k)$, respectively.
- **Smoothing method**. As discussed in Section 2, stage II and III are problematic and difficult to be correctly classified because od their similarity. In this paper we propose additional smoothing method applied only on stage III gene expression data. Hereupon, we used Moving Average smoothing method, a lowpass filter, to remove the short term fluctuations.
- **Hypothesis testing**. Once we used the previous methods our data is ready for the generative modelling of the stages' prior distributions. In order to eliminate the possibility of randomly picking up the patients whose distributions does not represent the real stage's distribution, we choose the training set according to the skewness factor, i.e., the training set consists of the patients whose floored skewness factor is most common at the particular stage. The number of patients involved in the training set is nearly $\frac{3}{4}$ from the total number of patients in each stage. Our generative model fits to four types of distributions: Normal, Lognormal, Gamma, and Extreme Value. The distribution parameters are estimated using the Maximum Likelihood Estimation

(MLE) method, with a confidence level of $\alpha = 0.01$. Then we perform the Chi-square goodness-of-fit test of the default null hypothesis that the data in the tissue (vector) comes from the particular distribution with mean and variance estimated from the MLE method, using the same significance level of $\alpha = 0.01$. Once we have obtained the probabilities from the testing for each gene distinctively, we choose the distribution whose probability is highest and we assign it to the particular gene in each of the four stages.

3.4 Multiclass Bayesian Classification

As we modelled the prior distributions of all four CRC stages, we are now able to use them in the Bayes' theorem and to calculate the posterior probability for each patient to belong to each of the four classes. Given the prior distributions we can calculate the class conditional densities, $p(\boldsymbol{x}|C_i)$, as the product of the continuous probability distributions of each gene from $\boldsymbol{x}$ distinctively:

$$p(\boldsymbol{x}|C_i) = \prod f_1 f_2 ... f_n \tag{2}$$

Since we have unequal number of patients in all four classes, considering the total number of 657 tissues, we defined the prior probabilities $P(C_i)$, to be $P(C_1) = 0.2085$, $P(C_2) = 0.3912$, $P(C_3) = 0.2770$ and $P(C_4) = 0.1233$. We consider these prior probabilities to be *test case I*. In order to assume equality in the probability of patient to be diagnosed with any of the four stages, we define *test case II*, where the prior probabilities are $P(C_1) = P(C_2) = P(C_3) = P(C_4) = 0.25$. Therefore, we calculate the posterior probability $P(C_i|\boldsymbol{x})$, as:

$$p(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i) * P(C_i)}{\sum_{1}^{4} p(\boldsymbol{x}|C_i) * P(C_i)} \tag{3}$$

The tissue $\boldsymbol{x}$ is classified according to the rule of maximizing the a posteriori probability (MAP):

$$C_i = \max p(C_i|\boldsymbol{x}) \tag{4}$$

4 Experiments and Results

In this section we present the experiments and the obtained results.

In Section 3.2 we presented the methodology for biomarkers revealing from 32 carcinogenic and 32 healthy tissues whose gene expression was measured using the Affymetrix microarray technology. Comparing the two types of tissues, 138 genes showed significant changes in their gene expressions. Since, they showed great ability in distinguishing CRC from healthy patients, we used them in this paper to test whether the same precision will be obtained when classifying different CRC stages.

Once we retrieved gene expression data from patients diagnosed with different CRC stages we excluded all genes except the 138 biomarkers. Following the

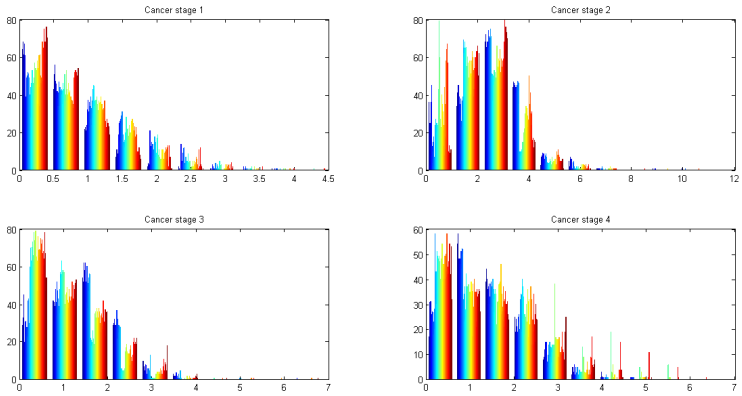

Fig. 1. The four stages after normalization

Round-up threshold method explained in Section 3.3, we handled the negative gene expression values. The results in Table 1 are from the classification of the CRC stages using the Bayesian classifier we developed in [3]. *Test Case I* and *II* refer to the prior probabilities we defined in our research [3] for both carcinogenic and healthy class. The results show that all CRC stages are classified as carcinogenic with high percent of correctness. Hereafter, our aim is to design a highly accurate Bayesian classifier with the ability to classify between CRC stages.

In order to emphasize the stages prior distributions we used the normalization method presented in (1). The results presented in Figure 1 show that stage I and stage IV have many similarities in common, as well as stage II and stage III. This is not an unexpected phenomena, since we presented some problematic classifications in Section 2. At the beginning of this research, the classification results, presented in Table 2, didn't show any problems in discriminating between stage I and stage IV; however, stage II and stage III could not be properly recognized. As a solution to this problem, we propose additional smoothing method, applied only on gene expression data from stage III. Figure 2 presents the visual changes in the distribution of stage III data.

Table 1. Bayesian Binary Classification Sensitivity

Input	**Test Case I**	**Test Case II**
Stage I	0.971	0.846
Stage II	0.969	0.876
Stage III	0.967	0.83
Stage IV	0.988	0.84

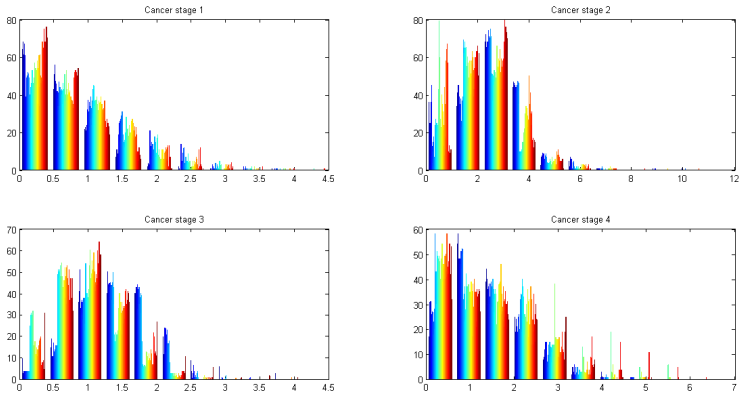

Fig. 2. The four stages after smoothing

Eventually, using carefully chosen training set of patients, we applied our generative approach for modeling the prior distributions of each class as described in Section 3.3. Applied the class conditional densities in the Bayes' theorem as defined in Section 3.4, produced the results from the Bayesian multiclass classification presented in Table 3. *Test Case I* and *II* refer to the different probabilities we defined in Section 3.4. The comparison of the percentage of correctly classified patients against the other classes is presented in Table 4.

5 Conclusion

In this paper we confronted the challenge of designing a multiclass Bayesian classifier, capable of discrimination between different colorectal cancer stages. The medical analysis shows that it is very important to discriminate between the cancer stages, in order to give the right treatment to the patient. For that purpose we used the revealed CRC biomarkers, and performed series of preprocessing procedures to produce applicable data for Bayesian classification. The results showed that Bayes' theorem can be used for problems where even details determine the class.

Table 2. Classification results before the smoothing procedure

Input/Class	**Stage I**	**Stage II**	**Stage III**	**Stage IV**
Stage I	**71.53%**	0.73%	18.25%	9.49%
Stage II	11.28%	**69.65%**	15.18%	3.89%
Stage III	26.37%	**37.36%**	**35.16%**	1.09%
Stage IV	20.99%	7.41%	8.64%	**62.96%**

Table 3. Bayesian Multiclass Classification Sensitivity

Input	Test Case I	Test Case II
Stage I	0.74	0.76
Stage II	0.54	0.51
Stage III	0.73	0.71
Stage IV	0.64	0.69

Table 4. Classification results after the smoothing procedure

Input/Class	Stage I	Stage II	Stage III	Stage IV
Stage I	**73.72%**	2.18%	13.86%	10.21%
Stage II	9.33%	**53.69%**	34.24%	2.72%
Stage III	4.94%	14.83%	**72.52%**	7.69%
Stage IV	20.98%	7.40%	7.40%	**64.19%**

The main contribution of this paper is the solution of the distinguishing problem between the critical CRC stages, that remained unsolved in the published materials of [5, 6, 9] - stage I with stage IV, and stage II with stage III. We applied a 'smoothing procedure' over the data from the third stage, which succeeded to distinguish the probabilities between the aforementioned stages. The developed Bayesian classification methodology is a result of a sound mathematical and statistical theory implementation and the produced results are reliable.

In our future work we aim to test the methodology presented in this paper on gene expression data obtained from other microarray technologies, and therefore, derive general conclusion over the multiclass classification.

References

1. GLOBOCAN: (2008), `http://globocan.iarc.fr/factsheets/cancers/colorectal.asp`
2. Jain, K.: Applications of biochips: From diagnostics to personalized medicine. Current opinion in drug disc. & develop. 7(3), 285–289 (2004)
3. Simjanoska, M., Bogdanova, A.M., Popeska, Z.: Bayesian posterior probability classification of colorectal cancer probed with affymetrix microarray technology. In: Proceedings of the 36th International Convention, MIPRO, CIS Intelligent Systems (2013)
4. NCI: Colon cancer treatment (2013), `http://www.cancer.gov/cancertopics/pdq/treatment/colon/Patient/page2`
5. Eschrich, S., Yang, I., Bloom, G., Kwong, K.Y., Boulware, D., Cantor, A., Coppola, D., Kruhøffer, M., Aaltonen, L., Orntoft, T.F., et al.: Molecular staging for survival prediction of colorectal cancer patients. Journal of Clinical Oncology 23(15), 3526–3535 (2005)
6. Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., et al.: Gene expres-

sion signature to improve prognosis prediction of stage ii and iii colorectal cancer. Journal of clinical oncology 29(1), 17–24 (2011)
7. Donada, M., Bonin, S., Barbazza, R., Pettirosso, D., Stanta, G.: Management of stage ii colon cancer-the use of molecular biomarkers for adjuvant therapy decision. BMC gastroenterology 13(1), 1–13 (2013)
8. Ahmed, F.E.: Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular cancer 4(1), 29 (2005)
9. Frederiksen, C.M., Knudsen, S., Laurberg, S., Ørntoft, T.F.: Classification of dukes' b and c colorectal cancers using expression arrays. Journal of cancer research and clinical oncology 129(5), 263–271 (2003)
10. Laibe, S., Lagarde, A., Ferrari, A., Monges, G., Birnbaum, D., Olschwang, the COL2 Project, S.: A seven-gene signature aggregates a subgroup of stage ii colon cancers with stage iii. OMICS: A Journal of Integrative Biology 16(10), 560–565 (2012)
11. Tsukamoto, S., Ishikawa, T., Iida, S., Ishiguro, M., Mogushi, K., Mizushima, H., Uetake, H., Tanaka, H., Sugihara, K.: Clinical significance of osteoprotegerin expression in human colorectal cancer. Clinical Cancer Research 17(8), 2444–2450 (2011)
12. Hong, Y., Downey, T., Eu, K.W., Koh, P.K., Cheah, P.Y.: A metastasis-pronesignature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. Clinical & experimental metastasis 27(2), 83–90 (2010)
13. Jorissen, R.N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L.A., Arango, D., Kruhøffer, M., et al.: Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer. Clinical Cancer Research 15(24), 7642–7651 (2009)
14. Schlicker, A., Beran, G., Chresta, C.M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D.A., et al.: Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC medical genomics 5(1), 66 (2012)
15. OConnell, J.B., Maggard, M.A., Ko, C.Y.: Colon cancer survival rates with the new american joint committee on cancer sixth edition staging. Journal of the National Cancer Institute 96(19), 1420–1425 (2004)
16. Gene expression omnibus (2013), `http://www.ncbi.nlm.nih.gov/geo/`
17. MayoClinic: Colon cancer (2013), `http://www.mayoclinic.com/health/colon-cancer/DS00035/DSECTION=tests-and-diagnosis`
18. Wu, Z., Aryee, M.: Subset quantile normalization using negative control features. Journal of Computational Biology 17(10), 1385–1395 (2010)
19. Hui, Y., Kang, T., Xie, L., Yuan-Yuan, L.: Digout: Viewing differential expression genes as outliers. Journal of Bioinformatics and Computational Biology 8(supp01), 161–175 (2010)
20. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences 96(6), 2907–2912 (1999)