Event Models for Tumor Classification with SAGE Gene Expression Data

Xin Jin¹, Anbang Xu¹, Guoxing Zhao^{2,3}, Jixin Ma³, and Rongfang Bie^{1,*}

¹ College of Information Science and Technology,
Beijing Normal University, Beijing 100875, P.R. China

² School of Mathematical Sciences, Beijing Normal University, Beijing 100875, P.R. China

³ School of Computing and Mathematical Science,
University of Greenwich, London SE18 6PF, U.K
xinjin796@126.com, anbangxu@mail.bnu.edu.cn,
G.Zhao@gre.ac.uk, j.ma@gre.ac.uk,
rfbie@bnu.edu.cn

Abstract. Serial Analysis of Gene Expression (SAGE) is a relatively new method for monitoring gene expression levels and is expected to contribute significantly to the progress in cancer treatment by enabling a precise and early diagnosis. A promising application of SAGE gene expression data is classification of tumors. In this paper, we build three event models (the multivariate Bernoulli model, the multinomial model and the normalized multinomial model) for SAGE data classification. Both binary classification and multicategory classification are investigated. Experiments on two SAGE datasets show that the multivariate Bernoulli model performs well with small feature sizes, but the multinomial performs better at large feature sizes, while the normalized multinomial performs well with medium feature sizes. The multinomial achieves the highest overall accuracy.

1 Introduction

Tumor classification from gene expression data is meaningful because there is high correlation between certain classes of cancers and respective gene expression levels. Traditional cancer classification methods are based on *microarray* gene expression data [9,13], which cannot tell us the expression levels of unknown genes (they may be useful for cancer prediction).

SAGE is a recently developed method which can monitor both known and unknown gene expression levels [2,5]. SAGE identifies a short mRNA tag from each individual transcript and concatenates them into long DNA molecules, which are then sequenced. By counting these tags one can estimate the gene expression levels in a cell. The information gained from performing this technique on a tissue sample is called a SAGE library.

In this study, we build three different event models for SAGE data. The multivariate Bernoulli model captures the information of which tags are found in a library, but

* Corresponding author.

not the number of times each tag appears. The multinomial model captures the information about how many times a tag appears in each library. The normalized multinomial model not only captures tag frequency but also takes into account library length.

The remainder of this paper is organized as follows: We first describe the gene selection method in Section 2. Then, in Section 3, we present the three event models for SAGE data. In Section 4 we evaluate the performance of the three models. Conclusions are presented in Section 5.

2 Feature Selection

Gene selection, or feature selection, can reduce over-fitting to the training samples and to speed up the classification process for gene expression data [3].

In our study, we use information gain to rank and select *useful* genes for SAGE data classification. Information gain is the expected entropy reduction caused by partitioning the data according to an attribute [7].

Let S be a set of SAGE samples, $E(S)$ be the entropy of S , the information gain $I(S; t)$ of a tag (or feature) t is defined as:

$$I(S, t) = E(S) - \sum_{v \in V(t)} \frac{|S_v|}{|S|} E(S_v), \quad (1)$$

where $|S|$ is the cardinality of S , S_v is the subset of S for which tag has value v , and $V(t)$ is set of all possible value of tag t .

3 Three Models for SAGE Data Classification

Suppose that SAGE data is generated by a mixture model parameterized by θ . Since the true parameters θ of the mixture model are not known, we need to estimate the parameters θ' from labeled training libraries. Taking into account that one SAGE library only belongs to one class (type of cancer), for a new library l we search for a class c_i that maximizes the posterior probability $P(c_i | l; \theta')$, by applying Bayes rule [11,12,6]. Note that $P(l | \theta')$ is the same for all classes, thus l can be classified by computing:

$$c_l = \arg \max_{c_i \in C} P(c_i | \theta') P(l | c_i; \theta') \quad (2)$$

The estimation of $P(l | c_i; \theta)$ depends on the choice of generative models which are described below.

3.1 Multivariate Bernoulli Event Model

Under the multivariate Bernoulli event model, a SAGE library $l_j \in L$ (L is the library set, $|L|$ is the number of tags appearing in the set) is generated by a series of $|L|$ Bernoulli experiments, one for each tag. The outcome of each experiment determines whether the corresponding tag will be included in the library. Thus a SAGE library is

a binary vector over the space of tags. Dimension k of the vector for library l_j is written B_{ik} , and is either 0 or 1, indicating whether tag t_k occurs in the library. We can also represent each library as $|l|+1$ variables $V_0, V_1, \dots, V_{|l|}$. V_0 takes values in $\{c_1, \dots, c_M\}$ and presents the class of the library. $V_1, \dots, V_{|l|}$ take values in $\{0, 1\}$ and represent whether the particular tags appear in the library. Fig. 1 presents this model as a graphical Markov model (or Bayesian network).

Fig. 1. The multivariate Bernoulli event model

We assume that the $|L|$ trials are independent on each other, that is, the probability of each tag occurring in a library is independent of the occurrence of other tags in a library. Thus we can estimate the probability of a library given a class from the probabilities of the tags given the class,

$$P(l_j | c_i; \theta) = \prod_{k=1}^{|L|} (B_{jk} P(t_k | c_i; \theta) + (1 - B_{jk})(1 - P(t_k | c_i; \theta))) \quad (3)$$

The multivariate Bernoulli event model does not capture the number of times each tag occurs, however, it includes the non-occurrence probability of tags that do not appear in the library but appear in the whole library set.

The parameters $\theta_{t_k | c_i} = P(t_k | c_i; \theta)$ of the mixture component c_i can be estimated as the fraction of training documents in c_i that contain t_k :

$$\theta'_{t_k | c_i} = P(t_k | c_i; \theta') = \frac{\sum_{i=1}^N B_{jk} P(c_i | l_j)}{\sum_{i=1}^N P(c_i | l_j)} \quad (4)$$

3.2 Multinomial Event Model

Multinomial approach to modeling tag frequency is to treat tags for a length $|l_j|$ (the number of tags of the library) SAGE library l_j as resulting from $|l_j|$ drawing events on a d -valued multinomial variable V . Each library can then be represented by a set of random variables $V_0, V_1, \dots, V_{|l|}$. V_0 takes values in $\{c_1, \dots, c_M\}$ and presents the class of the library. $V_1, \dots, V_{|l|}$ take values in $\{0, 1, \dots\}$ and represent the number of occurrences of particular tags in the library. Fig. 2 presents a graphical Markov model representation where a single node represents a SAGE library's vector of tag counts. The multinomial model captures tag frequency information in SAGE libraries.

Fig. 2. The multinomial event model

We assume that the lengths of libraries are independent of class. We also assume that the draws on V are independent, each tag of the library is generated independently from every other. Thus, each library l_j is drawn from a multinomial distribution of tags with as many independent trials as the length of l_j . If N_{jk} denotes the count of the number of times tag t_k occurs in library l_j . Then, the probability of a library given a class is the multinomial distribution:

$$P(l_j | c_i; \theta) = P(l_j | l) | l_j | ! \prod_k \frac{P(t_k | c_i; \theta)^{N_{jk}}}{N_{jk}!} \quad (5)$$

where $k=1, 2, \dots, T$, T is the number of tags in library l_j .

The parameters of the generative component for each class c_i are the probabilities for each tag, written $\theta_{t_k | c_i} = P(t_k | c_i; \theta)$, where $\sum_k \theta_{t_k | c_i} = 1$. It can be estimated from the training libraries:

$$\theta'_{t_k | c_i} = P(t_k | c_i; \theta') = \frac{\sum_j N_{jk} P(c_i | l_j)}{\sum_s \sum_j N_{js} P(c_i | l_j)} \quad (6)$$

where $s=1, \dots, T$, T is the total number of tags in library l_j .

3.3 Normalized Multinomial Event Model

Normalized multinomial event model for SAGE data not only capture tag frequency but also take into account library length. We normalize tag counts by transforming the tag frequencies according to

$$N'_{jk} = \frac{N_{jk}}{\sqrt{\sum_{s=1}^T (N_{js})^2}} \quad (7)$$

yielding a length 1 tag frequency vector for each library.

Normalization is common within the gene expression data clustering community because the probability of assigning a sample to a cluster is estimated by calculating distance across samples; in such a case it is not meaningful to use some of the distance functions such as the Euclidean distance to compare the libraries to each other, since their tag frequencies are not on the same scale [1,5]. For classification, however,

because comparisons are made across classes, and not across libraries, the benefit of such normalization is subtler, especially as the multinomial model accounts for length very naturally [8,10]. In experiments we find that normalized multinomial model performs better than standard multinomial for *breast* SAGE library set (classify tumor types), but the other way round for *brain* libraries (classification between tumor and normal tissue).

4 Experiments

We experiment on classification of SAGE data under three event models. We run 10 random splits with 50% training data and 50% testing data per class. The number of selected genes varies from 10 to all for each experiment. The performance results are measured by the common used *accuracy*, the percentage of correct predictions on the test sets.

4.1 Data and Preprocessing

The experiments are based on two SAGE data sets, the raw data is available on the NCBI SAGE website [4]. One problem with the raw SAGE data is that many tags in each library are expected to contain sequencing errors, and since these errors result in noise and increase the dimensionality of the data, error removal is a must need. Within one library, some tags have a frequency of 1; these unique tags are either sequencing errors or representations of very low expression level genes. In our experiment, we just remove these single frequency tags to filter out data noise.

Brain dataset: The dataset is based on 52 Hs (human sapiens) SAGE brain libraries. These libraries are made of samples from human brain and fall in to four categories: Astrocytoma (11 libraries), Ependymoma (9 libraries), Glioblastoma (8 libraries) and Medulloblastoma (24 libraries). There are 64558 tags (after noise removal) in the dataset. We used the dataset for multicategory classification experiments. We also used the biggest two categories, Astrocytoma and Medulloblastoma, for binary classification.

Breast dataset: The dataset is based on 26 Hs (human sapiens) SAGE breast libraries. These libraries are made of samples from human breast and fall in to two classes: Normal (10 libraries) and Cancer (16 libraries). There are 36087 tags (after noise removal) in the dataset. We used this dataset for binary classification experiments.

4.2 Results

Fig.3 shows the performance results of three event models for binary classification on the human *brain* dataset. The multinomial model achieves a maximum of 98.3% accuracy with the 3000 top-ranked genes; the multivariate Bernoulli model is best with 96.1% accuracy at 500 features; the normalized multinomial model reaches a maximum of 94.1% accuracy at 1000 selected genes.

Fig. 3. Three event models for *brain* SAGE libraries on different feature sizes: *binary classification* (astrocytoma or medulloblastoma)

Fig. 4. Three event models for *brain* SAGE libraries on different feature sizes: *multicategory classification*

Fig. 5. Three event models for *breast* SAGE libraries on different feature sizes: *binary classification* (tumor or normal)

Fig.4 shows the results of multicategory classification on the human *brain* libraries. The multinomial event model reaches a maximum of 90.4% accuracy when all genes are involved in the classifier. The multivariate Bernoulli model achieves a maximum of 88.5% accuracy at both 500 and 1000 features, while the normalized multinomial 85.3% at 3000 genes.

From Fig.3 and 4 we can see that for *brain* libraries, the three models do best when feature selection is used. The multivariate Bernoulli performs best with a small feature size and the multinomial performs best at a larger feature size, while normalized multinomial model do best at a moderate feature size. The multinomial model achieves higher accuracy than the multivariate Bernoulli model and the normalized multinomial model.

For classification on the human *breast* libraries, as shown in Fig.5, the best performance is achieved by the normalized multinomial event model with a maximum of 89.8% at both 1000 and 5000 features. The standard multinomial model reaches a maximum of 87.2% accuracy at 5000 features, while the multivariate Bernoulli model reaches is best with 77.0% accuracy at 100 features. Like the results for *brain* libraries, the multivariate Bernoulli performs best with a few features, the multinomial performs best with a larger feature size and normalized multinomial model do best with a moderate amount of selected genes. However, for *breast* libraries, the normalized multinomial model performs better than the multinomial and the multivariate Bernoulli model. In ongoing work we are exploring the reasons that the *breast* libraries show results different from the *brain* libraries. Note that for breast libraries the classification is done between tumor and normal samples, while for brain libraries it is between tumor types.

5 Conclusions

In this paper, we build and evaluate three event models (a multivariate Bernoulli model, a multinomial model and a normalized multinomial model) for SAGE data classification. We used information gain ranking function for gene selection.

In empirical results on publicly available SAGE libraries, we find that the multivariate Bernoulli performs well with a few features, the normalized multinomial model do well with a medium amount of selected genes and the un-normalized multinomial performs well with a larger feature size. The normalized multinomial model performs better than the multinomial for tumor-normal classification of *breast* SAGE libraries, but the multinomial is better for tumor-tumor *brain* libraries classification (both binary and multicategory). The multinomial achieves the highest overall accuracy.

Acknowledgments

This work was supported by the National Science Foundation of China under the Grant No. 60273015 and No. 10001006.

References

1. Raymond T. Ng, Jörg Sander, Monica C. Sleumer: Hierarchical Cluster Analysis of SAGE Data for Cancer Profiling. *BIOKDD*, 65-72 (2001)
2. Velculescu V. E., Zhang L., Vogelstein B., Kinzler K.W.: Serial Analysis of Gene Expression. *Science*, Vol. 270, Oct 20 484-487 (1995)
3. I. Guyon, J. Weston, S. Barnhill, V. Vapnik: Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 46(1/3) 389-422 (2002)

4. NCBI SAGE: <http://www.ncbi.nlm.nih.gov/SAGE> (2005)
5. J. Sander, R.T. Ng, M.C. Sleumer, M. Saint Yuen, S.J. Jones: A Methodology for Analyzing SAGE Libraries for Cancer Profiling. *ACM Transactions on Information Systems*, 23(1) 35-60 (2005)
6. Andrew McCallum, Kamal Nigam: A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, 41-48. AAAI Press (1998)
7. Cover, Thomas: *Elements of Information Theory*. Wiley & Sons (1991)
8. Jason D. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger: Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Twentieth International Conference on Machine Learning*. August 22 (2003)
9. Marcel Dettling: BagBoosting for Tumor Classification with Gene Expression Data. *Bioinformatics*, Vol. 20, No. 18, 3583-3593 (2004)
10. Lewis, D. D. (1998). Naive (Bayes) at forty: The Independence Assumption in Information Retrieval. *Proceedings of ECML98* (1998)
11. J. Hilden: Statistical Diagnosis Based on Conditional Independence Does not Require It. *Computational Methods in Biology and Medicine*, 14(4) 429-435 (1984)
12. J. Hellerstein, Jayram Thathachar, I. Rish: Recognizing End-user Transactions in Performance Management. In *Proceedings of AAAI-2000*, Austin, Texas, 596-602 (2000)
13. Helman P, Veroff R, Atlas, SR., Willman CL: A Bayesian Network Classification Methodology for Gene Expression Data. *Journal of Computational Biology*, 11(4) 581-615 (2004)