Data Driven Geometry for Learning

Elizabeth P. Chou(✉)

Department of Statistics, National Chengchi University, Taipei, Taiwan
eptchou@nccu.edu.tw

Abstract. High dimensional covariate information provides a detailed description of any individuals involved in a machine learning and classification problem. The inter-dependence patterns among these covariate vectors may be unknown to researchers. This fact is not well recognized in classic and modern machine learning literature; most model-based popular algorithms are implemented using some version of the dimension-reduction approach or even impose a built-in complexity penalty. This is a defensive attitude toward the high dimensionality. In contrast, an accommodating attitude can exploit such potential inter-dependence patterns embedded within the high dimensionality. In this research, we implement this latter attitude throughout by first computing the similarity between data nodes and then discovering pattern information in the form of Ultrametric tree geometry among almost all the covariate dimensions involved. We illustrate with real Microarray datasets, where we demonstrate that such dual-relationships are indeed class specific, each precisely representing the discovery of a biomarker. The whole collection of computed biomarkers constitutes a global feature-matrix, which is then shown to give rise to a very effective learning algorithm.

Keywords: Microarray · Semi-supervised learning · Data cloud geometry · biDCG

1 Introduction

Under the high dimensionality, it becomes unrealistic to build learning algorithms based on required smoothness of manifolds or distributions to typical real world datasets. After recognizing the fact of that, it is clearly essential to extract authentic data structure in a data-driven fashion. Ideally if such computed structures can be coherently embedded into a visible geometry, then the developments of learning algorithm would be realistic and right to the point of solving the real issues in hand.

Microarrays are examples of the high dimensional datasets. Microarrays provide a means of measuring thousands of gene expression levels simultaneously. Clustering genes with similar expression patterns into a group can help biologists obtain more information about gene functioning [5,10]. In addition, clustering subjects into groups by their gene expression patterns can help medical professionals determine people's clinical diagnosis status [3,4,14]. Machine learning

P. Perner (Ed.): MLDM 2015, LNAI 9166, pp. 395–402, 2015.
DOI: 10.1007/978-3-319-21024-7_27

has been discussed extensively in this setting because it can help researchers investigate medical data in a more efficient way. Therefore, many methods for classifying microarray data have been developed and reviewed by researchers [17,20,22,24].

Many studies have shown that the logistic regression approach is a fast and standardizable method for data classification [9,25]. Regardless of its extensive use, it might not be appropriate for dealing with gene expression data [19,23,26]. Since most of the microarray data are in a large p small n setting, a subset of the genes is selected through some methods and the regression prediction is performed with these genes. However, it is difficult to determine the size of the gene subset that will be chosen. If too few genes are included, the prediction error may be large. If too many genes are used, the model may be overestimated and either fail to converge or yield an unstable result. It is difficult to find a reliable method for both selecting the genes and performing logistic regression. Although logistic regression can be extended to a multi-class classification problem, a suitable method for multi-class classification with gene expression is needed [2,6,8,21].

Multicollinearity may be another problem in regression analysis on gene expression data. Since gene expression is highly correlated to the expression of other genes, the classification line that we obtain to separate the data may be unstable. Another problem may be sparseness. The regression model may not reach convergence under these conditions. When the sample size is too small, logistic regression may not provide enough power for performing the prediction.

Cross-validation is a measure for checking the performance of a predicted model. However, in such high dimensional microarray data, it may not be efficient and may yield a range of predicted results.

Two-way clustering was introduced to microarray clustering decades ago. Researchers tried to narrow down the numbers of genes and of subjects and found features for a small subset of genes and a small subset of subjects [1,13]. The two-way method overcomes the problems identified above and also decreases the noise from irrelevant data. Feature selections can improve the quality of the classification and clustering techniques in machine learning. Chen et al. [7] developed an innovative iterative re-clustering procedure biDCG through a DCG clustering method [12] to construct a global feature matrix of dual relationships between multiple gene-subgroups and cancer subtypes.

In this research, we attempt to take the accommodating attitude toward the high covariate dimensionality, and to make use computational approaches to uncover the hidden inter-dependence patterns embedded within the collection of covariate dimensions. The essential component is to include all covariate information when constructing the DCG tree geometry. It is important because the geometry pertaining to a subset of covariate data might be significantly different from the geometry pertaining to the whole. The DCG tree is better to be based all involving covariate information of labeled and unlabeled subjects. The computed pattern information would be used as the foundation for constructing learning algorithm. So that the theme of "machine learning" here is a data-driven discovery in the computational and experimental enterprize in contrasting to heavily handed statistical modeling endeavors. This data-driven discovery theme is detailed as follows.

Consider n subjects indexed by $i = 1, .., n$, and each subject is encoded with class-category number and is equipped with p-dimensional covariate information. Let a $n \times p$ matrix collectively record all covariate information available. Here we assume that an empirical distance among the n row vectors, and another empirical distance for the p column vectors are available. By using either one of empirical distances, we calculate a symmetric distance matrix. Then we apply the Data Cloud Geometry (DCG) computational algorithm, developed by Fushing and McAssey [11,12], to build an Ultrametric tree geometry $\mathcal{T}_S$ on the subject space of n p-dimensional row vectors, and another Ultrametric tree geometry $\mathcal{T}_C$ on the covarite space of p n-dimensional column vectors.

In our learning approach, we try to make simultaneous use of computed pattern information in the Ultrametric tree geometries $\mathcal{T}_S$ and $\mathcal{T}_C$. The key idea was motivated by the interesting block patterns seen by coupling these two DCG tree geometries on the $n \times p$ covariate matrix. The coupling is meant to permute the rows and columns according to the two rooted trees in such a fashion that subject-nodes and covariate-nodes sharing the core clusters are placed next to each other, while nodes belonging to different and farther apart branches are placed farther apart. This is the explicit reason why a geometry is needed in both subject and covariate spaces. Such a block pattern indicates that each cluster of subjects has a tight and close interacting relationship with a corresponding cluster of covariate dimensions. This block-based interacting relationship has been discovered and explicitly computed in [7], and termed a “dual relationship” between a target subject cluster and a target covariate cluster. Functionally speaking, this dual relationship describes the following fact: By restricting focus to a target subject cluster, the target covariate cluster can be exclusively brought out on the DCG tree as a standing branch. That is, this target covariate cluster is an entity distinct from the rest of the covariate dimensions with respect to the target subject cluster. Vice versa, by focusing only on the target covariate cluster, the target subject cluster can be brought out in the corresponding DCG tree.

Several real cancer-gene examples are analyzed here. Each cancer type turns out to be one target subject cluster. And interestingly, a cancer type has somehow formed more than one dual relationship with distinct target covariate (gene) clusters. If an identified dual relationship constitutes the discovery of a biomarker, then multiple dual relationships mean multiple biomarkers for the one cancer type. Further, the collection of dual relationships would constitute a global-feature matrix of biomarkers. A biomarker for a cancer type not only has the capability to identify such a cancer type, but at the same time it provides negative information to other cancer types that have no dual relationships with the biomarker. Therefore, a collection of dual-relation-based blocks discovered on the covariate matrix would form a global feature identification for all involved cancer types. An effective learning algorithm is constructed in this paper.

2 Method

2.1 Semi-supervised Learning

Step 1. Choosing a particular cancer type (which includes target labeled subjects and all unlabeled subjects) to cluster genes into groups.

Step 2. Classifying whole labeled and unlabeled subjects by each gene-subgroup. Finding a particular gene-subgroup that can classify the target cancer type. Repeating the procedures to whole the cancer types. These procedures yield the first dual relationship between the gene-subgroups and cancer subtypes. The cancer subtypes here may contain some unlabeled subjects within the cluster.

Step 3. Classifying genes again by a particular cancer subtype and the unknown ones that are in the same cluster as in step 2 yields the second gene-subgroups. Then, with these new gene-subgroups, classifying all subjects will yield the second dual-relationship.

Step 4. The calculation of

$$\frac{V_i V_{i'}}{||V_i|| \, ||V_{i'}||} = cos\ \theta_{ii'},\ \ i, i' = 1, .., n$$

is performed using the 2nd dual relationship to calculate. Here V_i is a vector for the unlabeled subject's data and $V_{i'}$ is a vector for the other target labeled subject's data..

Step 5. Plotting the density function of $cos\ \theta_{ii'}$ for each cancer subtype determines the classification with the function having the largest density mode.

By the method above, we can obtain clusters of the unlabeled data and labeled data. We will not lose any information from the unlabeled data. By repeating the re-clustering procedure, we can confirm that the unlabeled subjects have been correctly classified.

2.2 Datasets

We applied our learning algorithm to several datasets. The first dataset is the one from [7]. The dataset contains 20 pulmonary carcinoids (COID), 17 normal lung (NL), and 21 squamous cell lung carcinomas (SQ) cases. The second dataset was obtained from [18], containing 83 subjects with 2308 genes with 4 different cancer types: 29 cases of Ewing sarcoma (EWS), 11 cases of Burkitt lymphoma (BL), 18 cases of neuroblastoma (NB), and 25 cases of rhabdomyosarcoma (RMS). The third gene expression dataset comes from the breast cancer microarray study by [16]. The data includes information about breast cancer mutation in the BRCA1 and the BRCA2 genes. Here, we have 22 patients, 7 with BRCA1 mutations, 8 with BRCA2 mutations, and 7 with other types. The fourth gene expression dataset comes from [15]. The data contains a total of 31 malignant pleural mesothelioma (MPM) samples and 150 adenocarcinoma (ADCA) samples, with the expression of the 1626 genes for each sample. A summary of the datasets can be found in Table 1.

Table 1. Data description

Data	Number of labels	Number of subjects in each label	Dimensions
Chen	3	20 COID, 17 NL, 21 SQ	58×1543
Khan	4	29 EWS, 11 BL, 18 NB, 25 RMS	83×2308
Hedenfalk	3	7 BRCA1, 8 BRCA2, 7 others	22×3226
Gordon	2	31 MPM, 150 ADCA	181×1626

Table 2. Data description in semi-supervised setting

Data	Number of unlabeled subjects	Number of subjects in each label
Chen	15	15 COID, 12 NL, 16 SQ
Khan	20	23 EWS, 8 BL, 12 NB, 20 RMS
Hedenfalk	6	5 BRCA1, 6 BRCA2, 5 others
Gordon	20	21 MPM, 140 ADCA

Table 3. Accuracy rates for different examples - semi-supervised learning

Data set	Accuracy
Chen	15/15
Khan	1/20
Hedenfalk	4/4
Gordon	20/20

3 Results

We made some of the subjects unlabeled to perform semi-supervised learning. For the Chen dataset, we took the last 5 subjects in each group as unlabeled. For the Khan dataset, unlabeled data are the same as those mentioned in [18]. Since the sample size for Hedenfalk dataset is not large, we unlabeled only the last 2 subjects in BRCA1 and the last 2 subjects in BRCA2. We unlabeled 10 subjects for each group for the Gordon dataset. The number of labeled subjects and unlabeled subjects can be found in Table 2. The predicted results can be found in Table 3. However, we could not find the distinct dual-relationship for the second dataset.

4 Discussion

In the present study, we have proposed a semi-supervised data-driven learning rule based on the biDCG algorithm. Through our learning rule, we have efficiently classified most of the datasets with their dual relationships. In addition, we incorporated unlabeled data into the learning rule to prevent misclassification and the loss of some important information.

A large collection of covariate dimensions must have many hidden patterns embedded in it to be discovered. The model-based learning algorithm might capture the aspects allowed by the assumed models. We made use computational approaches to uncover the hidden inter-dependence patterns embedded within the collection of covariate dimensions. However, we could not find the dual relationships for one dataset, as demonstrated in the previous sections. For that dataset, we could not predict precisely. The reason is that the distance function used was not appropriate for a description of the geometry of this particular dataset. We believe that the measuring of similarity or distance for two data nodes plays an important role in capturing the data geometry. However, choosing a correct distance measure is difficult. With high dimensionality, it is impossible to make assumptions about data distributions or to get *a priori* knowledge of the data. Therefore, it is even more difficult to measure the similarity between the data. Different datasets may require different methods for measuring similarity between the nodes. A suitable selection of measuring similarity will improve the results of clustering algorithms

Another limitation is that we have to decide the smoothing bandwidth for the kernel density curves. A different smoothing bandwidth or kernel may lead to different results. Therefore, we can not make exact decisions. Besides, when the size of gene is very large, a great deal of computing time may be required.

By using the inner product as our decision rule, we know that, when two subjects are similar, the angle between the two vectors will be close to 0 and cos_θ will be close to 1. The use of cos_θ makes our decision rule easy and intuitive. The performance of the proposed method is excellent. In addition, it can solve the classification problem when we have outliers in the dual relationship.

The contributions of our studies are that the learning rules can specify gene-drug interactions or gene-disease relations in bioinformatics and can identify the clinical status of patients, leading them to early treatment. The application of this rule is not limited to microarray data. We can apply our rule of learning processes to any large dataset and find the dual-relationship to shrink the dataset's size. For example, the learning rules can also be applied to human behavior research focusing on understanding people's opinions and their interactions.

Traditional clustering methods assume that the data are independently and identically distributed. This assumption is unrealistic in real data, especially in high dimensional data. With high dimensionality, it is impossible to make assumptions about data distributions and difficult to measure the similarity between the data. We believe that measuring the similarity between the data nodes is an important way of exploring the data geometry in clustering. Also, clustering is a way to improve dimensionality reduction, and similarity research is a pre-requisite for non-linear dimensionality reduction. The relationships among clustering, similarity and dimensionality reduction should be considered in future research.

References

1. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. **96**(12), 6745–6750 (1999)
2. Bagirov, A.M., Ferguson, B., Ivkovic, S., Saunders, G., Yearwood, J.: New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. Bioinformatics **19**(14), 1800–1807 (2003)
3. Basford, K.E., McLachlan, G.J., Rathnayake, S.I.: On the classification of microarray gene-expression data. Briefings Bioinf. **14**(4), 402–410 (2013)
4. Ben-Dor, A., Bruhn, L., Laboratories, A., Friedman, N., Schummer, M., Nachman, I., Washington, U., Washington, U., Yakhini, Z.: Tissue classification with gene expression profiles. J. Comput. Biol. **7**, 559–584 (2000)
5. Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering gene expression patterns. J. Comput. Biol. **6**(3–4), 281–297 (1999)
6. Bicciato, S., Luchini, A., Di Bello, C.: PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinf. **19**(5), 571–578 (2003)
7. Chen, C.P., Fushing, H., Atwill, R., Koehl, P.: biDCG: a new method for discovering global features of dna microarray data via an iterative re-clustering procedure. PloS One **9**(7), 102445 (2014)
8. Chen, L., Yang, J., Li, J., Wang, X.: Multinomial regression with elastic net penalty and its grouping effect in gene selection. Abstr. Appl. Anal. **2014**, 1–7 (2014)
9. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inf. **35**(5–6), 352–359 (2002)
10. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95**(25), 14863–14868 (1998)
11. Fushing, H., McAssey, M.P.: Time, temperature, and data cloud geometry. Phys. Rev. E **82**(6), 061110 (2010)
12. Fushing, H., Wang, H., Vanderwaal, K., McCowan, B., Koehl, P.: Multi-scale clustering by building a robust and self correcting ultrametric topology on data points. PLoS ONE **8**(2), e56259 (2013)
13. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. USA **97**(22), 12079–12084 (2000)
14. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
15. Gordon, G.J., Jensen, R.V., Hsiao, L.L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. **62**(17), 4963–4967 (2002)
16. Hedenfalk, I.A., Ringnér, M., Trent, J.M., Borg, A.: Gene expression in inherited breast cancer. Adv. Cancer Res. **84**, 1–34 (2002)
17. Huynh-Thu, V.A., Saeys, Y., Wehenkel, L., Geurts, P.: Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics **28**(13), 1766–1774 (2012)

18. Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7**(6), 673–679 (2001)
19. Liao, J., Chin, K.V.: Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics **23**(15), 1945–1951 (2007)
20. Mahmoud, A.M., Maher, B.A., El-Horbaty, E.S.M., Salem, A.B.M.: Analysis of machine learning techniques for gene selection and classification of microarray data. In: The 6th International Conference on Information Technology (2013)
21. Nguyen, D.V., Rocke, D.M.: Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics **18**(9), 1216–1226 (2002)
22. Saber, H.B., Elloumi, M., Nadif, M.: Clustering Algorithms of Microarray Data. In: Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, pp. 557–568 (2013)
23. Shevade, S.K., Keerthi, S.S.: A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics **19**(17), 2246–2253 (2003)
24. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C.: Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics **22**(19), 2405–2412 (2006)
25. Wasson, J.H., Sox, H.C., Neff, R.K., Goldman, L.: Clinical prediction rules. Applications and methodological standards. New Engl. J. Med. **313**(13), 793–799 (1985). PMID: 3897864
26. Zhou, X., Liu, K.Y., Wong, S.T.: Cancer classification and prediction using logistic regression with bayesian gene selection. J. Biomed. Inform. **37**(4), 249–259 (2004)