Feature Selection for Brain-Computer Interface Using Nearest Neighbor Information

Yung-Kyun Noh
Department of Computer Science
KAIST
Daejeon, Korea
yungkyun.noh@gmail.com

Byoung-Kyong Min
Department of Brain and Cognitive Engineering
Korea University
Seoul, Korea
min_bk@korea.ac.kr

Abstract—**We consider the feature selection problem for a brain-computer interface (BCI). A BCI collects data from sensors, and the data are discriminated using information in a high-dimensional space. We show how relevant features in a high dimensional space can be selected using a simple nearest neighbor method for estimating an information-theoretic measure, Jensen-Shannon divergence. Conventional nonparametric estimation using nearest neighbors already works very well for the feature selection problem and outperforms many other methods. In this paper, we show how this nearest neighbor method can be further exploited by properly trimming the non-informative direction for a distance calculation, and estimate the Jensen-Shannon divergence more accurately. Through experiments with synthetic data, we show how the proposed method outperforms a conventional nearest neighbor method as well as other feature selection methods with a large margin.**

Keywords-component; feature selection; Jensen-Shannon divergence; information theory; nearest neighbor

I. INTRODUCTION

Communication using a brain-computer interface (BCI) faces many computational issues in extracting relevant information from sensor signals and in transforming the information into a code for communication. In general, the entire process depends on a data-driven analysis through learning, and many computational problems arise because of the high dimensionality of the data compared with the amount of data and restricted computational resources. Once a proper method is used for feature selection in the first stage, we can take advantages of various perspectives such as the accuracy, speed of communication, learning time, and memory.

In machine learning community, many information-theoretic approaches have been investigated for feature selection. For discriminating different signals in feature selection, it is important not to lose important information for discrimination, while other irrelevant information can be ignored. In addition, selecting redundant features is less preferred even when the features contain relevant information individually. A simple *t*-test and SNR (signal-to-noise-ratio) have previously been widely used as statistical criteria to select individual genes [1,2,3]. The top features identifed by these statistics are then concatenated together to form the relevant feature set for analysis. Unfortunately, this approach does not consider any dependency or causality between the identifed features. To obtain the set of features that are collectively the most informative, multivariate methods are necessary. The statistical criterion commonly discussed in this approach is the Jensen-Shannon divergence for two-class classification, which is defined as the mutual information between the data $\mathbf{x} \in \Re^d$ and labels $y \in \{1, 2\}$:

$$J_{JS}(X; y) = -\sum_{y=1}^{2} \int p(\mathbf{x}, y) \log \frac{p(\mathbf{x})p(y)}{p(\mathbf{x}, y)} d\mathbf{x}. \quad (1)$$

Here, $p(\mathbf{x}, y)$, $p(\mathbf{x})$, and $p(y)$ are the joint density function, the marginalized density of the data, and the probability of the labels. By selecting d number of sets which maximizes this Jensen-Shannon divergence, we can use a set of the most discriminative features from the perspective of the Shannon information.

In this paper, we explain a nearest neighbor method for estimating the Jensen-Shannon divergence, and how this method is significantly improved using generative information. By generative information, we mean a method for measuring the distance for the nearest-neighbor selection obtained from the class-conditional densities using generative models. For feature selection, we choose a set of candidate features using forward selection. After we calculate the Jensen-Shannon divergence using the nearest neighbor information along with the generative information, we choose the features one-by-one by recursively selecting additional features that increases the Jensen-Shannon divergence the most. When comparing the proposed method with a conventional method using only nearest neighbor information, as well as other well-known feature selection methods, the proposed method is shown to be superior in the selection of the relevant features. In addition, once the relevant features are selected through the proposed method, the selected features yield higher classification accuracies.

The rest of this paper is organized as follows. We briefly introduce our method for estimating the Jensen-Shannon divergence in Section II and our feature selection method in Section III. In Section IV, we present experiments comparing the classification accuracies and the numbers of relevant features between the proposed and other conventional methods. Finally, we provide some concluding remarks in Section V.

II. Jensen-Shannon Divergence Estimation Using Nearest Neighbor Estimation

Nearest neighbor of one datum is another datum with the shortest distance between the two. As an example, a method for estimating KL-divergence and Renyi α-divergence was proposed [6]. Similarly, we introduce a Jensen-Shannon divergence estimator using the nearest neighbors.

A. Conventional Nearest Neighbor Estimation

A nearest-neighbor estimator for a Jensen-Shannon estimator can be designed using a Monte-Carlo summation of the nearest neighbor information:

$$\widehat{J}_{JS}(X;y) = \frac{1}{N}\left(\sum_{i;\mathbf{x}_i\in C_1}\log\frac{u_1(\mathbf{x}_i)}{u(\mathbf{x}_i)} + \sum_{i;\mathbf{x}_i\in C_2}\log\frac{u_2(\mathbf{x}_i)}{u(\mathbf{x}_i)}\right) \tag{2}$$

where $u(\mathbf{x})$, $u_1(\mathbf{x})$, and $u_2(\mathbf{x})$ contain the nearest neighbor information, and each term is defined as $u(\mathbf{x}) = (N-1)l(\mathbf{x})^d$ with $l(\mathbf{x})$, the distance to the nearest neighbor of $\mathbf{x}\in\Re^d$; $u_1(\mathbf{x}) = (N_1-1)l_1(\mathbf{x})^d$ with $l_1(\mathbf{x})$, the distance to the nearest neighbor within class 1; and $u_2(\mathbf{x}) = (N_2-1)l_2(\mathbf{x})^d$ with $l_2(\mathbf{x})$, the distance to the nearest neighbor within class 2. Here, N, N_1, N_2 are the number of total data, the number of data belonging to class 1, and the number of data belonging to class 2, respectively, and C_1 and C_2 are the sets of class 1 and class 2.

The asymptotic convergence of this estimator can be proven using the convergence proofs in previous research for a KL-divergence estimator and Renyi α-divergence estimator [4, 6]; however, the proof is not presented in this paper.

B. Metric Learning Using Generative Information

The estimation method described in the previous section is not accurate in a high-dimensional estimation problem because of the inaccuracy of the distance. Instead of using the Euclidean distance, metric learning using the Mahalanobis-type distance was proposed in [5]:

$$d(\mathbf{x}_1,\mathbf{x}_2) = \sqrt{(\mathbf{x}_1-\mathbf{x}_2)^T A(\mathbf{x}_1-\mathbf{x}_2)}, \tag{3}$$

where the positive matrix A is learned using a continuous twice-differentiable generative model such as a Gaussian. We obtain the finite sampling bias of Eq. (2) up to the second order, and find the curvature dependency of this bias. We can appropriately control the curvature to minimize the bias. The details of minimizing the bias through metric control can be found in [5].

III. Feature Selection Using Jensen-Shannon Divergence

The feature selection can be performed by finding a predefined number of features having maximum Jensen-Shannon divergence. However, the combinatorial estimation is often formidable, and a forward selection method is often used. The forward selection incrementally adds new features showing the greatest increase of Jensen-Shannon divergence. For example, the ith feature is selected using the estimated divergence using previously selected $i-1$ features and new candidate features:

$$\mathbf{x}_i = \arg\max_{\mathbf{x}_k\notin\{\mathbf{x}_1,\ldots,\mathbf{x}_{i-1}\}} \widehat{J}_{JS}(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_k;y)$$

IV. Experiments

In this section, we show how the forward selection of the features can be improved using the proposed nearest neighbor Jensen-Shannon estimation with the appropriate metric control. We generated two sets of 1000-dimensional Gaussian data points using the random mean and covariance. The two sets use the same mean and covariance, but for the second set, the mean and covariance of the first thirty dimensions are slightly perturbed. When we classify two sets of data, the discriminative information is only within the first thirty dimensions. In this configuration, conventional classifiers barely succeed in capturing the discriminative information with around 1000 data points. We performed feature selection using our proposed method with the Jensen-Shannon divergence estimated using the nearest neighbor method exploiting the Gaussian generative model (JSGNN); Jensen-Shannon divergence using the nearest neighbor method (JSNN); a recently well-known feature selection method, mIMR; and an individual selection of features using the t-score.

In Fig. 1, we used four different configurations of two-class Gaussians by increasing the mean difference between the two sets while using the same covariance configuration. In a high-dimensional space, even when the mean difference is small, there is enough discriminative information that is consistently captured through the proposed feature selection. When the mean difference is small, the conventional mIMR and t-score do not choose informative features (the first 30 dimensionalities), while JSGNN and JSNN choose many informative features, and show high classification accuracies. Moreover, JSGNN outperforms JSNN in all cases with a large margin. Herein, we used twenty chosen features.

As shown in Fig. 2, JSGNN can choose informative features better than other methods. The conventional mIMR method barely captures the relevant features when the mean difference is small, despite the fact that mIMR also considers the Jensen-Shannon divergence. JSGNN consistently chooses around nine of the twenty chosen features. By adopting only generative information, the nearest neighbor estimation method has become significantly useful in the feature selection problem.

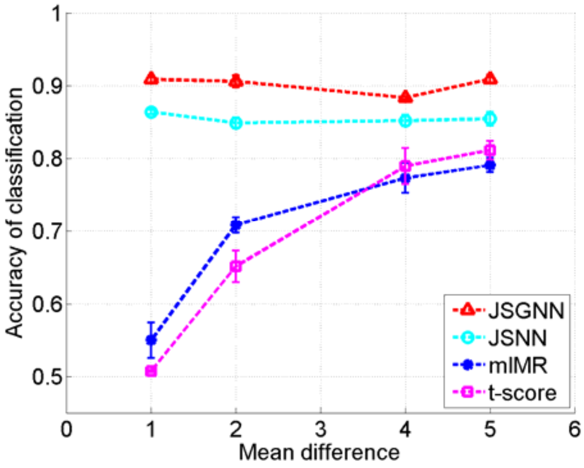

Figure 1. Accuracy of classification with selected features

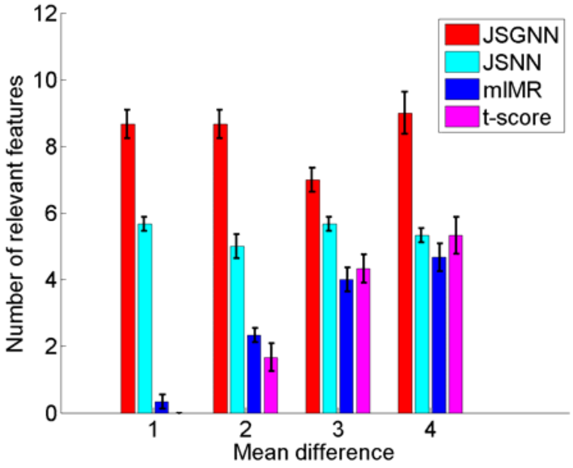

Figure 2. Number of relevant features selected

V. CONCLUSION

In this work, we introduced a novel feature selection method obtained from the nearest neighbor estimation of an information-theoretic measure, Jensen-Shannon divergence. We showed that generative information can improve the simple nearest neighbor method significantly using an appropriate metric for a reduction of the estimation bias.

An estimation of Jensen-Shannon divergence is inherently a multivariate approach considering the maximum accumulation of information excluding redundant information. However, a conventional approximation of Jensen-Shannon divergence such as mIMR does not perform well under all situations owing to an inaccurate estimation. The proposed method can be used in many BCI applications by estimating the correct Jensen-Shannon divergence.

ACKNOWLEDGMENT

This work was supported by the Basic Science Research Program (grant number 2012R1A1A1038358), which is funded by the Ministry of Education, Science, and Technology through the National Research Foundation of Korea.

REFERENCES

[1] 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., Bloomfeld, C. D., & Lander, E. S. (1999). Molecular classifcation of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, 286, 531–537.

[2] Haury, A.-C., Gestraud, P., & Vert, J.-P. (2011). The infuence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE* 6(12): e28210.

[3] Lai, C., Reinders, M., Veer, L. V., & Wessels, L. (2006). A comparison of univariate and multivariate gene selection techniques for classifcation of cancer datasets. *BMC Bioinformatics*, 7.

[4] Leonenko, N., Pronzato, L., & Savani, V. (2008). A class of R´enyi information estimators for multidimensional densities. *Annals of Statistics*, 36, 2153–2182.

[5] Noh, Y.-K., Zhang, B.-T., & Lee, D. D. (2010). Generative local metric learning for nearest neighbor classifcation. In *Advances in Neural Information Processing Systems 23*, 1822–1830.

[6] Poczos, B., & Schneider, J. (2011). On the estimation of alpha-divergences. *Proceedings of the International Conference on Artifcial Intelligence and Statistics*. 609–617.