Published online: 2020-05-07

A Method to Extract Feature Variables Contributed in Nonlinear Machine Learning Prediction

Mayumi Suzuki[1] Takuma Shibahara[1] Yoshihiro Muragaki[2]

[1] Hitachi, Ltd. Research and Development Group, Tokyo, Japan
[2] Faculty of Advanced Techno-Surgery, Institute of Advanced Biomedical Engineering and Science, Graduate School of Medicine, Department of Neurosurgery, Neurological Institute, Tokyo Women's Medical University, Tokyo, Japan

Address for correspondence Mayumi Suzuki, MEng, Hitachi, Ltd. Research and Development Group, 1-280 Higashi-Koigakubo, Kokubunji, Tokyo, Japan (e-mail: mayumi.suzuki.ov@hitachi.com).

Methods Inf Med 2020;59:1–8.

Abstract

Background Although advances in prediction accuracy have been made with new machine learning methods, such as support vector machines and deep neural networks, these methods make nonlinear machine learning models and thus lack the ability to explain the basis of their predictions. Improving their explanatory capabilities would increase the reliability of their predictions.

Objective Our objective was to develop a factor analysis technique that enables the presentation of the feature variables used in making predictions, even in nonlinear machine learning models.

Methods A factor analysis technique was consisted of two techniques: backward analysis technique and factor extraction technique. We developed a factor extraction technique extracted feature variables that was obtained from the posterior probability distribution of a machine learning model which was calculated by backward analysis technique.

Results In evaluation, using gene expression data from prostate tumor patients and healthy subjects, the prediction accuracy of a model of deep neural networks was approximately 5% better than that of a model of support vector machines. Then the rate of concordance between the feature variables extracted in an earlier report using Jensen–Shannon divergence and the ones extracted in this report using backward elimination using Hilbert–Schmidt independence criteria was 40% for the top five variables, 40% for the top 10, and 49% for the top 100.

Conclusion The results showed that models can be evaluated from different viewpoints by using different factor extraction techniques. In the future, we hope to use this technique to verify the characteristics of features extracted by factor extraction technique, and to perform clinical studies using the genes, we extracted in this experiment.

Keywords
- machine learning
- deep learning
- factor analysis
- genomics

Introduction

Recent advances in nonlinear machine learning models which are created by deep neural networks (DNNs) have enabled the generation of highly accurate predictions in the fields of image and speech recognition. Progress is also being made in the health care field, where attempts are being made to perform data analysis using nonlinear machine learning models, and to analyze diseases and the harmful effects of medicines. In general, when a nonlinear machine learning model is used, it may be possible to achieve highly accurate predictions, but it is impossible to disclose which features were used in this prediction due to the complicated structure of the model.

In earlier DNNs studies, models were validated based on the accuracy of their predictions.[1] It was therefore possible that even a model that had a high-prediction accuracy might

received
May 7, 2019
accepted after revision
December 18, 2019

DOI https://doi.org/10.1055/s-0040-1701615.
ISSN 0026-1270.

make predictions using features that were unrelated to the outcome. More reliable predictions can be obtained from models based on predictive features that are strongly related to the outcomes than from models based on unrelated predictive features.[2] In other words, when machine learning is used in situations where it predicts incidents, such as loss of life or failure of social infrastructure, people need to be able to evaluate the reliability of the prediction results.[3]

Objectives

We previously proposed a factor analysis technique to extract the feature variables used for predictions in nonlinear machine learning models.[4] With this factor analysis technique, it is possible to evaluate the reliability of predictions by validation using the meaning of the extracted features. It consisted of two techniques. First, backward analysis technique, based on Bayesian statistics, is used to calculate the posterior probability distribution for the machine learning model. Next, factor extraction technique based on Jensen–Shannon divergence (JS divergence) is applied to the posterior probability distribution calculated by backward analysis technique. The feature variables that show a large difference between outcomes are extracted and ranked.[5] A machine learning model for applying this factor analysis technique is created in advance by learning from sources, such as data aimed at distinguishing cancer patients from healthy subjects.

The selection of factors extracted as worthful feature variables in prediction depends on the factor extraction technique used. Since JS divergence evaluates the distribution of each feature variable, it is desirable to use other factor extraction techniques due to extract by other evaluation criteria. It is considered that multiple types of factor extraction technique are able to achieve feature variables extraction for use in prediction both multilaterally and comprehensively.

In this study, we aimed at improving the ability to explore the feature variables used in prediction by applying the backward elimination using Hilbert–Schmidt independence criteria (BAHSIC) as factor extraction technique, which can be used to evaluate nonlinear correlations.

Methods

Analysis Procedure

A block diagram of the analysis procedure is shown in ►Fig. 1. In this figure, data are represented by cylinder forms, processes are represented by square forms, and lists of feature variables are represented by scroll forms. First, we constructed a machine learning model to predict outcome by using analysis data. For example, a model can be constructed by training it with data on cancer and healthy subjects. When new patient data were input to this model, it predicted whether the patient had cancer. Next, we could obtain ranking list ranked in the order of importance of the feature variables used for the prediction by applying factor analysis technique to the machine learning model. As described above, the factor analysis technique consisted of two steps: backward analysis and factor extraction. In the backward analysis step, a posterior probability distribution was derived from the machine learning model applying by backward analysis technique. In the factor

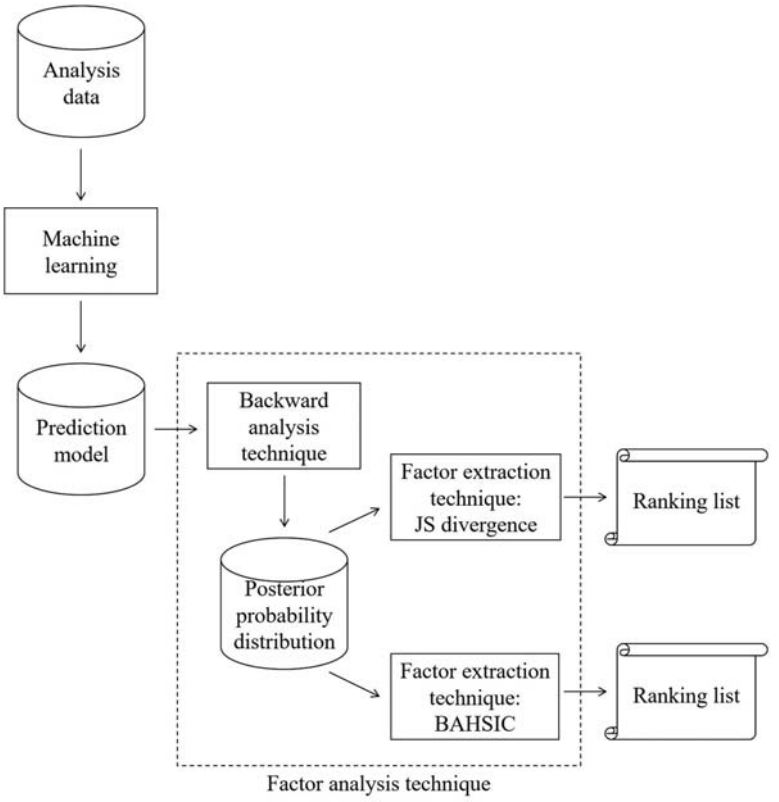

Fig. 1 Block diagram of analysis procedure. BAHSIC, backward elimination using Hilbert–Schmidt independence criteria; JS divergence, Jensen–Shannon divergence.

extraction step, a ranking list of feature variables was obtained from the posterior probability distribution applying by factor extraction technique. A different ranking list was obtained by using a different factor extraction technique. In this study, we applied BAHSIC as factor extraction technique instead of JS divergence of a previous report.

Backward Analysis Technique

By applying Replica Exchange Markov's Chain Monte Carlo Methods (RMC) to the constructed machine learning model, we could efficiently derive the factors that maximize or minimize the likelihood of the patient having cancer.[6] RMC method is a type of Expanded Ensemble Monte Carlo method and is based on the algorithm of the Metropolis–Hastings (MH) method for improving the sampling efficiency of Monte Carlo methods and Markov's chain Monte Carlo methods. RMC performs sampling based on the MH algorithm using N systems starting with different initial values, which makes it less likely that the results will be affected by the initial values. In this sampling based on the MH algorithm, the target distribution is given as a Boltzmann's distribution, and the sampled values are obtained according to the acceptance probability and a uniform distribution. Among them, $\boldsymbol{x}$ is feature variables, $f(\boldsymbol{x})$ is the machine learning model, and t is the acceptance probability. The posterior probability distribution $p(\boldsymbol{x})$ is given below:

$$p(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\frac{(f(\boldsymbol{x}) - t)^2}{T}\right) \quad (1)$$

where Z is a normalization constant, and T is thermodynamic temperature. We assumed $T = 2\sigma^2$ as the variance parameter of the distribution to intuitively control the sampling of $p(x)$.

To avoid the influence of the initial values, a certain percentage of samples is generally eliminated at the initial stage of sampling. This initial removal period is called the burn-out period. Using the RMC method, it is possible to calculate the posterior probability distribution with respect to the probability of being a cancer patient. Since the aim of this study was to analyze machine learning models, we did not evaluate, such as comparing the distributions with the analysis data.

We used Gelman–Rubin diagnostic to confirm the convergence of the results sampled using the RMC method. The closer the potential scale reduction factor (PSRF) $\widehat{\mathrm{R}}$ approaches 1.0 the better. If its value lies in the range of 1.1 to 1.05 or thereabouts, the results can be said to have converged. In this study, we evaluated two feature vectors calculated from different initial vectors up to the third decimal place.[7]

Factor Extraction Technique using Correlation Coefficients

For the posterior probability distribution obtained by backward analysis, we used BAHSIC to extract the feature variables that had a strong correlation between two quantitative variables. We can calculate the correlation coefficients between two variables by mapping data to a high-dimensional reproducing kernel Hilbert's space that maximizes the correlation coefficient using by BAHSIC. It can also be used to calculate nonlinear correlations.[8,9] If $\boldsymbol{x}$ is the feature variables, y is the outcome, $\mathcal{F}$ and $\mathcal{G}$ are the reproducing kernel Hilbert's spaces of x and y, p_{xy} is their joint probability measure and C_{xy} is their crosscovariance matrix, then the HSIC is defined as the square Hilbert–Schmidt norm of C_{xy} as shown by Eq. (2) below:

$$HSIC(p_{xy}, \mathcal{F}, \mathcal{G}) \coloneqq \|C_{xy}\|_{\mathcal{HS}}. \quad (2)$$

When Eq. (2) is zero, it indicates that $\boldsymbol{x}$ and y are independent, that is, uncorrelated. In BAHSIC, a comparison is made between the independence of all feature variables and outcomes, and the independence of a subset of feature variables and outcomes. By repeatedly deleting feature variables that do not affect independence, it is able to rank the features by independence based on the order in which they were deleted.

Experiment

Analysis Data

We used two types of analysis data, artificial data for verification of extraction accuracy of the proposed method, and clinical data for verification of extraction accuracy of real world data. First, we created the artificial data for indicating the validity of the extraction accuracy of the feature variables used for prediction by applying factor analysis technique. The artificial data were composed of feature variables that should be extracted because the probability of contributing to the prediction is high, and feature variables that should not be extracted because the probability of contributing to prediction is low. We decided to call the feature variables that should be extracted as S-features and feature variables that should not be extracted as N-features. S-features were composed of 16 feature variables whose prediction accuracy by linear support vector machine (SVM) using a linear kernel alone is 75% or more in clinical data. N-features were composed of 84 feature variables which generated by random values with an average of 0 and variance of 0.5. The total number of feature variables of the artificial data was 100. Considering that the number of available real world data samples is small in most cases, the number of samples was set to 92, same as the clinical data, 47 samples were positive cases and 45 samples were negative cases. To obtain the feature variables used for prediction explicitly, the prediction model was linear SVM. We used the ranking of the feature variables used for the prediction is defined as the descending order of the absolute value of the weight multiplied to each feature variables in linear SVM model. From bellow, we called this ranking as weighted ranking.

Next, we used clinical data on gene expression in prostate tumor patients and healthy subjects obtained from a paper by Singh et al.[10] The analysis data consisted of 1,000 feature variables from 92 samples, of whom 47 were positive cases (tumor patients) and 45 were negative cases (healthy subjects). In general, since each feature variable is expressed in

different units, there may have been cases in which there was a large difference in the absolute values of each feature variable. When constructing a machine learning model in such cases, there is a risk of falling into local minimums, where the result depends on numerical values with large values. We therefore normalized each value of feature variables for each feature variable.

Machine Learning Model

For the machine learning model, we used a DNNs and the optimal model was determined by cross validation. The search range of the hyperparameters is shown in ►Table 1. For each layer, settings, such as the types of activation function, were optimized based on random searching using the method of Bergstra and Bengio in the range shown in ►Table 1.[11] Also, drop-out were used to keep down the processing cost and suppress over fitting.[12] To evaluate the DNNs in terms of its prediction performance, we made predictions with a SVM using a polynomial kernel which was one of the nonlinear machine learning methods.

Backward Analysis Technique

The convergence of the backward analysis technique was evaluated on the basis of the value of $\widehat{R}$ in the Gelman–Rubin diagnosis. The parameters used in factor analysis technique consisted of 10 independent chains with values of σ spread uniformly over the domain $0.1 < \sigma < 0.5$, and each was sampled 100,000 times. As a burn-out period, we excluded the first 50% of samples from the data, and we adopted five independent chains of $0.25 < \sigma$ as empirical posterior probability distributions.

Ethical Considerations

The analysis data used in this experiment consisted of data on gene expression in prostate tumor patients and healthy subjects obtained from a paper by Singh et al.[10] This is an open data, and there are no special restrictions on its use.

Results

In the extraction accuracy verification result, the prediction accuracy was 100%, and the weighted ranking included 11 S-features to be extracted in the top 16 feature variables. In the extraction results using JS divergence, nine S-features were included in the top 16 feature variables and 10 feature variables which were ranked top 16 in weighted ranking were included in the top 16 feature variables. In the extraction results using BAHSIC, five S-features were included in the top 16 feature variables and six feature variables which were ranked top 16 in weighted ranking were included in the top 16 features. In the case of random extraction, the expected number of extraction is 2.6 in the top 16 ranks, and it was shown that the feature variables used for prediction can improve extraction accuracy by using proposed method.

►Table 2 shows the hyperparameters of the constructed machine learning model which is optimal DNNs model, and ►Table 3 shows the prediction accuracy of this model and SVM model. The prediction accuracy of the DNNs was approximately 5% better than that of the SVM.

The results of applying our factor analysis technique to this DNNs model are shown below. The value of $\widehat{R}$ indicates the convergence of the backward analysis technique, with a value of 1.00 indicating convergence. As a result of applying our factor analysis technique, the top five feature variables out of the feature variables used for the extracted prediction are listed in ►Table 4. The rate of concordance between the factors extracted in an earlier report using JS divergence and the factors extracted in this report using BAHSIC was 40% for

Table 1 Hyperparameter search range

Hyperparameter	Optimization range
Number of layers	4, 5
Activation function	Sigmoid, tanh, or ReLU
Number of hidden units	100 to 500
Dropout rate	0.1 to 0.5
Regularization function	L1, L2, or elastic-net

Table 2 Hyperparameter search results

Hyperparameter	Optimization results
Number of layers	4
Activation function	ReLU
Number of hidden units	143
Dropout rate	0.3
Regularization function	L1

Table 3 Prediction accuracy

	Accuracy	Precision	Recall	F-measure
DNNs	0.977	0.966	0.983	0.960
SVM	0.924	0.947	0.915	0.925

Abbreviations: DNNs, deep neural networks; SVM, support vector machine.

Table 4 Top five feature variables used for prediction using BAHSIC

Rank	Gene symbol	Gene description
1	*VDAC1*	Voltage-dependent anion channel 1
2	*CYP2C8*	Cytochrome P450 family 2 subfamily C member 8
3	*PHIP*	Pleckstrin homology domain interacting protein
4	*TRPM2*	Transient receptor potential cation channel subfamily M member 2
5	*ANGEL2*	Angel homolog 2

Abbreviation: BAHSIC, backward elimination using Hilbert-–Schmidt independence criteria.

the top five factors, 40% for the top 10 factors, and 49% for the top 100 factors.

Discussion

In the extraction accuracy verification result by artificial data, in the weighted ranking, five feature variables created by random values were included at the top 16 feature variables. Since the prediction accuracy is sufficiently high, it can be said that sufficient prediction was possible without using all 16 S-features that are known to contribute to the prediction. To extract all 16 S-features, we can use methods for selecting feature variables, such as the stepwise method before prediction.

In the extraction result by JS divergence, 10 out of 16 feature variables which used for prediction in linear SVM included in the top 16 feature variables. The extraction accuracy of factor analysis technique was 62.5%. The reason why the extraction accuracy did not reach 100% is thought that JS divergence is difficult to extract the feature variables which has small distribution even if it is able to identify the positive example and the negative example by these distribution differences. In the extraction result by BAHSIC, 7 out of 16 feature variables which used for prediction in linear SVM included in the top 16 feature variables. The extraction accuracy of factor analysis technique was 43.8%. The reason why the extraction accuracy did not reach 100% is thought that BAHSIC have strength to extract nonlinearity. There are some studies to accomplish biologically meaningful gene selection from microarray data.[13] It concluded the feature variables extracted by BAHSIC were reach high levels of accuracy and robustness when compared with other feature selection techniques, especially if strong nonlinearity are present in the data then nonlinear kernels can be more suitable. In other words, in the case of nonlinear model and the feature variables that is nonlinearly related to the outcome, BAHSIC would show its strength.

The top five feature variables identified as feature variables used for prediction were not mentioned as important in the paper from which we obtained the data. The top ranked variable, VDAC1, has been pointed out that it is a positive regulator of the exogenous pathway of apoptosis in prostate tumor cells.[14,15] The second ranked one, CYP2C8, is related to drug metabolizing enzymes present in the liver. The expression of this gene is dominant in liver tumors.[16] The third ranked one, PHIP, is a tumor-related factor that has been shown to be useful for identifying malignant melanomas.[17] The fourth ranked one, TRPM2, was known to cause the cell death responsible for conditions including Parkinson's disease, Alzheimer's disease, familial bipolar disorder, and ischemic brain disease.[18] The fifth ranked one, ANGEL2, has been pointed out that it is related to prostate tumors as evidenced by a specimen showing moderate staining in tests using prostate tumor tissue samples for human tissue immunostaining antibodies as part of the Human Protein Atlas project.[19]

Although these five top-ranked genes are not explicitly related to prostate tumors, studies have suggested a correlation between two of them and prostate tumors: CYP2C8 (rank 2) and TRPM2 (rank 4). Since CYP2C8 belongs to the same CYP family as CYP3A4, which is said to be associated with the onset of prostate tumors and leukemia, CYP2C8 may be related to prostate tumors. It has been pointed out that *TRPM2* is related to cell death, which is thought to play an important role in the regression mechanism of the prostate gland. Benign prostatic hyperplasia occurs in one of the diseases associated with prostate tumors. The prostate is an organ that grows and develops due to the effects of the male hormone testosterone. It also exhibits a strong relationship with testosterone, such as undergoing atrophy in castrated males. Testosterone is said to have no bearing on the onset of prostate tumors, but it is known to affect their malignancy. There were also similarities in the patient backgrounds. For example, like prostate tumors, benign prostatic hyperplasia is more common in men aged 60 years and over. Therefore, although there is currently no direct indication of a relationship between prostate tumors and benign prostatic hyperplasia, they may be indirectly related via testosterone. It thus seems that the top five results from the factor extraction technique using BAHSIC are associated with prostate tumors, with the exception of PHIP (rank 3).

To compare the results of previous work using JS divergence with those obtained here using BAHSIC, we list in ►Table 5 the top five genes extracted using JS divergence. The top five genes were *TIAL1*, followed by *PHIP*, *PCNA*, *VDAC1*, and *NEBL*. In previous paper, the importance of the top five feature variables was not mentioned in that paper which we obtained data from. The details of the second and fourth ranked genes are described above and are omitted here. The top ranked gene, *TIAL1*, is a cytotoxicity-related protein and is expressed in cytotoxic T lymphocytes, which have been shown to be associated with tumors.[20] The third ranked gene, *PCNA*, is a very useful factor for judging the biological malignancy and proliferation ability of tumor cells, devising treatment plans for malignant tumors, and studying the prognosis.[21] The fifth ranked gene, *NEBL*, is associated with conditions, such as dilated cardiomyopathy. Its relevance to tumors is unknown. Four out of the top five genes have been cited as having a possible link to tumors other than prostate tumors.

We focused on the reason for the extraction of the fifth ranked gene, *NEBL*, which is the most worth of special

Table 5 Top five feature variables used for prediction using JS divergence

Rank	Gene symbol	Gene description
1	*TIAL1*	TIA1 cytotoxic granule-associated RNA binding protein-like 1
2	*PHIP*	Pleckstrin homology domain interacting protein
3	*PCNA*	Proliferating cell nuclear antigen
4	*VDAC1*	Voltage-dependent anion channel 1
5	*NEBL*	Nebulette

Abbreviation: JS, Jensen–Shannon.

mention result of the extraction. Of the five top ranked genes extracted using JS divergence and BAHSIC, *NEBL* is the only one not related to the prostate gland or tumors. Dilated cardiomyopathy is a disease that especially occurs widely in men aged 60 years and over, and is suspected to be familial in nature. Prostate tumors also exhibit a remarkable increase in morbidity in men aged 60 years and over, with age and familiality being mentioned as factors of morbidity. That is, patients with dilated cardiomyopathy and prostate tumors have similar background information. It is therefore possible that most prostate tumor patients have dilated cardiomyopathy while most healthy subjects do not. In other words, the distribution of prostate tumor patients and healthy subjects might be similar to the distribution of dilated cardiomyopathy patients and healthy subjects. That is why the evaluation of distributions using JS divergence resulted in the extraction of *NEBL* as one of the feature variables having a relationship with prostate tumors.

Further, we focused on the 49% rate of concordance for the top 100 feature variables extracted using JS divergence and BAHSIC. Feature variables with distributions that differ widely were also highly likely to be extracted using BAHSIC. For example, for diseases that have high morbidity in people who are excessively underweight or overweight, there is a strong nonlinear correlation between morbidity and body weight. Also, in the distribution of feature variable values, there was a detectable difference between patients and healthy subjects. In some cases, this could result in a large difference between distributions as well as a strong correlation between them. It thus seems that there was a 49% match between the results obtained using BAHSIC and those obtained using JS divergence among the top 100 extracted results. As for the 51% that did not match, it seems that there were two types of relationships between distribution difference and correlation magnitude. The first type was extracted only when using BAHSIC, that is, feature variables having a small distribution difference and a strong nonlinear correlation. For example, in the case of diseases that have high morbidity in people aged 20 years and under or 40 to 60 years, there is a strong nonlinear correlation between morbidity and age and a small distribution difference because of the periodic distribution. Feature variables that have such distributions are thought to correspond to feature variables that have escaped noticed due to the similarity of their distributions. The second type was extracted only when using JS divergence, that is, feature variables having a large distribution difference and a weak nonlinear correlation. *NEBL* corresponds to this type. For example, for diseases that have low morbidity in people whose gene expression is excessively high or low and for those that have unknown morbidity in people whose gene expression is nonnoteworthy, there are large differences in distribution and a weak nonlinear correlation between morbidity and the gene expression level. Feature variables having such distribution are thought to correspond to feature variables that have escaped noticed due to the specificity for some patients between distributions. In other words, both JS divergence and BAHSIC are thought to capable of extracting feature variables that have escaped noticed by statistical analysis.

Here, to deepen biological considerations, gene ontology (GO) analysis is performed on the feature variables used for prediction using two types of factor extraction techniques. GO analysis is statistically able to detect GO term (annotation information defined in GO) that is densely contained in a gene list using the *p*-value of hypergeometric distribution. One of the GO analysis method is using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and a database provided by the National Institute of Allergy and Infectious Diseases (NIAID). It is possible to analyze functional trends for each cluster by classifying the given DNA into clusters for each gene group having similar functions using by Functional Annotation Clustering in DAVID. In this analysis, it is calculated Enrichment Score for each cluster which is considered that it is significantly different over 1.3.

The parameters (default settings) of Functional Annotation Clustering for the top 100 feature variables used for prediction using JS divergence and the top 100 feature variables used for prediction using BAHSIC is shown in ►Table 6. ►Table 7 shows the clustering results which enrichment score of cluster was over 1.3 or top three enrichment scores. In the clustering result of JS divergence, the enrichment score is over 1.3 in the top three clusters, it is said to be significantly included in the top 100 feature variables. In addition, *p*-value of all the annotations included in the cluster 1 are under 0.05, it seems that the cluster has particularly large significant differences. The cluster contains annotations about gene mutations, such as DNA damage, DNA repair, and DNA replication. The third ranked one identified as feature variables used for prediction, PCNA, also has this annotation. The cluster 2 contains annotations about Myb/SANT DNA → SANT/Myb domain at appropriate locations. The cluster 3 contains difficult annotations to identify functions, such as nucleus and transcription. These results show that the top 100 feature variables used for prediction using JS divergence contain many genes related to the structure and design of DNA. In other words, it can be suggested that extraction results using JS divergence contain significantly genes related to the cancerization.

Table 6 Parameters of DAVID analysis

Parameters		Settings
DAVID version		6.8
Classification stringency		Medium
Kappa similarity	Similarity term overlap	3
	Similarity threshold	0.50
Classification	Initial group membership	3
	Final group membership	3
	Multiple linkage threshold	0.50
Enrichment thresholds	EASE	1.0
Display		Benjamini

Abbreviation: DAVID, Database for Annotation, Visualization, and Integrated Discovery.

Table 7 Functional annotation clustering results which enrichment score of cluster was over 1.3 or top 3 enrichment score

(A) For the extraction results using JS divergence		
Cluster 1	**Enrichment score: 1.08**	***p*-Value**
UP_KEYWORDS	DNA repair	1.9E-3
UP_KEYWORDS	DNA damage	4.7E-3
GOTERM_BP_DIRECT	DNA repair	3.4E-2
UP_KEYWORDS	DNA replication	6.5E-2
Cluster 2	Enrichment score: 1.07	
INTERPRO	Myb domain	8.4E-4
UP_SEQ_FEATURE	DNA-binding region: H-T-H motif	6.0E-3
INTERPRO	SANT/Myb domain	1.9E-2
SMART	SANT	2.1E-2
INTERPRO	Homeodomain-like	2.3E-1
Cluster 3	Enrichment score: 0.95	
GOTERM_CC_DIRECT	Nucleoplasm	8.8E-3
UP_KEYWORDS	Nucleus	1.5E-2
GOTERM_CC_DIRECT	Nucleus	4.5E-2
UP_KEYWORDS	Transcription	1.7E-1
(B) For the extraction results using BAHSIC		
Cluster 1	Enrichment score: 1.08	p-value
KEGG_PATHWAY	Natural killer cell mediated cytotoxicity	1.0E-2
GOTERM_CC_DIRECT	Cell surface	5.9E-2
GOTERM_BP_DIRECT	Regulation of immune response	2.5E-1
Cluster 2	Enrichment score: 1.07	
GOTERM_MF_DIRECT	Potassium channel activity	1.6E-2
UP_KEYWORDS	Ion channel	8.2E-2
GOTERM_BP_DIRECT	Potassium ion transmembrane transport	1.4E-1
GOTERM_BP_DIRECT	Chemical synaptic transmission	1.4E-1
UP_KEYWORDS	Ion transport	1.7E-1
Cluster 3	Enrichment Score: 0.95	
KEGG_PATHWAY	GABAerqic synapse	9.8E-2
KEGG_PATHWAY	Morphine addition	1.1E-1
KEGG_PATHWAY	Retrograde endocannabinoid signaling	1.3E-1

Abbreviations: BAHSIC, backward elimination using Hilbert-–Schmidt independence criteria; JS, Jensen-–Shannon.

In the clustering result of BAHSIC, the enrichment score of all clusters are under 1.3. The cluster 1 contains annotations about autoimmune responses, such as natural killer cell mediated cytotoxicity is a match with ►**Table 7**. The cluster 2 contains annotations about important membrane proteins responsible for various physiological phenomena related to ions, release of neurotransmitters, secretion of hormones, such as potassium channel activity and ion channel. The cluster 3 contains annotations about synaptic transmission and morphine poisoning. These results show that the top 100 feature variables used for prediction using BAHSIC contain many genes related to the function of protecting the body when ill, such as the autoimmune system and nerve transmission. In other words, it can be suggested that extraction results using BAHSIC contain significantly genes related to the avoid cancerization. It can be considered that the top of the extraction results by applying the factor analysis technique to prostate tumor patients data contain significantly genes related to the cancer, regardless of factor extraction technique from these GO analysis results,

These results show that models can be evaluated from different viewpoints by using different factor extraction techniques. We have improved the ability to explore the feature variables used for prediction in factor analysis technique by using BAHSIC as new factor extraction technique. It is desirable to select factor analysis techniques for analysis data and aim which is based on the characteristics of factor analysis techniques. In the future, it will be necessary to perform benchmarking using a dataset to verify the extraction characteristics of each technique. In addition, prospective studies are needed to investigate the relevance of the feature variables presented in this experiment to actual clinical tumor cases.

Conclusion

We have presented new feature variables used for prediction in nonlinear machine learning models based on a factor extraction technique using BAHSIC. It was hoped that future work will include investigating methods for verifying the characteristics of feature variables extracted by the factor extraction technique, and applying the extracted feature variables to new clinical research.

Authors' Contributions

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. M.S. and T. S., in particular, contributed equally to this work.

Conflict of Interest

None declared.

Acknowledgment

This article has been translated into English with the permission of the Japan Association for Medical

Informatics (JAMI). We would like to thank Human Global Communications Co., Ltd. for native English check.

References

1 Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. Neural Information Processing Systems (NIPS). Available at: https://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf. Accessed January 8, 2019

2 Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier Published at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135–1144; Doi: https://doi.org/10.1145/2939672.2939778

3 Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Available at: http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf. Accessed January 8, 2019

4 Suzuki M, Shibahara T, Muragaki Y. Factor Analysis Technique to Extract Feature Contributed to Prediction by Machine Learning. The 37th Joint Conference on Medical Informatics (JCMI). 2017 November;37:854–856. Japan

5 Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 1991;37(01):145–151

6 Hukushima K, Nemoto K. Exchange monte carlo method and application to spin glass simulations. J Phys Soc Jpn 1996;65(06):1604–1608

7 Gelman A. John B. Carlin, Hal S. Stern, Donald B. Rubin, Bayesian data analysis. 3rd ed.Vol. 2. Boca Raton, FL: Taylor & Francis; 2014

8 Gretton A, Bousquet O, Smola A, Scholkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. International Conference on Algorithmic Learning Theory. Published at: ALT'05: Proceedings of the 16th international conference on Algorithmic Learning Theory. 2005:63–77; Doi: https://doi.org/10.1007/11564089_7

9 Le Song JB, Borgwardt KM, Gretton A, Smola A. The BAHSIC family of gene selection algorithms. Available at: https://pdfs.semanticscholar.org/7def/cd626cbd7a1bbe662b50583262e0f323ccd0.pdf. Accessed January 8, 2020

10 Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1(02):203–209

11 Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(01):281–305

12 Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. Available at: https://arxiv.org/pdf/1207.0580.pdf. Accessed Janurary 17, 2019

13 Song L, Bedo J, Borgwardt KM, Gretton A, Smola A. Gene selection via the BAHSIC family of algorithms. Bioinformatics 2007;23(13):i490–i498

14 De Pinto V, Messina A, Lane DJ, Lawen A. Voltage-dependent anion-selective channel (VDAC) in the plasma membrane. FEBS Lett 2010;584(09):1793–1799

15 Thinnes FP. Neuroendocrine differentiation of LNCaP cells suggests: VDAC in the cell membrane is involved in the extrinsic apoptotic pathway. Mol Genet Metab 2009;97(04):241–243

16 The human protein atlas. Available at: https://www.proteinatlas.org/. Accessed Janurary 17, 2019

17 De Semir D, Nosrati M, Bezrookove V, et al. Pleckstrin homology domain-interacting protein (phip) as a marker and mediator of melanoma metastasis. Proc Natl Acad Sci U S A. 2012;109(18):7067–7072

18 Xie Y-F, Macdonald JF, Jackson MF. TRPM2, calcium and neurodegenerative diseases. Int J Physiol Pathophysiol Pharmacol 2010;2(02):95–103

19 The human protein atlas: ANGEL2. Available at: https://www.proteinatlas.org/ENSG00000174606-ANGEL2/pathology. Accessed Janurary 17, 2019

20 Rimkus C, Friederichs J, Boulesteix AL, et al. Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. Clin Gastroenterol Hepatol 2008;6(01):53–61

21 Cardoso WP, Denardin OVP, Rapoport A, Araújo VC, Carvalho MB. Proliferating cell nuclear antigen expression in mucoepidermoid carcinoma of salivary glands. Sao Paulo Med J 2000;118(03):69–74