Subspace-Based Method to Improve Classification Accuracy of High-Dimensional Data

Vu-Dinh Minh and Masaomi Kimura

***Abstract*—Machine learning has increasingly attracted the attention of healthcare service providers due to its capacity to collect and analyze huge volumes of data to facilitate effective predictions and treatments. Disease data are high-dimensional data that include noise and irrelevant attributes, and the computational time and accuracy of classification algorithms used in machine learning have been a major concern. Various methods have been proposed to reduce data dimensionality by removing redundant attributes, such as step-wise backward selection and subspace clustering methods. However, removing redundant attributes may affect the accuracy of the algorithm. Based on an observation that finding hidden features, such as the relationships among attributes, can improve classification efficiency, a new hidden subspace clustering model is proposed in this paper. Experiment results show that the proposed method can reduce data dimensionality and improve the accuracy of the classification method.**

***Index Terms*—High-dimensional data, subspace clustering, random projection, classification.**

I. Introduction

Machine learning has played an integral role in the evolution of medical diagnostics. In fact, early and accurate disease prediction will improve the ability to treat patients. Therefore, over the last decade, a large number of machine learning-based methods have been proposed to support early and accurate prediction of diseases, such as cancer, diabetes, and particularly dengue fever. Dengue fever is one of the most serious infectious disease can be found in tropical and subtropical regions. In fact, there are numerous clinical symptoms caused by dengue infection, such as mild fever, and, in the worst cases, patients may die. Aedes mosquitoes, including Aedes aegypti [1] and Aedes albopictus [2], serve as the main transmission vector of dengue viruses. Due to its simple transformation, dengue has become a public health problem in tropical regions. Currently, the World Health Organization has set a goal to reduce global dengue mortality by 50% by 2020 [3]. With dengue fever, patients high recovery chance if they are identified early, i.e., within the first three days of infection. However, dengue fever is difficult to recognize because the symptoms are similar to other diseases such as roseola or fever virus.

In addition to traditional methods, machine learning techniques can improve the detection and treatment of dengue fever. For example, Tanner *et al*. [4] used the data of 1200 patients within 72 hours of fever onset to distinguish patients with dengue illness using a decision tree algorithm. Another study [5] conducted an experiment using the data of 5726 children with fever (less than 72 hours from onset). They applied statistic methods and obtained better sensitivity accuracy than traditional methods, like nonstructural protein 1(NS1) or nonstructural protein antigen (NS1 Ag) strip test.

The Dengue NS1-Ag assay is a rapid test to determine the NS1 antigen in dengue plasma or serum from patients. NS1-Ag appears in the blood from day one to day nine of the disease; thus, it is used as a tool to help doctors perform diagnosis. Despite being the gold standard diagnostic, NS1-Ag it is difficult to apply to patients because this requires specific skills. However, in developing countries, which face a shortage of well-trained doctors and high patient volumes, using Dengue NS1-Ag is a serious problem that must be considered.

The greatest obstacle machine learning techniques must deal with is high-dimensional data. Such data include irrelevant attributes and noise, which results in high computational costs. A common solution is dimensional reduction to remove noise, sparse outlying entries, and missing entries. This approach can be recognized as one of the feature transformation and feature selection techniques. By creating combinations of the original attributes, feature transformation techniques [6] summarize a dataset in fewer dimensions. In feature selection techniques, [7] only the most relevant dimensions from the dataset are selected to reduce dimensionality. With such approaches, irrelevant dimensions and redundant attributes are removed such that computational time and memory requirements are reduced without affecting accuracy. However, with such methods, the meaning of subspaces may be overlooked.

Thus, we propose the Hidden Subspace Clustering (HSC) model to reduce data dimensionality to improve classification accuracy and reduce computational costs. Consequently, dengue fever can be detected with minimal symptoms. The proposed model finds hidden features in the data.

The remainder of this paper is organized as follows. In Section II, the proposed HSC model is defined formally. Section III describes an application of the proposed model to a dengue dataset of 5726 children from hospitals in Vietnam. Experimental results are discussed in Section IV. Finally, open issues and potential future work are discussed in Section V.

II. Methodology

In this section, the proposed HSC model is described. The operations of the model are summarized in Fig. 1 and Table I. Since the raw data may contain the records, coupled with

Manuscript received September 12, 2018; November 1, 2018.

V. D. Minh and M. Kimura are with Department of Information Science and Engineering, Shibaura Institute of Technology, Tokyo, Japan (e-mail: masaomi@sic.shibaura-it.ac.jp, nb17502@shibaura-it.ac.jp).

DOI: 10.7763/IJCTE.2018.V10.1222

missing values, the dataset is needed to be preprocessed. This process comprises of replacing the missing values with the mean of attributes, performing the normalization, and randomly taking the data for training process. Then, irrelevant dimensions are removed using the sparse subspace clustering (SSC) algorithm [8], [9]. In the next step, hidden features are identified and combined with the data. Finally, classification algorithms are applied to the dataset to obtain results.

The proposed model comprises the following tasks.

Fig. 1. The procedure of HSC model.

TABLE I: THE HSC MODEL

Model: Hidden Subspace Clustering (HSC)

Input: A union of data points $\{y_i\}_{i=1}^{N}$ which lie in a high dimension

1. Cleaning data: replace missing data,
2. Reduce dimension of data by sparse subspace clustering (SSC) algorithm
3. Find the hidden dimensions
   - Project subspaces to new areas by random projection(RP) algorithm
   - Cluster data points of new areas and choose the hidden features based on the distance of clusters.
4. A classifier with methods such as logistic regression(LR), Support Vector Machine(SVM), Random Forest(RF) to find the optimal prediction algorithm

Output: label of class

A. Dimension Reduction

High-dimensional data typically include irrelevant attributes and noise. This increases the computational costs and negatively impacts classification accuracy. To address this issue, dimension reduction is considered a crucial solution that can be performed by applying one of the following subspace clustering approaches: iterative, algebraic, statistical and spectral clustering. Each class has its own advantages; however, spectral clustering-based algorithms are used more commonly because of their ability to solve noise and outliers in data. In addition, knowing the dimensions and number of subspaces is not required. The proposed model can handle noise, missing data for dimension reduction by using a spectral clustering algorithm [10], [11], namely SSC algorithm. This approach is appropriate for the target dengue dataset because this dataset is high-dimensional with significant amounts of information and some irrelevant attributes. In addition, this dataset can be reduced without knowing the size of the separated dimensions.

In the SSC algorithm [8], [9], the number of dimensions is reduced by dividing subspaces into two segmentations. This algorithm includes two steps. In the first step, a sparse optimization program is used to find numerous other points that belong to the same subspace S_ℓ. In the second step, spectral clustering is applied to the similar graph to realize data segmentation.

Here, the dimension reduction operation finds the cut line between a pair of clusters, where $\{\mathbf{S}_\ell\}_{\ell=1}^{n}$ is a set of subspaces of $\mathbb{R}^D$ of dimension $\{d_\ell\}_{\ell=1}^{n}$ and data points $\{y_i\}_{i-1}^{N}$ in the union of n subspaces. In order to find the data points in the same subspace, assume y is a linear combination of N_i data points in the same subspace S_ℓ. From that, we define data point y_i as follows:

$$y_i = Y_i c_i . \quad (1)$$

where Y_i is a collection of data point which is a representation of S_ℓ, and c_i is a vector which includes the elements are linear values.

However, the representation of y_i is not unique in the subspaces. Thus, finding the set of appropriate subspaces is a necessary optimal problem. In which, the value of $\|c_i\|$ will be minimized and utilized to recognize a data point in the subspace. More concretely, $\|c_i\| > 0$ mentions that a data point lying in subspace S_ℓ and vice versa.

After solving the optimization program, a weighted graph G = (V, E, W) is utilized to organize the sparse points, where V stands for the collection of N nodes of the graph corresponding to N data points, the set of edges connecting them denoted by E, W is a symmetric non-negative and similarity matrix. In the second step, the SSC algorithm applies spectral clustering to matrix W to segment the data.

B. Find the Hidden Dimensions

As mentioned previously, the proposed HSC model reduces data dimensionality and finds meaningful hidden features in the dataset. Thus, the proposed model promisingly outperforms existing methods because doctors are not required to investigate all symptoms to make a diagnosis. Instead, they can make a precise diagnosis based on only several crucial symptoms. In the statistical point of view, the traditional methods can be expressed as follows.

$$Y = f(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n . \quad (2)$$

A patient is determined to have dengue based on the value of Y.

- Dengue(1) if Y > threshold
- No dengue (0) if Y < threshold

To improve the accuracy of diagnosis, the proposed model combines detecting a small group of patients with disease and traditional classification methods. This group will be identified by finding a pair of appropriate clusters in the subspaces. For example, in a given subspace, data points area divided into two clusters and these clusters are represented by a binary vector $\vec{h}$, which includes value 1 for a cluster of patients with disease and value 0 for a cluster of undefined

patients. In order to make the diagnosis, the processing described in Fig. 2.

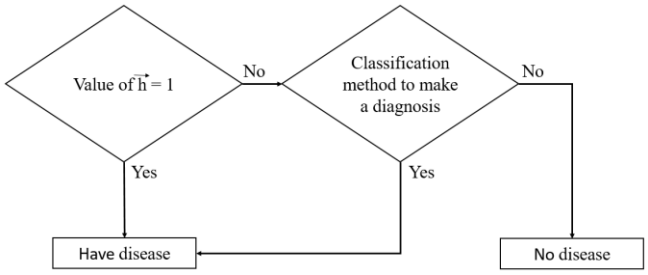


Fig. 2. The processing of making a diagnosis.

From (2), the patient will be distinguished as having disease as follows:

$$y_i = g\left(f\left(x_i\right)\right) = f\left(x_i\right) + h_i\left(threshold - f\left(x_i\right)\right) \quad (3)$$

a) y_i = threshold (disease) if $h_i = 1$

b) $y_i = f(x_i)$ if $h_i = 0$

Here, y_i is a data point of the disease data $\{y_i\}_{i=1}^{N}$, h_i is an element of a vector $\vec{h}$, and $f(x_i)$ is a function to predict disease. The advantage of this approach is that it can identify the patients who truly have dengue, thereby, improving the performance of diagnoses. More concretely, the number of patients *I* can be detected based on traditional classification as follow:

$$I = n \times a \qquad \text{s.t.} \quad a < 1. \quad (4)$$

where n is the number of patients who potentially have dengue and a is the accuracy of classification method.
By identifying the m patients who truly have dengue, (4) can be transformed as follows:

$$I = m \times a + (n - m) \times a\,. \quad (5)$$

However, the number of patients I is detected based on the hidden features as follows:

$$I = m + (n - m) \times a\,. \quad (6)$$

where the m patients will be determined by the values of vector $\vec{h}$. Assume that we can find the group of patients who truly have dengue, determining the patients have disease based on (6) provides better result than the (5). Thus, the performance of diagnosis will be improved.

To find vector $\vec{h}$, we propose a two-step solution. The first step is to find new areas that are a projection of data points with k dimension less than the original data. In the second step, the data points in the new space are clustered. Here, the primary goal is to collect clusters and evaluate the distance between clusters in order to find hidden useful dimensions.

To get a new subspace, we employ the random projection (RP) algorithm [12]-[14] which is a practical and effective method to project data points to another area. Accordingly, a lower dimensional $m \times k$ of subspace P is obtained by the multiplication between a $d \times k$ random projection matrix R (where $k < d$) with the $n \times d$ original data matrix Y.

$$P_{n \times k} = Y_{n \times d} R_{d \times k}\,. \quad (7)$$

The output of this step is new areas with different dimensions.

After finding the new area, the cluster operation is performed. Here, K-means [15], [16] is employed for clustering because it can plot high-dimensional data and can be calculated based on the Euclidean distance. This algorithm begins with each centroid defining one of the clusters. Then, each data point x is assigned to its nearest centroid based on the squared Euclidean distance:

$$\arg\min_{c_i \in C} dist(c_i, x)^2\,. \quad (8)$$

Here, c_i is the collection of centroids in set C, and dist(.) is the standard (L2) Euclidean distance.

Next, the centroids are recomputed by calculating the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x \in S_i} x\,. \quad (9)$$

Here, S_i is a cluster in the union of K clusters $\{S_i\}_{i=1}^{K}$

The results obtained by K-means are a pair of clusters which can be the hidden features in the dataset. From that, the hidden features can be selected based on the correlation between the values of vector $\vec{h}$ and the values of the result class in training set. This correlation also is the accuracy of detecting disease based on vector $\vec{h}$ which higher than the accuracy of classification methods. Therefore, it improves the performance of diagnosis.

C. Apply to Classification Methods

As the final task, a dataset that is the output of the previous task is given as the input to a classification algorithm., such as Logistic Regression(LR) [17], [18], Support Vector Machine(SVM) [19], [20], and Random Forest(RF) [21], [22], to identify an optimal prediction algorithm. In addition, the proposed model utilizes cross-validation techniques to restrict classification overfitting.

III. Experimental Results

We applied the proposed HSC model to a dataset that includes 5726 children with the following criteria.

a) Fever less than 72 hours from onset
b) Attending physician identifies dengue as a possible diagnosis.

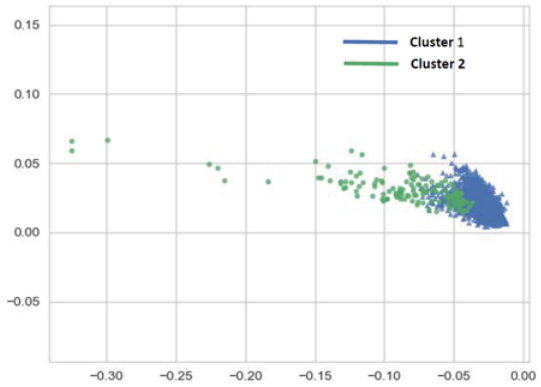
After preprocessing the data, dimension reduction was initially performed by the SSC algorithm. Here, the input data are the clinical symptoms on the day of admission to hospital, such as temperature, vomiting, etc. The SSC algorithm was used to obtain a new dataset, i.e., a subspace of the original data. Table II shows that the SSC algorithm divided the attributes of the original dataset into two segmentations (SSC dataset) in which Segment 2 is removed. This means that the

dimensions of the original dataset were reduced successfully which results in the reduction of the computational time. Besides, Table II also demonstrates that SSC did not change the classification accuracy between the original and SSC datasets.

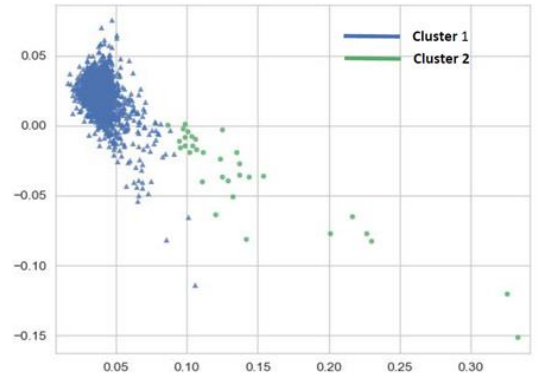
TABLE II: THE RESULT OF SSC ALGORITHM

	Original dataset	Apply SSC to dataset
Attributes	Day disease, Age, Sex, BMI, Temp, Vomiting, Skin bleeding, Mucosal bleeding, Abdominal, Rash, Flush, Injection, WBC, HCT, Platelet, ALB, AST, CK.	- *Segmentation 1*: Day disease, Age, Sex, BMI, Temp, Vomiting, Skin bleeding, Mucosal bleeding, Abdominal, WBC, HCT, Platelet, ALB, AST, CK. -*Segmentation 2*: Rash, Flush, Injection
Accuracy / Sensitivity of classification (Logistic regression)	80.07% / 73.01%	80.03% / 72.94%

BMI: body mass index; WBC: white blood cell count; HCT: hematocrit; ALB: albumin; AST: aspartate aminotransferase; CK: creatine kinas



(a) Projection is no meaning because the distribution of data points is not separate.



(b) Projection has meaningful

Fig. 3. Meaning of the projection subspaces base on distance and separation of clusters.

After reducing the dimensions, we attempted to find hidden features in the SSC dataset. The proposed algorithm selected hidden features as data points which lying in n dimensions of the subspace, where n < d. Based on the distance and separation of clusters, we selected appropriate features, as shown in Fig. 3. The subspace shown in Fig. 3b should be selected because the data points of Cluster 1 is distinguished from the data point of Cluster 2.

After combining useful features in the dataset, we applied classification algorithms to evaluate the accuracy of the proposed model.

TABLE III: THE RESULT OF HSC MODEL

	Original dataset + 1 hidden feature		Original dataset + 2 hidden features	
	Accuracy	Sensitivity	Accuracy	Sensitivity
Logistic Regression	81.15%	74.6%	81.54%	75.4%
SVM	81.3%	77.5%	81.52%	78%
Random Forest	81.19%	74%	80.98%	73.4%

TABLE IV: HSC MODEL IN COMPARISON WITH PREVIOUS MODEL

	HSC Model	Previous Model
Sensitivity	78%	74.8%

Table IV also shows that, relative to sensitivity, the proposed model outperforms a previous method (74.48 %) in [5]. The better sensitivity indicates that HSC model can find the patient more exactly. Therefore, the patients will be treated better.

IV. DISCUSSION

The primary purpose is to find underlying features in the subspaces. This is realized by calculating the distance between clusters in new areas. The dimensions include separating clusters that can represent meaningful features. Based on the assumption that these hidden features are linear to the referenced result of dengue, the subspace were selected based on the correlation between features and a dependent variable. However, this is not the best approach because the distance between the two clusters is not really meaningful in high-dimensional data.

The second purpose of the proposed model is to reduce data dimensionality to decrease computational time without affecting classification accuracy. The experimental results indicate that the proposed model successfully reduced the dimensionality of the data. The attributes were divided into two separated clusters. When we applied a classification algorithm to each segmentation, the accuracy of Segment 2 (Flush, Rash, and Injection) was less than that of Segment 1. Therefore, the dimensions in Segment 2 was removed. Table II shows that three redundant attributes were removed from the original data while maintaining classification accuracy.

The results of this study contribute to the medical diagnosis of dengue fever. The results in Table III show that sensitivity calculated using the SVM was better than the other algorithms (78%), and LR was the best option in term of accuracy (81.54 %). Furthermore, the sensitivity of medical tests, such as NS1 and NS1 Ag strip, is only 70.4%, which is less than that of the proposed method. This indicates that the proposed model is effective at finding the percentage of dengue patients. As a result, the proposed model can help doctors make accurate diagnoses.

V. CONCLUSION

In this paper, in the context of dengue fever, we have proposed a model to reduce data dimensionality and improve the accuracy of classification methods. To achieve this, the

proposed model uses the SSC algorithm to first reduce data dimensionality. Then, RP finds a new projection from subspaces, and the K-means algorithm clusters data points in new areas. We then find hidden features based on the distribution and distance among the points in the clusters. Finally, classifications algorithms, such as logistic regression, SVM, and RF, are applied to optimize the results.

However, the efficiency of the proposed HSC model with this dataset is not as high as expected due to the limitations of the method used to evaluate hidden features. It comes from the fact that the symptoms of dengue are similar to the other diseases which cause much confusion for diagnosis. In future, the proposed model will be applied to the other diseases to evaluate the performance of HSC model.

REFERENCES

[1] L. Rosen *et al*., "Transovarial transmission of dengue viruses by mosquitoes: Aedes albopictus and Aedes aegypti," *The American Journal of Tropical Medicine and Hygiene*, vol. 32, no. 5, pp. 1108-1119, 1983.

[2] Paupy, Christophe *et al*., "A chikungunya outbreak associated with the vector Aedes albopictus in remote villages of Gabon," *Vector-Borne and Zoonotic Diseases*, vol. 12, no. 2, pp. 167-169, 2012.

[3] World Health Organization, Special Programme for Research, Training in Tropical Diseases, Department of Control of Neglected Tropical Diseases, Epidemic, & Pandemic Alert, "Dengue: guidelines for diagnosis, treatment, prevention and control," *World Health Organization*, 2009.

[4] L. Tanner *et al*., "Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness," *PLoS Neglected Tropical Diseases*, vol. 2, no. 3, p. 196, 2008.

[5] N. M. Tuan *et al*, "Sensitivity and specificity of a novel classifier for the early diagnosis of dengue," *PLoS Neglected Tropical Diseases*, vol. 9, no. 4, p. 0003638, 2015.

[6] K. Fukunaga, "Introduction to statistical pattern recognition," *Academic Press*, 2013.

[7] R. Agrawal, I. Tomasz, and S. Arun, "Mining association rules between sets of items in large databases," *ACM Sigmod Record*, vol. 22, no. 2, 1993.

[8] E. Elhamifar and V. Rene "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 570-578, July 1993.

[9] L. Parsons, H. Ehtesham, and L. Huan, "Subspace clustering for high dimensional data: a review," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 90-105, 2004.

[10] Y. Andrew, I. J. Michael, and W. Yair, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, 2002.

[11] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007.

[12] Fern, Z. Xiaoli, and E. B. Carla, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. the 20th International Conference on Machine Learning (ICML-03)*, 2003.

[13] W. B. Johnson and L. Joram, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189-206, 1984.

[14] A. Hinneburg, K. Daniel, and W. Markus, "Using projections to visually cluster high-dimensional data," *Computing in Science & Engineering*, vol. 5, no. 2, pp. 14-25, 2003.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967.

[16] J. Yadav and S. Monika, "A review of k-mean algorithm," *International Journal of Engineering Trends and Technology*, vol. 4, no. 7, pp. 2972-2976, 2013.

[17] S. H. Walker and B. D. David, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1-2, pp. 167-179, 1967.

[18] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215-242, 1958.

[19] M. A. Hearst *et al*., "Support vector machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18-28, 1998.

[20] A. Ben-Hur *et al*., "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125-137, 2001.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[22] A. Liaw and W. Matthew, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18-22, 2002.

Vu Dinh Minh is a Ph.D. student in the Department of Information Science and Engineering of Shibaura Institute of Technology, Japan. His research interests include data mining, machine learning, and natural language processing. Contact him at Dept. of Information Science and Engineering, Shibaura Institute of Technology. 3-7-5 Toyosu, Koto-Ku, Tokyo 135-8548, Japan; nb17502@shibaura-it.ac.jp.

M.Kimura is a full professor and the head of Department of Information Science and Engineering at Shibaura Institute of Technology, Japan. His research interests include text mining, databases, data mining, information extraction, natural language processing and machine learning. Contact him at Dept. of Information Science and Engineering, Shibaura Institute of Technology. 3-7-5 Toyosu, Koto-Ku, Tokyo 135-8548, Japan; masaomi@sic.shibaura-it.ac.jp.