The 4th Joint Symposium on Computational Intelligence (JSCI 4), Bangkok, Thailand

2 February 2018

IEEE Computational Intelligence Society Thailand Chapter

KMITL พระจอมเกล้าลาดกระบัง

KMUTT King Mongkut's University of Technology Thonburi

KMUTNB

Prince of Songkla University

The 4th Joint Symposium on Computational Intelligence, Bangkok, Thailand

Title:	Proceedings of the 4th Joint Symposium on Computational Intelligence (JSCI 4)
Editor:	Kitsuchart Pasupa
Published by:	IEEE Computational Intelligence Society Thailand Chapter
Printing House:	King Mongkut's Institute of Technology Ladkrabang
Edition:	1
Month and Year:	February 2018
ISBN:	978-616-455-375-0

Message from Chair

The Joint Seminar on Computational Intelligence (JSCI) is a biannual event which was first organised in 2016. The event was initiated by IEEE Computational Intelligence Society Thailand Chapter (IEEE-CIS Thailand), that aims to support research students and young researchers, to create a place enabling participants to share and discuss on their research prior to publish their works. The event is open to all researchers who want to broaden their knowledge in the field of computational intelligence.

This is the fourth time and the name of the event is changed to the Joint Symposium on Computational Intelligence (JSCI) and will be co-locating with Deep Learning and Artificial Intelligence Winter School (Feb 1-4, 2018)–supported by Asia Pacific Neural Network Society (APNNS), IEEE-CIS Thailand, and IBM. The symposium will feature eight paper presentations by researchers.

Kitsuchart Pasupa
Chair

Organizing Committee

Advisory

Chanboon Sathitwiriyawong	King Mongkuts Institute of Technology Ladkrabang

Chair

Kitsuchart Pasupa	King Mongkuts Institute of Technology Ladkrabang

Technical Program Committee

Chanboon Sathitwiriyawong	King Mongkuts Institute of Technology Ladkrabang
Jonathan H. Chan	King Mongkuts University of Technology Thonburi
Kitsuchart Pasupa	King Mongkuts Institute of Technology Ladkrabang
Kiyota Hashimoto	Prince of Songkla University
Kuntpong Woraratpanya	King Mongkuts Institute of Technology Ladkrabang
Phayung Meesad	King Mongkuts University of Technology North Bangkok
Vithida Chongsuphajaisiddhi	King Mongkuts University of Technology Thonburi

Program

Time	**Title**/*Authors*
15:00–15:15	**Sentiment Analysis of the Burmese Language using the Distributive Representation of n-gram-based Word** Myat Lay Phyu and Kiyota Hashimoto
15:15–15:30	**Development of Hybrid Deep Learning in Sentence Classification** Thiptanawat Phongwattana, Praisan Padungweang, and Jonathan H. Chan
15:30–15:45	**Detection of Personal Vehicles Stopping on the Road in a No Parking Area Using Support Vector Machine** Eakbodin Gedkhaw, Manussawee Piyaneeranart, and Mahasak Ketcham
15:45–16:00	**A Comparision of Iteration-Free Bi-Dimensional Mode Decomposition and Empirical Monocomponent Image Decomposition** Donyarut Kakanopas and Kuntpong Woraratpanya
16:00–16:15	**SNP selection for Porcine breed classification by a hybrid information gain and genetic algorithm** Wanthanee Rathasamuth, Kitsuchart Pasupa, and Sissades Tongsima
16:15–16:30	**Subnetwork Identification based on Dissimilarity Profiles of Gene Co-Expressions** Thanyathorn Thanapattheerakul, Narumol Doungpan, and Jonathan H. Chan
16:30–16:45	**Development of an Automated Biological Tool for Visualizing Dissimilarity within Gene Co-Expression Networks in Hierarchical Clustering** Prissadang Suta, Panissara Thanapol, Jonathan H. Chan, and Thiptanawat Phongwattana
16:45-17:00	**An Adaptive Learning System Based on Proportional VARK to Enhance Learning Achievement Concept** Beesuda Daoruang, Suthida Chaichomchuen, and Anirach Mingkhwan

Contents

Sentiment Analysis of the Burmese Language using the Distributive Representation of n-gram-based Word

Myat Lay Phyu
Master of Science in Information Technology,
College of Computing, Prince of Songkla University,
Phuket Campus, Thailand.
Email: myatlayphyu096@gmail.com

Kiyota Hashimoto
Faculty of Technology and Environment,
Prince of Songkla University, Phuket Campus, Thailand.
Email: kiyota.h@phuket.psu.ac.th

***Abstract*—Less resourced languages are still difficult to treat due to the unavailability of annotated big corpora and basic natural language processing tools. This paper proposes a new method to use a character-based variable-length *n*-gram word model with distributive word representation techniques to reduce the number of *n*-gram words in data. We employed this method for news article sentiment analysis and achieved a better result than CRF-based ordinary word segmentation with a small size of supervised data. This enables to treat less resourced languages without focusing on language specific characteristics.**

I. INTRODUCTION

Exponential increase of available textual and speech data needs and enables advanced natural language processing techniques, but special attention is still necessary for less resourced languages with which big annotated corpora and basic tools like word segmentation tools are not available. In particular, languages whose writing concatenates words without spaces, such as Burmese, Lao, and Thai, have more difficulty because word segmentation as the initial step is not trivial at all. In this case, there are two possible approaches: to develop a word segmentation technique for a specific language with limited resources, and to develop a pseudo-word segmentation technique without paying much focus on language specific characteristics. In this paper, we propose a new method along the line of the second approach and compared both approaches for news article sentiment analysis as an example task.

Our proposal employs character-based variable-length *n*-gram words as words. By definition, all words, or pseudo-words, are regarded precisely as words, but the number of the words is much larger than the usual words contained in the same data. Thus we adopted character-based variable-length *n*-gram word estimation and distributive word representation models, word2vec [4] and GloVe [1] as word grouping methods to reduce the number of *n*-gram words in the data.

Sentiment analysis, or more specifically positive/negative value estimation of texts have been widely investigated. There are roughly two approaches: binary classification and cumulative calculation, though the combination of these is also possible. The first considers the task as a simple binary classification but as such, a supervised method. Thus a good amount of supervised data must be available, which is not the case in most low resourced languages. The second considers the task as a cumulative sentiment value calculation of all sentiment-valued words in each text on which final classification is based [2], but thus a sentiment dictionary must be constructed. Many methods for sentiment dictionary construction have been proposed, and we followed SO-LSA (Semantic Orientation-Latent Semantic Analysis) method. [7].

We conducted comparative experiments between an ordinary word segmentation, which is also proposed by the authors [5] using a character clustering method [6] and CRF [3] with a small size of annotated data, and our character-based variable-length *n*-gram word segmentation with and without word grouping before constructing a sentiment dictionary. The evaluation is based on the final sentiment value estimation of news articles. The result shows that our pseudo-word approach achieved almost the same result as an ordinary word approach, which enables to investigate less resourced languages without developing language specific tools and techniques.

The rest of this paper describes the research methodology in section 2 and our comparative experiments in section 3, finished by conclusion in section 4.

II. RESEARCH METHODOLOGY

The overview of our research methodology, as in Fig. 1, is sentiment value estimation of Burmese news articles based on a newly created sentiment dictionary with character-based variable-length *n*-gram words grouped with distributive word representation vector similarity.

Data consists of 100 news articles from the opinion section of *7Day Daily*, one of the most popular newspapers in Myanmar, and the sentiment value of each article is manually assigned based on a questionnaire survey. Among them, 52, 11 and 37 are positive, neutral and negative news respectively, but only positive and negative articles are used for classification.

Our character-based variable-length *n*-gram word recognition is conducted by first starting with 15-gram to choose very

978-616-455-375-0

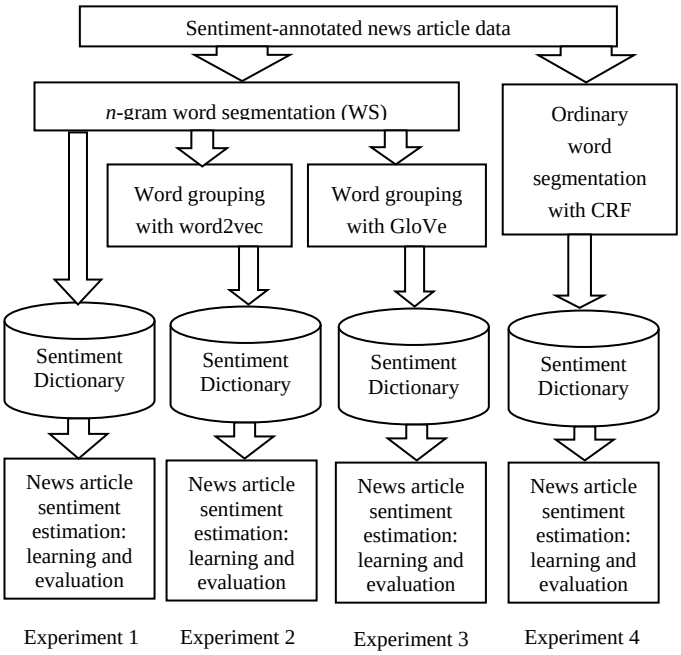


Fig. 1 Overview and experimental conditions

frequent 15-gram sequences as words, and iteratively repeating the same procedure to 3-gram (Experiment 1-3). In order to reduce the number of variable *n*-gram words, distributive word representation techniques, word2vec (Experiment 2) and GloVe (Experiment 3) are employed to make groupings of words based on cosine similarity. For these methods, we employed 1,280 unannotated articles (approximately 10 million characters) because these methods need larger sizes of data. CRF-based word estimation is also employed for comparison (Experiment 4). CRF is a supervised method, which needs word segmentation annotated data. Thus, we created a small manually annotated dataset for CRF learning.

Sentiment dictionaries are then constructed based on these four different word segmentations, using the SO-LSA method [7]. We selected 8 positive and 8 negative seed words and the calculation is based on the following equation:

$$\text{SO-LSA(word)} = \sum_{\text{pword} \in \text{Pwords}} \text{LSA (word, pword)} - \sum_{\text{nword} \in \text{Nwords}} \text{LSA (word, nword)}$$

Here, pwords is the set of positive seed words and nwords is the set of negative seed words. Note that an *n*-gram word containing a seed word but not containing a negative word is also regarded as the same seed word.

Finally, classification is performed for evaluation.

III. EVALUATION

A. Experiment

We conducted four comparative experiments to construct a sentiment dictionary and news article classification as shown in Fig.1. News article sentiment classification is conducted with 100 articles.

B. Result

The summary of the result is shown in Table 1. According to F-measure, the variable-length *n*-gram word achieved the highest result while variable-length *n*-gram words with GloVe-based word groupings achieved as high as that. According to accuracy, on the other hand, CRF-based ordinary word segmentation achieved the highest.

Table 1 Experimental results

	Accuracy	Precision	Recall	F-measure
Experiment 1	73.03%	66.7%	100%	80%
Experiment 2	65.2%	59%	96%	73%
Experiment 3	70.78%	67.57%	96.15%	79.34%
Experiment 4	76.4%	81.25%	75%	78%

C. Discussion

Experiment 2 performed worst. This seems to be caused mainly by our small data size. On the other hand, Experiment 3 achieved a much higher result in accuracy, precision, and F-measure. It seems that GloVe is more suitable than word2vec partly because GloVe is based on sentence structure while word2vec is based on simple word occurrence. However, our result is based on a small size of data, and thus evaluation with a larger data is waited.

IV. CONCLUSION

Our proposed method shows that variable-length *n*-gram word model with similarity-based grouping is a promising method for less resourced languages, because our method does not need a large annotated data or elaborately developed word segmentation tools. However, the results seem to be affected by our small data size. Thus the next step is to test our method with much bigger data, which is being pursued.

ACKNOWLEDGMENT

This work was also supported by the Higher Education Research Promotion and the Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission.

REFERENCES

[1] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representations", EMNLP, pp. 1532–1543, 2014.

[2] M. Godsay, "The Process of Sentiment Analysis: A Study", International Journal of Computer Applications, 126(7), pp. 26-30, 2015.

[3] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Seqmenting and Labeling Sequence Data", Proc. Of 18th ICML, pp.282-289, 2001.

[4] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", ICLR Workshop, 12 pages, 2013.

[5] M. L. Phyu and K. Hashimoto, "Burmese Word Segmentation with Character Clustering and CRFs", Proc. of 14th JCSSE, pp. 1-6, 2017.

[6] T. Thanaruk, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan, "Character Cluster Based Thai Information Retrieval", Proc. Of IRAL '00, pp. 75-80, 2000.

[7] P. D. Turney and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association", ACM Tran,sactions on Information Systems, 21-4, pp. 315–346, 2003.

Development of Hybrid Deep Learning in Sentence Classification

Thiptanawat Phongwattana, Praisan Padungweang, and Jonathan H. Chan*
School of Information Technology,
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand
thiptanawat.p@mail.kmutt.ac.th, praisan.pad@sit.kmutt.ac.th, jonathan@sit.kmutt.ac.th

***Abstract*— This paper explores the combination of two deep learning techniques that consists of convolutional neural networks (CNN) and long short-term memory recurrent neural networks (LSTM-RNN) as a hybrid approach to sentence classification. The technique used CNN in feature extraction, followed by LSTM-RNN to build a classifier. In the text mining field, the performance mostly depends on word features that we need to utilize the distinctive point of CNN to re-organize the feature set for training in the LSTM-RNN layer. In feature initialization, we used a pre-trained GloVe dataset of Common Crawl to reproduce our existing dataset that includes research articles, which were extracted into a set of sentences of success factors relationship statements. We hypothesize that using the hybrid deep learning technique, CNN and LSTM-RNN, with word embedding, GloVe, can improve the performance of sentence classification in terms of recall and precision from the prior work that used only a single CNN technique.**

I. Introduction

Currently, machine learning is increasingly being utilized in research and industry fields. Much interest has been developed in various established machine learning techniques as computational means have improved in form of faster parallelized CPUs and GPUs. Convolutional neural networks (CNN) is one of the most popular techniques that is implemented in image recognition problems. An example is a classification work that contributed a CNN model for classifying 1.2 million high-resolution images from the ImageNet dataset into a thousand different classes [1]. Feature selection is a nontrivial part of CNN that can be utilized to extract images by their pixels and color channels into a set of multi-dimensional vectors. Thereafter we use the features set as inputs to build a classifier by using multi-layer perceptron (MLP) neural networks, and the outputs in terms of accuracy, precision, as well as recall, are exceptional. Therefore, we would like to utilize the benefits of CNN in terms of feature extraction in the text mining domain as the classifiers' effectiveness is significantly affected by the proper selection of a set of appropriate features. In text mining, many tools are available, for example GENIA, BANNER, Turku Event Extraction System (TEES), Stanford natural language processing (NLP); however, there haven't been tools utilizing a neural network technique yet.

In this paper, we would like to build a set of word features in each sentence by using a combination of CNN and pre-trained word vectors. Thereafter we use another kind of neural networks, long short-term memory recurrent neural networks (LSTM-RNN) for training a classifier in sentence classification. By doing so, we define both deep learning techniques that are used in the same pipeline as a novel hybrid deep learning technique. As for the research domain, we used the corpus from Krathu et al. [2] as our base line for comparison. According to the previous work, they used many kinds of techniques for feature representation and classification, for example Naïve Bayes, Binary, Term frequency-inverse document frequency (TFIDF), Logistic Regression. None of the techniques produced good performance in both recall and precision. Consequently, we would like to explore the use of deep learning in sentence classification for this prior work in order to improve the recall and precision values simultaneously.

In particular, we utilized glove vector (GloVe) [3], which was developed by Stanford, for word representation. For the GloVe dataset, we used pre-trained word vectors from Common Crawl that includes 840 billion tokens, 2.2 million vocabularies as well as 300 dimensional vectors. The dataset was used as an input to convolutional neural networks (CNN) to build the second layer after the word representation layer in order to extract useful word features. Thereafter, LSTM-RNN was implemented for training a classifier. The corpus that we used was manually exported from a database that consists of many research papers in a business-related domain. Each research paper was already extracted into sentences, and each sentence was assessed by 2 domain experts for identifying the success factor statements within each article.

According to the assessment, the identification was categorized into binary classes comprising positive, which refers to a success factor, and negative, which is not. The corpus is separated into 2 sub-corpora with 80 percent for training and another 20% for testing. We elaborately build the corpora by balancing between 2 classes in order to minimize biases that could be an issue when training a classifier.

II. Methodology

According to Fig. 1, we used the existing corpus from Krathu et al. [2] in order to compare our technique with their results. First, we have to access to Hyporet KMUTT database that belongs to School of Information Technology,

978-616-455-375-0

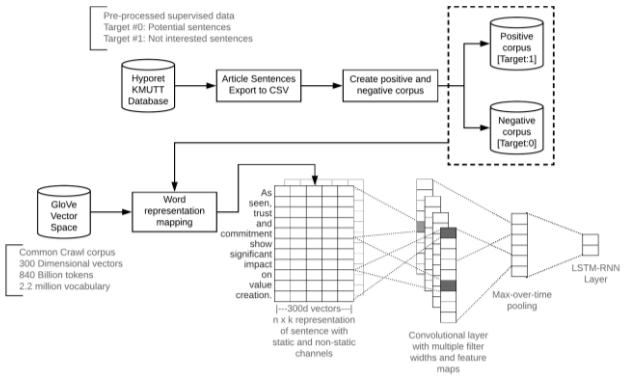


Figure 1. Model architecture with two deep learning techniques in sentence classification

King Mongkut's University of Technology Thonburi in order to retrieve the raw data of business success factor statements as a CSV file. As the raw data was already pre-processed as binary classes, we were able to classify the sentences by using their attributes. For the target equals to "1", we define those sentences as the positive sentences, whereas the negative sentences are assigned the target of "0". Thereafter, we separated the dataset into two files comprising the positives and the negatives as potential statements and non-interested statements respectively. Second, an input of convolutional neural networks must be vectors. Normally if we build a text mining model from scratch, any texts or words cannot be used directly for training. Consequently, we have to convert each token in each sentence to a dimensional vector. By doing so, we use a word embedding technique that it is called GloVe, which stands for Global Vector, which was developed by a Stanford team. There are several ways for building a set of word embedding as a vector space, for example Word2Vec, CBOWs, but we found that in a variety of research domains that use word embedding, GloVe is one of the most popular and it can provide a better result. The difference of GloVe from others is it builds a co-occurrence matrix for a corpus as the prior step to factorize it to yield matrices for word vectors. According to a prior work [4] that also used CNN, Word2Vec was used as the word representation for the work; however, we proposed a change from Word2Vec to GloVe in order to improve score of the accuracy without any hyper-parameters tuning. The pre-trained GloVe dataset that we use is Common Crawl corpus, which consists of 840 billion tokens and 2.2 million vocabularies. The dataset was trained and extracted its features to 300 dimensional vectors. In this step, we use the GloVe dataset to convert the input corpora to vectors before inputting the vector space to the CNN layer afterwards. Third, convolutional layer is implemented with multiple filter widths and feature maps that are built by feature extraction, which is the distinctive point of CNN. Fourth, max-over-time pooling is enabled for the architecture. By doing so can capture the most important feature, which is one with the highest value for each feature map. Hence the pooling scheme will be utilized for dealing with variable sentence lengths. Finally, in the classification stage, we use long short-term memory (LSTM) neural networks technique in training a classifier since the technique is the-state-of-the art in text mining, which out-performs, particularly, a dataset that relates to time series such as sentences. In practical use, we also prefer Dropout technique in the part of regularization that can improve the classifier for predicting unseen data by guarding against over-fitting.

Hyper-parameter	Value	Hyper-parameter	Value
ALLOW_SOFT_PLACEMENT	True	EVALUATE_EVERY	100
BATCH_SIZE	64	FILTER_SIZES	3, 4, 5
CHECKPOINT	100	L2_REG_LAMBDA	0.0
DROPOUT_KEEP_PROB	0.5	NUM_EPOCHS	200
EMBEDDING_DIM	128	NUM_FILTERS	128

Table 1. A Set of Hyper-Parameters of CNN Architecture

According to the training corpus information, the vocabulary size is 20,982, and the hyper-parameters are listed in Table 1, along with the tuned values.

III. Results

Fig. 2 shows the preliminary result that we evaluated by using a loss function. With the first 100 steps of training, the model was optimized and dropped by 50%. Hence, given more steps and epochs can help to define its global optimum that leads the model to be more satisfactory in sentence classification.

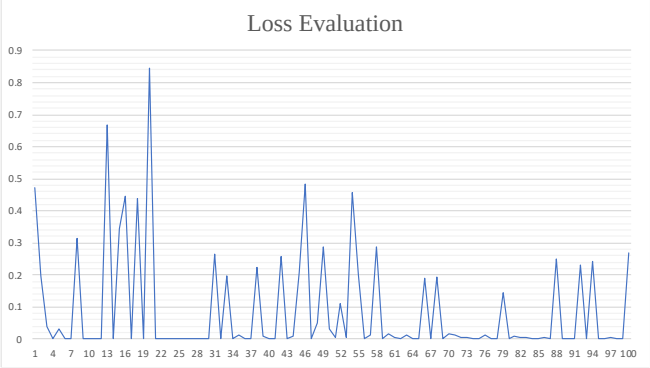


Figure 2. Loss evaluation within 100 steps.

References

[1] Alex Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks", NIPS proceedings, 2012.

[2] Worarat Krathu et al., "Data Mining Approach for Automatic Discovering Success Factors Relationship Statements in Full Text Articles" 8th International Conference on Advanced Computational Intelligence, Chiang Mai, Thailand; February 14-16, 2016.

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014.

[4] Yoon Kim, "Convolutional Neural Networks for Sentence Classification" New York University, 2014.

Detection of Personal Vehicles Stopping on the Road in a No Parking Area Using Support Vector Machine

Eakbodin Gedkhaw*, Manussawee Piyaneeranart†, and Mahasak Ketcham‡
Innovation and Technology Management Research Center
Faculty of Information Technology, King Mongkuts University of Technology North Bangkok, Bangkok, Thailand
Email: *eakbodin.g@chandra.ac.th, †s5907011910045@email.kmutnb.ac.th, ‡mahasak.k@it.kmutnb.ac.th

***Abstract*—Traffic jam is an important issue in the lives of people in the capital city. To reduce this problem, the researchers are interested in developing a personal vehicle recognition system that is stopping on the road in a no parking area. With the application of Support Vector Machine by receiving the signal from the CCTV camera to improve the image and find the specific characteristics of the car stopping on the road in a no parking area. Then learn to recognize the parking behavior of a personal car, that stopping on the road in a no parking area, using the Linear Regression. The experiments showed that recognition of personal cars parked on the road in a no parking area had the accuracy at 87.56 % which could implement to detect the car parking in a no parking area.**

I. Introduction

According to the Global Traffic Scorecard Report of the Year 2016 Global Traffic Report [1] found that Thailand is the world's number one of traffic jam. This traffic problem in Thailand had cause from many problems for example people do not drive discipline, stopping on the road in a no parking area, the lack of lane, also the sale of goods obstructing traffic [2]. In this article, the researchers proposed a vehicle recognition system that stopping on the road in a no parking area. To detection of personal cars parked, the process begins with memorizes the system about scenario of the local road and shape of car then train the characteristic of car parking with window slide technique. Next use the Local Binary Patterns (LBP) and the Histograms of Oriented Gradients (HOG) to classify the characteristics of personal cars stopping on the road in a no parking area. A Support Vector Machine (SVM) is used to group all available areas in the window slide process. Finally, the method used by the majority vote is to evaluate the car stopping on the road in a no parking area. Using binary results defined by SVM. The major difference with the previous approach was the use of the HOG feature with the SVM to recognize the car stopping on the road in a no parking area. In addition, the researcher's system was designed to be used for traffic images recorded by CCTV cameras.

II. Related Works

Techniques for analyzing accidental images in the analysis of road accidents. For example, accidental image analysis using the Hidden Markov Model (HMM) to classify events, various accidents [3]. It provides 4 types of roadside monitoring techniques, including parked vehicles, slow-moving vehicles, vehicle parts and lane changes [4]. However, Ikeda's research has not yet been able to detect car accidents on the road using a technique called "Image Tracking" [5] and HMM is used to detect collisions of cars, [6] which HMM used to detect unusual events on the road. In addition, research on car brand recognition (logo) with the LPR (License Plate Recognition) system (Intelligent Transportation Systems: ITS) is becoming more important. The current driving supervision of the car will control the speed. The license plate must be checked to verify the car properly [7]. Vehicle Logo Recognition that identifies the vehicles that can differentiate a vehicle's logo with Vehicle Manufacturers Recognition (VMR) [8]. In recognition of the car's logo. Detect vehicles from stereo cameras to detect vehicle models and create models of recognition using the HOG, LBP, and Haar features [9]. Bringing stereo cameras and HOG, LBP and Haar features together will allow real time vehicle classification.

Fig. 1. Personal car data set. [10]

III. Methodology

This section can be divided into three stages: Preprocessing, Feature Extraction and Classification.

A. Preprocessing

Preparation of information before the recognition of the format to distinguish image data types. At this stage, the image of the car will be selected. Experimented with a series of images of personal cars prepared in fixed light conditions. There are 1087 free background images. Reduce the image size to 180 × 200 to practice in the recognition system shown in Fig. 1.

978-616-455-375-0

B. Feature Extraction

In the extraction step, the HOG method is used to store the gradient values 0°-180°using the direction value of 6 bin and set the Grid value to 8 × 8 as shown in Fig. 2.

Fig. 2. Separation of image features. [11]

C. Classification

The image identification will be tested with the prepared personal car data set. Randomly selected data in each group of 1087 data sets, using 761 images for training and 163 images for testing. Selected by SVM method and Linear Regression. Then evaluate the efficiency of the classification rate which is based on the percentage accuracy of correctly classified images relative to the total number of images. The experiment was conducted for 15 cycles and the mean was obtained. Example of discrimination as shown in Fig. 3.

IV. Experimental and Result

Experiment on recognizing a personal car parked illegally. The Linear Regression algorithm in SVM has been implemented and received from CCTV cameras. Improve picture quality to suit your application and adjust contrast and noise. Bring images to distinguish features or region of interest. Bring images to distinguish features or areas of interest. Bring the features you want to practice and learn in the recognition system to classify. The training and test results shown in Table I.

TABLE I
Training and Testing of Car Recognition.

Training Confusion Table		
Output/Target	***Non-vehicle***	***Vehicle***
Non-vehicle	321 (42.2 %)	18 (2.4 %)
Vehicle	9 (1.2 %)	413 (54.3 %)
Test Confusion Table		
Output/Target	***Non-vehicle***	***Vehicle***
Non-vehicle	65 (39.9 %)	12 (7.4 %)
Vehicle	7 (4.3 %)	79 (48.50 %)

From table I, the experiment used 1,087 images which 761 images in training and 263 images in testing. The result showed that the learning image between image with car and without car (only lane scenario) had the accuracy at 42.2 % (correct 321 images from total 761) and the object learning which is car, the accuracy rate was 54.3 % (correct 413 image from total 761). In the part of training for recognition and in the test section, it can detect the road or in the absence of the correct car 65 out of 163 images or 39.9 % and correctly check the car 79 images or 48.5 %. Detecting a car stopping on the road in a no parking area. There must be additional features a car with a space in front of the car about 1 meter and the car was parked in the same place for 10 minutes is considered a personal car parked on the road in the park. Table II shows that the experiment using the illegally parking 225 cars can be detected cars a total of 197 cars that unlawful parking at 87.56 % of accuracy.

Fig. 3. Identification of individual cars.

TABLE II
System Performance Analysis.

Total vehicle	***Vehicle detected***	***Accuracy (%)***
225	197	87.56

V. Conclusion

Performance evaluation results for personal car illegally parking recognition on the road by applied SVM to receive CCTV image. Improved image clarity to distinguish specific features and was processed using the Linear Regression. From the results found that the system was able to recognize the personal car unlawful parking on the road, with the accuracy rate of 87.56 %. Can be applied to enforce traffic law effectively. Verification is provided to allow the driver to check his driver's license and verify other information to make the system more clear, such as using HMM, CNN, ANN, 3DCNN or RNN.

References

[1] INRIX. INRIX global traffic scorecard. [Online]. Available: http://inrix.com/scorecard (Accessed October 6, 2017).
[2] BBC. Thai cars ranked 1 in the world - survey shows no way to permanently eliminate the problem. [Online]. Available: http://www.bbc.com/thai/thailand-39038498 (Accessed October 6, 2017).
[3] Y. Zou, G. Shi, H. Shi, and Y. Wang, "Image sequences based traffic incident detection for signaled intersections using HMM," in *Proceedings of the 9th International Conference on Hybrid Intelligent Systems*, 2009, pp. 257–261.
[4] H. Ikeda, Y. Kaneko, T. Matsuo, and K. Tsuji, "Abnormal incident detection system employing image processing technology," in *Proceedings of the IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems*, 1999, pp. 748–752.
[5] C. P. Lin, J. C. Tai, and K. T. Song, "Traffic monitoring based on real-time image tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2003, pp. 2091–2096.
[6] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Transactions of Intelligent Transportation Systems*, vol. 1, no. 2, pp. 108–118, 2000.

[7] C.-N. E. Anagnostopoulos, I. E. A. annd I. D. Psoroulas, V. Loumos, and E. Kayafas, "License plate recognition from still images and video sequences: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, pp. 377–391, 2008.

[8] A. Psyllos, C. N. Anagnostopoulos, and E. Kayafas, "Vehicle model recognition from frontal view image measurements," *Comput. Stand. Interfaces*, vol. 33, no. 2, pp. 142–151, 2011.

[9] D. Neumann, T. Langner, F. Ulbrich, D. Spitta, and D. Goehring, "Online vehicle detection using Haar-like, LBP and HOG feature based image classifiers with stereo vision preselection," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 773–778.

[10] T. Tangkocharoen and A. Srisuphab, "Vehicle detection on a pint-sized computer," in *Proceedings of the 9th International Conference on Knowledge and Smart Technology*, 2017, pp. 40–44.

[11] N. Sedaghat and T. Brox, "Unsupervised generation of a viewpoint annotated car dataset from videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 40–44.

978-616-455-375-0

A Comparison of Iteration-free Bi-dimensional Mode Decomposition and Empirical Monocomponent Image Decomposition

Donyarut Kakanopas*, and Kuntpong Woraratpanya†
Faculty of Information Technology
King Mongkuts Institute of Technology Ladkrabang
Bangkok, Thailand
Email: *k.donyarut@gmail.com, †kuntpong@it.kmitl.ac.th

***Abstract*—This article describes the difference between iteration-free bi-dimensional mode decomposition and empirical monocomponent image decomposition, which are two recently published and very interesting papers, in terms of concept and their applications.**

I. Introduction

Image decomposition plays an important role in image analysis such as image denoising, image filtering, and related applications. The trend of recently proposed decomposition techniques has been focused on data-driven methods; i.e., these techniques do not need any prior functions for decomposition. Two recently published and very interesting papers were iteration-free bi-dimensional empirical mode decomposition (iBEMD) [1] and empirical monocomponent image decomposition (EMID) [2]. The iBEMD was originated from the classical empirical mode decomposition (EMD) introduced by Huang et al. [3]. It was strongly improved performance in terms of the computation time and decomposed image quality, thus making it practical for real applications. On the other hand, the EMID was originated from the empirical wavelet transform (EWT) proposed by Gilles [4] in 2013. It mainly improves quality of decomposed images or monocomponent images. Although both iBEMD and EMID are image decomposition techniques, they are rooted from the different concepts. Therefore, they always produce the different decomposed images and perhaps lead to different solutions for problem-solving. The following section describes the concept of both methods.

II. Analysis of iBEMD and EMID

This section briefly describes the evolutions of EMD and EWT, and analyze their extensions to the newest improved algorithms.

A. Iteration-free Bi-dimensional Mode Decomposition

An empirical mode decomposition (EMD) was first introduced by Huang et al. [3] in 1996 for non-linear and non-stationary signal analysis. Since then its extensions has been proposed for both 1D and 2D EMDs. The 2D EMD is also known as bi-dimensional empirical mode decomposition (BEMD). The most improvement of modified BEMDs: BEMD-RBF [5], FBEMD [6], FaBEMD [7], and MEMD [8] were proposed for different purposes. However, the computational cost of those BEMDs is very high [1]. Recently, iBEMD has been proposed by Titijaroonroj et al. for speeding up the computing time and for enhancing the decomposed image quality. The iBEMD was based on locally partial correlation for principal component analysis (LPC-PCA) to directly estimate mean surface from bi-dimensional signals without using iteration technique for extracting bi-dimensional intrinsic mode functions (BIMFs), which are decomposed images, from an original image. Fig.1 shows an original image and a set of BIMFs decomposed by BEMD-RBF, FBEMD, FaBEMD, MEMD, and iBEMF, respectively. The iBEMD method achieves in fast computation of algorithm and high quality of decomposed image.

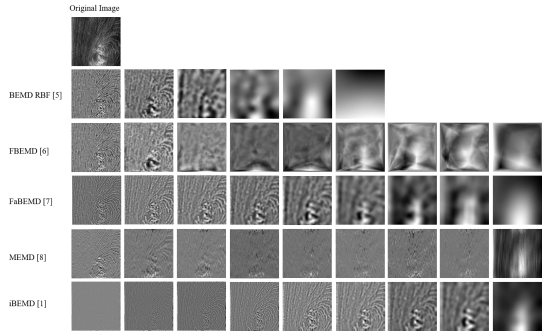


Fig. 1. A set of BIMFs decomposed by various approaches with wood texture image.

B. Empirical Monocomponent Image Decomposition

2D empirical wavelet transform (2D EWT) proposed by Gilles et al. [9] was the first adaptive image decomposition in frequency domain extended from the empirical wavelet transform [4] for signal analysis. The 2D EWT method adaptively decomposes an image by segmenting its spectrum in

978-616-455-375-0

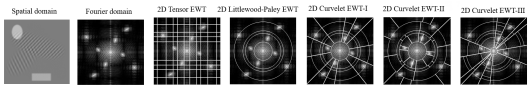


Fig. 2. Various segmentation methods lead to inaccurate spectrum segmentation.

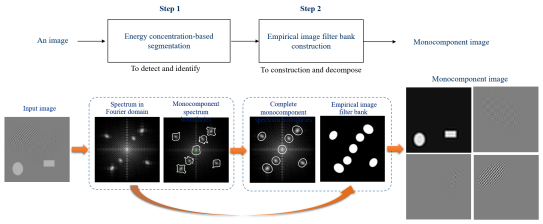


Fig. 3. Framework of empirical monocomponent image decomposition.

Fourier support. In Gilles's method, the spectrum segmentation is very important for monocomponent image decomposition. However, a variety of spectrum segmentations as shown in Fig. 2 presented by Gilles et al. still does not achieve the high quality of monocomponents. Newly, an EMID method was proposed by Suttapakti et al. [2]. The EMID method was based on (i) energy concentration-based segmentation and (ii) empirical image filter bank construction as shown in Fig. 3. In energy-based segmentation, 2D local maximum point detection is performed on an empirical mean plane that is created by principal component analysis (PCA) to detect the central frequencies of candidate mono-component spectra. Then a 2D local minimum boundary detection algorithm is used to detect a component boundary from each detected central frequency. After that, an actual monocomponent identification algorithm is used to identify actual monocomponent spectrum boundaries. In empirical image filter bank construction, all filter banks are constructed in accordance with monocomponent spectrum boundaries by means of ellipse and Gaussian functions, and is used to decompose an image into monocomponent images with fewer ringing artifacts. Like this, the proposed EMID method is able to achieve a high quality of monocomponent images.

C. Comparison

As mentioned in subsections II-A and II-B, we can summarize the difference between the iBEMD and EMID methods as follows. In iBEMD method, a decomposed image is a BIMF which can be represented more than one frequencies. Meanwhile, a decomposed image extracted by the EMID method is a monocomponent image which can be represented a single frequency. The different decomposed images of those methods are the different number of frequencies in each decomposed image. This leads to the different solutions for solving the same application. For example, in Thai text localization, the EMID can identify the text region on the image, since text-texture components in Fourier support are located in the same position, thus making it easy to segment text region from the background. However, it cannot eliminate the illumination effect on the input image, due to the fact that illumination component is one of non-linear and non-stationary signals which can be located in other positions of Fourier support. On the other hand, the iBEMD method can eliminate the illumination component from the input image easily, because it is designed for non-linear and non-stationary signal analysis. It can extract the BIMF which contains more than one frequencies and is also nearly similar oscillation, whereas the iBEMD method cannot directly segment the text region from the background, since more than one frequencies are there. Therefore, we can conclude that the EMID method is a powerful image decomposition for a linear signal, whereas the iBEMD method is effective in signal analysis for non-linear and non-stationary data.

D. Conclusion

This article has addressed two recently published papers in image decomposition techniques. Both techniques were rooted from different concepts. Therefore, they always produce the different decomposed images and perhaps lead to different solutions for problem-solving.

References

[1] T. Titijaroonroj and K. Woraratpanya, "Iteration-free bi-dimensional empirical mode decomposition and its application," *IEICE Trans. Information and System*, vol. E100-D, no. 9, pp. 2183–2196, 2017.

[2] U. Suttapakti, K. Pasupa, and K. Woraratpanya, "Empirical monocomponent image decomposition," *IEEE Access*, In Press.

[3] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hillbert spectrum for nonlinear and non-stationary time series analysis," *Proceeding of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[4] J. Gilles, "Empirical wavelet transform," *IEEE Trans. Signal Process*, vol. 61, no. 16, pp. 3999–4010, 2013.

[5] J. C. Nunes, Y. Bouaoune, E. Delechelle, O. Niang, and P. Bunel, "Image analysis by bidimensional empirical mode decomposition, image and vision computing," *IEEE Trans. Signal Process*, vol. 21, no. 12, pp. 1019–1026, 2003.

[6] C. Damerval, S. Meignen, and V. Perrier, "A fast algorithm for bidimensional EMD," *IEEE Signal Process Lett.*, vol. 12, no. 10, pp. 701–704, 2005.

[7] S. M. A. Bhuiyan, R. R. Adhami, and J. F. Khan, "A novel approach of fast and adaptive bidimensional empirical mode decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 1313–1316.

[8] J. C. Lee, P. S. Huang, T. M. Tu, and C. P. Chang, "Recognizing human iris by modified empirical mode decomposition," *Advances in Image and video Technology, Lecture Notes in Computer Science*, vol. 4872, pp. 298–310, 2007.

[9] J. Gilles, G. Tran, and S.Osher, "2D empirical transforms wave-lets, ridgelets, and curvelets," *SIAM J. Imaging Sci.*, vol. 7, no. 1, pp. 157–186, 2014.

978-616-455-375-0

SNP selection for Porcine breed classification by a hybrid information gain and genetic algorithm

Wanthanee Rathasamuth*, Kitsuchart Pasupa†, and Sissades Tongsima‡
*Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
†National Center for Genetic Engineering and Biotechnology (BIOTEC),
National Science and Technology Development Agency (NSTDA), Pathum Thani 12120, Thailand
Email: *rathasamuth.wan@gmail.com, †kitsuchart@it.kmitl.ac.th, ‡sissades@biotec.or.th

***Abstract*—Single Nucleotide Polymorphism (SNP) is a variability of DNA sequence that connects to a unique trait of an organism. A good SNP selection can provide a good porcine breed that grows fast with high yield. SNP selection can be done by a computerized feature selection method and classification technique. At present, an effective classification model can only handle a small number of features efficiently. Too large a number may cause an over-fitting problem in the classification. Therefore, SNPs or features need to be reduced to an optimum number for an effective porcine SNP analysis. This paper proposes an approach to reducing the number of features in porcine SNP analysis to an optimum number by a hybrid of Information Gain (IG) and genetic algorithm (GA) techniques. A performance test demonstrated that this approach was able to select a minimum number of features (at 1.51% of the total number of features) that provided an average classification accuracy of 94.02%, as compared to 95.28% provided by the total number of features.**

***Index Terms*—Feature selection, Bioinformatics, Machine learning, Single Nucleotide Polymorphisms.**

I. Introduction

China might be the first country that has started to selectively bred wild pigs 5,000 years ago. Pig breeding in Thailand was heavily influenced by the Chinese immigrants to this country. Today, pigs are an important economic animal in Thailand, hence selecting the right breed for the geographical location is a very important issue. Diverse physical traits of pigs are the results of the differences in DNA base sequences which are called single nucleotide polymorphism (SNP). A thorough porcine SNP analysis can determine the SNPs that provide good growth and reproduction. The issue is that there are millions of SNPs for a single organism, and so a manual SNP analysis by an expert is out of the question, not to mention the huge amount of other kinds of resources needed. Today, a good way to address this issue is to use bioinformatics, an integration of computer science, biology, mathematics, and engineering. Machine learning [1] has been applied to genomics, proteomics, microarray, and system biology for classification of genes. In [1], several classification techniques for bioinformatics are presented such as support vector machine (SVM), decision tree, neural networks, Bayesian classifiers, and nearest neighbors. Since these classification techniques cannot effectively support too large a number of features that may cause a commonly encountered over-fitting problem–high accuracy when used with training dataset but low accuracy with testing dataset, reducing the number of features into a subset of optimum features can make a classification attempt successful. Papers that deal with this issue are such as [2] which is a review of feature selection applying to bioinformatics. The paper reports three types of feature selection: filter methods such as i-test, and information gain (IG); wrapper methods such as genetic algorithms (GA) and other nature inspired algorithms; and embedded methods such as random forest, and decision tree. In [3], a review of several nature-inspired algorithms that were used to perform feature selection was presented. The main point was how to increase selection efficiency and reduce prediction error. The most common problems found were too large a number of features and too small a training dataset. These are some of the challenges in doing a classification analysis.

The conceptual frameworks of feature selection by filter, wrapper, and embedded methods are quite different. Simple and efficient, filter methods select features that have high index values, independent of the classification method. Wrapper methods rely on the classification method to select an optimal subset of features that provide high classification accuracy– each round of feature evaluation includes a classification step; therefore, a large number of features results in high computation time. Embedded methods are not very different from wrapper methods but include a feature reduction step that reduces their computation time. Wrapper and Embedded methods rely on a classification step to select an optimal subset of features hence the selected features can facilitate better learning of training dataset but their predictions may suffer from an over-fitting problem. On the other hand, filter methods tend to have less over-fitting problem.

This paper proposes using a hybrid feature selection technique that combines IG with GA (IG+GA) for the purpose of selecting the best porcine SNPs for classification of pig breeds. The next section presents a description of datasets, followed by experimental framework and its results.

II. Datasets

The dataset used in this study consists of SNP data from 677 pig samples of 22 breeds, 356 samples from the dataset of porcine colonization of the Americas [4] and 321 samples from the dataset of the Project of Porcine Breed Improvement by Selection according to Whole Genome SNP Data supported

978-616-455-375-0

by the National Center for Genetic Engineering and Biotechnology (BIOTEC), for a total of 16,579 SNPs. All of the data had been through data cleansing already. However, there were some missing values that were estimated by a single imputation method that replaced the values with the mode of the whole individual feature data. The combined dataset was random-seeded into 10 datasets; each dataset contains a training set (80% of the data) and a test set (20% of the data).

III. Experimental Framework

Among all features, some features may not exert any significant influence on the constructed learning model, hence they can only waste computation time; therefore, it is essential to select only significant features that strongly influence the learning model that can make accurate prediction. The experiments in this study tested the comparative feature-selection performances of IG, GA, and IG+GA when used with a linear-kernel SVM classifier. In this study, the hyperparameter of SVM, C, was specified to be in the range of $10^{-6} \sim 10^{6}$, and five-fold cross-validation was used to obtain an optimal parameter to construct an acceptably-accurate model.

IG is a feature selection technique of the filter type [2] that calculates feature indexes from the relationships between features. It has been successfully applied to selection of text, microarray, and SNPs yielding a small set of significant features. GA is another feature selection technique that can reduce the number of features to a small subset of significant features, but it has a problem of getting trapped at local optimums. IG+GA was proposed by a previous study [5] and demonstrated to provide a smaller subset of significant features that were able to give more accurate predictions than IG or GA alone. In this study, we used IG to rank the significant levels of features, then used an elbow method to find the cut-point for feature selection. This cut-point determined the number of genes of each chromosome to be constructed in GA. For instance, if the cut-point was 300, the number of constructed genes in each GA chromosome would be a number no higher than 300. The average cut-point we found by IG from our datasets was 409, so we roughly specified the number of genes in GA as 400. GA was used to select features within the range of the first 400 ranked features obtained from IG. The population size of GA was 20 chromosomes. The selection method in GA was a roulette wheel method. The crossover step was a multi-point crossover with a crossover probability of 0.8. The mutation step was a bit-flip mutation, and the maximum number of generations was 10.

To remedy the problem of solutions getting trapped at local optimums, we set the mutation probability of a '1 to 0' flip to 0.3 and that of '0 to 1' flip at 0.7. This strategy of setting different mutation probabilities for different kinds of bit-flipping was proposed in [6] for avoiding local optimums.

IV. Experimental Results

Results of the reduction of number of features by IG, GA, and IG+GA are presented in Table I. In the same table, accuracies in porcine breed classification by using the features from these 3 methods and from using the whole original features are presented.

TABLE I
Comparative results of average classification accuracies and the number of selected SNP.

Methods	Accuracy (%)	#SNP
All Features	95.28 ± 1.23	$16,579$
IG	93.96 ± 0.60	409 ± 0.09
GA	94.80 ± 1.40	$2,357.9 \pm 3.20$
IG+GA	94.02 ± 1.19	250.1 ± 0.01

It can be seen that every approach to feature selection that we had attempted gave nearly the same value of average classification accuracy: the using of the whole original SNPs approach yielded an accuracy of 95.28% while the using of the features from GA selection alone yielded the highest accuracy at 94.80%; that from IG alone yielded 93.96%; and that from IG+GA yielded 94.02%. The average percentages of the number of SNPs reduced were 85.78%, 97.53%, 98.49% achieved by GA, IG, and IG+GA, respectively. The approach with GA alone was more accurate but used a larger number of SNPs than the IG alone and IG+GA approaches, while the IG+GA approach yielded nearly the same accuracy as that provided by GA but used significantly fewer SNPs.

V. Conclusion

SNP selection for porcine breed classification can be done by several feature selection methods and classifiers. This study used IG alone, GA alone, and IG+GA approaches to select an optimal set of SNPs for SVM to classify and found that the IG+GA approach not only was able to reduce the highest number of features, at 98.49%, but also provided a classification accuracy that was very nearly equal to the highest accuracy provided by the GA alone approach.

References

[1] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Brief. Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[3] H. Frohlich, O. Chapelle, and B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm," in *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 142–148.

[4] W. Burgos-Paz, C. A. Souza, H. J. Megens, Y. Ramayo-Caldas, M. Melo, C. Lemús-Flores, E. Caal, H. W. Soto, R. Martínez, L. A. Álvarez, L. Aguirre, V. Iñiguez, M. A. Revidatti, O. R. Martínez-López, S. Llambi, A. Esteve-Codina, M. C. Rodríguez, R. P. M. A. Crooijmans, S. R. Paiva, L. B. Schook, M. a. M. Groenen, and M. Pérez-Enciso, "Porcine colonization of the Americas: A 60k SNP story," *Heredity*, vol. 110, no. 4, pp. 321–330, 2013.

[5] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm," in *Proceeding of the 2012 International Conference on Computer Science and Electronics Engineering*, 2012, pp. 355–358.

[6] G. Mahdevar, J. Zahiri, M. Sadeghi, A. Nowzari-Dalini, and H. Ahrabian, "Tag SNP selection via a genetic algorithm," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 800–804, 2010.

978-616-455-375-0

Subnetwork Identification based on Dissimilarity Profiles of Gene Co-Expressions

Thanyathorn Thanapattheerakul*, Narumol Doungpan†, and Jonathan H. Chan‡
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Email: *thanyathorn.tha@sit.kmutt.ac.th, †narumol.dou@sit.kmutt.ac.th, ‡jonathan@sit.kmutt.ac.th

***Abstract*—Given a large dataset of microarray gene expression data, an important problem is the identification of biomarkers linked to a disease. Gene-set based analysis would tend to group genes based on biological relevance. Gene co-expression networks can then be constructed from these gene-sets using a topological overlap-based dissimilarity measure. While a typical method of analysis involves searching for functional modules of highly connected genes, this work proposes a method that involves profiling the entire dissimilarity weight distribution within a gene-set. This grouped profile analysis is then used with functional analysis tools to identify potential subnetwork biomarkers for a disease. A case study with lung cancer is illustrated.**

***Index Terms*—Gene co-expression network, dissimilarity, topological overlap, biomarker, gene-set profile, microarray, ANOVA, WGCNA**

I. Introduction

Given a large dataset of microarray gene expression data, an important problem is identifying biomarkers linked to a disease. Gene-set-based analysis groups genes based on biological pathways. Gene co-expression networks can be constructed from these gene-sets using the topological overlap-based dissimilarity measure. While one method of analysis involves searching for functional modules of highly connected genes, we propose a method that involves profiling the entire dissimilarity weight distribution within a gene-set.

II. Methods

A. Data sets

1) Microarray Gene Expression Data: Lung cancer gene expression data (GSE10072) was downloaded from the online Gene Expression Omnibus (GEO) database [1]. The GSE10072 dataset is composed of a total of 107 samples, of which 58 are adenocarcinoma samples and 49 are non-tumor samples used as control [2]. The samples were taken from tissue samples of adenocarcinoma paired with non-involved lung tissue from current, former and non-smokers.

2) Gene-set Data: Gene-set data was collected from PathwayAPI [3], which is the database containing the consistent biological pathway information from KEGG, Ingenuity and Wikipathways, for gene-set integrative analysis purposes.

B. Methodology

1) Data Preprocessing: After eliminating probe sets that represent more than one gene, the data is normalized by using Z-Score Transformation for each gene across all samples. Gene-sets that contains three genes or less are removed since these gene-sets are meaningless to do gene-pair correlation. The remaining gene-sets are then used to identify the statistical significance of each gene labelled as either *significant* ("sig") or *non-significant* or "non") by using the Analysis of Variance (ANOVA) statistical tests with a *p*-value of 0.05.

2) Construction of Dissimilarity Matrix: For each gene-set, the WGCNA package in R was used to create and transform the correlation matrix to a scale-free topology [3]. The topological overlap measure was then determined, where a_{ij} is the weighted expression correlation between gene i and gene j:

$$\omega_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min\left\{\sum_u a_{iu}, \sum_u a_{uj}\right\} + 1 - a_{ij}} \tag{1}$$

By subtracting equation (2) from unity, we determined the dissimilarity weight or measure for each gene pair [5].

$$d_{ij} = 1 - \omega_{ij} \tag{2}$$

3) Gene-set Profiles: Profiles of each gene-set were created by plotting dissimilarity weights for all pairs in the set. After labelling each gene from the ANOVA tests, each gene pair was categorized as *"sig-sig", "sig-non", "non-non"* and assigned colors *blue, yellow, and grey*, respectively, for plotting. Each profile shows the relative proportions of dissimilarity weights and how dissimilarity varies across the gene-set.

4) Validation: To validate our hypothesis, we made use of biological pathway databases for this purpose, including Genetic Association Database (GAD) - a database of genetic association data [6], Kyoto Encyclopedia of Genes and Genomes (KEGG) - a web-based biological database [7], Catalogue of Somatic Mutations in Cancer (COSMIC) - a database and website storing human genetic data [8] and Database for Annotation, Visualization, and Integrated Discovery (DAVID) - a web-based program that integrates functional genomic annotations with graphical summaries [9].

978-616-455-375-0

TABLE I
AVERAGE PERCENTAGE OF LUNG CANCER RELATED GENES FOUND ON GAD, KEGG, DAVID AND COSMIC.

Shape	GAD	KEGG	DAVID	COSMIC
Normal	17.547	9.937	23.859	5.604
Reverse-S	16.308	8.047	20.612	4.991
Quarter	12.166	10.990	20.893	3.317

III. RESULTS AND DISCUSSION

The results are shown in two types of graphs in Fig. 1 and Fig. 2 as the plot of sample gene set processed by default parameters and tuning parameters of WGCNA, respectively. The top shows the distribution of dissimilarity weights, which was measured by using the different soft threshold values. According to [5], the soft threshold is the minimum value, in between 0 to 30, used for converting the correlation to be a scale-free network. The soft threshold should be the optimal one that gives a proper goodness-of-fit between the scale-free network and the correlation. The bottom shows the clustering of weights in each gene-pair category. The plot displays median, mean and standard deviation of weights in black, bold red and red, respectively.

Biological database [6] - [9] were used to explore differences between the three shapes of gene-sets, which were categorized manually, shown in Fig. 2. According to the results shown in Table I, the analysis of each gene-sets involvement in the lung cancer phenotype was found that Normal profiles tend towards higher percentage of cancer genes in GAD, DAVID and COSMIC for lung-cancer than other shapes. Quarter shaped profiles have a highest percentage of cancer genes in KEGG. In contrast, Reverse-S shaped profiles do not have a high percentage of cancer genes as the authors have hoped for.

IV. CONCLUSION

Gene-set profiles are a useful way of visualizing co-expression and identifying functional trends in gene-sets. Grouped profile analysis can be used with other tools to identify potential biomarkers for a disease. Furthermore, the results would be the features for clustering gene-set profile. The developed concept tools will be further tested with more disease datasets and more intelligent method for clustering the gene-set profiles. In addition, the method will be compared to existing methods in order to show the validity.

REFERENCES

[1] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. (2005). "NCBI GEO: mining millions of expression profiles-database and tools," *Nucleic Acids Research,* vol. 33, suppl 1 (2005).

[2] M. Landi, T. Dracheva, M. Rotunno, J. Figueroa, H. Liu, A. Dasgupta, F. Mann, J. Fukuoka, M. Hames, A. Bergen, S. Murphy, P. Yang, A. Pesatori, D. Consonni, P. Bertazzi, S. Wacholder, J. Shih, N. Caporaso and J. Jen, "Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival," *PLoS ONE,* vol. 3, no. 2, p. e1651, 2008.

[3] D. Soh, D. Dong, Y. Guo and L. Wong, "Consistency, comprehensiveness, and compatibility of pathway databases," *BMC Bioinformatics,* vol. 11, no. 1, p. 449, 2010.

[4] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics,* vol. 9, no. 1, p. 559, 2008.

[5] S. Prom-on, A. Chanthaphan, J. Chan and A. Meechai, "Enhancing Biological Relevance Of A Weighted Gene Co-Expression Network For Functional Module Identification," *Journal of Bioinformatics and Computational Biology,* vol. 09, no. 01, pp. 111-129, 2011.

[6] K. Becker, K. Barnes, T. Bright and S. Wang, "The Genetic Association Database," *Nature Genetics,* vol. 36, no. 5, pp. 431-432, 2004.

[7] M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research,* vol. 28, no. 1, pp. 27-30, 2000.

[8] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. Futreal, M. Stratton and R. Wooster, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *British Journal of Cancer,* vol. 91, no. 2, pp. 355-358, 2004.

[9] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane and R. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biology,* vol. 4, no. 9, p. R60, 2003.

978-616-455-375-0

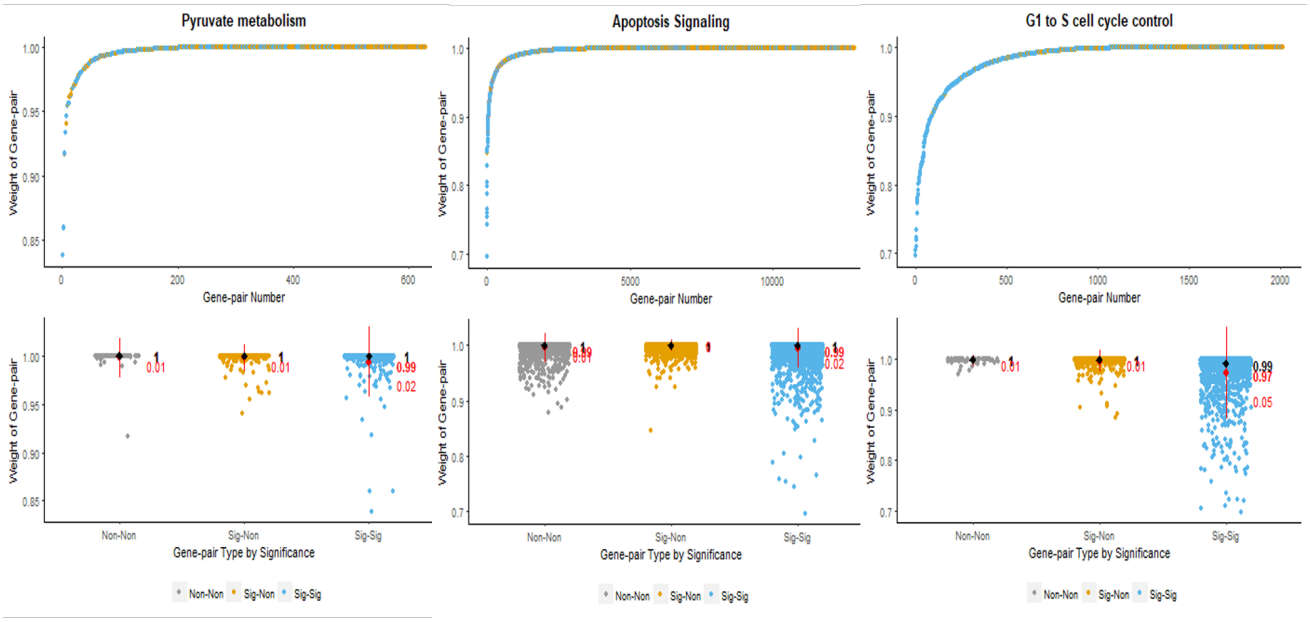


Fig. 1. Two types of plots for sample gene-sets processed by default parameters. Top: Distribution of dissimilarity weights, measured by using the default value of soft threshold in WGCNA package ($\beta = 6$) [6], over rank of gene-pairs.

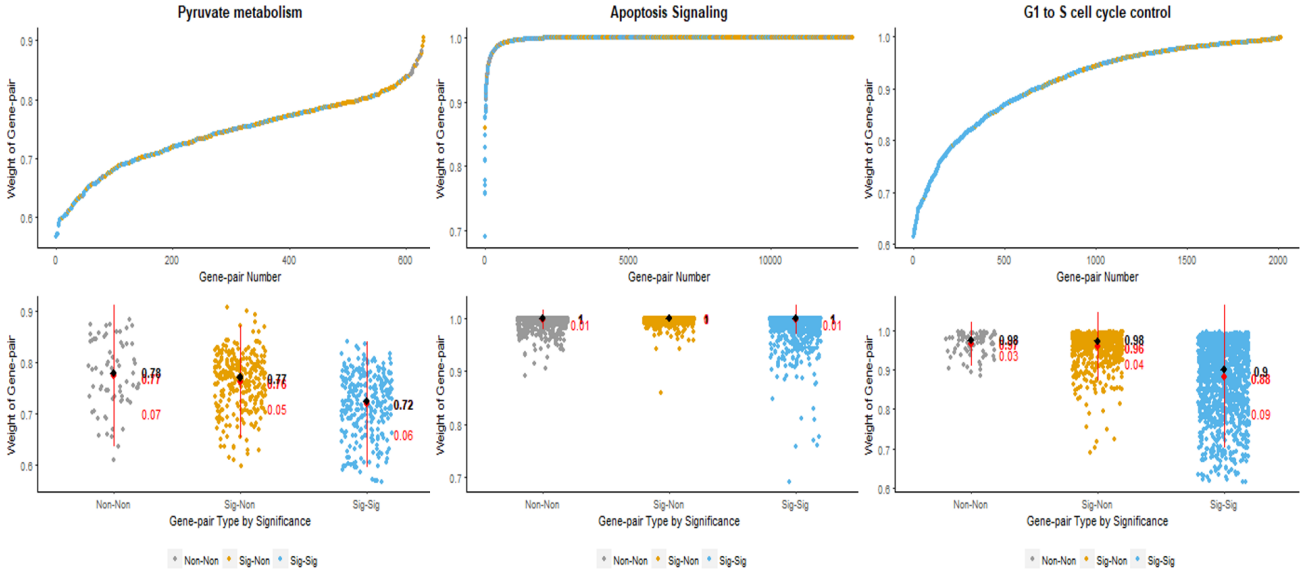


Fig. 2. Two types of plots for sample gene-sets processed by tuning parameters. Top: Distribution of dissimilarity weights, measured by using the soft threshold producing the maximum of R2, over rank of gene-pairs. There are three shapes of gene-set profiles, which are "Reverse-S", "Normal" and "Quarter" respectively.

978-616-455-375-0

Development of an Automated Biological Tool for Visualizing Dissimilarity within Gene Co-Expression Networks in Hierarchical Clustering

Prissadang Suta*, Panissara Thanapol†, Jonathan H. Chan‡, and Thiptanawat Phongwattana§
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
Email: *prissadang.sut@mail.kmutt.ac.th, †panissara.tha@mail.kmutt.ac.th, ‡jonathan@sit.kmutt.ac.th,
§thiptanawat.p@mail.kmutt.ac.th

***Abstract*—We present an automated biological tool in order to visualize gene co-expression from a gene expression dataset in form of a dendrogram of hierarchical clusters. Our proposed tool calculates dissimilarities within gene-sets of biological pathways using Topological Overlap Measure algorithm. There are two aspects in our motivation comprising an automated biological tool and algorithms that can be utilized for researchers who are interested in this area because it can help to save time in the data preparation stage. Currently, WGCNA that is developed in R language is still limited in some processes that a researcher has to do manually, for instance data pre-processing and visualization. However, the library can be utilized in gene pair correlation computation of a gene co-expression that is calculated between each gene pair in a gene-set. Moreover, the popular library is able to transform the networks into scale-free networks in order to calculate the dissimilarity weights that are used to create a gene-set profile. In this work, we combined all of the necessary steps in form of an automated tool. Furthermore, our approach also distinguishes between gene pairs consisting of one, both, or no statistical significant genes, based on ANOVA testing of a set of features. In conclusion, our automated tool provides a means of clustering visualization in terms of biological pathway, as well as how gene dissimilarity is linked to the mutual significance of gene pairs within a gene-set that can help researchers in relevant fields to analyze their data.**

I. Introduction

At present, there are many software libraries implemented in the biological domain, especially in form of molecular data such as genes. A popular library for gene network analysis that we found is called "WGCNA" [1], which stands for weighted gene co-expression network analysis, developed in R language. The library can be utilized to find gene co-expressions by using Pearson correlation algorithm. Furthermore, it is able to create a scale-free network between genes. At the end of WGCNA process, it outputs a set of dissimilarities of each gene pair with in a gene-set that is part of a biological pathway, which refers to a disease. According to the prior research, we found that the outputs can be plotted into a graph in order to observe its characteristic that may lead to identification of potential biomarkers. The WGCNA library provides the core process but there is still a need to manually perform tasks such as data pre-processing and visualization. In the data pre-processing, source data, which is obtained from gene expression omnibus (GEO), which is a database repository of high throughput gene expression data and hybridization arrays, chips, and microarrays, is in a form of a probe dataset that is necessary to be annotated to generate a gene expression dataset. This may be a time-consuming step for researchers. Also, visualization is a necessity in biological field that the prior research needs to plot manually for each pathway of interest. According to the abovementioned issues, our proposed tool can help researchers to perform more efficient analysis as the tool just needs a raw probe dataset and the annotation file for its initialization. Then it can automatically execute the pipeline to produce visualization output in form of a hierarchical dendrogram. Moreover, we use a statistical technique called "ANOVA" in order to distinguish between gene pairs that consist of significant and non-significant label for each gene in a gene-set. By doing so, a set of the statistical outputs can be utilized as potential features for research. We also provide hyper-parameters for researchers in order to perform fine-tuning of the correlation algorithms, e.g. the number of power term for scale-free networks, complex disease pathways, and so on. For our contribution, researchers can leverage our developed tool for reducing their analysis time for to obtain analytic results since the tool has merged all necessary techniques and algorithms into a single pipeline. And the researchers can utilize our tool for parameter tuning in order to find some significances in the area of interest more easily.

II. Methodology

Lung cancer gene expression data (GSE10072) is downloaded from Gene Expression Omnibus (GEO) database [2]. GSE10072 dataset consists of 107 samples that can be categorized into 2 groups comprising Adenocarcinoma, which has 58 samples, and Non-tumor, which has 49 samples that are classified as control [3]. The samples were taken from tissue samples of Adenocarcinoma that was paired with non-involved lung tissue from current, former and non-smokers.

In Fig. 1, we provide the proposed method which is able to identify biomarkers by using gene co-expression datasets.

978-616-455-375-0

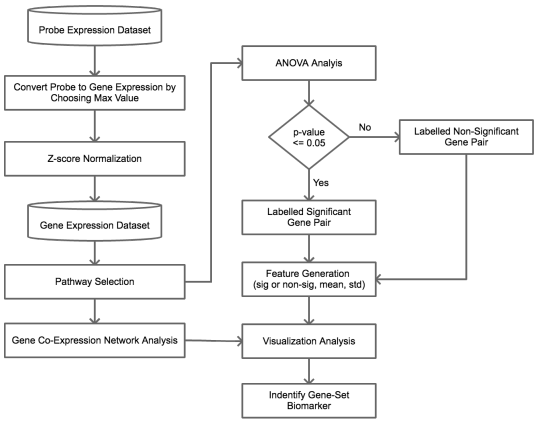


Fig. 1. Flowchart for identify potential biomarker.

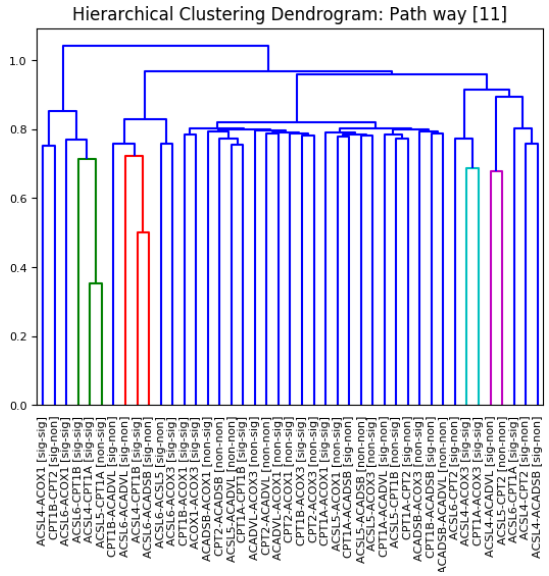


Fig. 2. Dendrogram of Hierarchical Clustering.

To identify each gene or biological pathway for gene-set transformation, it is divided into two groups by reference to ANOVA in order to tag labels significant (sig) or non-significant (non) in each gene-pair and categorize as sig-sig, sig-non and non-non, which is based on a threshold that is defined p-value $<= 0.05$. Next, we perform the gene co-expression network analysis. It includes Pearson's correlation, that generates a power term which the WGCNA is used as a default, i.e. beta equals to 6 that works well to analyze in gene co-expression networks [4]; it then creates a scale-free network topology [5], computes dissimilarity between genes using topological overlap measure dissimilarity [6], generates hierarchical clusters of genes, divides clustered genes into modules, and merges very similar modules. Finally, the measured dissimilarity will be used for graph plotting and it can be utilized for analyzing least errors that can be described how gene-pairs are visualized within gene co-expression networks. The visualization illustrates a graph characteristic and this result will be used to analyze by domain experts for biomarker identification. The result can identify biomarkers that are automatically obtained from gene co-expression visualization tool.

III. Results

In Fig. 2, we demonstrate a hierarchical dendrogram for the Fatty Acid Metabolism pathway. It consists of gene pairs that are categorized into each cluster. The value of each gene pair is based on weighted dissimilarity calculation that we use for clustering. However, some set of values in some pathway may have negligible variance, for example, they are very close to 0 or 1 mostly, and we would like to increase their spatial distribution in order to analyze clustering more effectively. The consequence is adding some exponential number to all of the values before categorizing into clusters. We also find a set of significant genes in each gene profile, which uses ANOVA, in order to analyze the significant genes (e.g. significant-significant, nonsignificant-significant, and nonsignificant-nonsignificant) in each cluster of dissimilarity of gene pairs for finding some potential relevance. Nonetheless, the statistical information in each cluster is still ambiguous. So, we would leverage the features in our future work for biomolecular analysis. Furthermore, we will colorize the clusters to highlight the hierarchical clustering visualization.

IV. Conclusion

This research paper contributes a useful biological tool that helps researchers to reduce lead time in research, especially in bio-data pre-processing. The automated tool can perform gene expression annotation, significance tagging, gene correlation analysis, weighted scale-free network analysis, topology overlap measure calculation, dissimilarity of gene co-expression network analysis as well as hierarchical clustering visualization within a single pipeline. In each stage, we also provide hyper-parameters tuning for convenience.

Acknowledgment

We would like to thank Narumol Dougpan, who gave us very useful advice and information for developing our automated tool that is based on WGCNA.

References

[1] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[2] T. Zeng, S. Y. Sun, Y. Wang, H. Zhu, and L. Chen, "Network biomarkers reveal dysfunctional gene regulations during disease progression," *FEBS Journal*, vol. 280, no. 22, pp. 5682–5695, 2013.

[3] T. Barrett, "NCBI GEO: mining millions of expression profiles–database and tools," *Nucleic Acids Research*, vol. 33, pp. D562–D566, 2004.

[4] M. T. Landi and et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, 2008.

[5] A. Barabási, E. Ravasz, and T. Vicsek, "Deterministic scale-free networks," *Physica A: Statistical Mechanics and its Applications*, vol. 299, no. 3–4.

[6] A. M. Yipand and S. Horvath, "The generalized topological overlap matrix for detecting modules in gene networks," pp. 1–19, 2005.

An Adaptive Learning System Based on Proportional VARK to Enhance Learning Achievement Concept

Beesuda Daoruang*, Suthida Chaichomchuen†, and Anirach Mingkhwan‡
Department of Computer Education
Faculty of Technical Education
King Mongkuts University of Technology North Bangkok
Bangkok, Thailand
Email: *beesuda.p@fitm.kmutnb.ac.th, †suthida.c@fte.kmutnb.ac.th, ‡anirach@ieee.org

***Abstract*—This research aims to study the conceptual framework of learning by selecting the Personal Learning Activity Log to classify learner's learning styles. Adaptive Learning is able to organize content as a percentage of classifying learning patterns. Moreover, the developed adaptive learning system tends towards enhance learner achievement, which can effectively adapt their learning style throughout the end of a course content.**

I. Introduction

Adaptive learning proposes a learning approach to tailor content directing towards learners based on individual learning abilities. The different perception of each learner influences to their effective learning. Characteristics of a learner with different learning styles is extremely important for the instructor to find out the appropriate contents. When learners obtain suitable content that match their preferences, they can finally have self-confidence and enjoyment in learning course.

Therefore, researchers decide to study the conceptual framework of learners learning by using the Personal Learning Activity Log to classify learners learning styles and provide adaptive learning systems that adapt learning patterns to fit learners learning patterns in the VARK. Learners do not have to take the quiz to classify academic aptitude as well as the methods of VARK questionnaire.

II. Literature Review

Fleming and Mills [1] proposed varying style of learning by preference or aptitude, dividing into four categories: Visual, Aural, Read/Write, and Kinesthetic, known as VARK Model or VARK Learning Styles. There are 16 questionnaires for the classification of four individual learning styles. Also, Hasibuan et al. [2] have developed a model for detecting learning patterns based on Agent Learning, known as Adaptive Dynamic Responsive (ADR). VARK learning model is selected as detected learning style. It can be adapted to meet the requirements of learner. Capability of this proposed model is that content is adjusted with the ability of learner to complete the course. The learning style can achieve of maintaining motivation of the learner with online learning. Tashtoush et al. [3] developed the Adaptive E-learning system for teaching English by taking into account data mining techniques. Cross-validation in the format of Jackson's Learning styles offers content that is adaptable to the learner's learning style. For example, videos, content presentations, and quizzes can be adapted depending on the learner's learning style. The results showed that learners had the high achievement (87.4 %). Abdullah et al. [4] studied the adaptive e-learning model that matches the learning patterns between teachers and learners by using machine learning. Matched Educator-Student learning can match the learning patterns of teachers and learners by using Bayes Classification Techniques, Decision Tree, and Support Vector Machine. Also, learner achievement was compared by using Mat-ES and traditional learning styles. This research introduced Kolb learning styles for teaching process and VARK for the learning process. Results showed that Mat-ES learning model can improved learning performance. In addition, the J48 classifier provides the highest accuracy in classifying data.

From literature review above, researchers are able to gain an understanding of hypothesis suggestion that if there is no personal information acquired from learner via pre-questionnaire before using the learning system, it will be convenient to use for learner.

III. The Conceptual Model

The concept development of adaptive learning system based on proportional VARK to enhance learning achievement is shown in Fig. 1. The modules can be divided into five modules as follows:

A. Content Module

A module stores the content of the lesson, pre-test, exercise during the course, the post-test, and the exam at the end of each chapter. The content of the lesson is created and stored in the VARK learning model, which consists of V (Visual), A (Aural), R (Read/write), and K (Kinesthetic), see Fig. 2a.

B. Adaptive Module

This module obtains results from the Expert module being able to classify a proportional VARK. It will then acquire the

978-616-455-375-0

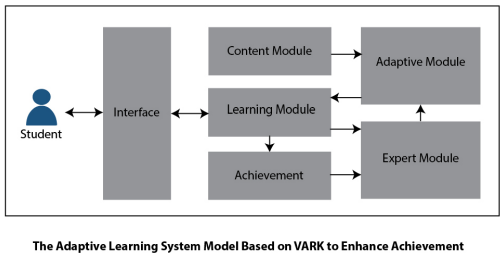
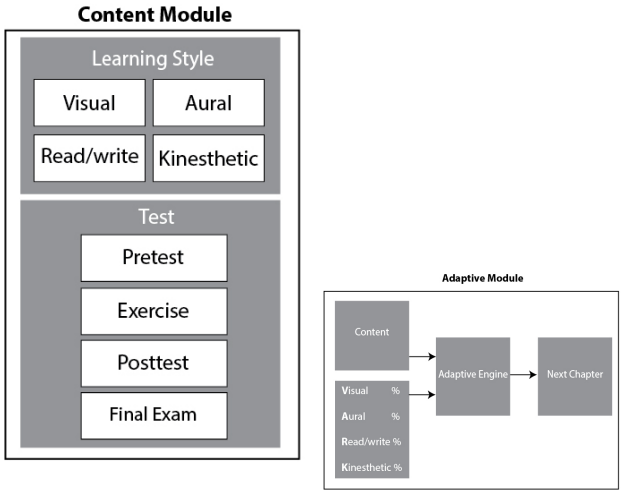
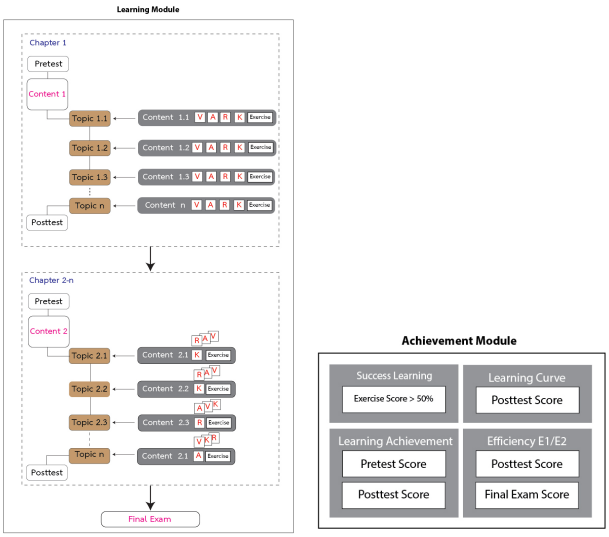


Fig. 1. The concept of developing an adaptive learning system based on proportional VARK to enhance learning achievement.



(a) (b)

Fig. 2. (a) Content Module, (b) Adaptive Module.



(a) (b)

Fig. 3. (a) Learning Module, and (b) Achievement Module.

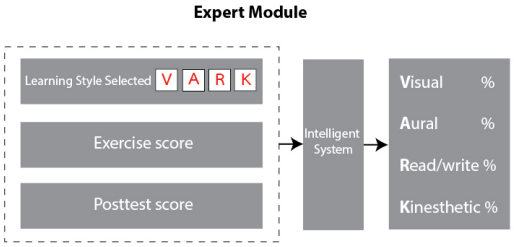


Fig. 4. Expert Module.

learning style and test from the content module to generate the next chapter for the learner, see Fig. 2b.

C. Learning Module

It is a learning module for learners, representing course materials, exercise during the course, pre-test and post-test scores for each lesson, and the final exam score after completing all contents as shown in Fig. 3a.

D. Achievement Module

This module is a collection of behaviour data from learners who use the adapted learning system. The information will be forwarded to the Expert Module for further analysis of learner learning and create a summary report of the system used by the learner as shown in Fig. 3b.

E. Expert Module

This section analyzes behavior of learners who have chosen the learning style in the lesson. Exercise score and post-test score are analyzed as classification of appropriate learning styles for the learner in the next chapter as shown in Fig. 4. As for the Intelligent System part, it uses learner behaviour data mining to classify learning style of the learner such as C4.5 or support vector machines algorithm.

IV. The Expected Result

Researchers strongly suggest that the developed adaptive learning system is able to improve learner achievement because the course content can be adjusted depending on the learner's ability. This research compares the pattern with the samples using the VARK questionnaire and the group using the classifying learning model in the adaptive learning system.

References

[1] VARK Learn. Introduction to VARK. [Online]. Available: http://vark-learn.com (Accessed January 9, 2018).

[2] M. Hasibuan, L. Nugroho, P. I. Santosa, and S. S. Kusumawardani, "A proposed model for detecting learning styles based on agent learning," *Internation Journal of Emerging Technologies in Learning*, vol. 11, pp. 65–69, 2016.

[3] Y. M. Tashtoush, M. Al-Soud, M. Fraihat, W. Al-Sarayrah, and M. A. Alsmirat, "Adaptive e-learning web-based English tutor using data mining techniques and Jackson's learning styles," in *Proceedings of the 8th International Conference on Information and Communication Systems (ICICS'17), Irbid, Jordan, April 4-6*, 2017, pp. 86–91.

[4] M. Abdullah, A. Y.Bayahya, E. S. B. Shammakh, K. A. Altuwairqi, and A. A. Alsaadi, "A novel adaptive e-learning model matching educator-student learning based on machine learning," in *Proceedings of the International Conference on Communication, Management and Information Technology (ICCMIT'16), Prague, Czech Republic, April 25-27*, 2016, pp. 773–782.

978-616-455-375-0

Index