Further Experiments on A Combination of Linear SVM Weight and ReliefF for Dimensionality Reduction

Wipawan Buathong
Department of Information Technology
Phuket Rajabhat University
Phuket, Thailand
w.buathong@pkru.ac.th

Pita Jarupunphol
Department of Information Technology
Phuket Rajabhat University
Phuket, Thailand
p.jarupunphol@pkru.ac.th

***Abstract*— This research further investigated how dimensional data could be efficiently downsized using a multilayered technique based on a combination of two major feature selections, including Linear SVM Weight and ReliefF together with classifier namely Support Vector Machine (SVM). Two datasets, including SRBCT and USPS, were used for the experiment. The results show that the proposed technique is more efficient than using either Linear SVM Weight or ReliefF alone for dimensionality reduction. The dimensional data could be downsized from 2,308 to 8 attributes where the accuracy rate could reach 100 percent in SRBCT. The experimental result of SBRCT was also consistent with that of USPS in which the dimensional data could be downsized from 256 to 55 attributes with the accuracy of 95.76 percent.**

Keywords- Dimensionality Reduction; Multilayered; Feature Selection; ReliefF; Linear SVM Weight

I. INTRODUCTION

Data mining has long been recognised among researchers and practitioners as a promising technique for selecting a representative sample (data dimension) from a considerable amount of data [18]. In various cases, determining data dimension confronts excessively large data causing several issues in machine learning modelling, e.g., understandability, learnability, accuracy, and resource consumption issues [19]. In the meantime, dimensionality reduction has also been perceived as a key process to prepare the data [14] that is, the default data is reduced by losing the key characteristics of the smallest data and losing the most accurate results because each data is important to group the data. With careful data selection techniques, useful information can be selected and used as a representation of most of the information. Since data dimension problems are usually found in large data, reducing the data dimension is also required. Data dimensionality reduction often acts in two perspectives: 1) feature selection (reducing the information describing the data to the rest of the key feature attributes in the group) and 2) feature extraction (reducing the number of data items to a list of information that represents the large group data). In addition to feature selection and feature extraction, sampling and clustering are extensively used for dimensionality reduction [1].

Nowadays, a number of research scholars have conducted research on feature selection for enhancing the efficiency of dimensionality reduction. For example, [2] compared the efficiency of dimensionality reduction using the Correlation-Based Feature Selection (CBFS), Gain Ratio, and Information Gain. The results were used as inputs for vector support vector machine (SVM). It was discovered that the Gain Ratio and Information Gain methods outperformed CBFS. This is consistent with [3] who studied the classification of document categories using the simulation model, together with the two feature selections, including Gain Ratio and Chi Squared, to classify the documents using Bayesian Data Classifier, SVM, and Decision Tree. The results showed that the selection of information gain and the use of vector data-classifiers provided the best performance [4]. The efficiency of feature selection techniques for data dimensionality reduction was measured in Buathong [16]. The Information Gain, Gain Ratio, and Linear SVM Weight were measured based on four classifiers (i.e., k-NN (k-Nearest Neighbor), Naive Bayes, SVM (Support Vector Machine) and Classification Tree). The author summarised that the Linear SVM Weight with the SVM classifier was the most efficient technique for dimensionality reduction. In terms of Web mining, using ReliefF was faster and more accurate than Hidden Naïve Bayes for dimensionality reduction of Content-Based Image [5]. In addition, a number of data dimensionality reduction techniques have been proposed in various articles. Some of them were based on a single layer or a single technique, which might ignore useful dimensions.

In Buathong and Meesad [17], a combined Linear SVM Weight and ReliefF was proposed as an effective technique for data dimensionality reduction. Nevertheless, the authors suggested that more experiments on different datasets would be required to ensure that the efficiency of the proposed technique was trustworthy, not biased. In particular, the experimented datasets attributes should be varied to measure the proposed technique efficiency in different datasets. As such, this research aims to conduct further experiments on a multilayered feature selection technique using a combination of ReliefF and Linear SVM weight. The combined technique was compared with a single feature selection, including Linear SVM weight and ReliefF. Support Vector

Machine was adopted as a classification technique for addressing potential issues that may arise from ignoring useful data dimensions usually discovered in a single layered or a single feature selection. The multilayered feature selection based on a combination of ReliefF and Linear SVM weight can be used to determine what kind of data dimensionality reduction is appropriate for the proposed technique.

II. Literature Reviews

A. Support Vector Machine

SVM is a type of data classification that uses the widest margin principle with a dual-class data classification solution to find the hyperplane to make a decision [6]. SVM divides the data into two parts using a linear equation to divide two different groups of fields and find the best results learning from data statistics based on finding the maximum margin of the decision hyperplane to divide the training data from each other. SVM maps the input space to the feature space and creates a similarity measurement function namely kernel function on the feature space as shown in Figure 1. This type of classification is intended to minimise predictive errors along with the maximized margin, which is different from common techniques such as artificial neural networks (ANN), which is intended to reduce predictive error only. SVM is suitable for data that has a large amount of data dimension.

In case two groups of data cannot be divided by using SVM because the data may be clustered in different positions, a tool to keep the data according to a particular sequence in a higher dimension space is required. A group of data from a multiple-dimensional plane is divided by using the kernel function to provide the better performance of dimensionality reduction. As such, SVM has become a widely recognised technique to simulate a machine learning model due to its high accuracy for data classification [7][8][9][10].

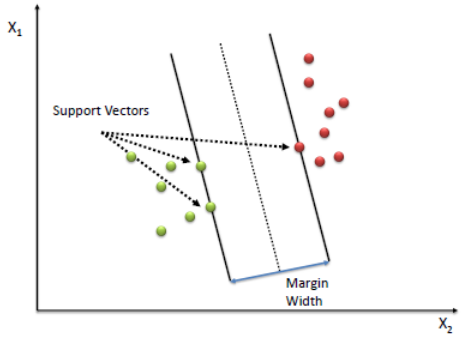

Fig. 1 Data Classification using Linear SVM.

B. Linear SVM Weight

Linear SVM Weight is a method of feature selection, which can be applied to all classifiers [11]. From the algorithm described in Figure 2, the input was a training data and the output has a structured data attribute based on certain steps: 1) uses a grid search to obtain the best C parameter value; 2) builds the model using the function L2-loss linear with data based on the best C parameter values from 1; and and 3) sorts the data attributes based on the absolute values of the available weight [15].

Algorithm 1 Feature Ranking Based on Linear SVM Weights
Input : Training sets, (X_i, Y_i), i =1,…..m.
Output : Sorted feature ranking list.
1. Use grid search to find the best parameter C.
2. Train aL2-loss linear SVM model using the best C.
3. Sort the features according to the absolute values of weights in the model.

Fig. 2 Linear SVM Weight algorithm.

C. ReliefF

ReliefF was developed from Relief, where there is a disadvantage that can be performed with only two types of data, including nominal and numeric data. Since ReliefF is a method of selecting a feature that can be done with data with more than two classes, it is available for all types of data and resistant to incorrect and incomplete data [12]. The ReliefF algorithm was invented by [13] is another feature selection algorithm that calculates the weight from data considerations similar to the random data. K-NN techniques are also taken into account. To find K values, the algorithm will start searching within the same class. The ReliefF algorithms are described in Figure 3.

Algorithm ReliefF
Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes
1. set all weights W[A] :=0.0;
2. **for** i:=1 **to** m **do begin**
3. randomly select an instance R_i;
4. find k nearest hits H_j ;
5. **for** each class $C \neq class(R_i)$ **do**
6. from class C find k nearest misses $M_j(C)$;
7. **for** A :=1 **to** a **do**
8. $w[A] := w[A] - \sum_{j=1}^{k} \frac{diff(A,R_i,H_j)}{(m.k)}$
$+ \sum_{C \neq class(R_i)} \frac{(P(C) * \sum_{j=1}^{k} diff(A,R_i,M_j(C)))}{(m.k)}$;
9. **end;**

Fig. 3 ReliefF algorithm.

D. Combined Linear SVM Weight and ReliefF

A combination of Linear SVM Weight and ReliefF feature selections was introduced by Buathong and Meesad [17] as a plausible technique for enhancing the dimensionality reduction efficiency. Leukemia and DLBCL from UCI Machine Learning Repository were two datasets used in the experiment. It was discovered that the proposed technique was more efficient than using either Linear SVM Weight or ReliefF alone when original data dimensions of the Leukemia dataset were reduced from 5,147 to 20. Furthermore, all performance assessment criteria of the combined method could reach 100% for the Leukemia dataset. At the same data dimensions for the DLBCL dataset, the combined method also showed satisfactory results for all performance evaluation criteria, which were higher than those in other feature selection techniques. According to Buathong and Meesad [17], Linear SVM was still an efficient feature selection technique for dimensionality reduction, since its accuracy performance was higher than that of the combined method when data dimensions were downsized to 10. Figure 4 represented methodological steps in the experiments of Buathong and Meesad [17].

III. Methodology

Three methodological steps involved in this research are similar to those conducted by Buathong and Meesad [17].

1) Data Selection

There are two datasets obtained from http://orange.biolab.si/datasets.php, including SRBCT (4 classes, 2,308 attributes) and USPS (10 classes, 256 attributes). All of the selected datasets have no missing values and contain more than 100 attributes for the data to be suitable for dimensionality reduction, see Table 1.

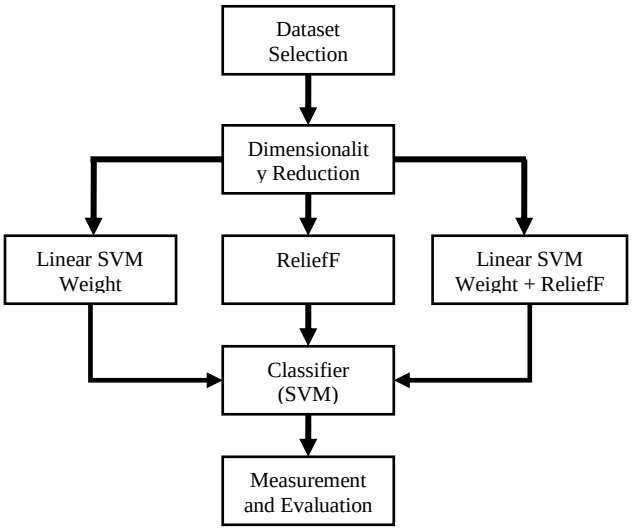

Fig 4. Methodological steps in the research.

TABLE I. Datasets for the experiment

Datasets	*Attributes*	*Classes*
SRBCT (Lymphoma)	2,308	4
USPS(handwritten digit recognition)	256	10

2) Dimentionality Reduction

The Orange Canvas version 2.72 was used to downsize data using feature selection techniques, including ReliefF, Linear SVM weight, and a combination of Linear SVM weight and ReliefF. The SVM was used to classify data together with the RBF kernel. The downsized data dimension was applied to build the machine learning model. For the multilayered technique based on a combination of Linear SVM and ReliefF, Linear SVM was applied in the first layer and followed by ReliefF assigned for the second layer.

3) Measurement and Evaluation

For dimensionality reduction measurements, the multilayered technique based on a combination of SVM and ReliefF was compared with two single layered techniques SVM and ReliefF separately. The data were classified using the 10-folds Cross-validation to test the performance of the machine learning model by dividing the series into k series equally. By using the number of k-1 series to create the learning model, the reserved one piece of data is used to test the accuracy and the process will repeat until all of the divided data is tested for the machine learning model accuracy. The accuracy values and the errors of each round will be summarised and calculated for the average to reflect the learning model efficiency. For the model performance assessment, precision, recall, and f-measure were evaluation criteria. The accuracy result of each round is calculated for the average precision of the equation (1-4).

$$\text{Precision} = (\text{Precision(TP)}+\text{Precision(TN)})/N \quad (1)$$

$$\text{Recall} = (\text{Recall(TP)}+\text{Recall(TN)})/N \quad (2)$$

$$\text{F-measure} = (N \times (\text{Recall} \times \text{Precision}))/(\text{Recall}+\text{Precision}) \quad (3)$$

$$\text{Accuracy} = (TP+TN)/(\text{All Data}) \quad (4)$$

Given that N = the number of classes
TP = the value of true positive
TN = the value of true negative

IV. Results and Discussion

The proposed multilayered technique based on a combination of Linear SVM Weight and ReliefF was the most efficient for dimensionality reduction. The data dimensions were reduced to 60 attributes. While several

research studies indicated that the most effective range of data dimensionality reduction is between 20 and 60 attributes. From the SRBCT dataset experiment, the multilayered technique could yield a maximum of 100 percent accuracy at 8 attributes. While the single-layered feature selection using Linear SVM Weight generated a maximum accuracy of 100 percent at 11 attributes, ReliefF provided the accuracy at 98.89 percent at 31 attributes. With the USPS dataset, the multilayered feature selection also provided more accurate classification than the single feature selection providing 95.76% accuracy at 55 dimensions from 256 dimensions. The results can be displayed in Figures 5 and 6, respectively.

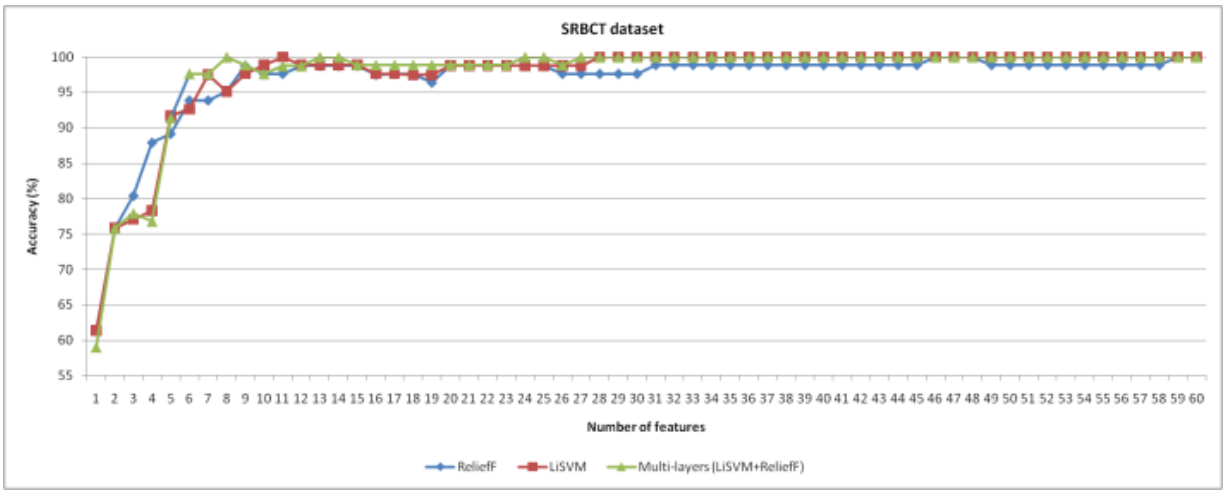

Figure 5. Feature Selections on SRBCT dataset.

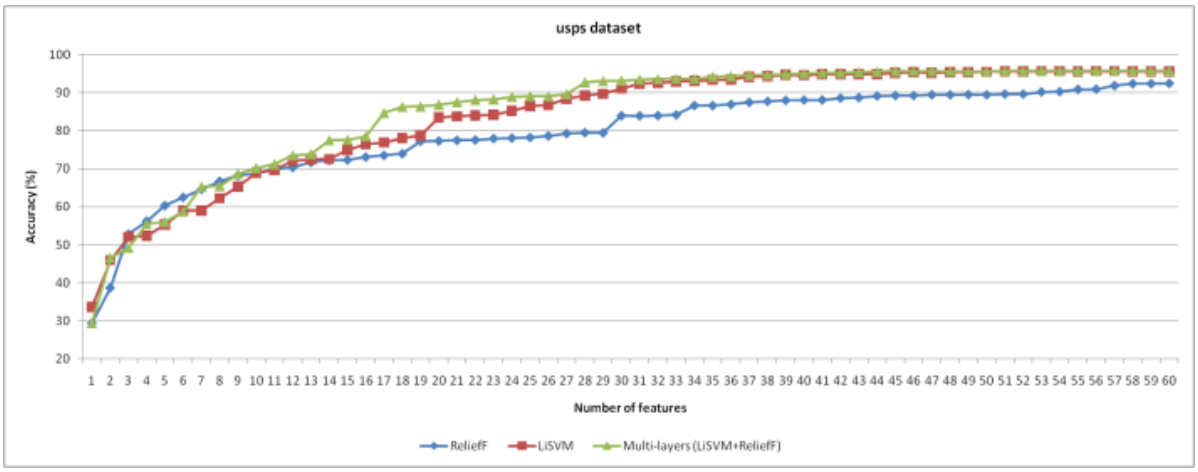

Figure 6. Feature Selections on USPS dataset.

V. CONCLUSIONS

The article has illustrated how the high dimensional data with different data classes could be downsized using the multilayered feature selection based on Linear SVM and ReliefF. It is reasonable to summarise that the multilayered technique could provide better data classification accuracy than the chosen single techniques, including Linear SVM weight and ReliefF. The proposed technique could also work well with large data, as it did not require a lot of time to calculate. In this case, this technique would be beneficial to any disciplines associated with large and high dimensional data, e.g., medical data. geographical data, etc.

REFERENCES

[1] Mohd Afizi Mohd Shukran Omar Zakaria, Noorhaniza Wahid Ahmad and Mujahid Ahmad Zaidi., "A Classification Method for Data Mining Using Svm-Weight And Euclidean Distance," Australian Journal of Basic and Applied Sciences, pp. 2053-2059, 2011.

[2] P. Saengsiri, S. Na Wichian and P. Meesad, " Classification of Leukemia Data Using Ranking and Support Vector Machine," Khon Kean University Research Journal, pp.10–17, April - June, 2018.

[3] W. Sriurai, P. Meesad and C. Haruechaiyasak, "Feature-Based Content Modeling for Document Classification," 5th National Conference on Computing and Information Technology, Bangkok, pp. 146-151, May, 2009.

[4] Abdolhossein Sarrafzadeh, Habibollah Agh Atabay, Mir Mosen Pedram, Jamshid Shanbehzadeh, "ReliefF Based Feature Selection In Content-Based Image Retrieval," in Proceeding of International MultiConference of Engineers and Computer Scientists , Hong Kong, 2012.

[5] Xin Jin, Rongyan Li, Xian Shen, Rongfang Bie, "Automatic Web Pages Categorization with ReliefF and Hidden Naïve Bayes," ACM: Associate for Computing Machinery, pp. 617-621, 2007.

[6] Hsu, C.-w., Chang, C.-c., & Lin, C.-j , "A practical guide to support vector classification," Department of Computer Science and Information Engineering, National Taiwan, Taiwan, 2010.

[7] Huang, C.-L., Chen, M.-C., & Wang, C.-J, "Credit scoring with a data mining approach based on support vector machines," Expert Systems with Applications, pp. 847-856, 2007.

[8] Xuehua, L., & Lan,S, "Fuzzy Theory Based Support Vector Machine Classifier," Fuzzy Systems and Knowledge Discovery, 2008.

[9] Sweilam, N. H., Tharwat, A. A., & Abdel Moniem, N. K, "Support vector machine for diagnosis cancer disease: A comparative study," Egyptian Informatics Journal, pp. 81-92, 2010.

[10] Chen, H-L., Yang, B., Liu, J., & Liu, D.-Y, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Systems with Applications, pp. 9014-9022, 2011.

[11] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V, "Gene selection for cancer classification using support vector machines," Machine learning, pp. 389-422, 2002.

[12] A, Sarrafzadeh., H.A, Atabay., M.M, Pedram., and J., Shanbehzadeh, "ReliefF Based Feature Selection In Content-Based Image Retrieval," in Proceeding of International MultiConference of Engineers and Computer Scientists, 2012.

[13] Kononenko, Igor, and Edvard Simec, "Induction of decision trees using RELIEF," Mathematical and Statistical Methods in Artificial Intelligence, pp. 363-375, 1995.

[14] Tan P. N., Steinbach, M., and Vipin, K. , Introduction to Data Mining, United State of America: Addison Wesley, 2005.

[15] Yin-Wen Chang, Chih-Jen Lin, "Feature Ranking Using Linear SVM," in Workshop and Conference Proceedings, 2008.

[16] Buathong, W. "TLiSVM (Triple Linear SVM Weight) for Dimensionality Reduction" The 11th International Conference on Computer Science & Education (ICCSE 2016). Nagoya, Japan, 2016. pp. 273-278.

[17] Buathong, W. and Meesad, P. "Enhancing the Efficiency of Dimensionality Reduction using a Combined Linear SVM Weight with ReliefF Feature Selection Method." The 9th International Conference on Computing and Information Technology (IC2IT 2013). Bangkok, Thailand, 2013. pp. 125-134.

[18] Larose, Daniel T, Discovering knowledge in data : An Introduction to Data Mining, New Jersey: John Wiley & Sons, 2014.

[19] J. Dean, Big Data, Data Mining, and Machine Learning, New Jersey: John Wiley & Sons, 2014.

Please fill in the Authors' background:

Position can be chosen from:

~~**Prof.**~~ **/** ~~**Assoc. Prof.**~~ **/** ~~**Asst. Prof.**~~ **/ Lect. /** ~~**Dr.**~~ **/** ~~**Ph. D Candidate**~~ **/** ~~**Postgraduate**~~ **/** ~~**Ms.**~~

Full Name	Email address	Position	Research Interests	Personal website (if any)
Wipawan Buathong	w.buathong@pkru.ac.th	Lecturer of Department of Information Technology	Data Mining	-

*This form helps us to understand your paper better; the form itself will not be published.