A Method to Clustering the Feature Ranking on Data Classification Using an Ensemble Feature Selection

Nuntawut Kaoungku, Kittisak Kerdprasop, and Nittaya Kerdprasop

***Abstract*—The aim of this paper is to improve the predictive performance of the classification process by means of building multiple data classification models based on the output from feature selection methods that use ensemble strategy to find the optimal set of features. Currently, the data volume has grown at an extreme rate causing a variety of problems. The big data situation has made automatic analysis tasks such as data classification facing low performance and high computational time problems when dealing with big data that are huge in both volume and dimensions. In this research work, we tackle the big data problem in the high dimensionality aspect. We propose an ensemble method to reduce data dimension by means of feature clustering to rank the potential features and also return suitable subset of features for further classifying the training data. The two paradigms of feature selection based on ensemble strategy are proposed and evaluated. Experimental results confirm the efficacy of our proposed feature ensemble method.**

***Index Terms*—Feature selection, ensemble learning, clustering, classification.**

I. Introduction

Traditional data classification seems to be an easy and straightforward task when applying a single classification model to predict future data. Currently, electronic equipments are ubiquitous and extensively used, thus, causing a variety of data forms such as numeric, categorical, time series, images, and so on. It is difficult to build a single model from these data to make a high performance classifier for accurately predicting future or unseen data. The basic solution idea is to build multiple models from the same dataset and then combine the predicted results from those multiple models to output a final prediction. This technique is called an ensemble learning.

Ensemble learning is basically a technique to use multiple models or multiple learning algorithms to predicted future data with the major purpose of better classification in terms of accuracy. Which combining results from multiple models built from various methods, the popular result combining method is a simple voting [1]. Typically, ensemble learning can be achieved from a wide range of methods, but the popular methods are bagging [2] and booting [3]. The two ensemble methods have long applied by many researchers, and they have been proven to provide better classification performance.

From the continuous and increasing advancement of software and hardware technologies, new structured and unstructured data have been generated every day. It is difficult to analyze and build a model from these mixed type data, even with the aid of ensemble learning method, because these data are high in dimensionality. The technique to solve this problem is the use of filter of find and extract only the optimal set of features for building classification model. The filtering techniques can be generally divided into 2 groups: feature selection and feature extraction. The research focusing on feature selection method uses some measures to calculate weight and then choosing a subset of features ordered by the weight [4], [5]. The feature selection methods can be further divided into 2 sub-groups: those that automatically return optimal set of features, and those that return weight of features. It is, however, difficult to choose the optimal weight of features for data classification.

Therefore, many researchers try to solve the optimal feature selection problem by proposing the ensemble feature selection method. Bolón-Canedo et al. [6] have shown that data classification using an ensemble of filters by using five groups of different feature selection methods for building instance-based learning (IB1) model [7] and support vector machine (SVM) model [8]. Seijo-Pardo et al. [9] propose technique to select optimal set of features by using several different threshold values, such as fisher discriminant ratio, $\log_2(n)$, and top percent of features, for ensemble feature selection. These research works [6]-[9] report a promising performance of ensemble feature selection strategy to increase classification accuracy.

This research, thus, aims at proposing a method to improve data classification accuracy by means of an ensemble feature selection. We propose a hybrid ensemble feature selection method by both automatically return optimal set of features and return weight of features. Our proposed method selects the optimal set of the feature from the return weight of features reported by the clustering method using k-Means algorithm [10], [11].

The contributions of this paper are as follows:

- With the proposed method, k-Means clustering can be applied as feature selection tool to select the optimal subset of the features.
- The proposed method can be applied to ensemble learning using a variety of learning algorithms and can

Manuscript received February 15, 2017; revised April 10, 2017. This work was supported in part by grants from Suranaree University of Technology through the funding of Knowledge and Data Engineering Research Units.

N. Kaoungku is with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nuntawut@sut.ac.th).

K. Kerdprasop is with the School of Computer Engineering. He is also with Knowledge Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: kerdpras@sut.ac.th).

N. Kerdprasop is with the School of Computer Engineering. She is also with Data Engineering Research Unit, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: nittaya@sut.ac.th).

doi: 10.18178/ijfcc.2017.6.3.494

increase the predictive accuracy.

II. Materials and Methods

A. Ensemble Learning

Ensemble learning is a technique to build multiple models from training data. The main purpose of this technique is to increase the model accuracy. Fig. 1 shows the main concept of ensemble learning. The ensemble process starts by taking the training data to build multiple models using either the same algorithm, or different algorithms. Then, combine the results from all the models to generate a single output. There are various strategies to combine results, but the most applicable one is a majority vote.

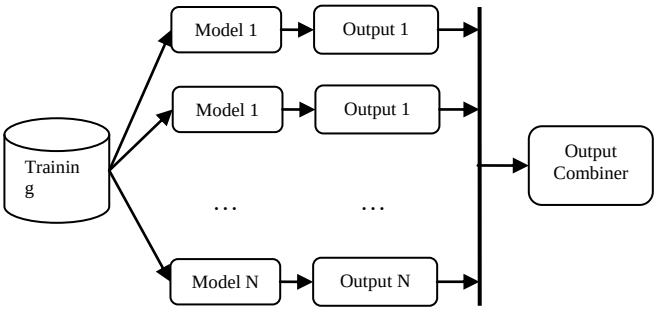

Fig. 1. The concept of ensemble learning.

Ensemble learning can be divided further into three classes of techniques [1]:

- Vote ensemble. It performs ensemble learning by building multiple models from one training dataset. To classify new data, it uses a majority vote to predict class of the new data.
- Bagging. It starts ensemble learning by dividing data, using random sampling technique, into several equal subsets. Each data subset is used to build the model. All built models are then used for classifying new data based on a majority vote.
- Random forest. It is ensemble learning method that is similar to bagging technique but it selects some of the features to each data subset.

B. Feature Selection Method

Feature selection is a method to handle high dimensional data by reducing the data features based on some selection criteria. This method can reduce data dimensions and at the same time can increase the model accuracy. The examples of criteria for selecting feature subsets and returning the feature ranking score are as follows:

- Association rule mining-based feature selection (AFS) [12]. It is a method based on association analysis for analyzing features that are most influencing the class attribute. The calculation of frequent features from association rules is shown in equation (1). If the feature has the highest *FrequentFeature* score, that feature is the most influencing factor to the class attribute.

$$FrequentFeature(A) = \frac{AppearFrequency(A)}{\#Rules} \quad (1)$$

- Information Gain (IG) [13]. It selects features by measuring entropy, which is the measurement for purity of data with the same class. The computation of IG is shown in equations (2) and (3). The feature with high value of IG means the high potential of that feature on classifying data into class c_1 to c_n.

$$InfoGain = Entropy(initial) - [P(c_1) \times Entropy(c_1) + ... + P(c_n) \times Entropy(c_n)] \quad (2)$$

where

$$Entropy\ (c_1, c_2, ..., c_n) = -P(c_1)\log_2 P(c_1) - P(c_2)\log_2 P(c_2) - ... - P(c_n)\log_2 P(c_n) \quad (3)$$

C. k-Means Clustering

k-Means clustering [10], [11] is an unsupervised learning algorithm for partitioning data into groups such that data subsets sharing similar attributes are assigned to be in the same group. This algorithm groups data into clusters by measuring the distance between data points. The most popular measure is Euclidean distance [14], as shown in equation (4).

$$dist(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (4)$$

Fig. 2 presents the detail of the k-Means algorithm, which consists of five steps.

Step 1: at line 1, define the number of clusters (K) and the initial centroid, or central point, of each cluster.

Step 2: from lines 2 to 3, assign all data points to the closest centroid by measuring the distance between a data point to each centroid.

Step 3: at line 4, recompute the centroid of each cluster by calculating the average attribute value among all the points in each cluster.

Step 4: repeat steps 2 and 3 until the centroid does not change.

Algorithm k-Means

1. Select K point as the initial K centroids.
2. Repeat
3. Form K clusters by assigning all points to the closet centroid.
4. Recomputed the centroid of each cluster.
5. Until the centroid does not change

Fig. 2. k-Means algorithm.

III. Proposed Work

In this section, we present the proposed process of clustering the feature ranking on data classification using an ensemble feature selection. The idea is that we use the k-Means algorithm to find the best cluster of the features from feature ranking scores and use these results to build the model for data classification. The objectives are to reduce the data dimensions and to increases the predictive accuracy.

The method of clustering the feature ranking on data classification using an ensemble feature selection is graphically shown in Fig. 3 and 4.

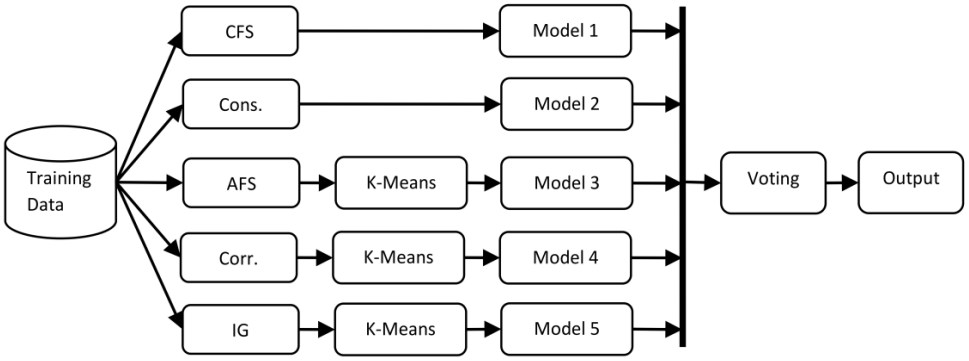

Fig. 3. The concept of ensemble 1.

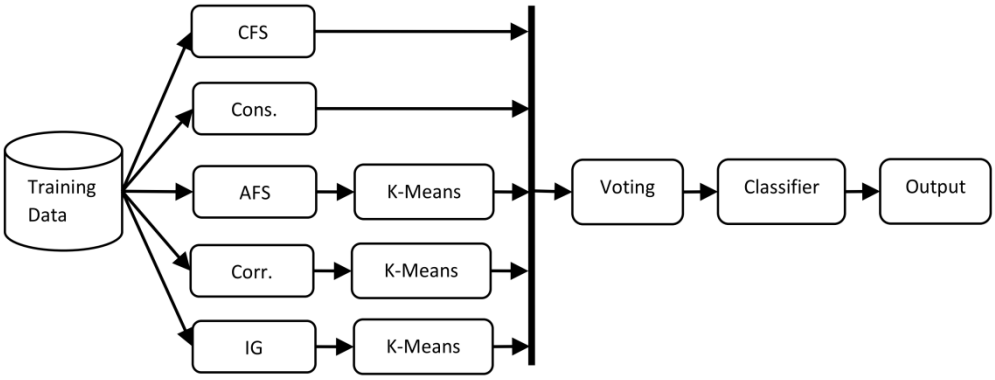

Fig. 4. The concept of ensemble 2.

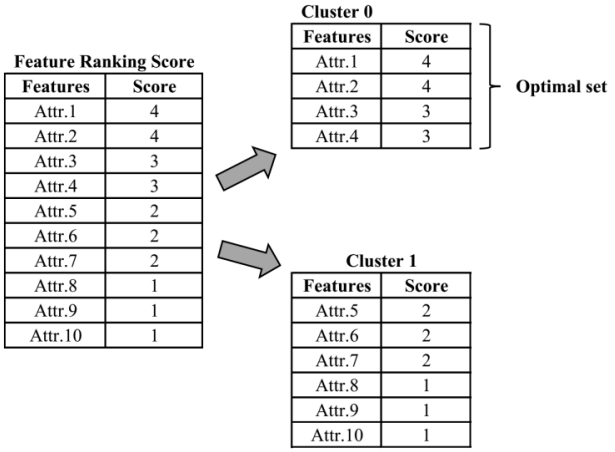

Fig. 5. The concept of cluster the feature ranking score.

Our method consists of two parts: ensemble 1 and ensemble 2. Fig. 3 shows the steps in ensemble 1, which consists of three phases. The first phase of ensemble 1 feature selection method is reducing dimensions of the training data by using 5 feature selection methods including the correlation-based feature selection (CFS) [15], the consistency-based filter (Cons.) [16], the association rule mining-based feature selection (AFS), the correlation-based filter (Corr.), and the information gain (IG).

Ensemble 1 phase 2 is the clustering of feature ranking scores with the k-Means algorithm. The three ranking score methods used for clustering are the scores from the AFS, Corr, and IG methods. Fig. 5 shows running example for phase 2 of ensemble 1. The feature weight in Attr.1 to Attr.10 are clustered by k-Means (set k=2, user can increase k when the optimal set of feature is needed to be small size). The optimal set of the feature is cluster 0, which contains a set of features to be used for building a classification model.

Ensemble 1 phase 3, build the model with an optimal set of the feature from phase 2 with any data classification algorithm. The final step of ensemble 1 is the combining of the outputs from multiple models using a majority vote scheme to predicted class of data.

Fig. 4 shows concept of ensemble 2, which shares similar main idea to ensemble 1, but the ensemble 2 build a single classifier. At phase 3 of the ensemble 2 method, a majority vote of the feature from several feature selection methods generate a single set of optimal features. Then the optimal set of features is used for classification model building. The classification algorithm can be any one such as SVM and C4.5.

IV. Experimental Results

The proposed ensemble feature selection methods have been experimented with data taken from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Table I shows details of the five data sets used in our experimentation. Each of these datasets has been divided into training dataset (70%) and test dataset (30%). We use the C4.5 and SVM algorithms for classification and use five feature selection methods, which are correlation-based feature selection (CFS), consistency-based (Cons.), association rule mining-based feature selection (AFS), correlation-based (Corr.), and information gain-based (IG).

Table II shows comparative results of classification accuracy and error. It can be seen that the ensemble 1 on C4.5

algorithm can improve the performance of accuracy on Spambase (92.72%) and Arrhythmia (66.91%) data sets. The proposed ensemble 1 and 2 on SVM algorithm can improve the performance of accuracy on Splice (96.68%) data set when compared to raw data set with no feature selection method and other feature selection algorithms.

Table III shows comparative results of average classification accuracy and error. It can be seen that the ensemble 1 performs well on the C4.5 algorithm (87.66%) when compared to ensemble 2 that is good on SVM algorithm (88.43%). When compared against other feature selection algorithms, it can be seen that our proposed ensemble feature selection algorithms using k-Means to cluster feature ranking can improve the performance of accuracy on C4.5 (IG = 86.63%) and SVM (Corr. = 88.39%). Table IV shows comparative results of the number of features selected by five feature selection algorithms. It can be seen the all five feature selection methods can reduce data dimensions.

TABLE I: DETAILS OF DATASETS

Datasets	# Instances	# Features
Spambase	4601	58
Splice	3190	62
Opt digits	5620	65
Ozone	2534	74
Arrhythmia	452	280

TABLE II: COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY AND ERROR

Methods	Spambase		Ozone		Splice		Arrhythmia		Opt digits	
	Acc.	Err.	Acc.	Err.	Acc.	Err.	Acc.	Err.	Acc.	Err.
Raw Data										
Raw + SVM	92.65%	7.35%	**94.00%**	6.00%	66.35%	33.65%	60.29%	39.71%	**98.92%**	1.08%
Raw + C4.5	92.13%	7.87%	92.09%	7.91%	93.57%	6.43%	62.50%	37.50%	89.70%	10.30%
SVM										
CFS + SVM	88.75%	11.25%	92.36%	7.64%	**96.57%**	3.43%	63.97%	36.03%	98.73%	1.27%
Cons. + SVM	89.71%	10.29%	93.86%	6.14%	93.03%	6.97%	60.29%	39.71%	86.33%	13.67%
AFS + SVM	89.63%	10.37%	93.45%	6.55%	94.75%	5.25%	63.77%	36.23%	**98.98%**	1.02%
Corr. + SVM	90.59%	9.41%	94.00%	6.00%	96.14%	3.86%	62.50%	37.50%	98.73%	1.27%
IG + SVM	88.60%	11.40%	93.45%	6.55%	96.46%	3.54%	62.50%	37.50%	98.61%	1.39%
C4.5										
CFS + C4.5	91.91%	8.09%	91.27%	8.73%	93.57%	6.43%	**66.18%**	33.82%	89.88%	10.12%
Cons. + C4.5	**92.06%**	7.94%	91.95%	8.05%	93.25%	6.75%	61.76%	38.24%	80.48%	19.52%
AFS. + C4.5	**92.06%**	7.94%	93.04%	6.96%	93.68%	6.32%	62.50%	37.50%	90.42%	9.58%
Corr. + C4.5	91.47%	8.53%	**94.13%**	5.87%	94.00%	6.00%	60.29%	39.71%	90.48%	9.52%
IG + C4.5	91.91%	8.09%	93.18%	6.82%	94.00%	6.00%	65.44%	34.56%	88.61%	11.39%
Ensembles										
Emsemble1 + SVM	90.07%	9.93%	**93.45%**	6.55%	95.61%	4.39%	63.24%	36.76%	**98.67%**	1.33%
Emsemble1+ C4.5	**92.72%**	7.28%	93.86%	6.14%	93.68%	6.32%	**66.91%**	33.09%	91.14%	8.86%
Emsemble2 + SVM	89.78%	10.22%	**93.04%**	6.96%	**96.68%**	3.32%	63.97%	36.03%	**98.67%**	1.33%
Emsemble2 + C4.5	91.54%	8.46%	92.77%	7.23%	94.00%	6.00%	65.44%	34.56%	89.28%	10.72%

TABLE III: COMPARATIVE RESULTS OF NUMBER OF FEATURES BY FIVE FEATURE SELECTION ALGORITHMS

Datasets	Raw	A	B	C	D	E
Spambase	58	16	18	15	21	**13**
Splice	62	20	11	**8**	24	23
Opt digits	65	36	**10**	42	40	33
Ozone	74	15	23	26	23	**13**
Arrhythmia	280	32	**19**	50	46	25

TABLE IV: COMPARATIVE RESULTS OF AVERAGE ACCURACY AND ERROR

Methods	SVM		C4.5	
	Accuracy	Error	Accuracy	Error
Raw data	82.44	17.56	86.00	14.00
Features Selection				
CFS	88.08	11.92	86.56	13.44
Cons.	84.64	15.36	83.90	16.10
AFS	88.12	11.88	86.34	13.66
Corr.	**88.39**	**11.61**	86.07	13.93
IG	87.92	12.08	**86.63**	**13.37**
Ensembles				
Ensemble 1	88.21	11.79	**87.66**	**12.34**
Ensemble 2	**88.43**	**11.57**	86.61	13.39

V. CONCLUSION

This research aims at studying a method to clustering the feature ranking on data classification using an ensemble feature selection. The problem of learning efficient model from data with high dimensionality can cause trouble to most algorithms. Thus, we propose to use the ensemble method at the feature selection step prior to the application of learning algorithm in order to increase accuracy and reduce learning problem due to dimensionality. We present clustering method using the k-Means algorithm to cluster the feature ranking scores for choosing an optimal set from feature ranking score.

From experimental results, it has been revealed that the proposed ensemble feature selection method can increase the accuracy of data classification, and can reduce high dimensional data problem by obtaining a small set of features. However, in some datasets our proposed ensemble method shows lower accuracy than the raw dataset with no feature selection applied. Even though the proposed method can

reduce data dimensions and hence expected to remedy the over-fitting problem, it still needs further improvement to perform well on every dataset. Such improvement is obviously planned as our future work.

REFERENCES

[1] T. G. Dietterich, "Ensemble methods in machine learning," *International Workshop on Multiple Classifier Systems*, pp. 1-15, June 2000.
[2] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
[3] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197-227, 1990.
[4] I. Guyon, "Practical feature selection: from correlation to causality," *NATO Science for Peace and Security*, vol. 19, pp. 27-43, 2008.
[5] H. G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. the Eleventh International Conference on Machine Learning*, pp. 121-129, 1994.
[6] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Data classification using an ensemble of filters," *Neurocomputing*, vol. 135, pp. 13-20, 2014.
[7] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37-66, 1991.
[8] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods," Cambridge University Press, 2000.
[9] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos, "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp. 124-139. 2017.
[10] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
[11] M. R. Anderberg, "Cluster analysis for applications," *Academic Press*, 1973.
[12] N. Kaoungku, K. Kerdprasop, and N. Kerdprasop, "Data classification based on feature selection with association rule mining," *The International Multi Conference of Engineers and Computer Scientists 2017*, 2017.
[13] M. A. Hall, "Smith, Practical feature subset selection for machine learning," *Comput. Sci*, vol. 98, pp. 181–191, 1998.
[14] M. M. Deza and E. Deza, "Encyclopedia of distances," *Encyclopedia of Distances*, pp. 1-583, 2009.
[15] M. A. Hall, "Correlation-based feature selection for machine learning (Ph.D. thesis)," *Citeseer*, 1999.
[16] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell*, pp, 155–176, 2003.

Nuntawut Kaoungku is currently a lecturer at School of Computer Engineering, Suranaree University of Technology, Thailand. He received his doctoral degree in computer engineering from Suranaree University of Technology, Thailand, in 2014, the bachelor degree in computer engineering from Suranaree University of Technology, Thailand, in 2012, and the master degree in computer engineering from Suranaree University of Technology, Thailand, in 2013. His current research includes data mining and semantic web.

Kittisak Kerdprasop is an associate professor and chair of the School of computer engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and the doctoral degree in computer science from Nova Southeastern University, U.S.A., in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.

Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, the master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and the doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, logic programming, and intelligent databases.