Received: 4 October 2019 | Revised: 8 January 2020 | Accepted: 6 March 2020

DOI: 10.1111/exsy.12553

ORIGINAL ARTICLE

Expert Systems WILEY

Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results

Chih-Wen Chen[1,2,3] | Yi-Hong Tsai[1] | Fang-Rong Chang[1,4,5] | Wei-Chao Lin[6,7,8]

[1]Graduate Institute of Natural Products, College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

[2]Department of Pharmacy, Kaohsiung Municipal Chinese Medical Hospital, Kaohsiung, Taiwan

[3]Department of Nursing, Fooyin University, Kaohsiung, Taiwan

[4]Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung, Taiwan

[5]National Research Institute of Chinese Medicine, Ministry of Health and Welfare, Taipei, Taiwan

[6]Department of Information Management, Chang Gung University, Taoyuan, Taiwan

[7]Healthy Aging Research Center, Chang Gung University, Taoyuan, Taiwan

[8]Department of Thoracic Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan

Correspondence
Wei-Chao Lin, Department of Information Management, Chang Gung University, Linkou, Taoyuan, Taiwan.
Email: viclin@gap.cgu.edu.tw

Abstract

Feature selection is a process aimed at filtering out unrepresentative features from a given dataset, usually allowing the later data mining and analysis steps to produce better results. However, different feature selection algorithms use different criteria to select representative features, making it difficult to find the best algorithm for different domain datasets. The limitations of single feature selection methods can be overcome by the application of ensemble methods, combining multiple feature selection results. In the literature, feature selection algorithms are classified as filter, wrapper, or embedded techniques. However, to the best of our knowledge, there has been no study focusing on combining these three types of techniques to produce ensemble feature selection. Therefore, the aim here is to answer the question as to which combination of different types of feature selection algorithms offers the best performance for different types of medical data including categorical, numerical, and mixed data types. The experimental results show that a combination of filter (i.e., principal component analysis) and wrapper (i.e., genetic algorithms) techniques by the union method is a better choice, providing relatively high classification accuracy and a reasonably good feature reduction rate.

KEYWORDS

ensemble, feature selection, data mining, dimensionality reduction, feature selection, medial datasets

1 | INTRODUCTION

Feature selection (or variable selection) is a very important data pre-processing step in data mining and pattern recognition problems aimed at filtering out unrepresentative features or selecting a subset from a given training dataset. The advantages of performing feature selection include minimization of the curse of dimensionality, the avoidance of overfitting problems, reduction in the training time for model construction, and enhancement in the generalization ability of constructed models (Chandrashekar & Sahin, 2014; Guyon & Elisseeff, 2003; Liu & Yu, 2005).

In many medical domain problems, such as medical imaging, biomedical signal processing, and DNA microarray data, the collected datasets usually contain very high feature dimensions. To deal with this high dimensionality problem, related literature have shown the positive effect of considering feature selection on various medical domain datasets (Huang et al., 2019; Huang, Chen, Lin, Ke, & Tsai, 2017; Remeseiro & Bolon-Canedo, 2019; Shilaskar & Ghatol, 2013; Zhu et al., 2015).

In general, feature selection methods can be classified into three categories, namely, filter, wrapper, and embedded or hybrid methods (Bolon-Canedo, Sanchez-Marono, & Alonso-Betanzos, 2013; Chandrashekar & Sahin, 2014; Kumar & Minz, 2014; Liu & Yu, 2005; Saeys, Inza, & Larranaga, 2007). Kumar and Minz (2014) pointed out a number of related works which have focused on comparing several feature selection methods for different domain problems. Others have proposed novel feature selection algorithms featuring filter (Lim, Lee, & Kim, 2017; Wang,

Wei, Yang, & Wang, 2017), wrapper (Das, Das, & Ghosh, 2017; Zhang et al., 2016), and embedded (Liu, Huang, Meng, Gong, & Zhang, 2016; Zhu, Zhu, Hu, Zhang, & Zuo, 2017) techniques. However, many of these novel algorithms have been developed based only on one type of selection technique, filter, wrapper, or embedded feature selection processes. In other words, the individual selection criteria have limited their ability to identify relevant feature subsets.

Recently, ensemble learning based feature selection or ensemble feature selection methods have been developed, which take advantage of the idea of ensemble learning to overcome the potential local optima problem of single algorithms. The aim of ensemble learning based feature selection is to generate diverse feature subsets from a given training dataset and then aggregate the different feature subsets into a final output. The superiority of ensemble learning based feature selection to single feature selection methods has been shown in many studies (Guan, Yuan, Lee, Najeebullah, & Rasel, 2014).

However, the ways in which the different diverse feature subsets can be generated have not been fully explored. According to a literature review conducted by Guan et al. (2014), most studies have focused only on using different filter-based feature selection algorithms to generate diverse feature subsets. Recently, Seijo-Pardo, Porto-Diaz, Bolon-Canedo, and Alonso-Batanzos (2017) attempted to combine multiple feature subsets generated by filters and embedded feature selection methods. In Seijo-Pardo, Bolon-Canedo, and Alonso-Batanzos (2019), a new automatic threshold is proposed to select the combinations of multiple ranking based feature selection methods. Particularly, four filter and two embedded methods are considered.

In short, one major limitation of related works is that the three different categories of feature selection methods are not all considered in an ensemble feature selection. Moreover, ensemble feature selection has not been examined in related medical domain problems. Therefore, our objective is to examine the performance of ensemble feature selection methods along with different combinations of filter, wrapper, and embedded feature selection methods over different types of medical datasets, including categorical, numerical, and mixed data types. The contribution of this paper is twofold. First, the findings allow us to understand which combination of different types of feature selection methods can perform better than the others for what kind of data types. Second, the best combination can be used as a representative baseline for novel feature selection algorithms proposed in the future.

The rest of this paper is organized as follows. Section 2 briefly describes the three types of feature selection methods. Section 3 describes the procedures for ensemble feature selection including the generation of diverse feature subsets and the aggregation of multiple feature subsets. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper.

2 | LITERATURE REVIEW

2.1 | Filter techniques

In the filter methods, the relevance of features is assessed. Feature relevance is scored by calculating a ranking criterion and low-scoring features that fall below a specified threshold are removed. The well-known statistically oriented filter-based feature selection methods include information gain, stepwise regression, and principal component analysis (PCA) methods, to name a few. The advantages of this type of technique are that they are computationally efficient and independent of the classification/clustering algorithms.

For the example of PCA, it computes new orthogonal variables called principal components by maximizing the variance of the data. These variables are obtained as linear combinations of the original variables. In addition, the values of these new variables are called factor scores and are interpreted geometrically as the projections of the observations onto the principal components. Therefore, for any factor, high loadings in absolute value indicate that corresponding variables contribute more to the factor than other variables (Guo, Wu, Massart, Boucon, & de Jong, 2002; Morchid, Dufour, Bousquet, Linares, & Torres-Moreno, 2014).

2.2 | Wrapper techniques

Unlike the filter techniques, wrapper methods use a predictor (usually based on some supervised learning algorithm) as a black box and the predictor performance is used as an objective function to evaluate the representativeness of the feature subset. In particular, during the search procedure, various feature subsets are generated and evaluated. The evaluation of a specific feature subset is based on training and testing the predictor. Some widely used wrapper-based feature selection methods are based on evolutionary algorithms, such as genetic algorithms (GA) and particle swarm optimization.

For the example of GA, a population of strings (called chromosomes), which encode candidate solutions (called individuals) to an optimization problem, evolves for better solutions. In general, the genetic information (i.e., chromosome) is represented by a bit string (such as binary strings of 0 s and 1 s) and sets of bits encode the solution. Then, genetic operators are applied to the individuals of the population for the next generation (i.e., a new population of individuals). There are three main genetic operators, which are selection, crossover, and mutation. The selection operator

selects individuals in the population based on the individual's fitness. Higher-fitness individuals have more probability to be selected than lower-fitness individuals. A couple of selected individuals are then crossovered to exchange their information with each other to generate new individuals or offspring. Then, the mutation operator is performed by flipping the value 0/1 of the bit of selected individuals. Furthermore, a fitness function is used to measure the quality of an individual in order to increase the probability that the single bit can survive throughout the evolutionary process (Goldberg, 1989).

The advantages of the wrapper methods include the interaction between feature subset searches and predictor selection, and the ability to take into account feature dependencies. However, wrapper methods are usually computationally expensive since training and testing the predictor requires a certain computational cost.

2.3 | Embedded techniques

With the embedded methods, the search for an optimal feature subset is built into the classifier construction, which can reduce the computation time taken up for reclassifying different subsets which are needed in the wrapper methods. In particular, the feature subset search process is incorporated as part of the classifier training process. Some well-known classifiers for embedded feature selection methods are the support vector machine (SVM), artificial neural network, and decision tree (DT) methods.

For the example of the DT, it can select from a large number of explanatory variables that are most important in determining the response variable to be explained. In general, a DT is binary that contains two branches for each node. Particularly, the root node t is separated into two samples based on some condition. The samples that fit the condition will be separated into the left nodes (t_l), and the others will be separated into the right nodes (t_r). P_L and P_R are the paths that the node t goes through t_l and t_r. In particular, a DT is based on the entropy theory that the attribute (or feature) with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the non-leaf node (Quinlan, 1986). As a result, the decision nodes (t, t_l, and t_r) can be regarded as representative features over a given dataset.

3 | ENSEMBLE FEATURE SELECTION PROCEDURES

In this paper, three different types of feature selection methods, filter, wrapper, and embedded methods, are combined. Two ensemble feature selection strategies are used. In the first, two different types of feature selection methods are combined, and in the second, three types of feature selection methods are combined. Figures 1 and 2 show the two ensemble feature selection strategies.

Given a training set containing M dimensional features, each type of feature selection is used to generate a reduced feature set whose dimensions are lower than M. It is a fact that different feature selection methods will produce different reduced feature sets. Union and intersection methods can be applied for the aggregation of different reduced feature sets that contain different selected features. For example, all the selected features in A and B are used in the union of reduced feature sets A and B (i.e., A ∪ B). On the other hand, the intersection of reduced feature sets A and B (i.e., A ∩ B) is only based on overlapping features that appear in both A and B.

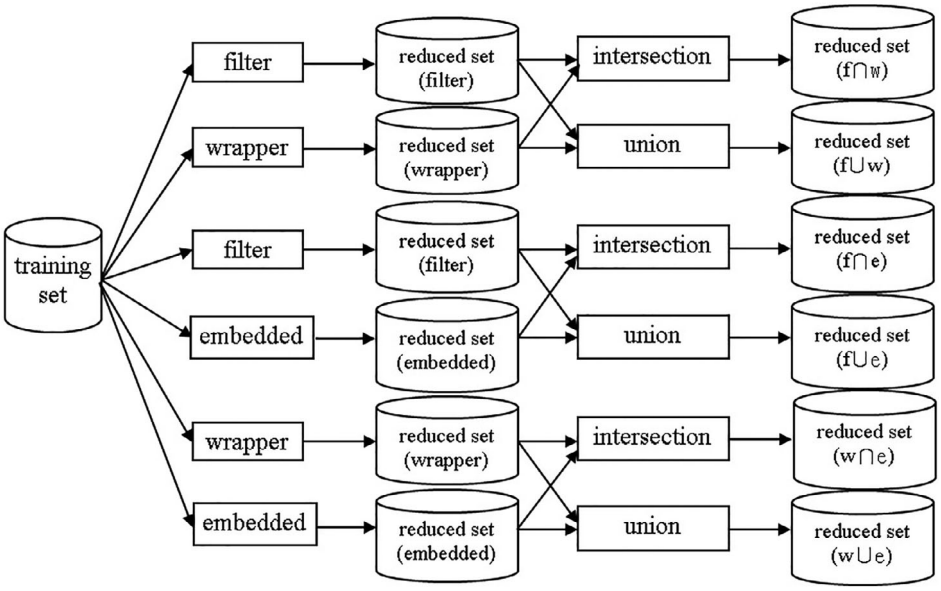

FIGURE 1 Combination of two types of feature selection methods

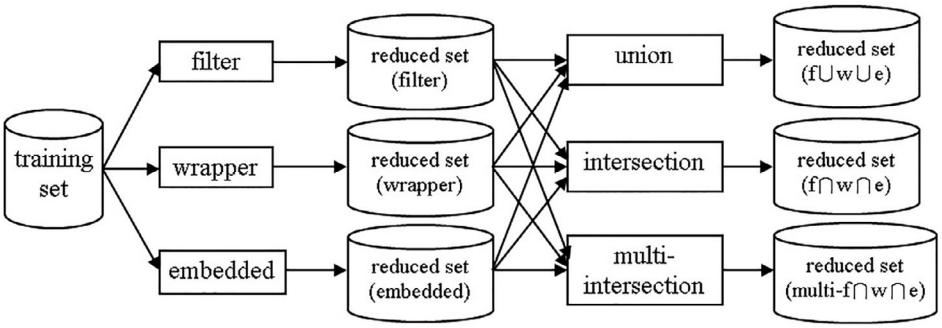

FIGURE 2 Combination of three types of feature selection methods

A third aggregation approach, the multi-intersection method, is also examined for the combination of three different reduced sets. The result is obtained based not just from the intersection of reduced feature sets A, B, and C (i.e., A ∩ B ∩ C), but also the intersection of A and B, A and C, and B and C.

4 | EXPERIMENTAL STUDIES

4.1 | Experimental setup

In this paper, two experimental studies are described. In the first study, six small scale medical datasets from the UCI Machine Learning Repository are used, with numbers of features ranging from 9 to 325. Specifically, there are two categorical, one numerical, and three mixed data types of datasets. The aim of this study is to identify which combination of multiple feature selection methods performs the best. In contrast, in the second experiment, three very high dimensional datasets containing 2,000, 4,322, and 5,966 features, respectively, are used, in order to understand whether the findings are consistent with those obtained in the first study.

The performance of these feature selection methods is evaluated by examination of the classification accuracy and feature reduction rates. According to related literature, the SVM method can provide better classification performance than many other classification techniques in various pattern recognition problems (Byun & Lee, 2003; Chandra & Bedi, 2018; Huang et al., 2017; Mountrakis, Im, & Ogole, 2011; Reddy, Kavya, & Ramadevi, 2014). Therefore, the SVM method based on fivefold cross validation is used in the design of the classifier. Particularly, the SVM classifier is constructed based on the Weka machine learning software with the default parameters of LIBSVM (Chang & Lin, 2011).

Feature selection is performed by PCA, GA, and DT, representing the filter, wrapper, and embedded techniques, respectively, using C4.5. In literature, many related works have considered these feature selection methods for different domain problems (Chaikla & Qi, 1999; Guo et al., 2002; Kazemitabar, Amini, Bloniarz, & Talwalkar, 2017; Morchid et al., 2014; Sindhiya & Gunasundari, 2014; Sugumaran, Muralidharan, & Ramachandran, 2007; Tsai & Hsiao, 2010). With PCA, three different selection criteria are compared, keeping the top 80, 65, and 50% of feature variances. In this study, we found that keeping the top 80% of feature variances performs the best. Therefore, we only report the results obtained using this selection criterion. The parameters for GA are based on Grefenstette (1986) where the number of iterations, population size, crossover rate, and mutation rate are 30, 30, 0.9, and 0.01, respectively.

4.2 | Experimental results

4.2.1 | Small scale datasets

Table 1 lists the classification performance results obtained with SVM using different single feature selection algorithms and ensemble feature selection methods over the six small scale datasets. (Note that E, F, and W indicate embedded, filter, and wrapper approaches, respectively.) In addition, the numbers in brackets indicate the number of selected features. In addition, the numbers following each dataset give its original feature size.

For the single feature selection methods, on average, GA outperforms PCA and C4.5, which can make the SVM classifier produce the highest rate of classification accuracy, that is, 0.8. Based on the Wilcoxon signed-rank test (Demsar, 2006), the performance differences between GA, PCA, and C4.5 are significant ($p < .05$).

However, it is found that the ensemble feature selection methods used in combination with the intersection method do not perform better than the single best feature selection method, that is, GA. In contrast, ensemble feature selection techniques using multi-intersection perform

TABLE 1 Classification accuracy of support vector machine and number of selected features obtained by different feature selection methods over the six small scale datasets

				Ensemble feature selection								
	Single feature selection			Intersection				Multi-intersection	Union			
Dataset	E	F	W	E + F	E + W	F + W	E + F + W	E + F + W	E + F	E + W	F + W	E + F + W
Dermatology (35)	0.94 (18)	0.90 (26)	0.96 (21)	0.96 (15)	0.96 (13)	0.97 (16)	0.96 (10)	0.96 (24)	0.96 (29)	0.97 (26)	0.96 (31)	0.96 (31)
Heart (75)	0.83 (13)	0.80 (10)	0.84 (8)	0.77 (10)	0.83 (8)	0.78 (5)	0.78 (5)	0.83 (13)	0.83 (13)	0.83 (13)	0.83 (13)	0.83 (13)
Liver (7)	0.67 (6)	0.63 (4)	0.69 (6)	0.64 (4)	0.67 (6)	0.64 (4)	0.64 (4)	0.67 (6)	0.67 (6)	0.67 (6)	0.67 (6)	0.67 (6)
Lymphography (19)	0.82 (18)	0.74 (14)	0.83 (11)	0.81 (14)	0.8 (11)	0.73 (8)	0.73 (8)	0.84 (17)	0.82 (18)	0.82 (18)	0.84 (17)	0.82 (18)
Lung_discrete (325)	0.63 (13)	0.59 (13)	0.73 (208)	0.51 (7)	0.63 (10)	0.5 (7)	0.5 (6)	0.69 (12)	0.65 (19)	0.9 (211)	0.9 (214)	0.9 (216)
Pima (9)	0.77 (8)	0.75 (6)	0.77 (5)	0.76 (6)	0.77 (5)	0.77 (3)	0.77 (3)	0.77 (8)	0.77 (8)	0.77 (8)	0.77 (8)	0.77 (8)
Avg. (78)	0.78 (13)	0.74 (12)	0.8 (43)	0.74 (9)	0.78 (9)	0.73 (7)	0.73 (6)	0.79 (13)	0.78 (16)	0.83 (47)	0.83 (48)	0.83 (49)

TABLE 2 Classification accuracy of support vector machine and number of selected features obtained by different feature selection methods over very high dimensional datasets

				Feature selection ensembles								
	Single feature selection			Intersection				Multi-intersection	Union			
Dataset	E	F	W	E + F	E + W	F + W	E + F + W	E + F + W	E + F	E + W	F + W	E + F + W
Colon (5,966)	0.83 (3)	0.85 (21)	0.84 (1,360)	—	0.88 (2)	—	—	0.88 (2)	0.92 (24)	0.92 (1,001)	0.92 (1,021)	0.92 (1,022)
RELATHE*** (2,000)	0.60 (2)	0.58 (11)	0.63 (818)	—	0.66 (2)	0.55 (6)	—	0.71 (8)	0.58 (13)	0.88 (818)	0.88 (823)	0.63 (823)
Prostate_GE (4,322)	0.69 (60)	0.81 (186)	0.36 (18)	0.55 (6)	0.55 (1)	0.55 (1)	—	0.55 (8)	0.78 (240)	0.71 (77)	0.73 (203)	0.78 (256)
Avg. (4,096)	0.70 (22)	0.75 (73)	0.61 (732)	0.55 (6)	0.70 (2)	0.55 (4)	—	0.71 (6)	0.76 (92)	0.84 (632)	0.84 (682)	0.78 (700)

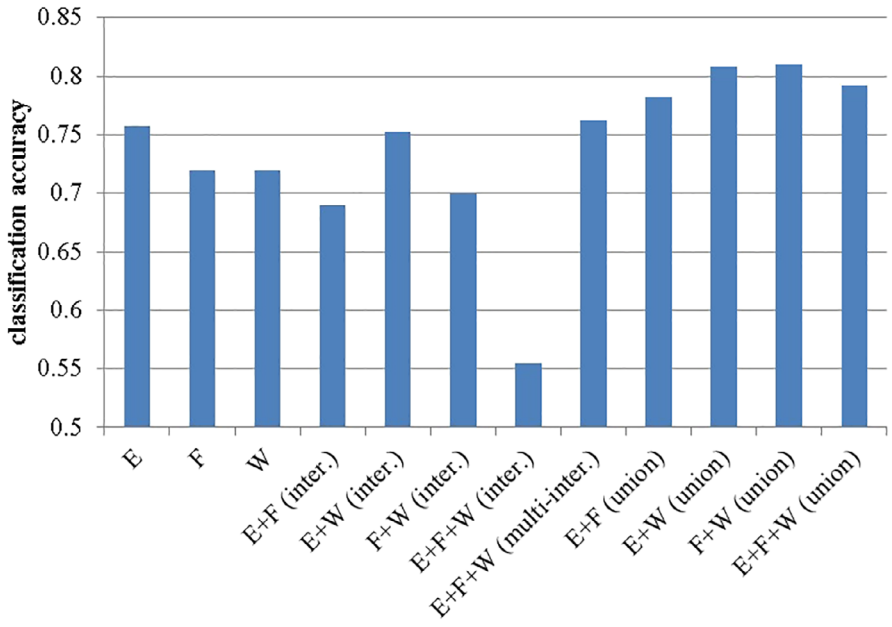

FIGURE 3 Average classification accuracy of different feature selection methods

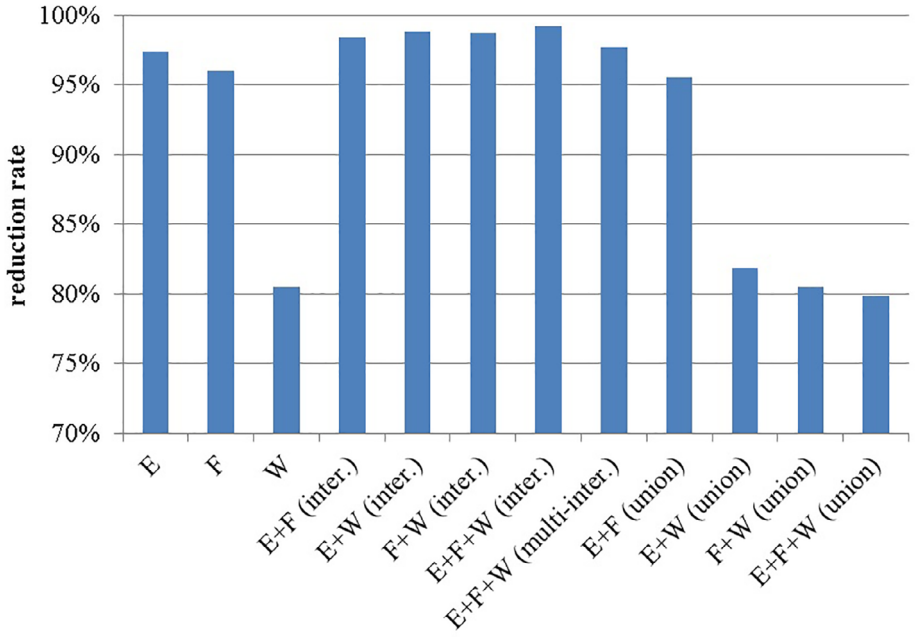

FIGURE 4 Average reduction rate of different feature selection methods

similar to GA and ensemble feature selection by the union methods consistently perform better than GA, except for embedded + filter (C4.5 + PCA), where there is no a significant level of performance difference.

In terms of the number of selected features, we can see that ensemble feature selection methods using the intersection method filter out the most features, keeping the smallest number of features. In comparison, the 43 features selected by GA produce a classification accuracy of 0.8. Over-selection is a problem with most feature ensemble-intersection methods because they make the SVM classifiers provide lower classification accuracy.

On the other hand, most features are kept when using ensemble feature selection approach with the union method, which avoids the risk of over selection. Specifically, C4.5 + GA and PCA + GA produce the highest rate of classification accuracy, keeping on average, about 60 and 62% of the original features, respectively, that is, 47 and 48 out of 78. The reduction rate ranges from 38 to 40%.

However, if both classification accuracy and the feature reduction rate are considered to have the same level of importance, ensemble feature selection techniques using multi-intersection would be the better choice. This is because the multi-intersection method can not only provide relatively better classification performance, that is, 0.79, which is similar to the single best method, that is, GA, without significant difference, but can

also produce a very good reduction rate, that is, about 83%, much better than the other ensemble feature selection approaches in combination with the union methods.

4.2.2 | Very high dimensional datasets

The performance of different feature selection methods is further examined using three very high dimensional datasets, containing 2,000, 4,322, and 5,966 dimensions, respectively. Table 2 lists the classification performance results obtained with SVM in combination with different single feature selection algorithms and ensemble feature selection methods. Note that for some ensemble feature selection methods using the intersection method, since there are no duplicated features that are selected by different feature selection methods, they cannot be used to train and test the SVM classifier. In this case, there is no classification accuracy.

The results show that the ensemble feature selection methods using the union method perform better than the others. In particular, C4.5 + GA and PCA + GA perform the best, allowing the SVM classifier to provide the highest rate of average classification accuracy, 0.84, with a significant level of performance differences from the other methods ($p < .05$).

For the feature reduction rate, C4.5 + GA and PCA + GA with the union method produce very good reduction rates for purposes of dimensionality reduction of about 83–85%. Although the single feature selection methods and ensemble feature selection methods using the intersection and multi-intersection methods can filter out much larger amounts of features than the feature ensemble methods using the union method, they do not provide similar classification accuracy.

The average classification accuracy of different feature selection methods for all of the datasets is shown in Figure 3. As we can see, ensemble feature selection techniques using the union and multi-intersection methods significantly outperform the single best feature selection method, C4.5 ($p < .05$). Specifically, a combination of filter and wrapper feature selection methods, that is, PCA + GA, is the better choice for obtaining relatively higher classification accuracy than the other feature selection methods. In addition, it also provides reasonably good feature reduction rates, as shown in Figure 4.

5 | CONCLUSION

Ensemble feature selection methods are able to solve the limitations of single feature selection algorithms that are based on different selection criteria, to produce the feature subsets by combining different feature selection results. In this paper, we focus on examining the classification performance obtained by combining different types of feature selection algorithms with ensemble feature selection techniques. The methods for combining multiple feature selection results include the union, intersection, and multi-intersection methods.

Different types of datasets are used including categorical, numerical, and mixed data types as well as very high dimensional datasets. In addition, principal component analysis (PCA), genetic algorithm (GA), and C4.5 DT techniques are employed to represent the filter, wrapper, and embedded feature selection techniques, respectively.

The experimental results show that on average, ensemble feature selection performed by the union and multi-intersection methods would allow the SVM classifier to provide higher classification accuracy than single feature selection algorithms. In particular, combining filter (PCA) and wrapper (GA) techniques with the union method would be better, producing the highest rate of average classification accuracy while maintaining a reasonably good feature reduction rate.

There are several issues that can be considered as the future research works for ensemble feature selection. First, since the feature selection result can be affected by the attribute data types including discrete and continuous values, numbers of feature dimensions, numbers of data samples, class imbalanced data, and so forth, it is worth examining related factors affecting the ensemble feature selection result. Second, as ensemble feature selection requires higher computation cost than single feature selection methods, some solutions for efficiently handling very large scale datasets need to be proposed.

ACKNOWLEDGEMENTS

The work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST 106-2410-H-182-024, in part by the Healthy Aging Research Center, Chang Gung University from the Featured Areas Research Center Program within the Framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan under Grant EMRPD1K0481, EMRPD1K0461, and in part by Chang Gung Memorial Hospital, Linkou under Grant CMRPD3I0031. The research project was also supported in part by grants from the Kaohsiung Municipal Chinese Medical Hospital, Department of Health, Kaohsiung City Government, Kaohsiung, Taiwan (KMCMH-10401) and from the Ministry of Science and Technology of Taiwan, awarded to Fang-Rong Chang (Grant Nos. 1 MOST 103-2320-B-037-005-MY2, 105-2628-B-037-001-MY3, 106-2320-B-037-008-MY2, 107-2911-I-037-502, and 108-2320-B-037-022-MY3), and awarded to Yang-Chang Wu (MOST 106-2622-B-037-003-CC2, 106-2320-B-037-007-MY3).

ORCID

Wei-Chao Lin https://orcid.org/0000-0002-5803-513X

ENDNOTE

[1] http://www.cs.waikato.ac.nz/ml/weka/

REFERENCES

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, *34*, 483–519.

Byun, H., & Lee, S.-W. (2003). A survey on pattern recognition applications of support vector machines. *International Journal of Pattern Recognition and Artificial Intelligence*, *17*(3), 459–486.

Chaikla, N., & Qi, Y. (1999). Genetic algorithms in feature selection. Paper presented at: IEEE International Conference on Systems, Man, and Cybernetics. pp. 538-540.

Chandra, M. A., & Bedi, S. S. (2018). Survey on SVM and their application in image classification. *International Journal of Information Technology*, *2*, 1–11. https://doi.org/10.1007/s41870-017-0080-1

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16–28.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Das, A. K., Das, S., & Ghosh, A. (2017). Ensemble feature selection using bi-objective genetic algorithm. *Knowledge-Based Systems*, *123*, 116–127.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Goldberg, D. E. (1989). *Genetic algorithms in search optimization and machine learning*. Boston, MA, United States: Addision Wesley.

Grefenstette, J. J. (1986). Optimization of control parameters of genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, *16*(1), 122–128.

Guan, D., Yuan, W., Lee, Y.-K., Najeebullah, K., & Rasel, M. K. (2014). A review of ensemble learning based feature selection. *IETE Technical Review*, *31*(3), 190–198.

Guo, Q., Wu, W., Massart, D. L., Boucon, C., & de Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, *61*, 123–132.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Huang, C., Du, J., Nie, B., Yu, R., Xiong, W., & Zeng, Q. (2019). Feature selection method based on partial least squares and analysis of traditional Chinese medicine data. *Computational and Mathematical Methods in Medicine*, *2019*, 9580126.

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS One*, *12*(1), e0161501.

Kazemitabar, S.J., Amini, A.A., Bloniarz, A., & Talwalkar, A. (2017). Variable importance using decision trees. Paper presented at: International Conference on Neural Information Processing Systems. pp. 425–434.

Kumar, V., & Minz, S. (2014). Feature selection: A literature review. *Smart Computing Review*, *4*(3), 211–229.

Lim, H., Lee, J., & Kim, D.-W. (2017). Optimization approach for feature selection in multi-label classification. *Pattern Recognition Letters*, *89*, 25–30.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *17*(4), 491–502.

Liu, P., Huang, Y., Meng, L., Gong, S., & Zhang, G. (2016). Two-stage extreme learning machine for high-dimensional data. *International Journal of Machine Learning and Cybernetics*, *7*, 765–772.

Morchid, M., Dufour, R., Bousquet, P.-M., Linares, G., & Torres-Moreno, J.-M. (2014). Feature selection using principal component analysis for massive retweet detection. *Pattern Recognition Letters*, *49*, 33–39.

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Reddy, R. R., Kavya, B., & Ramadevi, Y. (2014). A survey on SVM classifiers for intrusion detection. *International Journal of Computer Applications*, *98*(19), 38–44.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*, 103375.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Seijo-Pardo, B., Bolon-Canedo, V., & Alonso-Batanzos, A. (2019). On developing an automatic threshold applied to feature selection ensembles. *Information Fusion*, *45*, 227–245.

Seijo-Pardo, B., Porto-Diaz, I., Bolon-Canedo, V., & Alonso-Batanzos, A. (2017). Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems*, *118*, 124–139.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications*, *40*(10), 4146–4153.

Sindhiya, S., & Gunasundari, S. (2014). A survey on genetic algorithm based feature selection for disease diagnosis system. Paper presented at: IEEE International Conference on Computer Communication and Systems. pp. 164–169.

Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal Processing*, *21*(2), 930–942.

Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269.

Wang, J., Wei, J.-M., Yang, Z., & Wang, S.-Q. (2017). Feature selection by maximizing independent classification information. *IEEE Transactions on Knowledge and Data Engineering*, *29*(4), 828–841.

Zhang, T., Ren, P., Ge, Y., Zheng, Y., Tnag, Y. Y., & Chen, C. L. P. (2016). Learning proximity relations for feature selection. *IEEE Transactions on Knowledge and Data Engineering*, *28*(5), 1231–1244.

Zhu, M., Xia, J., Yan, M., Cai, G., Yan, J., & Ning, G. (2015). Dimensionality reduction in complex medical data: Improved self-adaptive niche genetic algorithm. *Computational and Mathematical Methods in Medicine*, *2015*, 794586.

Zhu, P., Zhu, W., Hu, Q., Zhang, C., & Zuo, W. (2017). Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, *66*, 364–374.

AUTHOR BIOGRAPHIES

Chih-Wen Chen received a Master's degree at Graduate Institute of Natural products, Kaohsiung Medical University, Taiwan in 1997. He is now a leader at the Department of Pharmacy, Kaohsiung Municipal Chinese Medical Hospital, Taiwan; also a teacher in the Department of Nursing, Fooyin University, Taiwan since 2019. He has been a PhD student since 2015. Besides the Chinese herbal medicine and natural products' knowledge. His studies also on data mining. He has published more than 30 professional publications. Some were in international symposiums and some were published in prestigious journals including: Expert Systems, Technology and Health Care.

Yi-Hong Tasi received a PhD at Graduate Institute of Natural products, Kaohsiung Medical University, Taiwan in 2015 and has been a post-doctoral research scholar since then. His current research focuses on Pharmacognosy. He has published many professional publications where some were published in prestigious journals including: Journal of Food and Drug Analysis, International Journal of Molecular Sciences, Frontiers in Pharmacology, Bioorganic & Medicinal Chemistry Letters, Phytochemistry, Food Research International, Toxicological Sciences, Bioorganic & Medicinal Chemistry Letters, Food and Chemical Toxicology.

Fang-Rong Chang was born at 1966, Kaohsiung. He obtained his pharmacognosy Ph.D degree at Kaohsiung Medical University, Taiwan in 1995; Honorary Doctor of Pharmacy, Uppsala University, Sweden in 2018; Doctor Honoris Causa, Faculty of Pharmacy, University of Szeged, Hungary in 2018 as well. He is now the director in National Research Institute of Chinese Medicine (NRICM), Ministry of Health and Welfare, Taiwan; also a professor in the Graduate Institute of Natural Products, KMU since 2005. He has been the Vice Dean of the Office of Global Affairs, KMU during 2013-2018, and the director in the Graduate Institute of Natural Products, KMU during 2006-2012. His research interests would be on anything related to Natural products, like TCMs, herbal medicines, pharmacy administrations and so on. He published more than 340 research articles in SCI refereed journals; more than 500 oral or poster presentation; reviewer of more than 120 different international journals; editorial board members of more than 10 international journals; authorship of several book chapters; more than 30 patents issued or in application; more than 20 industry-academic cooperation; more than 5 patent/tech transfer experiences (including one new drug R&D tech transfer). Prof. Chang won many awards and the national and international level ones would be Doctor Honoris Causa, University of Szeged, Hungary; Honorary Doctor in Pharmacy - Uppsala University, Sweden; The 9th and 12th National Innovation Award, Taiwan; The Faculty Gold Medal, Faculty of Pharmacy, University of Szeged, Hungary; Member of the Szent-Györgyi International Mentors, Szeged Scientists Academy, Foundation for the Future of Biomedical Sciences in Szeged, Hungary and so on.

Wei-Chao Lin received the PhD degree from the University of Sunderland, UK. He is currently an associate professor at the Department of Information Management, Chang Gung University, Taiwan. He has published over thirty journal papers from such as Information Sciences, Journal of Systems and Software, Applied Soft Computing, Neuro computing, etc. His current research interests include data mining, machine learning, and information retrieval.

How to cite this article: Chen C-W, Tsai Y-H, Chang F-R, Lin W-C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*. 2020;37:e12553. https://doi.org/10.1111/exsy.12553