Wisdom of Crowds: An Empirical Study of Ensemble-Based Feature Selection Strategies

Teo Susnjak[1(✉)], David Kerry[1], Andre Barczak[1], Napoleon Reyes[1], and Yaniv Gal[2]

[1] Massey University, Auckland, New Zealand
t.susnjak@massey.ac.nz
[2] Compac Ltd., Auckland, New Zealand

Abstract. The accuracy of feature selection methods is affected by both the nature of the underlying datasets and the actual machine learning algorithms they are combined with. The role these factors have in the final accuracy of the classifiers is generally unknown in advance. This paper presents an ensemble-based feature selection approach that addresses this uncertainty and mitigates against the variability in the generalisation of the classifiers. The study conducts extensive experiments with combinations of three feature selection methods on nine datasets, which are trained on eight different types of machine learning algorithms. The results confirm that the ensemble based approaches to feature selection tend to produce classifiers with higher accuracies, are more reliable due to decreased variances and are thus more generalisable.

Keywords: Ensemble feature selection · Dimensionality reduction · Machine learning · Classification · Data mining · Ensemble classifiers

1 Introduction

The main purpose of machine learning is to produce classifiers that generalise in their predictive accuracy beyond the datasets used to train them. To a large degree, their final accuracy is dependent on the descriptive strength and quality of the features that constitute the training dataset. It is often tempting to simply provide a machine learning algorithm with as many features as are available for a given dataset. However, doing so has been consistently shown to be associated with negative outcomes [15–17].

The inclusion of large feature numbers in a training dataset presents computational challenges that mostly arise during the training phase and can be prohibitive for some algorithms [28], but can also be a strain during the detection time for real-time systems processing high-volume data streams. Unnecessary and redundant features increase the search space for a machine learning algorithm. This in turn dilutes the signal strength of a true pattern and makes it more likely that due to the presence of noisy and irrelevant features, a spurious pattern will be discovered instead.

B. Pfahringer and J. Renz (Eds.): AI 2015, LNAI 9457, pp. 526–538, 2015.
DOI: 10.1007/978-3-319-26350-2_47

In general it is not known *a priori* which features are meaningful, and finding the optimal feature subset has been proven to be a NP-complete problem [2]. Nonetheless, it is still imperative that feature selection algorithms be applied to a dataset as a pre-processing step before training classifiers, in order to reduce feature dimensionality [14]. Not only are both the computational complexity and the generalisability improved by selecting the most concise subset, but the resulting model is more interpretable due to the fact that it is generated with the fewest possible number of parameters [10].

Research into feature selection has produced a wide array of techniques and algorithms. Each technique provides a different perspective on the data and thus a different assessment of how meaningful individual features are. Some techniques perform considerably better than others on different datasets, sample sizes, feature numbers and problem domains [12] and it is generally uncertain which technique will be most suitable for a problem at hand. Prior to performing machine learning, it is not uncommon in some domains where stability in feature subsets is important [1], to initially investigate the results from several feature selection algorithms before choosing the best feature selection technique which satisfies a required criterion [6].

Feature selection techniques can generally be divided into two broad categories. Filter methods are univariate techniques which consider the relevance of a particular feature in isolation to the other features and rank the features according to a metric. These algorithms are computationally efficient since they do not integrate the machine learning algorithm in its evaluation. However, they can be susceptible to selecting subsets of features that may not produce favourable results when combined with a chosen machine learning algorithm [30]. These methods lack the ability to detect interactions among features as well as feature redundancy. On the other hand, wrapper methods overcome some of these shortcomings. They explicitly use the chosen machine learning algorithm to select the feature subsets and tend to outperform filter methods in predictive accuracy [30]. However, these techniques exhibit bias in favour of a specific machine learning algorithm, and since they are computationally more intensive, they are also frequently impractical on large datasets.

Hybrid filter-wrapper methods have been a subject of recent research due to their ability to exploit the strengths of both strategies [17,18]. Hybrid approaches essentially allow any combination of filter and wrapper methods to be combined. Due to this, some novel and interesting hybrid approaches have recently been proposed such as: using the union of feature-subset outputs from Information Gain, Gain Ratio, Gini Index and correlation filter methods as inputs to the wrapper Genetic Algorithm [20], hybridization of the Gravitational Search Algorithm with Support Vector Machine [23] and using Particle Swarm Optimisation-based multi-objective feature selection approach in combination with k-Nearest-Neighbour [27]. Given their flexibility, hybrid approaches thus offer some degree of tuning the trade-offs between accuracy and performance. Nonetheless, devising a feature selection algorithm that is both highly accurate and computationally efficient is still an open question [10].

The ubiquity of data acquisition technologies and the affordability of ever increasing data storage capacities means that datasets are now larger in both sample numbers as well as feature vectors. In the age of Big Data, it is not uncommon to encounter datasets having many thousands of features [4,12] in a variety of problem domains. This presents considerable challenges and for these reasons, feature selection is an active and important part of ongoing research. The challenge is to some degree amplified since machine learning has entered into mainstream use and is becoming more frequently utilised in numerous industrial [24] as well as business sectors, where in-depth expertise in the intricacies of this field are not always readily available.

Motivation. Our motivation is to devise a strategy for performing feature selection which increases the likelihood of generating good and robust feature subsets that can effectively be combined with a wide range of machine learning algorithms and datasets spanning numerous domains. The aim is to address the need to formulate a strategy that generates a reliable feature subset in a timely manner, and can particularly be useful in industrial and business settings where machine learning is employed by practitioners who have not necessarily had expert training. The purpose is to automate the process of feature selection and to eliminate the possibilities of generating poor subsets, while foregoing the goals of finding the optimal solution due to its impracticality.

This research investigates combining outputs of multiple feature selection algorithms in order to produce an effective feature subset for machine learning. The inspiration is drawn from the theory of ensemble-based classifiers. Its foundational principle states that while any one classifier may perform more accurately than a combined classification of all available classifiers on a given dataset, across the space of all possible problems, the aggregate decisions of multiple classifiers will however outperform any one available individual classifier. Ensemble-based classifiers have demonstrated superior results compared to individual classifiers in a wide range of applications and scenarios [22]. In empirical studies, it has been shown that ensembles yield better results provided that there is significant diversity among the classifiers.

Our research builds and extends on previous work by Tsai and Hsiao [25] who experimented with combining the feature subset selections of Principal Component Analysis, Information Gain and the Genetic Algorithm, as inputs for the Neural Network classifier on the domain of stock price prediction. This research goes further and the key contributions lie in demonstrating how the ensemble-based feature selection strategy can be generalised to a much broader set of domains, and can be combined with a wider number of machine learning algorithms. We empirically show how this strategy performs in conjunction with eight machine learning algorithms using different combinations of the Information Gain (IG) [13], Linear Discriminant Analysis (LDA) and Chi^2 [19] techniques for generating feature subsets, using nine datasets for testing.

The extensive experiments in this research show that the ensemble-based strategy for feature selection does indeed generalise to multiple machine learning

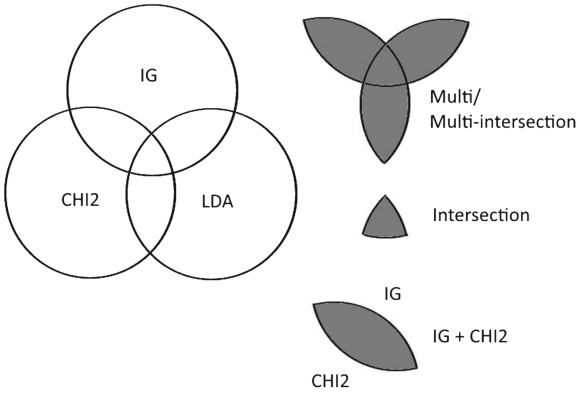

Fig. 1. Three different strategies for combining the outputs of the feature selection algorithms.

algorithms and different problem domains. The results confirm that relying on multiple sources for input on feature selection does outperform any one single feature selection algorithm in the long run. In addition, the results also provide some insights and rough rules-of-thumb for machine learning practitioners as to which individual feature selection strategies have a tendency to work well in combination with machine learning algorithms.

2 Experimental Design

IG, LDA and Chi^2 were selected for the experiments since they are readily available in most data mining software packages and, individually, are widely used for the purposes of feature selection [10,28,29]. These methods were also chosen due to the slightly different perspectives each one has on what constitutes 'meaningfulness' of a given feature. The success of ensemble-based decision making lies with the existence of disagreement amongst individual methods. LDA's strength lies in projecting the data cloud onto new axes which maximise the variance and in the process identify redundancy, while taking class membership into account. Meanwhile, Chi^2 and IG do not consider feature redundancy but rank the feature according to different criteria. Chi^2 tests the independence of each feature in respect to its class label, while IG similarly evaluates how commonly occurring a feature value is for a given class, compared to its frequency amongst all other classes.

Given that a sufficient degree of diversity exists within an ensemble system, it is then important to devise an appropriate aggregation strategy. Ensemble-based classifiers usually assign a weight to each of its constituent classifiers which reflects their discriminative strength, calculated during the training process. Since feature selection algorithms cannot easily be assigned a confidence weight, a different strategy must be applied. This research applies set theory in order

to aggregate the outputs of the feature selection algorithms, which has been shown to be effective by [25]. The strategy is depicted in Fig. 1, showing how features that are an (1) intersection of all, (2) multi-intersection of all, or (3) an intersection of two feature selection algorithms, can be combined. Different permutations of the intersections, together with the individual feature selection algorithms provided eight strategies for the experiment, plus the control which did not apply any feature selection.

Experiments were conducted using nine datasets whose properties are outlined in Table 1. The datasets originated from two sources, the first three fruit datasets were obtained from an industrial source[1]. The datasets represent fruit surface features from three different varieties. Though the datasets originated from entirely different fruit, they were all generated by their proprietary software and there is thus an element of risk that the datasets capture some artefacts of the feature extraction process, which may bias them towards certain feature selection algorithms. The remaining datasets were sourced online from the UCI Machine Learning Repository [3]. Procurement of datasets with a variety of feature and sample numbers as well as domains of origin was the goal.

Table 1. Dataset properties.

Dataset name	Classes	Instances	Features
Nectarines	4	587	13
Peaches	3	240	10
Plums	3	141	13
Waveform	3	5000	21
Fac profile	10	2000	216
Fourier	10	2000	76
Karhunen-Love	10	2000	64
Pixel avg	10	2000	240
Zernike Moments	10	2000	47

The various feature subsets were trained on eight different machine learning algorithms, listed together with their tunable parameters in Table 2. Given that a total of 6840 classifiers were trained across the entire experimental process (9 datasets $\times$ 5 folds $\times$ 19 thresholds $\times$ 8 machine learning algorithms), tuning the machine learning algorithms for optimal training parameters was not feasible. Therefore, most classifiers were trained with default parameters for each of their respective algorithms.

The experimental workflow is depicted in Fig. 2. Each dataset was passed into the feature selection stage where eight subsets were created. The first subset is

[1] These datasets were provided by Compac Sorting Ltd., a company that specialises in automated fruit sorting via image processing.

Table 2. Machine learning training parameters.

Classifier	Training settings	Implementation source
kNN	k = 3	scikit-learn [21]
SVM [26]	linear kernel, regularisation parameter C = 0.025	scikit-learn [21]
Decision Tree	maximum depth = 10	scikit-learn [21]
Random Forest [5]	maximum depth = 10, estimators = 20	scikit-learn [21]
AdaBoost [9]	number of estimators = 100	scikit-learn [21]
Naive Bayes	Gaussian default setting	scikit-learn [21]
AdaBoost.ECC [11]	100 boosting iterations	authors' C++ implementation
RIPPER [7]	2 rounds of optimisations with pruning enabled	authors' C++ implementation

the control containing all features. The next three subsets were created from applying Chi^2, IG and LDA to rank the features in order of how informative they are according to their respective feature evaluation criteria.

Subsequently, the classification stage trains and tests the classifiers on each feature subset using five fold cross-validation. The process continues with the thresholding stage. The feature selection algorithms rank the features in the order of their apparent usefulness but are not able to determine if a given feature is 'informative' or 'poor'; a threshold therefore needs to be selected as the cut-off for the percentage of features to keep in the training/testing subset. The testing was repeated exhaustively with a threshold range of 5 % to 95 % of the features accepted, with 5 % intervals. This enabled every combination of classifier, approach and dataset the opportunity to achieve its optimal feature subset size as a proportion of the ranked features.

The last stage in the process gathered the performance data for every combination of dataset, feature selection subset and classifier. The accuracy and the geometric-mean scores with the corresponding standard deviation were collected from the thresholding stage. Geometric-mean was calculated in addition to accuracy due to the greater ability of the geometric-mean to convey classifier generalisation on datasets with significant class imbalances. Some degree of class imbalance was present on the Nectarines, Peaches and Plums datasets; however, the negligible differences between the geometric-mean and accuracy scores showed that this was not at a significant level. For this reason, the geometric-mean scores are not reported in the results.

3 Results

The performance results from the experiments presented here involve several thousand classifiers. Space limitations preclude us listing all accuracy results in

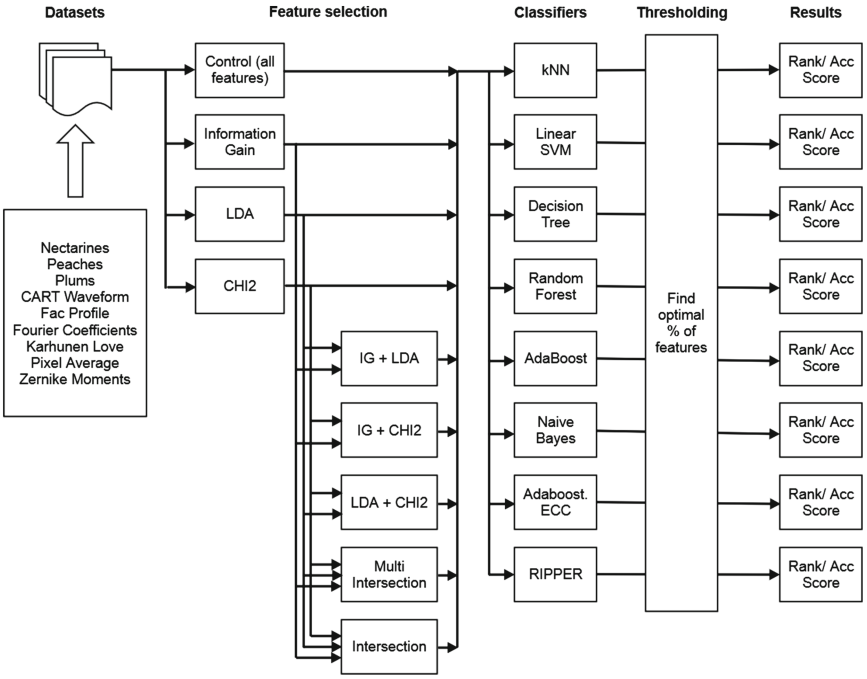

Fig. 2. The schematic representation of the stages in the experimental process together with the workflow.

their raw form. Instead, we provide small snapshots of the underlying accuracy results from each feature selection algorithm, while relying mostly on their summaries in the form of mean ranks as is acceptable practice [8]. In addition the non-parametric Friedman statistical test will also be employed in order to verify statistical significances in the findings.

Table 3 lists the accuracies of each feature selection algorithm across all datasets for the Random Forest classifier. The table is summarised in the form of mean ranks which aggregate the performances of all the feature selection algorithms for this particular classifier. The table shows that across all the datasets, the ensemble-based multi-intersection method outperformed all remaining methods in this study. Performing machine learning without first conducting feature subset selection has predictably generated the least generalisable classifiers.

A further seven tables in the same format as Table 3 were generated for the remaining classifiers used in these experiments. The mean rank summaries were extracted from the tables and all combined together in order to render a graphical depiction in Fig. 3. The mean ranks in Table 3 for the Random Forest classifier can be traced in this figure.

Figure 3 is a box-and-whisker plot, in which each of the feature selection methods are ordered based on the average of their mean ranks from each of the classifiers on all datasets. The median, inter-quartile and occasional outliers are

Table 3. Example of Random Forest classifiers' raw accuracy results on all datasets using each of the feature selection methods. The results in the table represent one of eight tables generated from which a set of mean ranks were calculated.

	Multi	Chi^2	LDA	LDA+ Chi^2	IG+ Chi^2	IG	LDA+ IG	Inters	Control
Nectarines	**0.625**	0.612	0.576	0.574	**0.625**	0.617	0.589	0.576	0.588
Peaches	**0.804**	0.796	0.78	0.788	0.775	0.78	0.776	0.792	0.759
Plums	0.623	0.579	0.623	0.579	0.595	0.621	**0.636**	0.566	0.589
Waveform	**0.844**	0.841	0.838	0.839	0.836	0.832	0.837	0.835	0.83
Fac profile	0.947	**0.949**	0.944	0.943	0.947	0.948	0.94	0.943	0.942
Fourier	**0.825**	0.817	0.821	0.817	0.774	0.723	0.784	0.779	0.714
Karhunen Love	0.922	**0.928**	0.919	0.923	0.843	0.808	0.838	0.857	0.852
Pixel Average	0.953	0.951	0.949	0.947	0.949	0.945	0.952	**0.956**	0.943
Zernike	**0.73**	0.727	0.729	0.714	0.727	**0.73**	0.729	0.722	0.724
Mean Ranks	**1.9**	3.6	4.4	5.7	5.3	5.4	5.2	5.9	7.7

displayed. As expected, the worst performing method is the control whereby no feature selection was performed. The best performing strategy is the ensemble-based multi-intersection method, followed closely by Chi^2 and LDA methods. The two clear outliers in the graph indicate the positive responsiveness of the

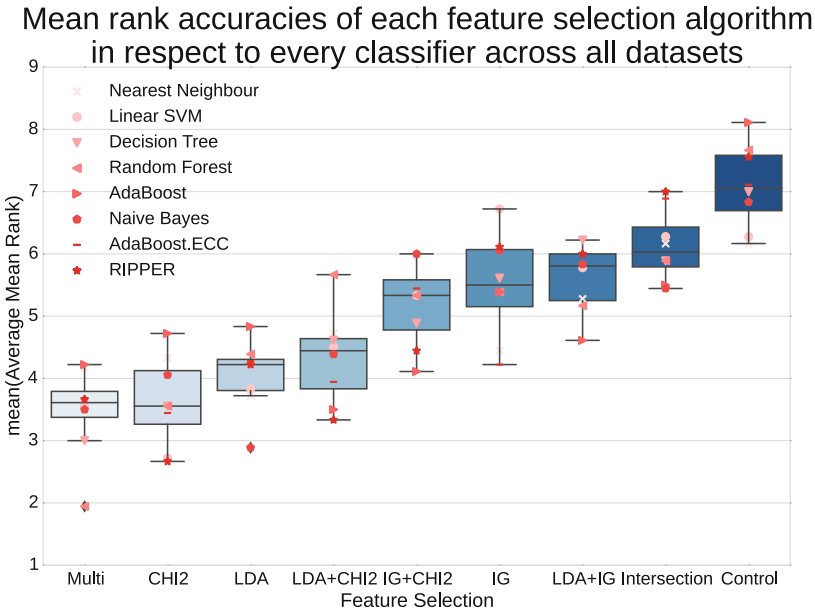

Fig. 3. Box-and-whisker plot showing the accuracies of each feature selection techniques in terms of their mean rank score for each machine learning algorithm, when combined across all datasets. The feature selection algorithms are listed in the order from best to worst, based on to the average of their mean rank scores.

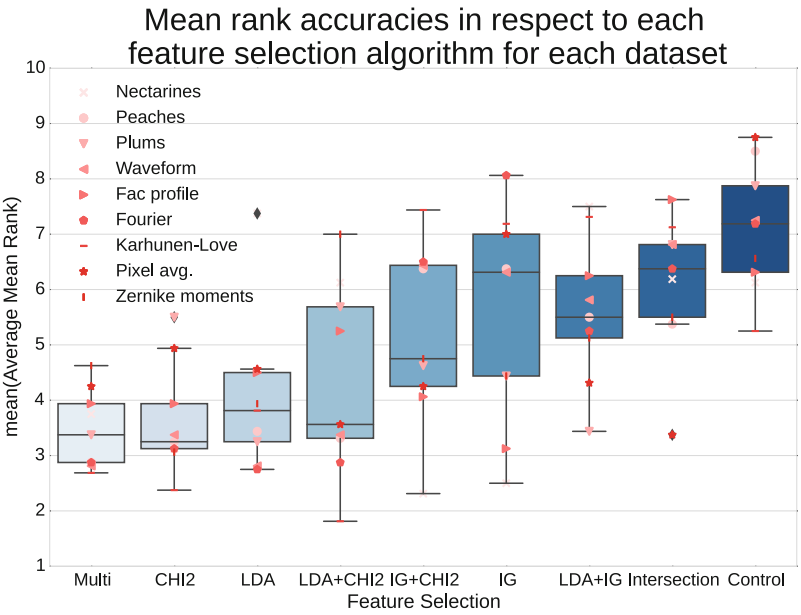

Fig. 4. Box-and-whisker plot showing the accuracies of each feature selection technique in terms of their mean rank score for each dataset, when combined across all machine learning algorithms. The feature selection algorithms are listed in the order from best to worst according the average of their mean rank scores.

Random Forest and Naive Bayes classifiers to the multi-intersection and LDA feature selection methods respectively.

Importantly, the figure also conveys the degree of variance for each of the feature selection algorithms. Smaller variability is desirable as it indicates more consistent and predictable performances. The multi-intersection method exhibits the lowest variance of all the strategies examined. Ensemble-based methods are known to reduce the variance and thus the results are not altogether surprising. However, it is noteworthy that variance is reduced, while comparatively the best accuracies are achieved, thus indicating that there has been no increase in bias.

In addition, the differences in the mean ranks in Fig. 3 are confirmed to be statistically significant. The critical value for the Friedman Rank Sum Test at $\alpha = 0.05$ is $\chi^2 = 15.507$. The test produces a test statistic $\chi^2_F(8) = 47.3$ and p-value $= 1.365e - 07$.

The effectiveness of feature selection algorithms is not only determined by their suitable combination with specific machine learning algorithms, but also by the actual underlying datasets. Presented in Fig. 4 is an alternative perspective on the result from the previous figure which illustrates the effectiveness of the feature selection algorithms on each of the datasets, using the combined accuracies of the classifiers. The overarching message from the results data has not changed from Fig. 3. The ensemble-based multi-intersection method is still the

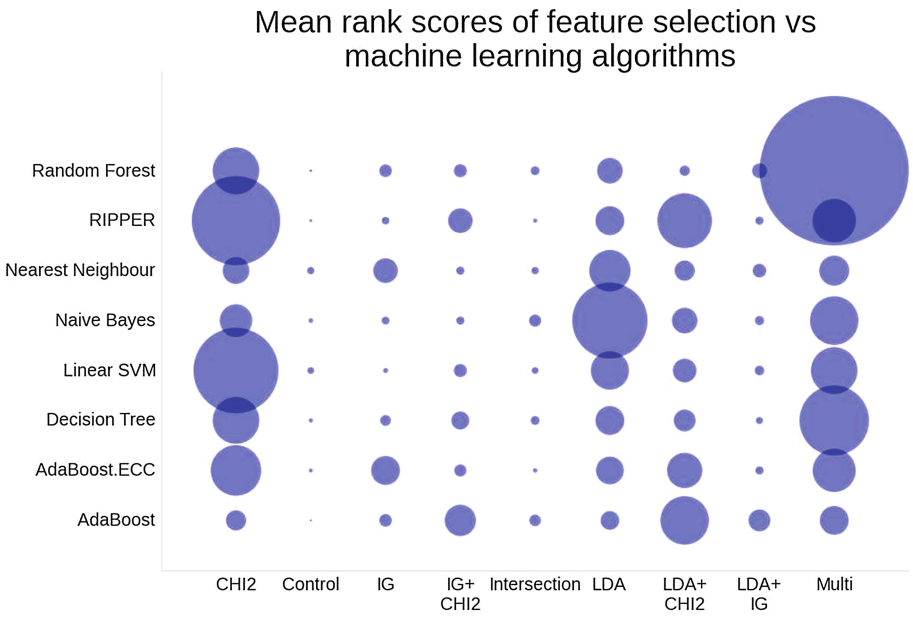

Fig. 5. Visual matrix of the effectiveness of each feature selection method in combination with every machine learning algorithm, across the aggregate of all datasets. Effectiveness is expressed in terms of mean ranks from accuracy scores and projected on to the exponential scale in order to emphasise strengths in the patterns.

best performing, while the ordering of the remaining methods is intact. Importantly, the findings indicate that the ensemble-based method is also invariant to the variety of datasets and domains, and not only to the type of machine learning methods used. Overall, Fig. 3 demonstrates the variances have increased across all feature selection algorithms. Most notably the combination of LDA+Chi^2 and IG+Chi^2, as well as for the IG, the variances have increased markedly. Of note is the negative effect of one dataset in particular (Nectarines) on the performance of LDA which once again demonstrates the point that while each feature selection algorithm will perform acceptably on some datasets, there also exist datasets on which a given method will perform very poorly.

Even though the number of datasets used in this research is limited and thus precludes us from making definitive claims, it is nonetheless useful to be able to draw out some insights and very general rules-of-thumb from these empirical findings as to which feature selection methods and machine learning algorithms have a tendency to work well in combination. Figure 5 attempts to graphically convey this information and draw out insights from the experimental data conducted here. The figure demonstrates the responsiveness of a classifier to the various feature selection methods using the mean ranks, to which the exponential function has been applied in order to emphasise patterns. Clearly the combination of multi-intersection and the Random Forest classifier stands

out as a suitable combination. A strong signal can be observed between the Naive Bayes classifier with LDA, as with the combination Chi^2 with Ripper and Linear SVM.

It should be noted that there are also very poor pairings of classifier and feature selection combinations for a given data set. An example of this is shown as Chi^2 generally performs very well, but performs poorly in combination with AdaBoost. A good starting point to find the optimum and avoid potential poor performance, therefore, would be to use the performance of an ensemble approach as a benchmark and use only combinations that outperform the benchmark for further performance refinement.

4 Conclusion

Feature selection is an indispensable component of machine learning. While there are numerous feature selection methods in existence, their robustness is affected by both the chosen machine learning algorithms they are intended to be used with, as well as the characteristics of the underlying datasets and the problem domains themselves. While some feature selection methods will generally tend to work well with certain machine learning algorithms, there are times when they will perform poorly in combination with given datasets which have specific properties. *A priori* knowledge of which combinations will work well together is usually inaccessible, and computationally deriving an optimal combination is a NP-complete problem.

This research considers the ensemble-based approach to selecting feature subsets. Instead of relying on a single feature-selection algorithm that might work well on some occasions but poorly on others, the ensemble-based approach advocates thoughtfully combining the outputs of several different methods. The ensemble-based approach does not guarantee that an optimal solution will be produced, but it does ensure that a very good solution will be found instead, and this is sufficient for many machine learning domains. Ensemble-based approaches do however guarantee that the provided solution will always be better than the weakest performing feature selection algorithms within its ensemble.

The study considered three frequently used filter methods: Chi^2, LDA and Information Gain. It applied different combinations of these methods on nine datasets from a wide range of domains, and tested the feature subsets against eight different classifiers from a broad range of machine learning algorithms.

Extensive experiments were conducted. The data shows that across a number of problem domains and different machine learning algorithms, the ensemble-based approach to feature selection tend to outperform the usage of single feature selection methods explored here. Ensemble-based approaches are more resistant to the variability that different machine learning algorithms and datasets bring to the classification accuracies, and are therefore generally more robust. Given the uncertainty as to which feature selection and machine learning algorithms will combine effectively, this research confirms the suitability of presented methods for domains which process broad varieties of datasets and require timely and consistently reliable solutions.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3), 392–398 (2010)
2. Albrecht, A.A.: Stochastic local search for the feature set problem, with applications to microarray data. Appl. Math. Comput. **183**(2), 1148–1164 (2006)
3. Asuncion, A., Newman, D.: UCI machine learning repository (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Bermejo, P., de la Ossa, L., Gámez, J.A., Puerta, J.M.: Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. Knowl.-Based Syst. **25**(1), 35–44 (2012)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
7. Cohen, W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123 (1995)
8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
10. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recogn. **43**(1), 5–13 (2010)
11. Guruswami, V., Sahai, A.: Multiclass learning, boosting, and error-correcting codes. In: Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, pp. 145–155. ACM, New York (1999)
12. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. **42**(3), 409–424 (2009)
13. Hunt, E.B., Marin, J., Stone, P.J.: Experiments in induction. Academic Press, New York (1966)
14. Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on pso and rough sets for medical diagnosis. Comput. Methods Programs Biomed. **113**(1), 175–185 (2014)
15. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artif. Intell. Med. **31**(2), 91–103 (2004)
16. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1), 273–324 (1997)
17. Kotsiantis, S.: Feature selection for machine learning classification problems: a recent overview. Artif. Intell. Rev. **42**, 1–20 (2011)
18. Leung, Y., Hung, Y.: A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **7**(1), 108–117 (2010)
19. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: TAI, p. 388. IEEE (1995)
20. Oreski, S., Oreski, G.: Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Syst. Appl. **41**(4), 2052–2064 (2014)
21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

22. Polikar, R.: Essemble based systems in decision making. IEEE Circuits Syst. Mag. **6**(3), 21–45 (2006)
23. Sarafrazi, S., Nezamabadi-pour, H.: Facing the classification of binary problems with a gsa-svm hybrid system. Math. Comput. Model. **57**(1), 270–278 (2013)
24. Susnjak, T., Barczak, A., Reyes, N.: On combining boosting with rule-induction for automated fruit grading. In: Kim, H.K., Ao, S.-L., Amouzegar, M.A. (eds.) Transactions on Engineering Technologies, pp. 275–290. Springer, Netherlands (2014)
25. Tsai, C.F., Hsiao, Y.C.: Combining multiple feature selection methods for stock prediction: union, intersection, and multi-intersection approaches. Decis. Support Syst. **50**(1), 258–269 (2010)
26. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
27. Xue, B., Zhang, M., Browne, W.N.: Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Trans. Cybern. **43**(6), 1656–1671 (2013)
28. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412–420 (1997)
29. Ye, J., Li, Q.: A two-stage linear discriminant analysis via QR-decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6), 929–941 (2005)
30. Zhu, Z., Ong, Y.S., Dash, M.: Wrapper-filter feature selection algorithm using a memetic framework. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **37**(1), 70–76 (2007)