



Dimensionality Reduction: Is Feature Selection More Effective Than Random Selection?

Laura Morán-Fernández^(✉)  and Verónica Bolón-Canedo 

CITIC, Universidade da Coruña, A Coruña, Spain
{laura.moranf, veronica.bolon}@udc.es

Abstract. The advent of Big Data has brought with it an unprecedented and overwhelming increase in data volume, not only in samples but also in available features. Feature selection, the process of selecting the relevant features and discarding the irrelevant ones, has been successfully applied over the last decades to reduce the dimensionality of the datasets. However, there is a great number of feature selection methods available in the literature, and choosing the right one for a given problem is not a trivial decision. In this paper we will try to determine which of the multiple methods in the literature are the best suited for a particular type of problem, and study their effectiveness when comparing them with a random selection. In our experiments we will use an extensive number of datasets that allow us to work on a wide variety of problems from the real world that need to be dealt with in this field. Seven popular feature selection methods were used, as well as five different classifiers to evaluate their performance. The experimental results suggest that feature selection is, in general, a powerful tool in machine learning, being correlation-based feature selection the best option with independence of the scenario. Also, we found out that the choice of an inappropriate threshold when using ranker methods leads to results as poor as when randomly selecting a subset of features.

Keywords: Dimensionality reduction · Feature selection · Filters · Classification

1 Introduction

Driven by recent advances in algorithms, computing power, and big data, artificial intelligence has made substantial breakthroughs in the last years. In particular, machine learning has great success because of its impressive ability to automatically analyze large amounts of data. One of the most important tasks in machine learning is classification, which allows to predict events in a plethora of applications; from medicine to

This research has been financially supported in part by the Spanish Ministerio de Economía y Competitividad (research project PID2019-109238GB-C2), and by the Xunta de Galicia (Grants ED431C 2018/34 and ED431G 2019/01) with the European Union ERDF funds. CITIC, as Research Center accredited by Galician University System, is funded by “Consellería de Cultura, Educación e Universidades from Xunta de Galicia”, supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014–2020, and the remaining 20% by “Secretaría Xeral de Universidades” (Grant ED431G 2019/01).

finances. However, some of the most popular classification algorithms can degrade their performance when facing a large number of irrelevant and/or redundant features. This phenomenon is known as *curse of dimensionality* and is the reason why dimensionality reduction methods play an important role in preprocessing the data.

One of such dimensionality reduction techniques is feature selection, which can be defined as the process of selecting the relevant features and discarding the irrelevant or redundant ones. There are considerable noisy and useless features that are often collected or generated by different sensors and methods, which also occupy a lot of computational resources. Therefore, feature selection performs a crucial role in the framework of machine learning of removing nonsense features and preserving a small subset of features to reduce the computational complexity.

There are several applications in which it is necessary to find the relevant features: in bioinformatics (e.g. to identify a few key biomolecules that explain most of an observed phenotype [5]), in respect to the fairness of decision making (e.g. to find the input features used in the decision process, instead of focusing on the fairness of the decision outcomes [9]), or in nanotechnology (e.g. to determine the most relevant experimental conditions and physicochemical features to be considered when making a nanotoxicology risk assessment [8]). A shared aspect of these applications is that they are not pure classification tasks. In fact, an understanding of which features are relevant is as important as accurate classification, as these features may provide us with new insights into the underlying system.

However, there is a large amount of feature selection methods available, and most researchers agree that the best feature selection method simply does not exist [3]. On top of this, new feature selection methods are appearing every year, which makes us ask the questions: do we really need so many feature selection methods? Which ones are the best to use for each type of data? In light of these issues, the aim of this paper is to perform an analysis of the most popular feature selection methods using the random selection as baseline in two scenarios: real datasets and DNA microarray datasets (characterized by having a much larger number of features than of samples). Our goal is to determine if there are some methods that do not obtain significantly better results than those achieved when randomly selecting a subset of features.

The remainder of the paper is organized as follows: Sect. 2 presents the different feature selection methods employed in the study. Section 3 provides a brief description of the 55 datasets used to reduce data dimensionality. Section 4 details the experimental results carried out. Finally, Sect. 5 contains our concluding remarks and proposals for future research.

2 Feature Selection

Feature selection methods have received a great deal of attention in the classification literature, which can be described according to their relationship with the induction algorithm in three categories [10]: filters, wrappers and embedded methods. Since wrapper and embedded methods interact with the classifier, we opted for filter methods. Besides, filter methods are a common choice in the new Big Data scenario, mainly due to their low computational cost compared to the wrapper or embedded methods. Below we describe the seven filters used in the experimental study.

- **Correlation-based Feature Selection** (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function [12].
- The **INTERACT** (INT) algorithm is based on symmetrical uncertainty and it also includes the consistency contribution [23].
- **Information Gain** (IG) filter evaluates the features according to their information gain and considers a single feature at a time [11].
- **ReliefF** algorithm (RelF) [13] estimates features according to how well their values distinguish among the instances that are near to each other.
- **Mutual Information Maximisation** (MIM) [15] ranks the features by their mutual information score, and selects the top k features, where k is decided by some predefined need for a certain number of features or some other stopping criterion.
- The **minimum Redundancy Maximum Relevance** (mRMR) [20] feature selection method selects features that have the highest relevance with the target class and are also minimally redundant. Both maximum-relevance and minimum-redundancy criteria are based on mutual information.
- **Joint Mutual Information** (JMI) [22] is another feature selection method based on mutual information, which quantifies the relevancy, the redundancy and the complementarity.

3 Datasets

In order to evaluate empirically the effect of feature selection, we employed 55 real datasets. 38 datasets were downloaded from the UCI repository [1], with the restriction of having at least 9 features. Additionally, 17 microarray datasets were used due to their high dimensionality [17]. Tables 1 and 2 profile the main characteristics of the real datasets used in this research in terms of the number of samples, features and classes. Continuous features were discretized, using an equal-width strategy in 5 bins, while features already with a categorical range were left untouched.

4 Experimental Results

The different experiments carried out consist of making comparisons between the application of the seven feature selection methods individually, as well as the random selection (represented as “Ran” in the tables/figures), which will be the baseline for our comparisons. While two of the feature selection methods return a feature subset (CFS and INTERACT), the other five (IG, ReliefF, MIM, JMI and mRMR) are ranker methods, so a threshold is mandatory in order to obtain a subset of features. In this work we have opted for retaining the top 10%, 20% and $\log_2(n)$ of the most relevant features of the ordered ranking, where n is the number of features in a given dataset. In the case of microarray datasets, due to the mismatch between dimensionality and sample size, the thresholds selected the top 5%, 10% and $\log_2(n)$ features, respectively. We computed 3×5 -fold cross validation to estimate the error rate.

According to the No-Free-Lunch theorem, the best classifier will not be the same for all the datasets [21]. For this reason, the behavior of the feature selection methods

Table 1. Characteristics of the 38 real datasets. It shows the number of samples (#sam.), features (#feat.) and classes (#cl.).

| Dataset | #sam. | #feat. | #cl. | Dataset | #sam. | #feat. | #cl. |
|------------------|-------|--------|------|---------------------|-------|--------|------|
| arrhythmia | 452 | 279 | 13 | molec-biol-promoter | 106 | 57 | 2 |
| bc-wisc-diag | 569 | 30 | 2 | molec-biol-splice | 3190 | 60 | 3 |
| bc-wisc-prog | 198 | 33 | 2 | musk-2 | 6598 | 166 | 2 |
| breast | 569 | 30 | 2 | optdigits | 5620 | 64 | 10 |
| coil20 | 1440 | 1024 | 20 | ozone | 2536 | 72 | 2 |
| congress | 435 | 16 | 2 | page-blocks | 5473 | 10 | 5 |
| conn-bench-sonar | 208 | 60 | 2 | parkinsons | 195 | 22 | 2 |
| connect-4 | 67557 | 42 | 2 | pendigits | 10992 | 16 | 10 |
| dermatology | 366 | 34 | 6 | satimage | 6435 | 36 | 6 |
| gisette | 7000 | 5000 | 2 | segmentation | 2310 | 19 | 7 |
| glass | 214 | 9 | 6 | semeion | 1593 | 256 | 10 |
| heart | 270 | 13 | 2 | sonar | 208 | 60 | 2 |
| hill-valley | 606 | 100 | 2 | soybeansmall | 47 | 36 | 4 |
| ionosphere | 351 | 35 | 2 | spect | 267 | 23 | 2 |
| isolet | 7797 | 617 | 2 | splice | 3175 | 60 | 3 |
| krvskp | 3196 | 36 | 2 | USPS | 9298 | 256 | 10 |
| landstat | 5435 | 36 | 6 | waveform | 5000 | 40 | 3 |
| libras | 360 | 90 | 15 | wine | 178 | 13 | 3 |
| low-res-spect | 531 | 100 | 9 | zoo | 101 | 17 | 7 |

will be tested according to the classification error obtained by five different classifiers, each belonging to a different family. The classifiers employed were: two linear (naive Bayes and Support Vector Machine using a linear kernel) and three nonlinear (C4.5, k -Nearest Neighbor with $k = 3$ and Random Forest). All five classifiers were executed using the Weka tool, with default values for the parameters.

4.1 Real Datasets

This section reports the experimental results achieved by the different feature selection methods over the 38 real datasets, depending on the classifier. In order to explore the statistical significance of our classification results, we analyzed the classification error by using a Friedman test with the Nemenyi post-hoc test. The following figures present the critical different (CD) diagrams, introduced by Demšar [6], where groups of methods that are not significantly different (at $\alpha = 0.10$) are connected. The top line in the critical difference diagram is the axis on which we plot the average ranks of methods. The axis is turned so that the lowest (best) ranks are to the right since we perceive the methods on the right side as better. As can be seen in Fig. 1, regardless of the classifier used, it seems that the most suitable feature selection methods for this type of datasets are CFS and INTERACT, which have the additional advantage that there is no threshold

Table 2. Characteristics of the 17 microarray datasets. It shows the number of samples (#sam.), features (#feat.) and classes (#cl.).

| Dataset | #sam. | #feat. | #cl. | Dataset | #sam. | #feat. | #cl. |
|---------------|-------|--------|------|-------------|-------|--------|------|
| 9-tumors | 60 | 5726 | 9 | gli85 | 85 | 22283 | 2 |
| 11-tumors | 174 | 12533 | 11 | leukemia-1 | 72 | 5327 | 3 |
| brain | 21 | 12625 | 2 | leukemia-2 | 72 | 11225 | 3 |
| brain-tumor-1 | 90 | 5920 | 5 | lung-cancer | 203 | 12600 | 5 |
| brain-tumor-2 | 50 | 10367 | 4 | ovarian | 253 | 15154 | 2 |
| CLL-SUB-111 | 111 | 11340 | 3 | smk | 187 | 19993 | 2 |
| CNS | 60 | 7129 | 2 | SRBCT | 83 | 2308 | 4 |
| colon | 62 | 2000 | 2 | TOX-171 | 171 | 5748 | 4 |
| DLBCL | 47 | 4026 | 2 | | | | |

for the number of features to select. In the case of ranker methods, which do need to set a threshold, in general it seems that the percentage of 20% is the best option.

We now compare the classification error achieved by the feature selection methods and our baseline, the random selection. It can be seen that for all the classification algorithms, the random selection, both with the logarithmic and 10% thresholds, is the one that obtains the worst results. However, we can also see that random selection, with the 20% threshold, achieves highly competitive results compared to certain feature selection methods. Due to the drawbacks of the traditional tests of contrast of the null hypothesis pointed up by [2], we have chosen to apply the Bayesian hypothesis test [14], in order to analyze the classification results achieved by “Ran-20” and the ranker methods. In this type of analysis, a previous step is needed, which consists in the definition of the Region of practical equivalence (Rope). Two methods are considered practically equivalent in practice if their mean differences given a certain metric are less than a predefined threshold. In our case, we will consider two methods as equivalent if the difference in error is less than 1%.

For the whole benchmark and each pair of methods, we calculate the probability of the three possibilities: (i) random selection (Ran) wins over filter method with a difference larger than *rope*, (ii) filter method wins over random selection with a difference larger than *rope*, and (iii) the difference between the results are within the *rope* area. If one of these probabilities is higher than 95%, we consider that there is a significant difference. Thus, Fig. 2 shows the distribution of the differences between each pair of methods using simplex graphs. It can be seen that, although random selection with the 20% threshold is not statistically significant with respect to the results obtained over several ranker methods, it always outperforms them. This means that applying the ranker methods (ReliefF, InfoGain and MIM) with an incorrect threshold produces results comparable to those obtained when randomly selecting the 20% of features. These results highlight the importance of choosing a good threshold, which is not a trivial task, especially because it usually depends on the particular problem to be solved (and sometimes, even the classifier that is subsequently used).

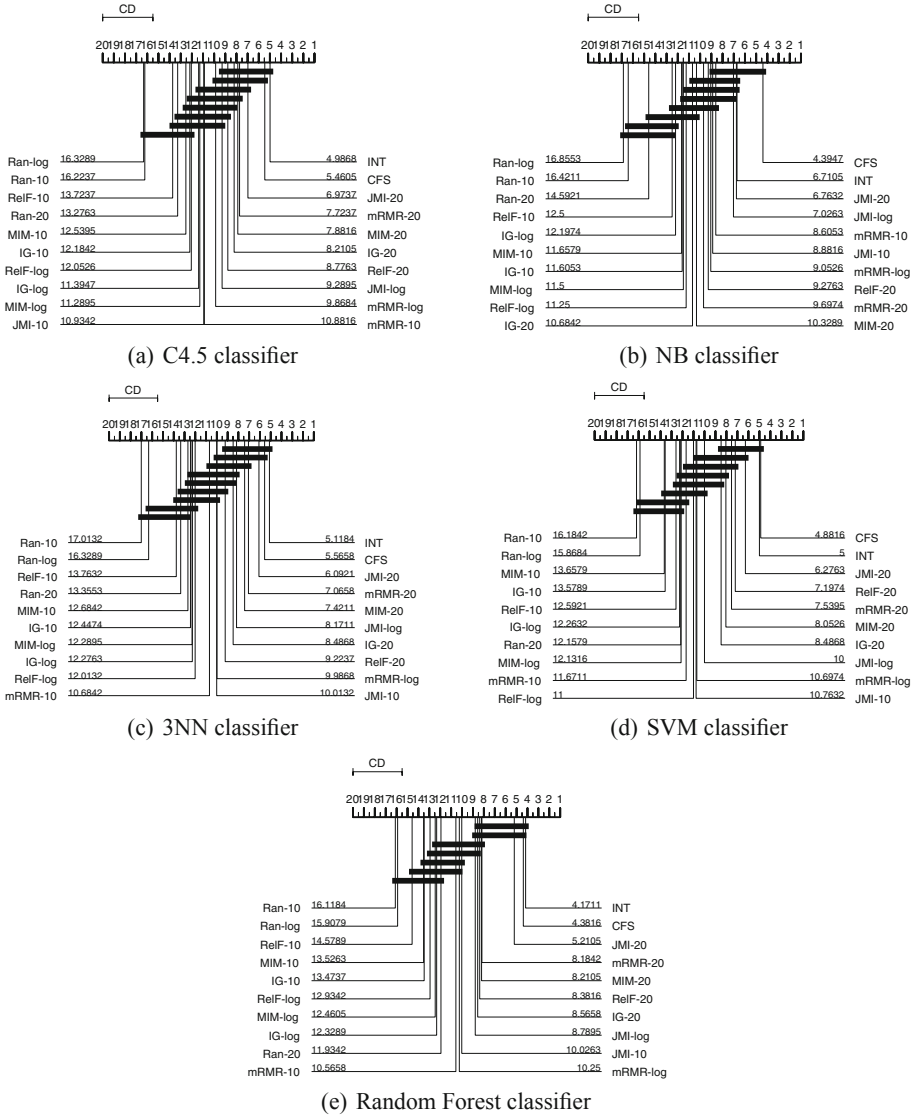


Fig. 1. Critical difference diagrams showing the ranks after applying feature selection over the 38 real datasets. For feature selection methods that require a threshold, the option to keep 10% is indicated by ‘-10’, the option to stay with 20% is indicated by ‘-20’, and the option ‘-log’ refers to use \log_2 .

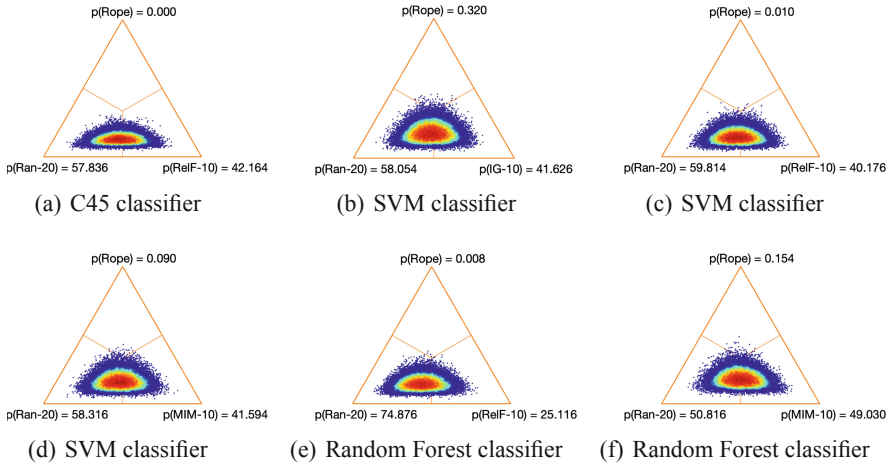


Fig. 2. Simplex graphs for pair comparison of each feature selection method and the baseline random selection (Ran) over the 38 real datasets using Bayesian hierarchical tests: random selection (left) and filter method (right).

Regarding the five different classifiers used, Table 3 shows the classification error obtained by the five classifiers and the eight feature selection methods—the seven filters and the random selection—over the 38 real datasets (lower error rates highlighted in bold). As can be seen, although the classification results obtained were not considerably different between the different feature selection methods used, it is notable that the results obtained with Random Forest outperformed those achieved by the other classifiers.

4.2 Microarray Datasets

The classification of DNA microarray has been viewed as a particular challenge for machine learning researchers, mainly due to the mismatch between dimensionality and sample size. Several studies have demonstrated that most of the genes measured in microarray experiment do not actually contribute to efficient sample classification [4]. To avoid this *curse of dimensionality*, feature selection is advisable so as to identify the specific genes that enhance classification accuracy.

Following the same study as for the previous datasets, and in order to analyze the ranks of the feature selection methods over the 17 microarray datasets, Fig. 3 presents the critical different diagrams for each classification algorithm. As can be seen, the feature selection method that performs best varies depending on the classifier. However, we can say that, in general, CFS is the best option. With regard to the different thresholds used by the ranker methods, the percentage that retains 5% of the features seems to be the most appropriate for these high dimensional datasets.

Table 3. Classification errors obtained by the five classifiers for the real datasets tested.

| | C4.5 | NB | 3NN | SVM | RF |
|----------|-------|-------|-------|-------|--------------|
| CFS | 15.17 | 18.05 | 14.83 | 14.85 | 13.06 |
| INT | 15.01 | 18.87 | 14.99 | 14.98 | 12.80 |
| IG-10 | 22.05 | 26.51 | 21.96 | 24.93 | 21.12 |
| IG-20 | 18.17 | 23.52 | 18.20 | 19.88 | 16.88 |
| IG-log | 21.65 | 27.30 | 21.96 | 25.84 | 20.92 |
| RelF-10 | 23.66 | 27.67 | 23.88 | 25.13 | 22.87 |
| RelF-20 | 19.86 | 24.39 | 19.84 | 20.33 | 18.11 |
| RelF-log | 23.57 | 28.12 | 23.40 | 26.27 | 22.67 |
| MIM-10 | 22.08 | 26.64 | 22.24 | 25.08 | 21.23 |
| MIM-20 | 18.13 | 23.55 | 17.92 | 19.88 | 16.69 |
| MIM-log | 21.88 | 27.37 | 22.23 | 26.04 | 20.98 |
| mRMR-10 | 20.79 | 24.15 | 20.64 | 23.19 | 19.56 |
| mRMR-20 | 18.10 | 23.35 | 17.88 | 19.66 | 16.57 |
| mRMR-log | 19.48 | 23.79 | 19.31 | 22.93 | 18.39 |
| JMI-10 | 20.34 | 23.29 | 19.95 | 22.44 | 19.02 |
| JMI-20 | 16.84 | 20.70 | 16.40 | 17.95 | 15.05 |
| JMI-log | 18.89 | 22.43 | 18.55 | 21.98 | 17.64 |
| Ran-10 | 30.34 | 34.87 | 30.87 | 32.08 | 29.45 |
| Ran-20 | 23.66 | 29.15 | 24.12 | 24.96 | 22.13 |
| Ran-log | 29.16 | 34.66 | 29.69 | 32.66 | 28.57 |

If we observe in depth the results provided by the statistical tests, we can also see that the random selection, both for the thresholds that retain 5 and 10% and for the logarithm, obtains the poorest classification accuracy in the C4.5, NB, 3NN and Random Forest classifiers. The SVM results show a particularly interesting behavior. It seems that this classification algorithm does not work too well when the number of features is low (compared to the original size of the dataset) [16]. Remember that, in the case that the threshold used by the ranker methods select the top $\log_2(n)$ features, the number of features used to train the model will be a maximum of 15 for these datasets (not even 1% of the number of features in the original microarray dataset). Analogously as with the real datasets, Fig. 4 shows the distribution of the differences between random selection—with 5% and 10% thresholds—and the ranker methods with the logarithm threshold using simplex graphs. As can be seen, the random selection performs better on average and with statistical significance over the ranker methods which retain the top $\log_2(n)$ features. Again, these results demonstrate, and in this case more prominently, that an incorrect choice of threshold when using ranker methods might lead to performance as poor as with a random selection of features. This problem is difficult to solve, as the only way to ensure that we are using the correct threshold is to try a significant number of them and compute the classification performance for that subset of features, which would result in inadmissible computation times.

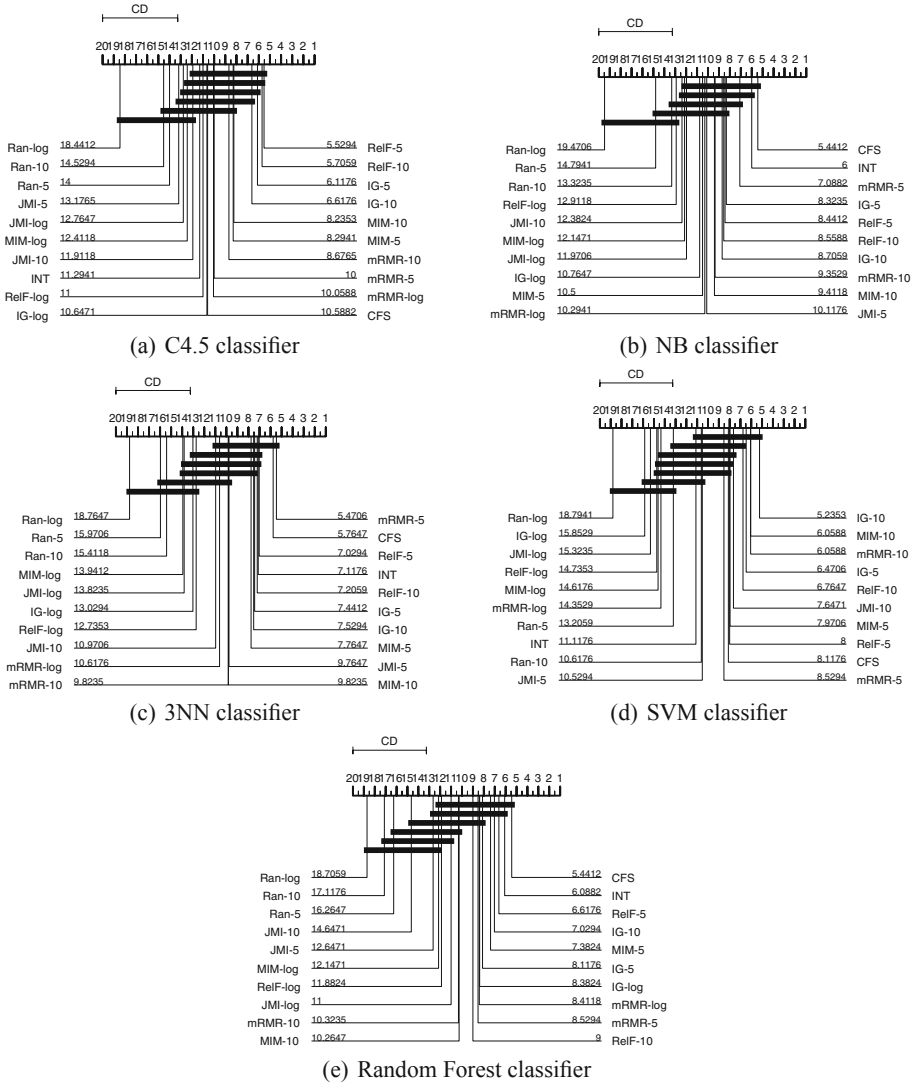


Fig. 3. Critical difference diagram showing the ranks after applying feature selection over the 17 microarray datasets. For feature selection methods that require a threshold, the option to keep 5% is indicated by ‘-5’, the option to stay with 10% is indicated by ‘-10’, and the option ‘-log’ refers to use \log_2 .

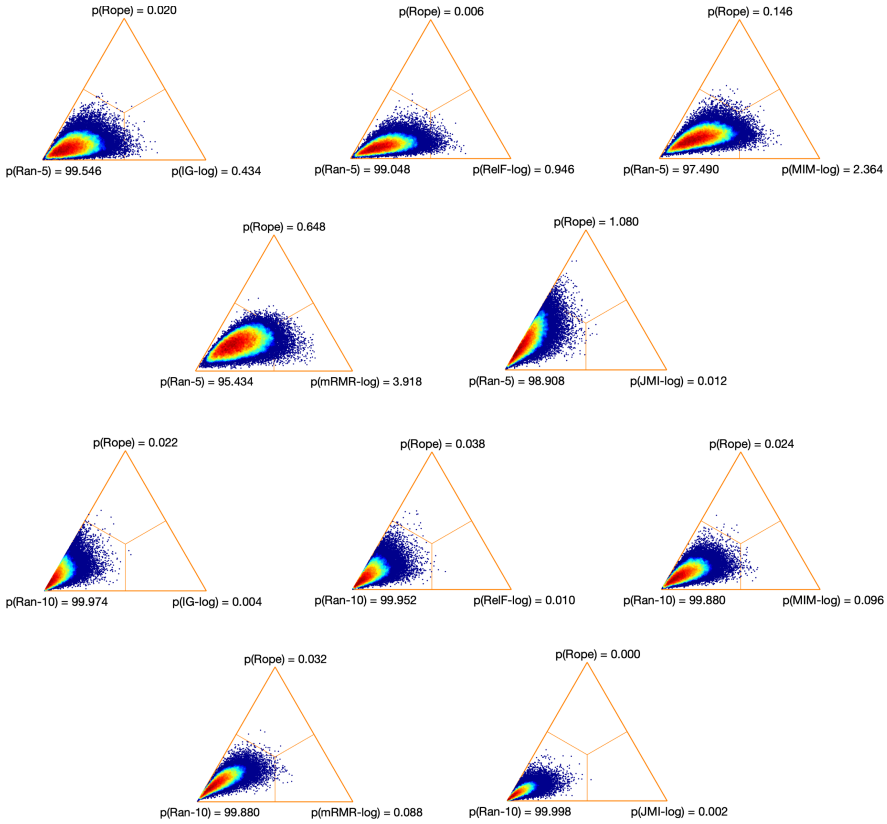


Fig. 4. Simplex graphs for pair comparison of each feature selection method and the baseline random selection (Ran) over the 17 microarray datasets for SVM classifier using Bayesian hierarchical tests: random selection (left) and filter method (right).

Table 4 shows the classification error obtained by the five classifiers and the eight feature selection methods over the 17 microarray datasets (the lowest error rates highlighted in bold). These results show the superiority in performance of SVM over other classifiers in this domain, as it is also stated in González-Navarro [19].

Table 4. Classification errors obtained by the five classifiers for the microarray datasets tested.

| | C4.5 | NB | 3NN | SVM | RF |
|----------|-------|--------------|--------------|--------------|--------------|
| CFS | 30.15 | 19.77 | 19.49 | 17.53 | 22.52 |
| INT | 30.40 | 20.26 | 19.56 | 18.46 | 22.56 |
| IG-5 | 27.10 | 21.98 | 20.08 | 15.88 | 23.73 |
| IG-10 | 27.52 | 22.05 | 20.55 | 15.73 | 23.52 |
| IG-log | 30.54 | 23.37 | 24.73 | 25.60 | 23.98 |
| RelF-5 | 27.46 | 22.99 | 19.00 | 16.90 | 23.16 |
| RelF-10 | 27.10 | 23.01 | 19.04 | 16.81 | 24.81 |
| RelF-log | 31.76 | 27.24 | 25.73 | 27.30 | 26.91 |
| MIM-5 | 29.08 | 23.73 | 20.37 | 16.70 | 24.40 |
| MIM-10 | 28.83 | 22.94 | 21.15 | 15.82 | 25.28 |
| MIM-log | 31.90 | 24.95 | 25.78 | 24.86 | 27.00 |
| mRMR-5 | 30.07 | 21.67 | 18.92 | 16.74 | 24.63 |
| mRMR-10 | 29.45 | 22.94 | 21.15 | 15.82 | 25.97 |
| mRMR-log | 30.33 | 23.56 | 23.71 | 24.31 | 24.84 |
| JMI-5 | 32.72 | 24.17 | 23.19 | 17.89 | 27.77 |
| JMI-10 | 32.06 | 25.19 | 23.68 | 16.72 | 29.36 |
| JMI-log | 32.51 | 25.91 | 27.21 | 26.28 | 27.16 |
| Ran-5 | 33.00 | 28.08 | 28.22 | 19.62 | 32.08 |
| Ran-10 | 32.69 | 26.66 | 28.11 | 17.83 | 32.96 |
| Ran-log | 43.70 | 43.00 | 41.62 | 41.47 | 41.35 |

5 Conclusions

The objective of this work is to study in an exhaustive way the most popular methods in the field of feature selection, making the corresponding comparisons between them, as well as to determine if there exist some methods that are not able to outperform those results obtained by the random selection. We performed experiments with 55 datasets (including the challenging family of DNA microarray datasets) and demonstrated that, in general, feature selection is effective and, in most of the cases, the feature selection methods are better than the random selection, as expected.

In particular, our experiments showed that CFS is a very good choice for any type of dataset. Therefore, in complete ignorance of the particularities of the problem to be solved, we suggest the use of the CFS method, which has the added advantage of not having to establish a threshold. Regarding the use of different thresholds, it seems that 20% is more appropriate for the normal datasets (although worse than the subset methods, which are the winning option for this type of datasets) and the 5% threshold for microarray datasets. Indeed, our experiments confirmed that the choice of threshold when using ranker feature selection methods is critical. In particular, for some thresholds, the results obtained were as poor as when just randomly selecting some features.

Besides, although the classification results obtained were not considerably different between the feature selection methods used—as discussed in Morán-Fernández et al. [18]—, we can conclude that Random Forest in the case of the real datasets and SVM in the case of the microarrays were those that obtained, in a general way over all the datasets used, the best results in terms of classification precision, as Fernández-Delgado et al. [7] concluded in their study.

As mentioned before, the study of an adequate threshold for ranker-type methods is a major problem in the field of feature selection that has yet to be resolved. Thus, as future research, we plan to test a larger number of thresholds, as well as develop an automatic threshold for each type of dataset.

References

1. Bache, K., Lichman, M.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. <http://archive.ics.uci.edu/ml/>. Accessed Dec 2020
2. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *J. Mach. Learn. Res.* **18**(1), 2653–2688 (2017)
3. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **34**(3), 483–519 (2013)
4. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **282**, 111–135 (2014)
5. Climente-González, H., Azencott, C.A., Kaski, S., Yamada, M.: Block HSIC lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics* **35**(14), i427–i435 (2019)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006)
7. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**(1), 3133–3181 (2014)
8. Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A.: Nanotoxicology data for in silico tools: a literature review. *Nanotoxicology* 1–26 (2020)
9. Grgic-Hlaca, N., Zafar, M.B., Gummedi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. *AAAI* **18**, 51–60 (2018)
10. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction: Foundations and Applications, vol. 207. Springer, Heidelberg (2008). <https://doi.org/10.1007/978-3-540-35488-8>
11. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning (1998)
12. Hall, M.A.: Correlation-based feature selection for machine learning (1999)
13. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
14. Kuncheva, L.I.: Bayesian-analysis-for-comparing-classifiers (2020). <https://github.com/LucyKuncheva/Bayesian-Analysis-for-Comparing-Classifiers>
15. Lewis, D.D.: Feature selection and feature extraction for text categorization. In: Proceedings of the workshop on Speech and Natural Language, pp. 212–217. Association for Computational Linguistics (1992)

16. Miller, A.: *Subset Selection in Regression*. CRC Press, Cambridge (2002)
17. Morán-Fernández, L., Bolón-Canedo, V., Alonso-Betanzos, A.: Can classification performance be predicted by complexity measures? a study using microarray data. *Knowl. Inf. Syst.* **51**(3), 1067–1090 (2017)
18. Morán-Fernández, L., Bolón-Canedo, V., Alonso-Betanzos, A.: Do we need hundreds of classifiers or a good feature selection? In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 399–404 (2020)
19. Navarro, F.F.G.: *Feature selection in cancer research: microarray gene expression and in vivo 1h-mrs domains*. Ph.D. thesis, Universitat Politècnica de Catalunya (UPC) (2011)
20. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005)
21. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. *Neural Comput.* **8**(7), 1341–1390 (1996)
22. Yang, H.H., Moody, J.: Data visualization and feature selection: new algorithms for non-gaussian data. In: *Advances in Neural Information Processing Systems*, pp. 687–693 (2000)
23. Zhao, Z., Liu, H.: Searching for interacting features in subset selection. *Intell. Data Anal.* **13**(2), 207–228 (2009)