Evaluation of Rule-Based Learning and Feature Selection Approaches For Classification

Fatima Chiroma[1], Mihaela Cocea[1], and Han Liu[2]

1 **School of Computing**
University of Portsmouth, United Kingdom
fatima.chiroma@port.ac.uk, mihaela.cocea@port.ac.uk
2 **School of Computer Science and Informatics**
Cardiff University, United Kingdom
LiuH48@cardiff.ac.uk

Abstract

Feature selection is typically employed before or in conjunction with classification algorithms to reduce the feature dimensionality and improve the classification performance, as well as reduce processing time. While particular approaches have been developed for feature selection, such as filter and wrapper approaches, some algorithms perform feature selection through their learning strategy. In this paper, we are investigating the effect of the implicit feature selection of the PRISM algorithm, which is rule-based, when compared with the wrapper feature selection approach employing four popular algorithms: decision trees, naïve bayes, k-nearest neighbors and support vector machine. Moreover, we investigate the performance of the algorithms on target classes, i.e. where the aim is to identify one or more phenomena and distinguish them from their absence (i.e. non-target classes), such as when identifying benign and malign cancer (two target classes) vs. non-cancer (the non-target class).

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases Feature Selection, Prism, Rule-based Learning, Wrapper Approach

Digital Object Identifier 10.4230/OASIcs.xxx.yyy.p

1 Introduction

The application of machine learning has been on the rise in recent years [10] as its various techniques have been applied to different problem domains successfully. For example, in medicine, machine learning techniques have been used to predict the effectiveness of drugs in patients with depression [16] while in finance it was used to detect fraudulent activities on credit cards [16], to mention a few. Furthermore, approaches used by machine learning techniques differ. For example, rule-based learning is a machine learning technique that makes its decisions based on a number of rules [15] - a popular rule-based algorithm is Prism [8, 5, 18, 3], which works with the concept of target class and is capable of selecting attributes based on their importance to a particular class.

Another machine learning technique is Feature selection. Its strategy is to select only the attributes that are relevant and effective from a large number of features or attributes in a data-set where the selected attribute determines the performance of the classification [6, 12]. These approaches will be explored in this study, especially their performance when applied to classification problems; more specifically, we will investigate the performance of Prism, which has implicit feature selection, in comparison with other feature selection approaches.

This paper is organized as follows: Section 1 introduces the background of the study; section 2 reviews the related works that have been carried out by various researchers; Section

Conference/workshop/symposium title on which this volume is based on.
Editors: Billy Editor and Bill Editors; pp. 1–6
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

3 discusses the experimental approach; Section 4 comprises of the results and discussion; and Section 5 concludes the study and presents the future work.

2 Related Work

Classification is one of the most popular machine learning tasks which typically involves the training of an algorithm to build a model which is subsequently used to identify the category of an unseen instance [8]. Various research relating to classification has been carried out using several approaches. Rule-based learning and Feature selection are some of the approaches that have been applied to classification problems.

Rule-based learning is an approach in which the model consists of a set of rules which were learned from the data [15]. For example, Prism, which is a rule learning algorithm learns from a set of rules that separates a specific class i.e. the target class from other classes [8, 13]. Prism has been used by several researchers in classification problems. For example, [5] who developed Prism, used it to identify types of contact lenses and the results have shown that Prism has a higher classification accuracy than ID3, a decision tree learning algorithm. Another study showed a 92% classification accuracy was achieved by Prism on image segmentation data set for multi-task feature selection [13].

Feature selection can be done by following one of three approaches, i.e. filter, wrapper or embedded approaches [17].The filter method does not require the application of a classification algorithm to evaluate the quality of the features selected while the wrapper method is the opposite [14], i.e. it is dependent on the classification algorithm to evaluate the quality of selected features. The embedded method, on the other hand, performs its feature selection as the optimal parameters are being learned [14].

A study was done by [17], where they compared the Naïve Bayes wrapper feature selection with other filter feature selection algorithms on Human Activity Recognition machine learning problem. The result of their study showed that the wrapper method outperformed all the filter algorithms and they were also able to discover that features selected by the wrapper method are efficiently usable with other machine learning algorithms.

The research that is most related to this study is the research done by [15]. They used feature selection with wrapper approach based on ensemble learning on 13 data-sets using two base learners: Decision Tree and Naïve Bayes. They were able to identify which wrapper approach has better classification accuracy and their results showed that the forward selection when applied to Decision Tree had the highest accuracy in their study.

Therefore, the aim of this study is to explore how the implicit feature selection within Prism compares with the wrapper feature selection approach.

3 Data and Experimental Setup

The experiment was carried out using seven classification data-sets which were acquired through the UCI Machine Learning repository [9] and Knowledge Extraction based on Evolutionary Learning (KEEL) data-set repository [4].

Table 1 lists the data-sets used, as well as their properties, i.e. the number of instances, the number of attributes, the type of attributes and the number of classes. We chose data-sets with at least 3 classes, as we focused our investigation on non-binary data-sets, where one or more target classes need to be distinguished from one or more non-target classes.

All data-sets in Table 1 are classification data-sets with numeric data, that have already been pre-processed before acquisition. However, due to the importance of pre-processing for

classification [8] additional pre-processing was done to ensure the data is clean, compatible and ready for classification. These pre-processing includes the conversion of the label or class attribute to string, filtration of attributes that are not relevant for the study, renaming of important attribute names for better readability and easier identification, and also the concatenation of data-sets with multiple files.

Table 1 Data-sets Description

S/N	Name	Instances	Attributes	Type	Classes
1	Balance Scale	625	4	Integer	3
2	Breast Tissue	106	10	Real	4
3	Forest Type	198	27	Integer, Real	4
4	Heart Disease Cleveland	303	75	Integer, Real	5
5	Lymphography	148	18	Integer	4
6	Soybean	47	35	Integer	4
7	Website Phishing	1353	9	Integer	3

These data-sets were classified using five machine learning algorithms: Prism, Decision Tree (DT), Naïve Bayes (NB), Library of Support Vector Machine (LibSVM) and K-Nearest Neighbors (KNN). Prism is the target-classifier for this experiment and the remaining four are subsequently going to be referred to as the other-classifiers.

Furthermore, the forward selection and backward elimination algorithms which are based on the wrapper feature selection method, were applied to the data-sets using the other-classifiers. The reason for using both the forward selection and the backward elimination algorithm is due to the fact that the forward selection algorithm is known to improve accuracy but only on some data-sets as it may not have any effect on others [10], while the backward elimination allows for backtracking when it removes features therefore allowing for the inclusion of previously eliminated features [15, 2].

For the evaluation, the 10-fold cross validation was applied to both the target-classifier and other-classifiers. This validation technique was applied due to its ability to limit the level of influence of randomly selected training sets on the overall results [8].

4 Results and Discussion

The results of the experiment have been presented in three tables for better comparison across the machine classifiers and data-sets. Thus, we present the results across all classes (Table 2) and across the target classes (Table 3), as well as the number of features selected (Table 4). In terms of the performance of the algorithms, we report the F-measure (which is the harmonic mean of precision [1] and recall [2]) rather than accuracy, as it is less influenced by an unbalanced distribution of instances across classes.

Table 2 shows the performance of the machine classifiers for each data-set. The results show that Prism has the highest performance on two of the data-sets: Website Phishing and Heart Disease Cleveland with an F-measure of 0.88 and 0.46, respectively. Moreover, for the Heart Disease Cleveland data-set which has the highest number of attributes (75) and classes (5), we notice that Prism outperforms all other wrapper approaches by a very high margin.

For the other 5 data-sets, the results show that: (a) LibSVM is best on three of the data-sets; (b) DT and NB are equally best on one data-set; (c) KNN and LibSVM are equally best on one data-set.

[1] Precision is the number of correctly identified instances from all instances
[2] Recall is the number of correctly identified instances from the subset of relevant instances

The results also show that on the used data-sets the forward selection performance is higher than the performance of the backward elimination approach – this is likely to be due to the simplicity of forward selection and the ability to add only feature with the highest performance[15].

Table 2 All Classes F-Measure Results

Data-sets	Prism	DT		NB		KNN		LibSVM	
		FS	*BE*	*FS*	*BE*	*FS*	*BE*	*FS*	*BE*
Balance Scale	0.62	0.59	0.44	0.63	0.43	0.62	0.37	0.86	0.43
Breast Tissue	0.82	1.00	1.00	1.00	1.00	0.78	0.73	0.78	0.73
Forest Type	0.88	0.95	0.73	0.93	0.55	0.94	0.54	0.98	0.99
Heart Disease Cleveland	0.46	0.28	0.22	0.31	0.14	0.23	0.14	0.30	0.17
Lymphography	0.43	0.47	0.36	0.45	0.36	0.49	0.18	0.59	0.36
Soybean	0.98	0.98	0.84	0.94	0.84	1.00	0.67	1.00	0.73
Website Phishing	0.88	0.87	0.56	0.65	0.56	0.81	0.23	0.61	0.55

Table 3 shows the performance results for the target classes i.e. the F-measure for only the target classes. It shows that Prism also has the highest performance for the Heart Disease Cleveland data-set with an F-measure of 0.37 but equal performance with Decision Tree for the Website Phishing with an F-measure of 0.86 as well as the soybean data-set which has an F-measure of 1.0 for Prism, Decision Tree, K-Nearest Neighbor and LibSVM. The 1.0 performance on the soybean data-sets obtained by the wrapper approaches was achieved using the forward selection algorithm.

For the other four data-sets, we observe the following: (a) LibSVM and NB are equally best on one data-set; (b) DT and NB are equally best on one data-set; (c) LibSVM is best for two data-sets.

Table 3 Target Class F-Measure Results

Data-sets	Prism	DT		NB		KNN		LibSVM	
		FS	*BE*	*FS*	*BE*	*FS*	*BE*	*FS*	*BE*
Balance Scale	0.93	0.87	0.66	0.95	0.64	0.92	0.56	0.95	0.675
Breast Tissue	0.83	1.00	1.00	1.00	1.00	0.72	0.44	0.67	0.39
Forest Type	0.89	0.95	0.74	0.94	0.47	0.96	0.48	0.98	0.69
Heart Disease Cleveland	0.37	0.17	0.09	0.18	0	0.12	0	0.17	0.03
Lymphography	0.58	0.62	0.62	0.67	0.48	0.65	0.65	0.79	0.79
Soybean	1.00	1.00	0.79	0.97	0.79	1.00	0.83	1.00	0.64
Website Phishing	0.86	0.86	0.41	0.53	0.41	0.76	0	0.47	0.40

For classification, attributes can be redundant, irrelevant or problematic [15]. Therefore, applying feature selection approaches ensures the selection of attributes that are relevant or important. Table 4 shows the total number of attributes selected to achieve the highest performance for each classifier. These selected attributes are considered to be the most relevant for the classification; however, these may vary across the different approaches.

Prism used the most number of attributes across all data-sets when compared with the wrapper approaches. This seems to be an advantage in some situations, e.g. on the Heart Disease Cleveland and the Website Phishing data-sets, but not in others. Thus, Prism performed better with data-sets that have large instances or high number of attributes.

Furthermore, the other-classifiers performed better with the forward selection algorithm than the backward elimination. Also, Prism had higher performance than all the classifiers for the backward elimination algorithm, except for LibSVM for the target classes, which has an F-measure of 0.98 for the Lymphography data-sets.

Table 4 Number of Attributes

Data-sets	Total Attributes	Prism	DT		NB		KNN		LibSVM	
			FS	*BE*	*FS*	*BE*	*FS*	*BE*	*FS*	*BE*
Balance Scale	4	4	4	4	4	4	4	4	4	4
Breast Tissue	10	9	2	2	2	2	7	4	2	4
Forest Type	27	22	4	4	10	11	8	6	17	9
Heart Disease Cleveland	75	13	4	3	4	3	4	9	4	7
Lymphography	18	17	4	8	10	10	16	11	11	15
Soybean	35	4	2	2	2	2	2	2	2	2
Website Phishing	9	9	6	7	4	8	9	9	9	5

Additionally, according to [15] one of the benefits of feature selection is the reduction of run-time for large and multidimensional data-sets as well as increased accuracy. However, on the used data-sets, LibSVM which is a library of Support Vector Machine [7], had the longest processing time.

5 Conclusion and Future Work

In this paper, we explored how the implicit feature selection within prism compares with the wrapper feature selection approach using four popular machine learning algorithms: Decision Tree, Naïve Bayes, LibSVM and K-Nearest Neighbour. The results of the experiments have shown that both Prism and the other-classifiers have varying performance. Therefore, we will further extend this study by exploring the same approach and algorithms on text data-sets to measure its performance for text classification. We will also further investigate what properties of data make Prism more suitable for some classification problems than others.

Acknowledgements The authors would like to extend their gratitude to the Petroleum Technology Development Fund for their support. Additionally, some of the data-sets used in this study: the lymphography data-set was obtained by M. Zwitter and M. Soklic from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia; the website phishing data-set was compiled by [1]; the forest type by [11] and the Heart Disease Cleveland data-set was provided by Robert Detrano, M.D., Ph.D. from the Cleveland Clinic Foundation.

References

1 Neda Abdelhamid, Aladdin Ayesh, and Fadi Thabtah. Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13):5948–5959, 2014.

2 Shigeo Abe. Modified backward feature selection by cross validation. In *ESANN*, pages 163–168, 2005.

3 Maher Aburrous, M Alamgir Hossain, Keshav Dahal, and Fadi Thabtah. Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert systems with applications*, 37(12):7913–7921, 2010.

4 Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. Keel data-mining software tool: data set repository,

integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

5 Jadzia Cendrowska. PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, 27(4):349–370, 1987.

6 Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

7 Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

8 Fatima Chiroma, Han Liu, and Mihaela Cocea. Suicide related text classification with prism algorithm. *International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6, 2018.

9 Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

10 Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

11 Brian Johnson, Ryutaro Tateishi, and Zhixiao Xie. Using geographically weighted variables for image classification. *Remote sensing letters*, 3(6):491–499, 2012.

12 Vipin Kumar and Sonajharia Minz. Feature selection. *SmartCR*, 4(3):211–229, 2014.

13 Han Liu, Mihaela Cocea, and Weili Ding. Multi-task learning for intelligent data processing in granular computing context. *Granular Computing*, 3(3):257–273, 2018.

14 Mehdi Naseriparsa, Amir-Masoud Bidgoli, and Touraj Varaee. A hybrid feature selection method to improve performance of a group of classification algorithms. *International Journal of Computer Applications*, 69(17):28–35, 2013.

15 Rattanawadee Panthong and Anongnart Srivihok. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. *Procedia Computer Science*, 72:162–169, 2015.

16 The Royal Society. Machine learning: the power and promise of computers that learn by example. Online, April 2017.

17 Jozsef Suto, Stefan Oniga, and Petrica Pop Sitar. Comparison of wrapper and filter feature selection algorithms on human activity recognition. In *Computers Communications and Control (ICCCC), 2016 6th International Conference on*, pages 124–129. IEEE, 2016.

18 Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.