Data Cleansing Meets Feature Selection: A Supervised Machine Learning Approach

Antonio J. Tallón-Ballesteros(✉) and José C. Riquelme

Department of Languages and Computer Systems,
University of Seville, Seville, Spain
atallon@us.es

Abstract. This paper presents a novel procedure to apply in a sequential way two data preparation techniques from a different nature such as data cleansing and feature selection. For the former we have experienced with a partial removal of outliers via inter-quartile range whereas for the latter we have chosen relevant attributes with two widespread feature subset selectors like CFS (Correlation-based Feature Selection) and CNS (Consistency-based Feature Selection), which are founded on correlation and consistency measures, respectively. Empirical results on seven difficult binary and multi-class data sets, that is, with a test error rate of at least a 10%, according to accuracy, with C4.5 or 1-nearest neighbour classifiers without any kind of prior data pre-processing are outlined. Non-parametric statistical tests assert that the meeting of the aforementioned two data preparation strategies using a correlation measure for feature selection with C4.5 algorithm is significant better, measured with roc measure, than the single application of the data cleansing approach. Last but not least, a weak and not very powerful learner like PART achieved promising results with the new proposal based on a consistency measure and is able to compete with the best configuration of C4.5. To sum up, bearing in mind the new approach, for roc measure PART classifier with a consistency metric behaves slightly better than C4.5 and a correlation measure.

Keywords: Data cleansing · Feature selection · Classification · Outlier detection · Inter-quartile range

1 Introduction

Knowledge Discovery in Databases (KDD)[9] is a multidisciplinary paradigm of computer science comprising challenging tasks to transform a problem into useful models for prediction such as dealing with raw data, analysing the problem, data preparation [25] and data mining [17].

Several machine learning approaches have tackled the classification problem. Roughly speaking, algorithms may be divided into strong and weak algorithms according to inner complexity of the classifier. There is a good number of families to get classification models like those based on decision trees, rules, nearest neighbours, support vectors and neural networks.

Data pre-processing is crucial due to some issues: a) real-world problems may be incomplete or noisy (with errors or outliers); b) the discovery of useful patterns depends on the starting data quality. Data cleansing [7] and feature selection [18] are two samples of data preparation approaches. The former aims to correct errors, to detect and analyse outliers, thus to purify data. The latter pursues to pick up the most important features in order to simplify the model and predict more accurately.

This paper goals to assess the potential usefulness of the ordered application in supervised machine learning problems of two very different data-preprocessing techniques like data cleansing and feature selection. Another additional aim is to improve the performance of the classification models.

The rest of this article is organized as follows: Sect. 2 describes some concepts about data cleansing and feature selection; Sect. 3 presents our proposal; Sect. 4 details the experimentation; then Sect. 5 shows and analyzes statistically the results obtained; finally, Sect. 6 states the concluding remarks.

2 Related Work

2.1 Data Cleansing

An outlier may be defined as a data point which is very different from the rest of the data based on some measure [1], or alternatively as a case that does not follow the same model as the remaining data and appears as though it comes from a different probability distribution [26].

The core question about outliers is to delete or not to delete them. The answer is unclear because there are contributions from both sides. On the one hand, some authors claim either that the deletion of outliers did not significantly change the overall error distribution, accuracy, ... [14], or that the elimination of instances which contain attribute noise is not a good idea, because many other attributes of the instance may still contain valuable information [26]. On the other hand, some works showed that dropping the outlier in the training set may be a beneficial action for the classifier [20].

2.2 Feature Selection

It can be defined as the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept. The different approaches for feature selection (FS) can be divided into two broad categories (i.e., filter and wrapper) based on their dependence on the inductive algorithm that will finally use the selected subset [16]. Filter methods are independent of the inductive algorithm, whereas wrapper methods use the inductive algorithm as the evaluation function. FS involves two phases: a) to obtain a list of attributes according to an attribute evaluator and b) to perform a search on the initial list. All candidate lists would be evaluated using a measure evaluation and the best one will be returned. Correlation-based Feature Selection (CFS) [12] and

Consistency-based feature selection (CNS) [6] are two of the most widespread feature subset selectors (FSS) and both work together with a search method such as Greedy Search, Best First or Exhaustive Search.

3 Proposal

A statistical outlier detection method based on the partial removal of outliers according the inter-quartile range of all the instances with the same class label was introduced in [22] with the name OUTLIERoutP. The framework can be reviewed in Figure 2 of the aforementioned work.

The current paper proposes to complete an additional data preparation stage after the data cleansing from a different perspective such as feature selection [16]. Typically, the application of feature selection has become a real prerequisite for model building in due to the multidimensional nature of many modeling task in some fields [21]. Figure 1 overviews the proposal. It is a generic methodology in the sense that there is no restriction in the kind of feature selection method or the number of classes that the classifier is able to operate with. As usually in data mining field, the data preparation techniques act on the training set and the test set stands unaltered and is evaluated by the first time once the classifier is trained. To the best of our knowledge, the main novelty of this work is to do a further data pre-processing phase by means of feature selection after the application of the data cleansing stage via an outlier detection method. According to the literature, the researches tackle data cleansing or feature selection in an isolated way.

4 Experimentation

Table 1 describes the data sets utilised together with the outlier level according to the taxonomy proposed in [22]. Most of them are publicly available in the UCI (University of California at Irvine) repository [4]. They come from real-world applications of different domains such as Finances, Physics, Life, Environment, and Analysis of Olive Oils. The following seven have been used: Cardiotocography (*CTG*), Statlog (*German* credit), MAGIC Gamma Telescope (*Magic*), Olive Oil (*Olitos*), Pima Indians Diabetes (*Pima*), *Tokyo* and *Water* Treatment Plant. Olitos problem is deeply explained in [3]. The size of the problems ranges from one hundred twenty to more than nineteen thousands. The number of features varies between eight and sixty one, while the number of classes is between two and four. The missing values have been replaced in the case of nominal variables by the mode or, when concerning continuous variables, by the mean, bearing in mind the full data set. The outlier level is computed once the imputation of the missing values has been carried out. These data sets contain a number of outliers that is between a low percentage (up to a 10%) and a moderate-high one (in the range 30-40%, for the *Tokyo* problem). It is important to stress that these outliers are originally present in the data set and we have not performed any artificial way to add them. The taxonomy of the problems depending on the

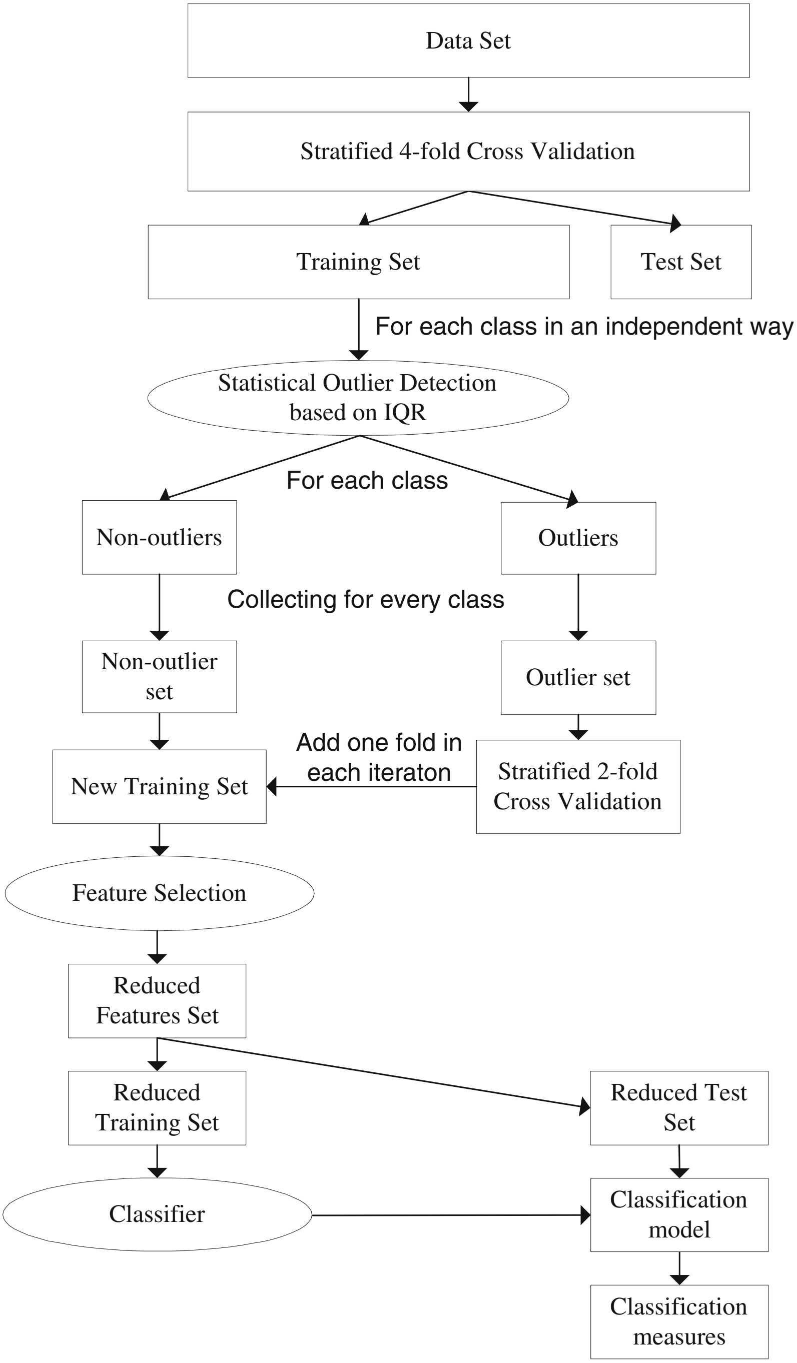

Fig. 1. Methodology OUTLIERoutP+FS

outlier level is based on the inter-quartile range (IQR) by classes. The common point of these data sets is that their important error rate in test phase without any kind of data pre-processing is about 10% or above with reference and robust classifiers such as 1-nearest neighbour (1-NN) [2] [5] or C4.5 [19]. In relation to the experimental design we have followed a stratified 4-fold cross validation [13], whereby the data set is divided into four parts and subsequently a partition is the test set and the three remaining ones are pooled as the training data. On the other hand, for the assessment of the classification models we have chosen the accuracy [15] and roc [8] measures. The former can be defined as the probability of correctly classifying a randomly select pattern. Sometimes, it is called as the number of successful hits [24]. The latter stands for receiver operating characteristic (ROC) and is the area under the ROC curve. We report both measures for the test set that is the performance with unseen data during the generalisation stage.

Table 1. Summary of the data sets

Data set	*Size*	*Features*	*Classes*	*Outlier level*
CTG	2126	23	2	*II*
German	1000	61	2	*I*
Magic	19020	11	2	*I*
Olitos	120	26	4	*II*
Pima	768	8	2	*I*
Tokyo	639	44	2	*IV*
Water	527	38	3	*II*
Average	3457.1	30.1	2.4	$I - II$

Table 2 depicts the data preparation methods concerning the different experiments that were conducted. The first one only includes the data cleansing, whereas the last two ones comprise the execution of the methodology OUTLIERoutP followed by a feature subset selection evaluated with a correlation or consistency measure. Last column defines an abbreviated name for each of them that often will be referred in next sections. We have chosen CFS and CNS as representative feature subset selection methods, because they are based on different kind of measures, have few parameters and have provided a good performance inside the supervised machine learning area. Often, BestFirst search is the preferable option by the researchers for both FSS algorithms. CFS is likely the most used FSS in data mining. CNS is also powerful, however the quantity of published works is more reduced [21]. As classification algorithms we have experienced with C4.5 [19] and PART [10]. For the two previous FSS and classifiers we have used the implementations provided by WEKA tool [11] with default parameters that are those suggested by the own authors of the algorithms.

Table 3 overviews the properties of the data sets with a stratified 4-fold cross validation in three moments: i) in the initial situation (see all the columns containing *Or.* word), ii) after the data pre-processing stage (refer to columns

Table 2. List of data preparation methods for the experimentation

Data cleansing	Attribute evaluator	Search method	Feature selector name	*Abb. Name*
$OUTLIERoutP$	–	–	–	$OutP$
$OUTLIERoutP$	CFS	$BestFirst$	$CFS\_BestFirst$	$OutP + FS1$
$OUTLIERoutP$	CNS	$BestFirst$	$CNS\_BestFirst$	$OutP + FS2$

Table 3. Number of instances and features with the data preparation methods with a 4-fold cross validation

Data set	*Or. Av. Tr. Sz.*	*Av. OutP Tr. Sz.*	♯ *Or. F.*	♯ *Av. F. FS*1	♯ *Av. F. FS*2
CTG	1594.5	1490.6	23	6.8	7.8
german	750.0	744.6	61	7.8	19.9
magic	14265.0	14058.5	10	4.0	10.0
olitos	90.0	80.4	25	12.3	12.3
pima	576.0	556.1	8	3.9	7.3
tokyo	719.3	619.6	44	11.0	11.4
water	390.8	362.9	38	10.8	11.0
Average	2626.5	2559.0	29.9	8.1	11.4

Or. = Original Av. = Average Tr. = Training Sz. = Size F. = Features

labelled with $OutP$) and iii) once the two data preparation approaches have been carried out (see two last columns).

5 Results

Tables 4 reports the accuracy and roc test results averaged with an outer stratified 4-fold cross validation over the original data set followed by an inner stratified 2-fold cross validation adding one fold of the outlier set into the training set in each iteration (see Fig. 1). Also, we have included the number of times that $OutP + FS1$ or $OutP + FS2$ result is better than the $OutP$ one. Finally, last row shows the average for each method with all the data sets. It is clear that the performance measured with accuracy for $OutP + FS1$ and $OutP + FS2$ is very similar to $OutP$ because the number of wins is lower than the half of the number of data sets. On the other hand, the results with roc measure require to be submitted to a deep analysis.

Table 5 shows the results of the non-parametric statistical analysis of OutP (baseline approach) versus $Out + FS1$ or $Out + FS2$. We have represented the average roc value for each method, their difference with the baseline case and its ranking. According to Wilcoxon signed-ranks test, since there are 7 data sets, the T value at $\alpha = 0.05$ should be less or equal than 2 (the critical value) to reject the null hypothesis. On the one hand, $OutP + FS1$ is significantly better than $OutP$. On the other hand, for $OutP + FS2$ the results are statistically in the line of $OutP$ but the $R+$ value is three times the $R-$ value, thus the performance of $OutP + FS2$ is promising in most of the cases.

Table 4. Classifier C4.5: Accuracy and roc test results averaged with a 4-fold cross validation

$Data\ set$	$C4.5$					
	$Accuracy$			Roc		
	$OutP$	$OutP + FS1$	$OutP + FS2$	$OutP$	$OutP + FS1$	$OutP + FS2$
CTG	89.49	86.50	88.88	0.8276	0.8379	0.8236
$german$	70.95	71.50	72.35	0.6170	0.6998	0.7168
$magic$	85.22	82.80	85.22	0.8691	0.8639	0.8691
$olitos$	65.42	68.75	68.75	0.7480	0.7730	0.7730
$pima$	74.86	74.60	74.66	0.7553	0.7645	0.7488
$tokyo$	90.62	91.71	91.03	0.9070	0.9290	0.9095
$water$	84.07	83.11	83.39	0.8035	0.8048	0.8199
$Wins$		3	3		6	4
$Average$	80.09	79.85	80.61	0.7896	0.8104	0.8087

Table 5. C4.5: Statistical tests for roc measure

$Data\ set$	$OutP$	$OutP + FS1$	$Difference$	$Ranking$	$OutP + FS2$	$Difference$	$Ranking$
CTG	0.8276	0.8379	0.0103	4	0.8236	−0.0040	3
$german$	0.6170	0.6998	0.0828	7	0.7168	0.0998	7
$magic$	0.8691	0.8639	−0.0052	2	0.8691	0.0000	1
$olitos$	0.7480	0.7730	0.0250	6	0.7730	0.0250	6
$pima$	0.7553	0.7645	0.0092	3	0.7488	−0.0065	4
$tokyo$	0.9070	0.9290	0.0220	5	0.9095	0.0025	2
$water$	0.8035	0.8048	0.0012	1	0.8199	0.0164	5
			$T = min\{26, 2\} = 2\ (*)$			$T = min\{21, 7\} = 7$	

5.1 Application of the New Proposal to Classifier PART

Once the new approach has been validated according to non-parametric statistical tests we extended it to the PART classifier. Generally speaking, CFS and CNS exhibited an intermediate performance as feature selectors operating directly on the original data [23]. In the current paper we do a previous data cleansing and after that the feature selection phase to evaluate the convenience or not to apply both data pre-processing strategies.

Tables 6 depicts the accuracy and roc test results for PART algorithm averaged with an outer stratified 4-fold cross validation and an inner stratified 2-fold cross validation as explained in the previous section. We have represented the number of times that $OutP + FS1$ is better than $OutP + FS2$ and the average for each method with all the data sets. We should remark that $OutP + FS2$ has an excellent performance for roc measure and wins 5 out of 7 times to $OutP + FS1$.

5.2 Statistical Comparison of the Two Best Classifiers with their Suitable Data Preparation Approach

This subsection compares the two best achievements that have been reported so far in the paper. Table 7 includes the results of the two best options and a

Table 6. Classifier PART: Accuracy and roc test results with OUTLIERoutP+FS for a 4-fold cross validation

Data set	*PART*			
	Accuracy		Roc	
	$OutP + FS1$	$OutP + FS2$	$OutP + FS1$	$OutP + FS2$
CTG	85.89	84.22	0.8210	0.8349
german	72.00	70.50	0.7058	0.6755
magic	82.90	84.71	0.8731	0.8980
olitos	67.50	67.50	0.7858	0.7858
pima	73.63	73.44	0.7571	0.7606
tokyo	91.71	91.29	0.9359	0.9543
water	83.30	83.49	0.8166	0.8443
Wins by pairs	4	2	1	5
Average	79.56	79.31	0.8136	0.8219

Table 7. Statistical comparison between C4.5 with OutP+FS1 and PART with OutP+FS2 for roc measure

Data set	$C4.5$	PART		
	$OutP + FS1$	$OutP + FS2$	*Difference*	*Ranking*
CTG2	0.8379	0.8349	−0.0030	1
german	0.6998	0.6755	−0.0243	4
magic	0.8639	0.8980	0.0341	6
olitos	0.7730	0.7858	0.0127	3
pima	0.7645	0.7606	−0.0039	2
tokyo	0.9290	0.9543	0.0253	5
water	0.8048	0.8443	0.0395	7
Wins	3	4		
		$T = min\{21, 7\} = 7$		

non-parametric statistical analysis via a Wilcoxon signed-ranks test of C4.5+ $OutP + FS1$ versus PART+$OutP + FS2$. Since the T value at $\alpha = 0.05$ is not less or equal than 2 the null hypothesis is accepted. Hence, both algorithms performs equally without significant differences. The good new is that PART behaves better according to the rankings; $R+$ value is three times the $R-$ value, thus PART with the proposed methodology should be consider as an interesting alternative to C4.5 classifier with our new approach.

6 Conclusions

An innovative methodology that performs two data preparation phases such as data cleansing via outlier detection and feature selection, in this order, was introduced. An empirical study on seven binary and multi-class classification problems with a test error rate of a 10% or above measured in accuracy was conducted. The experimentation shed light on that the roc measure is improved in global terms and the accuracy is increased in punctual cases. According to the non-parametric statistical tests, C4.5 with the new approach (OUTLIERoutP+FS) using a correlation measure overcame significantly the results for roc versus the framework OUTLIERoutP that was previously proposed. Moreover, C4.5 and a feature selector with a consistency measure, after the data cleansing stage, achieved better results than OUTLIERoutP in most of the data sets. Finally, the behaviour of a not very powerful classifier such as PART became excellent with the new approach until the extent that PART with a consistency measure reached slightly better results than the best setting of OUTLIERoutP+FS (significantly better than OUTLIERoutP) with C4.5 classifier.

Acknowledgments. This work has been partially subsidized by TIN2007-68084-C02-02 and TIN2011-28956-C02-02 projects of the Spanish Inter-Ministerial Commission of Science and Technology (MICYT), FEDER funds and P11-TIC-7528 project of the ”Junta de Andalucía” (Spain).

References

1. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: ACM Sigmod Record, vol. 30, pp. 37–46. ACM (2001)
2. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6(1), 37–66 (1991)
3. Armanino, C., Leardi, R., Lanteri, S., Modi, G.: Chemometric analysis of tuscan olive oils. Chemometrics and Intelligent Laboratory Systems 5(4), 343–354 (1989)
4. Bache, K., Lichman, M.: UCI machine learning repository (2013)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
6. Dash, M., Liu, H.: Consistency-based search in feature selection. Artificial Intelligence 151(1), 155–176 (2003)
7. Dasu, T., Johnson, T.: Exploratory data mining and data cleaning, vol. 479. John Wiley & Sons (2003)
8. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
9. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine 17(3), 37 (1996)
10. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization (1998)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
12. Hall, M.A.: Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato (1999)
13. Hjorth, J.S.U.: Computer intensive statistical methods: Validation, model selection, and bootstrap. CRC Press (1993)
14. Klawikowski, S.J., Zeringue, C., Wootton, L.S., Ibbott, G.S., Beddar, S.: Preliminary evaluation of the dosimetric accuracy of the in vivo plastic scintillation detector oartrac system for prostate cancer treatments. Physics in Medicine and Biology 59(9), N27 (2014)

15. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 14, 1137–1145 (1995)
16. Langley, P.: Selection of relevant features in machine learning. Defense Technical Information Center (1994)
17. Larose, D.T.: Discovering knowledge in data: an introduction to data mining. John Wiley & Sons (2014)
18. Liu, H., Motoda, H.: Computational methods of feature selection. CRC Press (2007)
19. Quinlan, J.R.: C4. 5: Programming for machine learning. Morgan Kauffmann (1993)
20. Shin, K., Abraham, A., Han, S.-Y.: Improving kNN text categorization by removing outliers from training set. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 563–566. Springer, Heidelberg (2006)
21. Tallón-Ballesteros, A.J., Hervás-Martínez, C., Riquelme, J.C., Ruiz, R.: Improving the accuracy of a two-stage algorithm in evolutionary product unit neural networks for classification by means of feature selection. In: Ferrández, J.M., Álvarez Sánchez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2011, Part II. LNCS, vol. 6687, pp. 381–390. Springer, Heidelberg (2011)
22. Tallón-Ballesteros, A.J., Riquelme, J.C.: Deleting or keeping outliers for classifier training? In: 2014 Sixth World Congress on Nature and Biologically Inspired Computing, NaBIC 2014, Porto, Portugal, July 30 - August 1, pp. 281–286 (2014)
23. Tallón-Ballesteros, A.J., Riquelme, J.C.: Tackling ant colony optimization metaheuristic as search method in feature subset selection based on correlation or consistency measures. In: Corchado, E., Lozano, J.A., Quintián, H., Yin, H. (eds.) IDEAL 2014. LNCS, vol. 8669, pp. 386–393. Springer, Heidelberg (2014)
24. Witten, I.H., Frank, E., Mark, A.: Data mining: Practical machine learning tools and techniques (2011)
25. Zhang, S., Zhang, C., Yang, Q.: Data preparation for data mining. Applied Artificial Intelligence 17(5-6), 375–381 (2003)
26. Zhu, X., Wu, X.: Class noise vs. attribute noise: A quantitative study. Artificial Intelligence Review 22(3), 177–210 (2004)