A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance

Theodoros Iliou
University of the Aegean
Department of Cultural Technology and Communication
University Hill, Mytilene, Greece
Phone: +302251036624
th.iliou@ct.aegean.gr

Christos-Nikolaos Anagnostopoulos
University of the Aegean
Department of Cultural Technology and Communication
University Hill, Mytilene, Greece
Phone: +302251036624
canag@aegean.gr

Marina Nerantzaki
Democritus University of Thrace
Medical School, 8100,
Alexandroupolis, Greece
Phone: +302551030503
marinera@med.duth.gr

George Anastassopoulos
Democritus University of Thrace
Medical School, 8100, Alexandroupolis, Greece
Phone: +302551030503
anasta@med.duth.gr

ABSTRACT
Data preprocessing describes any type of processing methods performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing methods transforms the data into a format that will be more easily and effectively processed for the classification algorithms. In this paper, a novel data preprocessing method is proposed and evaluated in three difficult classification data sets of the well known UCI Repository, in which various classifiers have average performance lower than 75%. The three UCI repository datasets that have been used are the Mammographic masses, Indian Liver and Contraceptive Method. The performance of our proposed data preprocessing method and Principal Component Analysis preprocessing method was evaluated using the 10-fold cross validation method assessing five classification algorithms, Nearest-neighbour classifier (IB1), C4.5 algorithm implementation (J48), Random Forest, Multilayer Perceptron and Rotation Forest, respectively. The classification results are presented and compared analytically. The results indicate that the generated features after our proposed preprocessing method implementation to the original dataset markedly improve the performance of the classification algorithms.

Categories and Subject Descriptors
• *Computing methodologies~Artificial intelligence*

16th EANN workshops, September 25-28, 2015, Rhodes Island, Greece

DOI: http://dx.doi.org/10.1145/2797143.2797155

Keywords
Data preprocessing; machine learning; data mining; classification algorithms.

1. INTRODUCTION
Real-world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve data quality, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making [1].

Data pre-processing includes data cleaning, normalization, transformation, feature extraction and selection. The output of a data pre-processing method is a new feature set that in the end will boost the classification performance. This is caused by the fact that the dimensionality of the data and the classification time is reduced, which allows classification algorithms to operate more effectively. In some way, accuracy or precision of classification can be improved, or the result is just a more compact, easily interpreted representation of the target concept [2].

In this paper, we suggest a novel data preprocessing method that through simple transformations from the field of Mathematics and especially from the field of Linear Algebra, achieves dimensionality reduction and redundancy removal of the original data set, decreasing the number of initial variables and generating a new set of features which contains useful information of the initial dataset to enhance classification performance. Our proposed method was experimentally evaluated in three difficult classification problems from UCI repository, the Mammographic masses, Indian Liver and Contraceptive Method.

Both the initial dataset features and the new generated features produced by our proposed method and by Principal Component Analysis (PCA) were evaluated by five classifiers. It is shown that the generated features of our propose method markedly improve the classification performance. The new features also significantly improve the precision of the resulting classifiers.

This paper is structured as follows: Section 2 describes our proposed Data Preprocessing Method. Section 3 presents the experimental results of this study, while Section 4 concludes this paper and describes future work.

2. THE PROPOSED DATA PRE-PROCESSING METHOD

The proposed method can substantially improve successful classification when applying machine learning techniques to data mining problems. It transforms the input data into a new form of data, which is more suitable and effective for the learning scheme chosen. Below follows the detailed description of the method.

2.1 Step 1

Let's assume that a dataset of a machine learning problem named dataset1 is chosen, with n instances (rows), k variables (columns) and m classes.

The differences between adjacent elements of every instance of dataset1 are calculated (see Equation 1), and the new k-1 variables are added in dataset1, creating a new dataset named dataset2 with k+(k-1) variables.

Dataset2= [X(2)-X(1), X(3)-X(2),…,X(k)-X(k-1)] (1)

For illustration process and due to space limitation, Table 1 is created assuming that n=m=3. After the application of step1, the initial dataset is transformed to dataset2, which now has k+2=5 variables (see Table 2).

2.2 Step 2

Assuming that the set of attributes for every instance is a vector whose elements are the coefficients of a polynomial in descending power, step 2 estimates the derivative of the vector. The result is a new vector (one element shorter than initial one), with the coefficients of the derivative in descending power. Then, this new vector in added in dataset2 forming a new dataset named dataset3 (Table 3).

2.3 Step 3

In the third step of the proposed method, a new set (called from now-on Basic-Set) is created randomly selecting 10% of data from dataset3, consisting of d instances and m classes. The remaining 90% of dataset3 is called Rest-Set. Then, matrix right division (or slash division) of every Basic Set instance (row) with the remaining rows of the Basic Set is computed (Slash or matrix right division B/A is roughly the same as B*inv(A), more precisely, B/A = (A'\B')'). Then, follows the calculation of mean and median values of the division result for every instance of each class with the rest instances of its class ($Mean\_class_m\_row_x$ and $Median\_class_m\_row_x$ respectively, see equation 2 and 3), producing totally m+m=2m new variables ($Total\_Mean_1$, $Total\_Mean_2$,…, $Total\_Mean_m$ and $Total\_Median_1$, $Total\_Median_2$, …, $Total\_Median_m$ for every row of the Basic Set. Hence, we have d values for $Total\_Mean_1$, d values for $Total\_Mean_2$, …, d values for $Total\_Mean_m$ and d values for Total_Median1, d values for Total_Median2, …, d values for $Total\_Median_m$. The $Mean\_class_m\_row_x$ and Median_classm_rowx values are calculated as shown in equation (2) and equation (3) respectively (m is the name of the class, x (from 1 to d) is the row of the Basic Set and m1, m2 ….mk is the first, second…and last row of m class of the Basic set.

Table 1. Dataset 1

Variable1	Variable2	Variable3
$X1$	$X2$	$X3$
$Y1$	$Y2$	$Y3$
$Z1$	$Z2$	$Z3$

Table 2. Dataset 2 (shaded columns indicate the results of step 1).

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
$X1$	$X2$	$X3$	$X2-X1$	$X3-X2$
$Y1$	$Y2$	$Y3$	$Y2-Y1$	$Y3-Y2$
$Z1$	$Z2$	$Z3$	$Z2-Z1$	$Z3-Z2$

Table 3

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
$X1$	$X2$	$X3$	$X2-X1$	$X3-X2$
$Y1$	$Y2$	$Y3$	$Y2-Y1$	$Y3-Y2$
$Z1$	$Z2$	$Z3$	$Z2-Z1$	$Z3-Z2$

Variable 6	Variable 7	Variable 8	Variable 9
$f'(X1*x^4)$	$f'(X2*x^3)$	$f'(X3*x^2)$	$f'[X(2)-X(1)*x^1]$
$f'(Y1*x^4)$	$f'(Y2*x^3)$	$f'(Y3*x^2)$	$f'[Y(2)-Y(1)*x^1]$
$f'(Z1*x^4)$	$f'(Z2*x^3)$	$f'(Z3*x^2)$	$f'[Z(2)-Z(1)*x^1]$

$$Mean\_class_m\_row_x = \sum_{i=1}^{k} Mean\ (row_x / row_i) \quad (2)$$

$$Median\_class_m\_row_x = \sum_{i=1}^{k} Median\ (row_x / row_i) \quad (3)$$

Apart from the above, the Total_Mean and Total_Median values

are calculated as shown in equation (4) and equation (5) respectively (m is the name of the class and d is the sum of Basic Set rows). Finally, m total_Mean and m total_MEDIAN values result, one for every class of the Basic set.

Equation 1

$$\text{Total\_Mean}_{m} = \sum_{i=1}^{d} \frac{(\text{Mean\_class}_{m}\text{\_row}_{i})}{d} \quad (4)$$

Equation 2

$$\text{Total\_Median}_{m} = \sum_{i=1}^{d} \frac{(\text{Median\_class}_{m}\text{\_row}_{i})}{d} \quad (5)$$

2.4 Step 4

Assuming that Rest-Set from step 3 has r instances (rows) and m classes, a similar to step 3 approach follows. Specifically, matrix right division of every single Rest-Set row with every single row of the Basic Set is performed. Then, the mean and median values of the division result of every row for each class are calculated ($RS\_Mean\_class_m\_row_j$ and $RS\_Median\_class_m\_row_j$ respectively, see equation 6 and 7), producing new m+m=2m variables for every row of the Rest Set. As a result, we have r values for $RS\_Mean\_class_m\_row_j$, and r values for $RS\_Median\_class_m\_row_j$ (Table 5). Similarly to step 3, we compute mean and medial values ($RS\_Mean\_class_m\_row_x$ and $RS\_Median\_class_m\_row_x$ respectively) for every class as shown in equations (6) and (7). The $RS\_Mean\_class_m\_row_j$ and $RS\_Median\_class_m\_row_j$ values are calculated as shown in equation (6) and equation (7) respectively (m is the name of the class, j (from 1 to r) is the row of the Rest Set and m_1, m_2 ….m_k is the first, second…and last row of m class of the Basic Set.

$$RS\_Mean\_class_{m}\_row_{j} = \sum_{i=1}^{k} \text{Mean}(row_{j}/row_{m_i}) \quad (6)$$

$$RS\_Median\_class_{m}\_row_{j} = \sum_{i=1}^{k} \text{Median}(row_{j}/row_{m_i}) \quad (7)$$

Apart from the above, the $Final\_Mean_m\_row_j$ and $Final\_Median_m\_row_j$ values are also calculated as shown in equation (8) and equation (9) respectively (m is the name of the class and j (from 1 to r) is the row of the Rest set, (Table 4).

$$Final\_Mean_{m}\_row_{j} = \text{total\_Mean}_{m\ (step\ 3)} - RS\_Mean\_class_{m}\_row_{j} \quad (8)$$

$$Final\_Median_{m}\_row_{j} = \text{total\_Median}_{m\ (step\ 3)} - RS\_Median\_class_{m}\_row_{j} \quad (9)$$

Finally, m $Final\_Mean_m\_row_j$ and $Final\_Median_m\_row_j$ values result, one for every class m and every row j of the Rest set (Table 4).

Table 4. Step 4

	row1 (of Rest set)	**...**	**r (of Rest set)**
Class_1	**$RS\_Mean\_class_1\_row_1$**	**...**	**$RS\_Mean\_class_1\_row_r$**
	$RS\_Median\_class_1\_row_1$	…	$RS\_Median\_class_1\_row_r$
	$Final\_Mean_1\_row_1$	…	$Final\_Mean_1\_row_r$
	$Final\_Median_1\_row_1$	…	$Final\_Median_1\_row_r$
Class_2	**$RS\_Mean\_class_2\_row_1$**	…	**$RS\_Mean\_class_2\_row_r$**
	$RS\_Median\_class_2\_row_1$	…	$RS\_Median\_class_2\_row_r$
	$Final\_Mean_2\_row_1$	…	$Final\_Mean_2\_row_r$
	$Final\_Median_2\_row_1$	…	$Final\_Median_2\_row_r$
	……	…	……
Class_m	**$RS\_Mean\_class_m\_row_1$**	…	**$RS\_Mean\_class_m\_row_r$**
	$RS\_Median\_class_m\_row_1$	…	$RS\_Median\_class_m\_row_r$
	$Final\_Mean_m\_row_1$	…	$Final\_Mean_m\_row_r$
	$Final\_Median_m\_row_1$	…	$Final\_Median_m\_row_r$

2.5 Step 5

The rows (variables) $RS\_Mean\_class_m\_row_j$, $RS\_Median\_class_m\_row_j$, $Final\_Mean_m\_row_j$ and $Final\_Median_m\_row_j$ for every class are selected from Table 4 and then are placed in a new table named Table 5.

2.6 Step 6

The method ends with the transposition of Table 5 and the final dataset is now ready to be forwarded in any classification schema. Concluding the description of the proposed method, it is evident that the final dataset consist of 4 variables, namely $RS\_Mean\_class_m\_row_j$, $RS\_Median\_class_m\_row_j$, $Final\_Mean_m\_row_j$ and $Final\_Median_m\_row_j$ for every class of the initial dataset. Thus, if the original dataset has m classes, the final dataset will have 4*m variables.

3. EXPERIMENTAL RESULTS

For our experiments we used three datasets from UCI Repository [3], the Mammographic Masses [4], the Indian Liver [5] and the Contraceptive Method Choice [6] datasets. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [3]. For the classification results the repeated 10-fold cross validation method was used [7] so as to assess generalization of our machine learning data preprocessing method. The classifiers developed in WEKA 3.6 data mining

software [8] by their default WEKA parameters. To measure the performance of the classifier, Precision, Recall, Kappa statistics and Weighted Avg ROC area metrics have been used. Tables 6,7,8 present the Precision, Recall, Kappa statistics and Weighted Avg ROC area measurement of the initial data, Principal Component Analysis and proposed data preprocessing method for each dataset.

Table 5. Step 5

$RS\_Mean\_class_1\_row_1$	**...**	**$RS\_Mean\_class_1\_row_r$**
$RS\_Median\_class_1\_row_1$	…	$RS\_Median\_class_1\_row_r$
$Final\_Mean_1\_row_1$	…	$Final\_Mean_1\_row_r$
$Final\_Median_1\_row_1$	…	$Final\_Median_1\_row_r$
$RS\_Mean\_class_2\_row_1$	…	**$RS\_Mean\_class_2\_row_r$**
$RS\_Median\_class_2\_row_1$	…	$RS\_Median\_class_2\_row_r$
$Final\_Mean_2\_row_1$	…	$Final\_Mean_2\_row_r$
$Final\_Median_2\_row_1$	…	$Final\_Median_2\_row_r$
……..	…	……..
$RS\_Mean\_class_m\_row_1$	…	**$RS\_Mean\_class_m\_row_r$**
$RS\_Median\_class_m\_row_1$	…	$RS\_Median\_class_m\_row_r$
$Final\_Mean_m\_row_1$	…	$Final\_Mean_m\_row_r$
$Final\_Median_m\_row_1$	…	$Final\_Median_m\_row_r$

Kappa statistics (Cohhen's kappa) is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that the classifier is doing better than chance.

As for the ROC area measurement, an "optimal" classifier will have ROC area values approaching 1, with 0.5 being comparable to "random guessing" (similar to a Kappa statistic of 0).

In equations 10 and 11 the used formulas for these metrics are presented. The True Positive (TP) is the number of items correctly labeled as belonging to the positive class. In the case that items which were not labeled as belonging to the positive class by the classifier but should have been, they are called False Negatives (FN). Finally, the items incorrectly labeled as belonging to the class, they are called False Positives (FP). Thus, the number of true positives, false negatives, true negatives, and false positives add up to 100% of the set.

$$\text{Precision} = (\text{true positives}) / (\text{true positives} + \text{false positives}) \quad (10)$$

$$\text{Recall} = (\text{true positives}) / (\text{true positives} + \text{false negatives}) \quad (11)$$

In Table 6, we can observe that both PCA and our proposed data preprocessing method have better classification results for each classifier than initial Mammographic Masses data, while our proposed method marginally outperform PCA.

In Table 7, the obtained results for Indian Liver dataset show that PCA did not improve the classification results of the initial data, while our proposed data preprocessing method has significantly improved the classification performance to each classifier except MLP.

In Table 8, it is obvious that our proposed data preprocessing method has noticeably improved the classification results of the initial Contraceptive Method Choice data and achieved almost double percentage of correct classification with each classification algorithm than PCA and initial data.

4. CONCLUSIONS

Data pre-processing is an important step in the data mining process. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection. The product of data pre-processing is the final training set. In this paper, we used both PCA and a new data preprocessing method in three well known UCI Repository datasets, Mammographic Masses, Indian Liver and Contraceptive Method Choice dataset respectively and they were assessed by 5 classification algorithms. Data preprocessing steps in our proposed method is obtained using trial-and-error technique. The experimental results reveal that our new proposed preprocessing method significantly improved the overall performance in all initial datasets while PCA method improved only the classification results of Mammographic Masses initial dataset. In our point of view, our proposed method can be used for markedly boost classification algorithms performance in every dataset.

In future work, it would be preferable to make the same experiments in more datasets using different classifiers. In addition, our proposed data preprocessing method could be modified or extended in order to become a classification algorithm.

5. ACKNOWLEDGMENTS

Part of this study was financially supported by the research project (code number 2522) "Synergy for the sustainable development and safe use of the Greek tourist beaches – Beachtour", which is implemented within the framework of the Action "Cooperation 2011 - Partnerships of Production and Research Institutions in Focused Research and Technology" of the Operational Programme "Competitiveness and Entrepreneurship" (OPCE II), (Action's Beneficiary: General Secretariat for Research and Technology- MIA-RTDI), and is co-financed by the European Regional Development Fund (ERDF) and the Greek State.

6. REFERENCES

[1] I. H. Witten, E. Frank, M. Hall, A. Mark, Data Mining: *Practical Machine Learning Tools and Techniques (3 ed.)*, Elsevier, 2011, ISBN 978-0-12-374856-0.

[2] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "*Data Preprocessing for Supervised Leaning*", World Academy of Science, Engineering and Technology, vol. 1, 2007, pp. 856-861.

[3] https://archive.ics.uci.edu/ml/about.html, [Accessed 10 May 2015].

[4] https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass, [Accessed 10 May 2015].

[5] https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29, [Accessed 10 May 2015].

[6] https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice, [Accessed 10 May 2015].

[7] R. Kohavi, *"A study of cross-validation and bootstrap for accuracy estimation and model selection*". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, 1995, pp. 1137–1143.

[8] Waikato Environment for Knowledge Analysis, Data Mining Software in Java, available online: http://www.cs.waikato.ac.nz/ml/index.html, [Accessed 10 May 2015].

Table 6. Mammographic masses dataset classification Results.

	Original data				PCA				Proposed method			
	Pre	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**
IB1	0.743	0.743	0.48	0.74	0.999	0.999	0.46	0.73	1	1	1	1
J48	0.839	0.837	0.67	0.87	0.97	0.977	0.63	0.87	0.992	0.992	0.98	0.99
Ran For	0.793	0.793	0.58	0.86	0.99	0.993	0.59	0.86	0.997	0.997	0.99	1
MLP	0.81	0.808	0.61	0.88	0.81	0.81	0.62	0.88	1	1	0.99	1
Rot For	0.836	0.836	0.67	0.89	0.98	0.983	0.66	0.89	1	1	1	1

Table 7. Indian Liver dataset classification Results.

	Original data				PCA				Proposed method			
	Pre	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**
IB1	0.66	0.65	0.17	0.59	0.69	0.687	0.242	0.623	0.949	0.949	0.87	0.922
J48	0.66	0.68	0.16	0.69	0.606	0.713	0.001	0.505	0.866	0.865	0.638	0.824
Ran For	0.68	0.70	0.20	0.74	0.683	0.713	0.20	0.706	0.902	0.899	0.734	0.9
MLP	0.67	0.70	0.18	0.71	0.653	0.686	0.137	0.716	0.605	0.705	0.007	0.674
Rot For	0.65	0.71	0.005	0.70	0.62	0.71	0.006	0.71	0.97	0.97	0.923	0.991

Table 8. Contraceptive Method Choice dataset classification Results.

	Original data				PCA				Proposed method			
	Pre	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**	**Pre**	**Rec**	**k**	**ROC**
IB1	0.436	0.433	0.126	0.564	0.44	0.438	0.13	0.567	0.999	0.999	0.99	0.999
J48	0.529	0.532	0.27	0.682	0.501	0.508	0.22	0.642	0.956	0.955	0.93	0.969
Ran For	0.517	0.519	0.248	0.7	0.502	0.509	0.23	0.664	0.99	0.989	0.98	0.999
MLP	0.562	0.545	0.3	0.724	0.571	0.558	0.32	0.722	0.986	0.985	0.97	1
Rot For	0.544	0.546	0.29	0.719	0.523	0.524	0.25	0.701	1	1	1	1