Bulletin of the *Transilvania* University of Braşov • Vol 12(61), No. 2 - 2019
Series III: Mathematics, Informatics, Physics, 481-486
https://doi.org/10.31926/but.mif.2019.12.61.2.25

EFFECTIVENESS ANALYSIS OF ZeroR AND J48 CLASSIFIERS USING WEKA TOOLKIT

Livia SANGEORZAN[1]

Abstract

A current problem is comparing the techniques for evaluating large datasets and interpretation of these data for making decisions in a better way. This paper compares two widely used classification algorithms (ZeroR, J48) from the point of view of several metrics with open source Weka tool. The experimental comparison was made on two datasets of different sizes and from different domains: business and life.

2000 *Mathematics Subject Classification:* 62-07, 62C99.
Key words: ZeroR classifier, J48 classifier, accuracy, cross validation.

1 Introduction

Classification is a task often required in data mining. Supervised learning classification techniques use a training dataset containing istances (observations) and their labels. Classification process build a model that is able to identify to which category a new observation belongs to.

In literature there exist many classification algorithms and in a data mining process one can use them, depending on their performance and classification accuracy. In [13] the authors present the classification from two perspectives: Supervised Learning and Unsupervised Learning. In supervised learning, the aim is to identify a class that a new observation belongs to, based on a training set of examples. In unsupervised learning algorithms we have only instances (observation) and the algorithms find themselves criteria to group the data and to build clusters (similar to classes in supervised classification).

In Supervised Learning an important issue is strong dependency of data. The No Free Lunch theorem in Machine Learning proves that there is not a specific classification method having a high performance for all problems, and all sets of data. In this paper we study the performance of two classification algorithms,

[1] Faculty of Mathematics and Informatics, *Transilvania* University of Braşov, Romania, e-mail: sangeorzan@unitbv.ro

ZeroR and J48. The J48 classifier that we consider in our paper is an open source Java implementation of the C4.5 algorithm. The C4.5 algorithm builds decision trees from a set of training data.

ZeroR is another classifier used in data mining. ZeroR classifier takes into account the target attribute and its possible values and does not include any rule. Data mining techniques are also used in information systems, [1]-[5], and mathematical modeling, [6]-[7]. We remark that decision making systems also use different supervized learning techniques for classification task. Applications of decision trees can be found in [8]-[10].

2 Data mining classifiers: ZeroR and J48

In [11], [12], [15], [16] the authors present the most used data mining algorithms, which are implemented in Weka tool, a software written in Java. Some features of this tool are: preprocessing, classification, clustering, association rules, attribute selection, visualization [17].

The systems that construct classifiers take as input a collection of instances, each belonging to one of a number of classes. Each instance is a vector of attributes values. The output is a classifier used to predict the class to which belongs an new instance.

The two classification algorithms, J48 and ZeroR are compared based on many metrics. We compute these metrics using two evaluation techniques: cross validation and percentage split. The experimental comparison was made on two datasets, "Absenteeism at work" and "Somerville Happiness Survey" [18].

In k-folds cross validation method, the data set is divided into k subsets of data of about the same size. From these subsets, a subset of data will be used as a test data set and the remaining k-1 subsets will be used as training data. The method allows all subsets to be used for both validation and training.

The percentage split method divides the database into two disjoint subsets, one for training and one for testing.

3 Case study using ZeroR and J48 classifiers

The datasets used for our case study are:

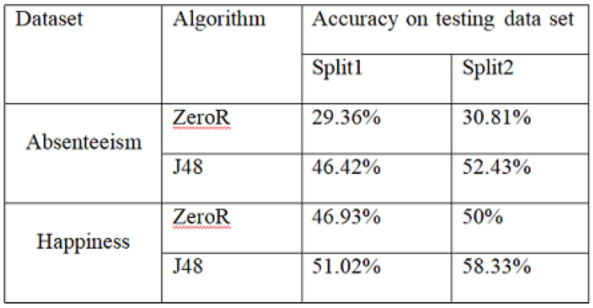
1. *Absenteeism at work*, having 21 attributes and 740 records;
2. *Somerville Happiness Survey*, having 7 attributes and 143 records.

3.1 Percentage Split Method for model evaluation

The datasets are randomly split in two disjoint parts, one for training and one for testing. We use two splits :

- Split1: 66% training and 34 % for testing;
- Split2: 75% training and 25% for testing.

In Figure 1 we present the accuracy on the testing data sets for the two splits.

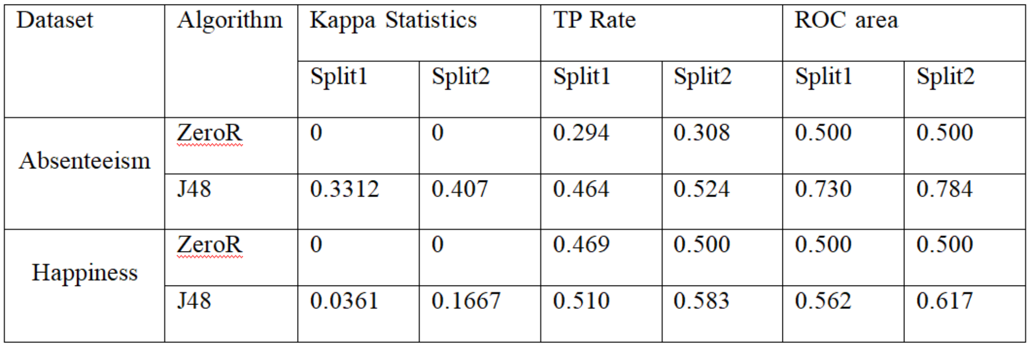


Dataset	Algorithm	Accuracy on testing data set	
		Split1	Split2
Absenteeism	ZeroR	29.36%	30.81%
	J48	46.42%	52.43%
Happiness	ZeroR	46.93%	50%
	J48	51.02%	58.33%

Figure 1: Comparison of Accuracy on testing data set, using Percentage Split Method for Absenteeism and Happiness

Figure 2 shows the different metrics of precision computed for the two classification techniques taken into account. We used as metrics Kappa statistics, True Positive Rate (TP Rate), Receiver Operating Characteristics (ROC) Area.

For the definition on this metrics you can see [14] and [15].



Dataset	Algorithm	Kappa Statistics		TP Rate		ROC area	
		Split1	Split2	Split1	Split2	Split1	Split2
Absenteeism	ZeroR	0	0	0.294	0.308	0.500	0.500
	J48	0.3312	0.407	0.464	0.524	0.730	0.784
Happiness	ZeroR	0	0	0.469	0.500	0.500	0.500
	J48	0.0361	0.1667	0.510	0.583	0.562	0.617

Figure 2: Accuracy Parameters for Absenteeism and Happiness evaluation

Conclusion is that the J48 algorithm has the best performance for all the three precision measures.

In Figure 3 we show the evaluation algorithms using other important precision metrics:

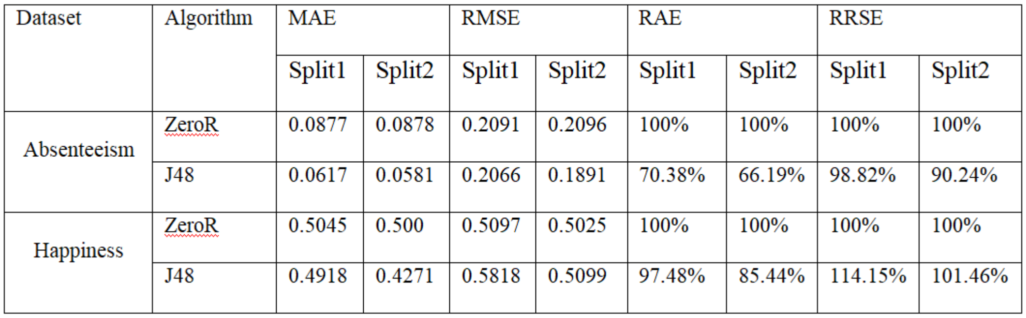
MAE - Mean Absolute Error;

RMSE - Root Mean –Squared Error;

RAE - Relative Absolute Error;

RRSE - Root Relative Squared Error.

For more details about metrics these see [14] and [15].



Dataset	Algorithm	MAE		RMSE		RAE		RRSE	
		Split1	Split2	Split1	Split2	Split1	Split2	Split1	Split2
Absenteeism	ZeroR	0.0877	0.0878	0.2091	0.2096	100%	100%	100%	100%
	J48	0.0617	0.0581	0.2066	0.1891	70.38%	66.19%	98.82%	90.24%
Happiness	ZeroR	0.5045	0.500	0.5097	0.5025	100%	100%	100%	100%
	J48	0.4918	0.4271	0.5818	0.5099	97.48%	85.44%	114.15%	101.46%

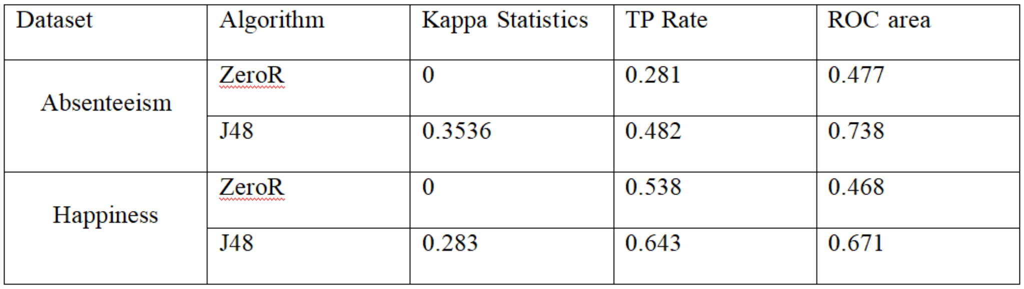
Figure 3: Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, Root Relative Squared Error for Absenteeism and Happiness evaluation

Figure 3 shows that J48 algorithm has the highest performance compared with ZeroR algorithm. If we have a smaller number of values like in Happiness dataset, the J48 algorithm has a very high error rate with poor performance [15].

3.2 Cross validation evaluation method

The statistical validation technique called "cross validation" decides on a fix number of folds, or partitions of data. If we use 10 folds, the data is divided (split) randomly into 10 approximately equal parts, where nine-tenth is for training and one-tenth for testing. One repeats the procedure 10 times, so that every instance will be used exactly once for testing. The mean accuracy reported on cross validation is then used for model evaluation.

Figure 4 shows the accuracy measure of the two classification techniques: Kappa statistics, True Positive Rate (TP Rate), Receiver Operating Characteristics (ROC) Area.

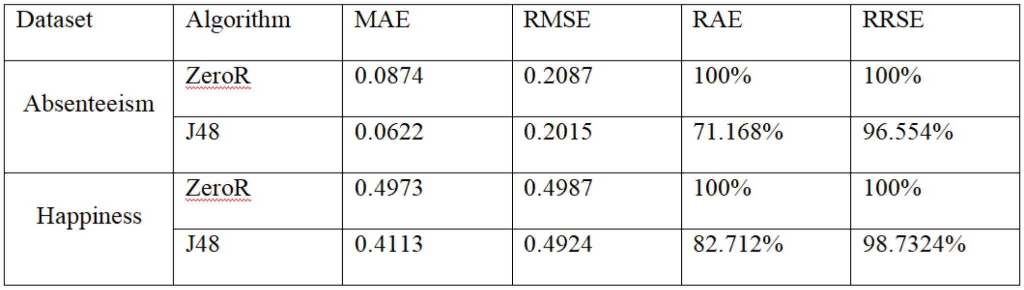


Dataset	Algorithm	Kappa Statistics	TP Rate	ROC area
Absenteeism	ZeroR	0	0.281	0.477
	J48	0.3536	0.482	0.738
Happiness	ZeroR	0	0.538	0.468
	J48	0.283	0.643	0.671

Figure 4: Accuracy Parameters for Absenteeism and Happiness evaluation

From Figure 4 results that J48 algorithm is good in terms of accuracy.

To evaluate the success of numeric prediction there are several alternative measures, like the following: Mean Absolute Error (MEA), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE). In Figure 5 we show the numeric prediction for the two evaluated datasets.



Dataset	Algorithm	MAE	RMSE	RAE	RRSE
Absenteeism	ZeroR	0.0874	0.2087	100%	100%
	J48	0.0622	0.2015	71.168%	96.554%
Happiness	ZeroR	0.4973	0.4987	100%	100%
	J48	0.4113	0.4924	82.712%	98.7324%

Figure 5: Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, Root Relative Squared Error for Absenteeism and Happiness evaluation

For both datasets, J48 algorithm has minimum error rate and good performance.

3.3 Conclusion

ZeroR and J48 have been evaluated on two different datasets, using Weka tool. The evaluation has been made using different metrics. For evaluation of accuracy we used two different split and also 10 fold cross validation.

Regardless of the number of data, J48 had the best performance. The two datasets used have different sizes and belong to different fields of data. All the experiments prove that J48 out performs ZeroR.

References

[1] Carstea, C., *Control and management in complex information systems*, Bulletin of the Transilvania University of Braşov, Series III Mathematics, Informatics, Physics, (2013), 73-87.

[2] Carstea, C., *Modeling System's Process for Control Of Complex Information Systems*, Proc. of IBIMA 2015, the 25th International Business Information Management Association Conference, Soliman KS (ed). Amsterdam, Netherlands, May 2015, 566-574, 2015.

[3] Carstea, C., *Optimization Techniques in Project Controlling*, Ovidius University Annals, Economic Sciences Series, Volume XIII Issue 1, pp. 428-432, Ovidius University Press, 2013.

[4] David, N., *Informational Systems and Informatic Applications in Businesses*, Lambert Academic , Saarbrücken, Germany, 2011.

[5] Enache-David, N., Sangeorzan, L., *An Application on Web Path Personalization*, Proc. of IBIMA 2016, the 27th International Business Information

Management Association Conference - Innovation Management and Education Excellence Vision 2020:From Regional Development Sustainability to Global Economic Growth, May 4-5, 2016, Milan, Italy, 2016, 2843-2848, 2016.

[6] Florea, O., Rosca, I., *The Mechanical Behavior and the Mathematical Modeling of an Intervertebral Disc*, Acta Technica Napocensis Series-Applied Mathematics Mechanics and Engineering, **58** (2015), 213-218.

[7] Florea, O., Rosca, I.C., *Stokes' Second Problem for a Micropolar Fluid with Slip*, PLOS ONE, **10**, Issue 7, 2015.

[8] Mandru, L., *How to Control Risks? Towards A Structure of Enterprise Risk Management Process*, Journal of Public Administration, Finance and Law (2016), 80-92.

[9] Popescu, M., Mandru, L., *Relationship between Quality Planning and Innovation*, Bulletin of the Transilvania University of Braşov, Series V, Economic Sciences, Vol. 9 (58) (2016), no. 2, 203-212.

[10] Mandru, L., *Managementul integrat calitate-risc pentru societăţile comerciale cu profil industrial*,Transilvania University Press, Brasov, 2011.

[11] Manning,C.D., Raghavan, P., Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[12] Robbins, H., Ryzin, J., *Introduction to statistics*, Science research associates Inc., 1975.

[13] Michie, D., Spiegelhalter, D.J., Taylor, C.C., *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Upper Saddle River, NJ, USA, 1994.

[14] Simian, D., Stoica, F., *Automatic optimized support vector regression for financial data prediction*, Springer Verlag London Ltd., 2019.

[15] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., *DATA MINING. Practical Machine Learning Tools and Techniques*, Elsevier, Morgan Kaufmann, 2017.

[16] Wu, X. et al, *Top 10 algorithms in data mining*, Knowledge and Information Systems **14** Springer-Verlag London Limited 2007, 2008.

[17] http://www.cs.waikato.ac.nz/ml/weka/.

[18] https://archive.ics.uci.edu/ml/index.php.