Research Paper
Abbas et al., V-8; I [2]: 104-110 [2020]

Corresponding Emails:
dr.mrabbas@gmail.com

DARSGAH-E-AHLEBAIT

Journal of natural sciences [ISSN: 2308-5061]

Effect of Data Normalization on the Performance of Classification Algorithms

Muhammad Rehan Abbas[1*], Malik Sajjad Ahmed Nadeem[1] and Syed Rizwan Abbas[2]

[1]Department of CS & IT, University of Azad Jammu and Kashmir, Muzaffarabad, AJK, Pakistan
[2]Department of Biological Sciences; Hunza Campus; Karakoram International University, Gilgit, Pakistan

Abstract
Data normalization is an elementary data preprocessing step for learning from data before feeding to some machine learning classifiers. We conducted an empirical study to improve the performance of machine learning classifiers by inducing data normalization during learning phase of the classification algorithms. Three data normalization techniques; Decimal Scaling (DS), Min-Max (MM), and Z-Score (ZS) were selected along with five machine learning classifiers (Support Vector Machine with linear kernel (SVML), Support Vector Machine with radial kernel (SVMR), Linear Discriminant Analysis (LDA), Random Forest (RF) and K-Nearest Neighbors (kNN)). The investigation has been carried out on five publicly available clinical cancer datasets. To evaluate the performances of classification algorithms, prediction accuracy, Mean Squared Error (MSE) and Improved Squared Error (ISE) are three factors that were taken into account. Performance comparison of different learning algorithms was made after applying to each normalization technique.

Keywords: Transformation, classification, data normalization, SVM, LDA, kNN, Random Forest

1. Introduction

Real-world data can be considered extremely complex for interpretation without pre-processing. Careful integration of data is now acceptable but needs to be transformed into forms suitable for mining useful information using ML algorithms. One of the most important steps in data preprocessing is the data transformation (CHRISTIE & ALFSEN, 1977).

Data transformation is a way to map the whole set of values of a given attribute to a new set of replacement values such that each old value can be recognized as one of the new values. Data transformation involves smoothing, a generalization of the data, attribute construction and normalization. Literature says, the transformation of data into appropriate form may increase the performance of classifiers built using ML algorithms (Al Shalabi et al., 2006; Dinç et al., 2014; Eftekhary et al., 2012; Jayalakshmi & Santhakumaran, 2011; Mustaffa & Yusof, 2011; Nayak et al., 2014; Ogasawara et al., 2010; Sechidis, 2011). Such classifiers perform better if the data which is being analyzed is normalized within a specific range such as 0.0 to 1.0 (Han, 2001). More precise modification of source data into various forms that: (i) facilitate easy use of ML algorithms (ii) improve the efficiency and effectiveness of ML algorithms (iii) that represent data in a simple and comprehensible form for people and machines (iv) provides data suitable for a particular analysis. The attribute information is adjusted to a certain extent.

There are various ways to transform data; one of the most effective methods used in this study is data normalization. Experiments say that data normalization can improve the accuracy and efficiency of ML algorithms. Normalization is especially useful for classification algorithms that include SVM, kNN, Artificial Neural Networks (ANN), and clustering classifiers. Such methods provide good results if the normalized data is used in the analyses, i.e. data is scaled to certain ranges, such as [0.0, 1.0] (Han, 2001). For distance-based methods, normalization assists in preventing attributes with originally large ranges from outweighing the attributes with initial smaller ranges (Han, 2001).

1.1 Data Normalization Techniques

There are many methods for normalizing data. In this article, we have selected three well known normalization techniques, i.e. Min-Max (MM) normalization, normalization by Z-Score (ZS) and normalization by the Decimal Scaling (DS).

1.1.1 Decimal Scaling Normalization

Normalizing data by using DS can be accomplished by a little computation. The decimal point is moved to an appropriate length which is actually the maximum absolute value of the attribute. Following equation (1) is the formula to transform a data point using DS:

$$v' = \frac{v}{10^j} \tag{1}$$

Where j is the lowest number such that Max $(|v'|) < 1$.

1.1.2 Min-Max Normalization

Min-Max (MM) normalization can be performed in a different way. A custom range can be specified as a minimum value and maximum value, such that all the transformed data points may lie in this range. First, minimum and maximum values of the attribute should be specified. Then new minimum and maximum are defined according to the needs. For an attribute A, let suppose min_A is the minimum value of attribute A and max_A is the maximum value of attribute A new_min_A is the new minimum value and new_max_A is the new maximum value of attribute A. Following equation (2) is the formula to compute MM normalization:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{2}$$

1.1.3 Z-Score Normalization

Z-Score (ZS) normalization is mostly used when the minimum and maximum values of an attribute are unknown. It is computed by subtracting the mean of attribute A by each value of attribute A, then divided by the standard deviation of A., As a result, the new transformed value of v' is obtained. ZS normalization is computed in the following way:

$$v' = \frac{v - \mu_A}{\sigma_A} \quad (3)$$

Where μ_A is used to represents the mean and σ_A for standard deviation.

Al Shalabi *et al.*,(2006) emphasized different types of normalization. Each was tested according to the ID3 method using the HSV data set. Comparisons were made between different learning methods, as they were applied to any normalization method. Empirical results show that MM normalization supersedes by producing highest accuracy than ZS and DS normalization methods. Sechidis (2011) reported how different preprocessing techniques and different feature selection algorithms affect the performance of different classifiers in cancer datasets. In the preprocessing step, two different normalizations (ZS and MM) techniques and one discretization (equal-width) were used. The results were compared using five different classifiers (kNN, LR, RF, BT, SVM). Jayalakshmi & Santhakumaran (2011) conducted a study to illustrate several methods used in backpropagation neural networks to increase the reliability of the learned network. The results showed that the performance of diabetes data model was dependent on normalization methods using neural networks. Mustaffa & Yusof (2011) used three common methods to predict the onset of dengue glow: MM, ZS, and decimal scaling. The techniques were used with the LS-SVM prediction model and the neural network (NNM), respectively. Results were compared based on prediction accuracy and mean squared error (MSE). The results show that LS-SVM is a better predictor then NNM. Ogasawara *et al.*,(2010) introduced a new way to normalize heterogeneous time series (with irregular fluctuations). This method, called Adaptive Normalization (AN), was tested on an artificial neural network (ANN) on three predictive problems. The results were compared with four conventional normalization methods and showed that AN, in both short and long-term predictions, improved the accuracy of ANN.

Eftekhary *et al.*,(2012) examined the impact of five data normalization methods and then computed the accuracies of the classification models before and after application of data normalization. For classification, the SVM algorithm was used because it was based on distance. Nayak *et al.*,(2014) picked two ANN models and two neuro-genetic hybrid models to predict Indian stock market closing prices. Different normalization techniques were applied on ANN model learned with gradient decent (ANN-GD), genetic algorithm (ANN-GA), and a functional link artificial neural network model learned with GD (FLANN-GD) and genetic algorithm (FLANN-GA). These model were used for assessing the daily closing price of Bombay Stock Exchange (BSE). Results revealed that performance of the learned models was strongly affected by using data preprocessing techniques.

Dinç *et al.*,(2014) evaluated the performance of the classification of protein crystallization images captured during the protein crystal growth process. The study was aimed to investigate the classifiers for best preprocessing methods, non crystal and likely leads datasets. To address this issue, the five classifiers were used along with specific methods for preprocessing data, such as Principal Component Analysis (PCA), MM and ZS normalization, to evaluate their impact on classifiers' performances for the non crystal and likely leads datasets. This experiment was conducted on 1606 non crystal and 245 likely leads images separately. They have reached 96.8% accuracy for non crystal datasets and 94.8% for likely leads datasets.

2. Meterials and Methods

2.1 Datasets

Five publicly available clinical cancer datasets were selected for this study. Four of them are from UCI machine learning repository (Bache & Lichman, 2013), and other is Lung Cancer from North Central Cancer Treatment Group (NCCTG) (Loprinzi et al., 1994).. All datasets have repeatedly appeared in the machine learning literature.

Table 1: Baseline characteristics of the datasets used in the study.

Dataset	#Instances	#Attributes	+ive class count	-ive class count
WPBC	194	33	46	148
WDBC	569	30	212	357
WBCD	683	9	239	444
Breast-cancer	286	9	85	201
NCCTG lung cancer	167	9	47	120

2.2 Model Learning

We have picked five machine learning classifiers i.e. Support Vector Machine with linear kernel (SVML), Support Vector Machine with radial kernel (SVMR), Linear Discriminant Analysis (LDA), Random Forest (RF) and K-Nearest Neighbors (kNN).

2.3 Performance Metrics

Following three popular performance metrics that are usually considered for evaluation of the techniques applied on imbalance class problems were used to evaluate the performances of classification algorithms.

2.3.1 Prediction Accuracy

Prediction accuracy speaks to the number of accurately predicted observations. On account of the imbalanced class issue, an excessively basic model predicting entire test samples labeling as negative class members may be able to produce higher accuracy. Hence, just accuracy is not sufficient performance metric for imbalance class problems. The accuracy of a classifier is computed by Equation 4.

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{4}$$

2.3.2 Mean Squared Error

The Mean Squared Error (MSE) represents the estimated probability error for the true y_i class label. A model that provides accurate probability estimates can minimize MSE. Although true probability for $p^T(y_i|x_i)$ is used as the answer to the original definition, $p^T(y_i|x_i)$ is generally not known in the real world data set. By contrast, in general, MSE is calculated by assuming $p^T(y_i|x_i) = 1$. Equation 5 is the formula to calculate MSE.

$$MSE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(1 - p(y_i|x_i)\right)^2 \tag{5}$$

2.3.3 Improved Squared Error

The MSE is commonly used for probability estimation. Regardless of whether the model predicts all class labels correctly, in any case, the MSE will still turn out to be high if the probability estimation is unstable. On the other hand, Fan et al. (2005) introduced a modified version of MSE known as Improved Squared Error (ISE) showing probability error in case of misclassification. Hence, the ISE consolidates the accuracy and the MSE into one value. It can be noted that for binary classification problems, the prediction threshold is fixed at 0.5.The modified form of Equation 5 is presented in Equation 6 to compute ISE.

$$ISE = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \left(1 - min\left(1.0, \frac{p(y_i|x_i)}{0.5}\right)\right)^2 \quad (6)$$

2.4 Experimental Setup

The experiments were carried out according to the following procedure:
Step 1: First we prepared our datasets by selecting and transforming relevant features, data cleaning and data integration.
Step 2: Then we trained and tested our models (SVML, SVMR, LDA, RF and KNN) on randomly sampled partitions (i.e., 10-foldcross-validation) and results were averaged over ten trials.
Step 3: Next we normalized the datasets by using three data normalization methods (i.e., DS, MM and ZS).
Step 4: After normalizing data, we trained and tested our five classification models on randomly sampled partitions (i.e., 10-fold cross-validation) and results were averaged over ten trials.
Step 5: Finally we have computed prediction accuracy, MSC and ISE for each model, normalization method and dataset.

3. Results and Discussions

This section presents the results based on application of different data normalization techniques on clinical cancer datasets. Selected classifiers were SVML, SVMR, LDA, RF and kNN. Three data normalization techniques, DS, MM and ZS were used on five clinical cancer datasets; WPBC, WDBC, WBCD, Breast-cancer and NCCTG Lung Cancer.

Table 1 summarizes the results based on the application of data normalization techniques on clinical cancer datasets. Accuracy, MSE and ISE were chosen as performance measure for classification algorithms.

For WPBC dataset in Table 1, significant improvement in the performance of kNN was found over all normalization methods for all measures except ISE in case of DS. Without normalization, accuracy of kNN was 71.5%, but after application of normalization methods, it enhanced up to 75.9%, 76.3% and 77.1% for DS, MM and ZS respectively. MM normalization superseded to other normalization methods by showing minimum scores of MSE and ISE. MM normalization also lead in case of SVML for accuracy measure. In case of WDBC dataset, SVML showed little improvement over all performance measures. LDA enhanced MSE for MM and ISE for all normalization techniques. Considerable improvement in case of kNN was found over all performance measures for MM and ZS. DS did not prove a good normalization method for kNN on WDBC dataset. No significant improvements in the performances of classifiers were found for WBCD dataset. Again DS did not prove a good technique for kNN by declining its performance. Little improvement in the performance of LDA was found over accuracy measure. Significant improvements could be seen for kNN over all normalization methods especially MM superseded for accuracy measure. Finally, for NCCTG lung cancer data set, without normalization kNN showed 68.1% accuracy. After application of normalization methods, kNN reached up to 70%, 72.2% and 71.9% for DS, MM and ZS normalization techniques respectively.

Table 2: Performance comparison of classifiers with and without data normalization.

Classifier	Normalization	WPBC			WDBC			WBCD			BC			NCCTG LC		
		ACC.	MSE	ISE	ACC.	MSE	ISE	ACC.	MSE	ISE	ACC.	MSE	ISE	ACC.	MSE	ISE
SVML	No Norm.	79.2	0.1458	0.0453	97.1	0.0236	0.0092	96.5	0.0256	0.0128	69.8	0.1992	0.0501	72	0.1963	0.0534
	DS	79.3	0.1456	0.0445	97.4	0.0229	0.0091	96.6	0.0251	0.0127	69.2	0.201	0.0503	72.1	0.1963	0.0533
	MM	79.7	0.1474	0.0457	97.2	0.0231	0.0088	96.5	0.0253	0.0127	69.1	0.2009	0.0506	71.8	0.1973	0.0553
	ZS	79	0.1465	0.0456	97.3	0.0231	0.0086	96.5	0.025	0.0124	69	0.201	0.0512	71.4	0.1991	0.056
SVMR	No Norm.	79.2	0.1589	0.0583	97.3	0.0208	0.0083	96.9	0.0251	0.0158	75.1	0.1822	0.0573	74.5	0.1833	0.0557
	DS	78.5	0.1595	0.0575	97.4	0.0207	0.0083	96.9	0.0249	0.0154	74.4	0.1833	0.0578	74.2	0.1829	0.0554
	MM	78.5	0.1603	0.0592	97.3	0.0211	0.0085	96.9	0.0251	0.0157	74.7	0.1823	0.0573	73.5	0.1857	0.0565
	ZS	78.9	0.1595	0.0593	97.4	0.0207	0.0084	97	0.0246	0.0154	74.8	0.1842	0.0588	74.3	0.185	0.0562
LDA	No Norm.	79	0.1494	0.0762	95.8	0.0335	0.0215	96	0.0345	0.029	72.5	0.1889	0.0663	74.4	0.1908	0.0684
	DS	79.3	0.1506	0.0777	95.6	0.0338	0.021	96	0.0348	0.0296	72.9	0.1892	0.0667	73.6	0.1922	0.0677
	MM	79	0.1528	0.0789	95.7	0.0332	0.021	96.1	0.0343	0.0291	72.9	0.19	0.0671	74.3	0.1916	0.0686
	ZS	78.9	0.152	0.0794	95.6	0.0336	0.0212	96.1	0.0347	0.0294	73.3	0.1891	0.067	73.9	0.1928	0.0687
RF	No Norm.	79.2	0.1686	0.0549	96.2	0.0305	0.008	97.3	0.0247	0.01	73.7	0.1941	0.0781	74.4	0.1746	0.0441
	DS	79.6	0.167	0.0531	96.2	0.0307	0.0084	97.2	0.025	0.01	73.1	0.1956	0.0792	74.4	0.1738	0.0423
	MM	79.8	0.1686	0.0544	96.2	0.0308	0.0086	97.3	0.0249	0.0102	73.5	0.1954	0.0792	73	0.1753	0.0434
	ZS	79	0.1678	0.0536	96.2	0.0306	0.0083	97.3	0.025	0.01	72.7	0.1973	0.0809	74.4	0.1739	0.0423
kNN	No Norm.	71.5	0.1964	0.0752	93.2	0.055	0.0298	97.3	0.0224	0.0104	69.7	0.2189	0.0924	68.1	0.197	0.0635
	DS	75.9	0.1912	0.1236	62.7	0.3726	0.3726	74.4	0.1831	0.1287	71.4	0.2042	0.0871	70	0.2329	0.0837
	MM	76.3	0.1867	0.0804	96.7	0.027	0.0118	97.2	0.0223	0.0102	73.5	0.2092	0.0991	72.2	0.2153	0.1006
	ZS	77.1	0.1907	0.0898	96.9	0.0272	0.0119	96.7	0.0258	0.0131	72.2	0.2154	0.1003	71.9	0.2115	0.0929

Literature says, classifiers which depend on distance calculations are affected from normalization but classifiers based on decision tree methodology; do not care about the range of the values. It was true but mostly for kNN as distance based classifiers. Other distance based classifiers i.e. SVM and LDA did not affect from normalization. Empirical results show that, overall kNN improved its performance for all datasets except WBCD. DS did not proved a good normalization technique as compared to others. Results reveal that, all data normalization techniques are not useful for all datasets and classifiers. It depends on nature of the data and classification algorithm.

5. Conclusion

An empirical study was conducted to see the effect of data normalization on the performances of different types of classification algorithms. The investigation has been carried out on six publicly available clinical cancer datasets. The overall impact of data preprocessing specifically data normalization has been evaluated systematically. Empirical results show that data normalization affects prediction accuracy to some extent; a single normalization method is not superior over all others. The literature says, classifiers which depend on distance calculations are affected by normalization but classifiers based on decision tree methodology; do not care about the range of the values. It was true but mostly

for kNN, other distance-based classifiers i.e. SVM and LDA did not affect from normalization for the datasets selected for this study. Empirical results show that, overall kNN improved its performance for all datasets except WBCD. DS did not prove a good normalization technique as compared to MM and ZS. Results reveal that, all data normalization techniques are not useful for all datasets and classifiers. It depends on the nature of the data and classification algorithm.

References

Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. Journal of Computer Science, 2(9), 735-739.

Bache, K., & Lichman, M. (2013). UCI machine learning repository.

CHRISTIE, O. H., & ALFSEN, K. H. (1977). Data transformation as a means to obtain reliable consensus values for reference materials. Geostandards and Geoanalytical Research, 1(1), 47-49.

Dinç, İ., Sigdel, M., Dinç, S., Sigdel, M. S., Pusey, M. L., & Aygün, R. S. (2014). Evaluation of normalization and pca on the performance of classifiers for protein crystallization images. Paper presented at the SOUTHEASTCON 2014, IEEE.

Eftekhary, M., Gholami, P., Safari, S., & Shojaee, M. (2012). Ranking normalization methods for improving the accuracy of SVM algorithm by DEA method. Modern Applied Science, 6(10), 26.

Fan, W., Greengrass, E., McCloskey, J., Yu, P. S., & Drammey, K. (2005). Effective estimation of posterior probabilities: Explaining the accuracy of randomized decision tree approaches. Paper presented at the Data Mining, Fifth IEEE International Conference on.

Han, J. (2001). Micheline Kamber amd Simon Fraser University "Data Mining Concepts and Techniques" Morgan Kaufmann Publishers: USA.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical Normalization and Back Propagationfor Classification. International Journal of Computer Theory and Engineering, 3(1), 89.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., . . . Klatt, N. E. (1994). NCCTG Lung cancer.

Mustaffa, Z., & Yusof, Y. (2011). A comparision of normalization techiques in predicting dengue outbreak. Paper presented at the International Conference on Business and Economics Research.

Nayak, S., Misra, B., & Behera, H. (2014). Impact of data normalization on stock index forecasting. Int. J. Comp. Inf. Syst. Ind. Manag. Appl, 6, 357-369.

Ogasawara, E., Martinez, L. C., De Oliveira, D., Zimbrão, G., Pappa, G. L., & Mattoso, M. (2010). Adaptive normalization: A novel data normalization approach for non-stationary time series. Paper presented at the Neural Networks (IJCNN), The 2010 International Joint Conference on.

Sechidis, K. (2011). Comparison of different preprocessing techniques and feature selection algorithms in cancer datasets.