

The Effect of Normalization in Violence Video Classification Performance

Ashikin Ali & Norhalina Senan

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Batu Pahat, Johor.

Corresponding author: gi150038@siswa.uthm.edu.my, halina@uthm.edu.my

Abstract. Basically, data pre-processing is an important part of data mining. Normalization is a pre-processing stage for any type of problem statement, especially in video classification. Challenging problems that arises in video classification is because of the heterogeneous content, large variations in video quality and complex semantic meanings of the concepts involved. Therefore, to regularize this problem, it is thoughtful to ensure normalization or basically involvement of thorough pre-processing stage aids the robustness of classification performance. This process is to scale all the numeric variables into certain range to make it more meaningful for further phases in available data mining techniques. Thus, this paper attempts to examine the effect of 2 normalization techniques namely Min-max normalization and Z-score in violence video classifications towards the performance of classification rate using Multi-layer perceptron (MLP) classifier. Using Min-Max Normalization range of [0,1] the result shows almost 98% of accuracy, meanwhile Min-Max Normalization range of [-1,1] accuracy is 59% and for Z-score the accuracy is 50%.

1. Introduction

Data collected from numerous resources and stored in data warehouse. These resources might include multiple databases, flat files, missing values and unranked attributes. Therefore, different issues is possible to rise during combination of data that we desired to have for mining, classification and so on. It is crucial to handle the process of data combination to avoid data redundancy, missing value issues, inconsistency that in return improves the accuracy and reliability to speed up the mining process. The crucial data combination needs to be transformed to suitable forms for mining. Data transformation inclusive of smoothing, generalization of the data, attribute construction and normalization [2]. Normalization is a scaling technique or a pre-processing stage [1]. Whereby, it can find new range from an existing range. Having that, it will be cooperative for the prediction, forecasting or classification purpose[4]. Essentially, aware that there are variety of existing way for classification, forecasting or prediction. In order to maintain the variations, Normalization technique is usually applied to conjunct these processes [20]. In violence video classification itself involve limitations that needs attention [17] [18] [19].



Normalization in data transformation will improve the accuracy and reliability of mining algorithms involving neural networks classifiers. Neural network provides better results if the data to be analyzed have been normalized, that is scaled to specific commonly used range such as (0 to 1) or (-1 to 1). Normalization which applied to input values of each attributes help improves the learning phase in back-propagation algorithm. Without normalization the process can be inaccurate, inefficient and might not produce expected outcome [3]. In Artificial Neural Network (ANN) and other data mining approaches normalization is essential to avoid the ill-condition of the network. This can guarantee stable convergence. In literature [13] [14] [15] [16], there are many methods of normalization, but this paper envisioned to examine the effect of normalization towards the VSD2014 datasets, two videos have been used, namely *Agony vs Violent 2-0* and *Short Peace: Ranko Tsukigime's Longest Day - PS3 - Possessions* using the Min-max normalization and Z-score normalization. Afterwards, the classification phase will be conducted using Multi-layer Perceptron (MLP) classifier apparently available in WEKA 3.8 a machine learning tool, which will be further discussed in next sections. The obtained accuracy rates is to represent the correctly and incorrectly classified violence content in selected video instances from the dataset.

The remaining part of this paper is organized as follows, Section 2 will discuss on the methodology of the experiments, inclusive of dataset, features, pre-processing and back-propagation. Section 3 elaborates the results and discussion of the outcome. Finally, Section 4 concluding the paper with future ideas.

2. Methodology

In this section involves few stages headed for completing the whole process. Beginning from the dataset acquisition, feature extraction, pre-processing and using back-propagation as the classifier. These phases have been completed using a machine learning tool, WEKA 3.8.

2.1 Dataset Acquisition

Benchmark VSD2014 dataset obtained from Technicolor Group [10] [11] has been used as a domain in this study. Two videos have been used, namely *Agony vs Violent 2-0* and *Short Peace: Ranko Tsukigime's Longest Day - PS3 - Possessions*.

2.2 Feature Extraction

Each video data from the benchmark VSD2014 dataset is segregated into 2 categories, which is audio features and visual features.

2.2.1 Audio Features. For the audio features, consist of 8 features. Each feature is provided by per-video-frame-basis. The 8 features are accordingly, amplitude envelope (AE), root-mean-square energy (RMS), zero-crossing rate (ZCR), band energy ratio (BER), spectral centroid (SC), frequency bandwidth (BW), spectral flux (SF) and mel-frequency cepstral coefficients (MFCC). For each window, 22 MFCC is computed while all other features are computed in 1-dimensional, by total of 29 features. The equations are as followings:

$$\begin{array}{ll} \text{Amplitude Envelope} & F(x, t) = \sin \left[2\pi \left(\frac{x}{\lambda - \Delta\lambda} - (f + \Delta f)t \right) \right] + \sin \left[2\pi \left(\frac{x}{\lambda - \Delta\lambda} - (f + \Delta f)t \right) \right] \\ \text{(AE)} & \end{array} \quad (2.1)$$

$$\begin{array}{ll} \text{Root-Mean-Square} & \\ \text{Energy (RMS)} & RMS = \sqrt{\frac{a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2}{n}} \end{array} \quad (2.2)$$

$$\begin{array}{ll} \text{Zero-Crossing Rate} & \\ \text{(ZCR)} & ZCR = \frac{1}{2N} \sum_{n=1}^N |sign(x[n]) - sign(x[n-1])| \end{array} \quad (2.3)$$

$$\begin{array}{ll} \text{Band Energy Ratio} & B(t) = \frac{\sum_{i=0}^{J-1} \sum_{m=0}^{M-1} s_i^2 (Mt + m) \cdot h(M - 1 - m)}{\sum_{i=0}^J \sum_{m=0}^{M-1} s_i^2 (Mt + m) \cdot h(M - 1 - m)} \\ \text{(BER)} & \end{array} \quad (2.4)$$

$$\begin{array}{ll} \text{Spectral Centroid} & \text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \\ \text{(SC)} & \end{array} \quad (2.5)$$

$$\begin{array}{ll} \text{Frequency} & BW = f_2 - f_1 \\ \text{bandwidth (BW)} & \end{array} \quad (2.6)$$

$$\begin{array}{ll} \text{Spectral flux (SF)} & F_r = \sum_{k=1}^{N/2} (|X_r[k]| - |X_{r-1}[k]|)^2 \\ & \end{array} \quad (2.7)$$

$$\begin{array}{ll} \text{Mel-frequency} & F_{mel} = \frac{1000}{\log(2)} \cdot \left[1 + \frac{F_{Hz}}{1000} \right] \\ \text{cepstral coefficients} & \\ \text{(MFCC)} & \end{array} \quad (2.8)$$

2.2.2 Visual Features. For the visual features, includes color naming histogram (CNH) with 99-dimensional, color moments (CM), local binary patterns (LBP) with 144-dimensional and histogram of oriented gradients (HOG). While, CM and HOG with 81-dimensional, by total of 405 features. The equations are as followings:

$$\begin{array}{ll} \text{Color Naming} & H = \cos^{-1} \frac{\frac{1}{2}[(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \\ \text{Histogram (CNH)} & \end{array} \quad (2.9)$$

$$\begin{array}{ll} & S = 1 - \frac{3[(\min(R, G, B))]}{R + G + B} \\ & V = \left(\frac{R + G + B}{3} \right) \\ \text{Color Moments} & E_i = \sum_{j=1}^{j=N} \frac{1}{N} P_{ij} \\ \text{(CM)} & \end{array} \quad (2.10)$$

$$\begin{array}{ll} \text{Local Binary} & LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \\ \text{Patterns (LBP)} & \end{array} \quad (2.11)$$

$$\begin{array}{ll} \text{Histogram of} & L2\text{-norm:} f = \frac{v}{\sqrt{||v||_2^2 + e^2}} \\ \text{Oriented Gradients} & \\ \text{(HOG)} & \end{array} \quad (2.12)$$

$$L1\text{-norm:} f = \frac{v}{(||v||_1 + e)}$$

$$L1\text{-sqr:} f = \sqrt{\frac{v}{(||v||_1 + e)}}$$

These equations were then used to extract the features from the data categories, either audio or visual. These enables to determine the values before proceeding to the preprocessing phase.

2.3 Preprocessing

This phase will involve 3 processes, namely data imputation, normalization and data splitting. All the processes will be done before the data loaded into classification phase.

2.3.1 Imputation of Missing values. The video data from VSD2014 have quite a sum of missing values. However, before proceeding to next phase, instances with missing values do not have to be removed, it can be replaced with some other meaningful values. This is called imputing missing values [12]. It is

common to impute missing values with the mean of the numerical distribution. It can be done easily in Weka 3.8 using the ReplaceMissingValues filter.

2.3.2 Normalization. In this phase, two normalization methods were considered, namely, Min-max normalization with two ranges, (0 to 1) and (-1 to 1) respectively and Z-score normalization. Min-max normalization performs a linear transformation to the original data. Let say, $\min_a$ and $\max_a$ are the minimum and the maximum values for attribute A. min-max normalization maps a value v of A to v' in the range $[new - \min_a, new - \max_a]$ by computing [5]:

$$v' = ((v - \min_a) / (\max_a - \min_a)) * (new - \max_a - new - \min_a) + new - \min_a \quad (2.13)$$

Meanwhile, in Z-score normalization, the values for attribute A are normalized based on the mean and standard deviation of A. A value of v of A is normalized to v' by computing $v' = (v - \bar{A}) / \sigma_A$, where $\bar{A}$ and σ_A are the mean and the standard deviation respectively of attribute A [5]. This normalization method is useful when the actual minimum and maximum of attribute A in its infancy. After the completion of this process.

2.3.3 Data Segregation. Afterwards, for classification phase, dataset will be segregated into training and testing set. In this study the ratio for training and testing is 60:40. Each video encoded with 25fps consist of 1612 instances.

2.4 Back-propagation

The back-propagation algorithm [8] is used in layered feed-forward MLP. The idea of the back-propagation algorithm is to reduce this error, until the MLP learns the training data [7]. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. The weighted sum of a neuron is written as following:

$$A_j(x, w) = \sum_{i=0}^n X_i W_{ji}, \quad (2.14)$$

where the sum of input X_i is multiplied by their respective weights, W_{ji} . The activation depends only on the inputs and the weights. The most used output function is sigmoid function [9]:

$$O_j(x, w) = \frac{1}{1 + e^{-A(x, w)}} \quad (2.15)$$

The sigmoid function is very close to one for large positive numbers and very close to zero for large negative numbers. The goal of the training process is to obtain a desired output when certain inputs are given. Since the error is the difference between the actual and desired output, the error depends on the weights and preferred to be adjusted in order to minimize the error. The error function for the output of each neuron can be defined as:

$$E_j(x, w, d) = (O_j(x, w) - d_j)^2 \quad (2.16)$$

The error of the network will simply be the sum of the errors of all the neurons in the output layer:

$$E(x, w, d) = \sum_j (O_j(x, w) - d_j)^2 \quad (2.17)$$

where O_j is the target output and d_j is the target or desired output. After finding this, the weights can be adjusted using the method of gradient descent:

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} \quad (2.18)$$

This equation inferred in the following way: the adjustment of each weight (Δw_{ji}) will be the negative of a constant eta (η), where η is the learning rate. Multiplied by the dependence of the previous weight on the error of the network, which is derivative of E in respect to w_{ji} . The size of the adjustment will depend on η , and on the contribution of the weight to the error of the function. This is, if the weight contributes a lot to the error. Equation (2.19) is used until appropriate weights with minimal error founded.

Henceforth, derivative of E in respect to w_{ji} discovered. First, calculate the error depends on the output, which is the derivate of E in respect to O_j from Equation (2.17).

$$\frac{\partial E}{\partial O_j} = 2(O_j - d_j) \quad (2.19)$$

The reliance of the output on the activation depends on the weights from Equation (2.15) and Equation (2.16). Can be seen that from Equation (2.17) and Equation (2.20):

$$\frac{\partial O_j}{\partial w_{ji}} = \frac{\partial O_j}{\partial A_j} \frac{\partial A_j}{\partial w_{ji}} = O_j(1 - O_j)x_i \quad (2.20)$$

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial O_j} \frac{\partial O_j}{\partial w_{ji}} = 2(O_j - d_j)O_j(1 - O_j)x_i \quad (2.21)$$

The adjustment to each weight will begin from Equation (2.19) and Equation (2.21):

$$\Delta w_{ji} = -2\eta(O_j - d_j)O_j(1 - O_j)x_i \quad (2.22)$$

Equation (2.22) can be used as it is for training ANN with two layers. For training the network with one more layer, some considerations are needed particularly on training time which can be affected by the architecture of the network [6].

3. Result and Discussion

Table 1 explains the effects of normalization to MLP classifier using 2 different approaches. This process have used 1612 instances of 2 videos from VSD2014 benchmark datasets. The data is segregated 60:40 for classification phase. The classification phase have been tested with different number of hidden nodes as well, 5, 10 and 20, respectively. The experiment is conducted based on trial and error method, to obtain the accurate out by running the simulations. The outcome is shown in table below:

Table 1: Accuracy rate from the two different approaches.

Approach	Min-max MLP, [0 to 1]	Min-max MLP, [-1 to 1]	Z-score MLP
Dataset	VSD2014	VSD2014	VSD2014
Accuracy (%)			
Hidden Nodes 5	97	57	49
Hidden Nodes 10	98	55	53
Hidden Nodes 20	97	59	50

From the results, hence proven that normalization plays an important role in classifying tasks. Most importantly choosing the right approach will be very much helpful in completing the desired tasks. By performing appropriate approach, it is believed that the mining process can be done without much complex issues. For this study, Min-max with [0 to 1] range gives the best classification accuracy as

compared from other 2 approaches. This results have been obtained from Figure 1 shows the experiment results visualized in graphical form.

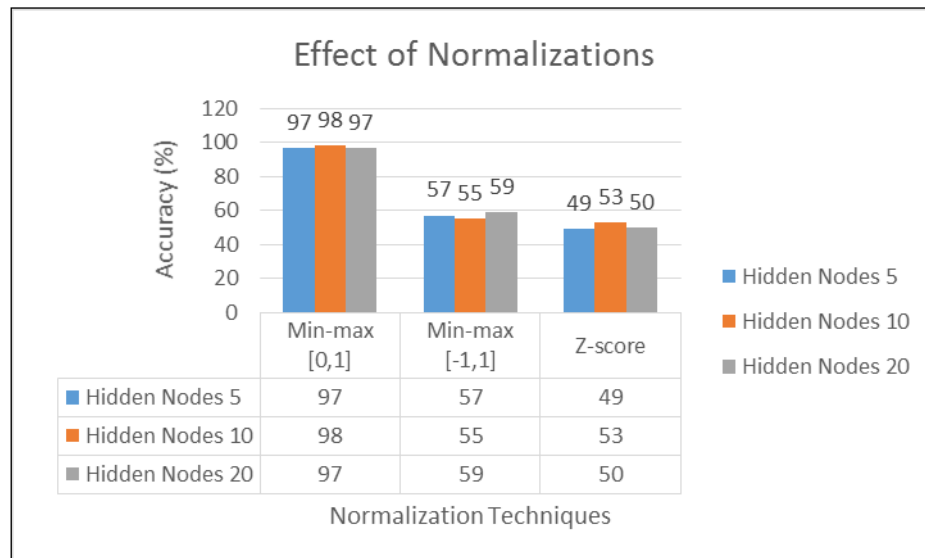


Figure 1. Result visualization for the experiment outcomes.

As seen in Table 1 and Figure 1, among Min-max with range [-1 to 1] and Z-score, Min-max with range [0,1] has outperformed. With hidden nodes 10 Min-max with range [0 to 1] have achieved 98% of accuracy to correctly classified instances, defeating the experiment using hidden nodes 5 and 10 by 1%. [1][15][16]. It is clear that dissimilar parameters, networks and architecture produces different desired output accordingly, depending on the data trends and network selections. However, this is meant for this particular dataset and classification algorithm. These results have been obtained by conducting trial and error based experiments in order to achieve desired output from the process.

4. Conclusion

This paper examines the effect of normalization for MLP in order to get best classification accuracy. It is proven that prior knowledge to datasets will give a vast effect in term of classification, especially in violence video classification. Data preparation have given an extensive effect in order to get desired output. The experimental results suggested choosing the min-max normalization data set as the best design for training data set. In all the experiments, the Min-max normalization data set always has the highest priority. In future, this study meant to fully utilize the VSD2014 using the Deep Neural Network approach with Min-max normalization as a prior knowledge to the dataset to correctly classify violence and non-violence content. This will remark the best outcome for the work.

Acknowledgments. This research funded by Ministry of Higher Education (MOHE) under the Fundamental Research Grant Scheme (FRGS) – Vot. No. 1608. Besides, partially supported by Office for Research, Innovation, Commercialization and Consultancy Management (ORICC), UTHM and Research Gates IT Solution Sdn. Bhd.

References

- [1] Al Shalabi L, Shaaban Z, Kasasbeh B. Data mining: A preprocessing engine. Journal of Computer Science. 2006 Sep;2(9):735-9
- [2] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011 Jun 9.

- [3] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2016 Oct 1.
- [4] Patro S, Sahu KK. Normalization: A Preprocessing Stage. arXiv preprint arXiv:1503.06462. 2015 Mar 19.
- [5] Kantardzic M. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons; 2011 Jan 5.
- [6] Ding C, Fan S, Zhu M, Feng W, Jia B. Violence detection in video by using 3D convolutional neural networks. In International Symposium on Visual Computing 2014 Dec 8 (pp. 551-558). Springer International Publishing.
- [7] Eyben F, Wenginger F, Lehment N, Schuller B, Rigoll G. Affective video retrieval: Violence detection in Hollywood movies by large-scale segmental feature extraction. PLOS one. 2013 Dec 31;8(12):e78506.
- [8] Ilonen J, Kamarainen JK, Lampinen J. Differential evolution training algorithm for feed-forward neural networks. Neural Processing Letters. 2003 Feb 1;17(1):93-105.
- [9] Gershenson C. Artificial neural networks for beginners. arXiv preprint cs/0308031. 2003 Aug 20.
- [10] Zhang B, Yi Y, Wang H, Yu J. MIC-TJU at MediaEval Violent Scenes Detection (VSD) 2014. In MediaEval 2014 Oct.
- [11] Schedi M, Sjöberg M, Mironică I, Ionescu B, Quang VL, Jiang YG, Demarty CH. Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos. In Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on 2015 Jun 10 (pp. 1-6). IEEE.
- [12] Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. Journal of machine learning research. 2007;8(Jul):1623-57.
- [13] Mironică I, Duță IC, Ionescu B, Sebe N. A modified vector of locally aggregated descriptors approach for fast video classification. Multimedia Tools and Applications. 2016 Aug 1;75(15):9045-72.
- [14] Zhang T, Jia W, Yang B, Yang J, He X, Zheng Z. Mowld: a robust motion image descriptor for violence detection. Multimedia Tools and Applications. 2017 Jan 1;76(1):1419-38.
- [15] Jayalakshmi T, Santhakumaran A. Statistical Normalization and Back Propagation for Classification. International Journal of Computer Theory and Engineering. 2011 Feb 1;3(1):89.
- [16] Pan J, Zhuang Y, Fong S. The Impact of Data Normalization on Stock Market Prediction: Using SVM and Technical Indicators. In International Conference on Soft Computing in Data Science 2016 Sep 21 (pp. 72-88). Springer Singapore.
- [17] Hu W, Ding X, Li B, Wang J, Gao Y, Wang F, Maybank S. Multi-perspective cost-sensitive context-aware multi-instance sparse coding and its application to sensitive video recognition. IEEE Transactions on Multimedia. 2016 Jan;18(1):76-89.
- [18] Lam V, Phan S, Le DD, Duong DA, Satoh SI. Evaluation of multiple features for violent scenes detection. Multimedia Tools and Applications. 2016:1-25.
- [19] Tan T, Li X, Chen X, Zhou J, Yang J, Cheng H, editors. Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings. Springer; 2016 Oct 21.
- [20] Attig A, Perner P. The Problem of Normalization and a Normalized Similarity Measure by Online Data. Tran. CBR. 2011;4(1):3-17.