

# A Robust Data Scaling Algorithm for Gene Expression Classification

Xi Hang Cao and Zoran Obradovic

**Abstract**—Gene expression data are widely used in classification tasks for medical diagnosis. Data scaling is recommended and helpful for learning the classification models. In this study, we propose a data scaling algorithm to transform the data uniformly to an appropriate interval by learning a generalized logistic function to fit the empirical cumulative density function of the data. The proposed algorithm is robust to outliers, and experimental results show that models learned using data scaled by the proposed algorithm generally outperform the ones using min-max mapping and z-score which are currently the most commonly used data scaling algorithms.

## I. INTRODUCTION

Genes can be expressed differently in different cells, allowing huge variety in creation of proteins [1]. In medicine and biology, gene expression analysis has become a very powerful way to understand underlying biological processes. Microarray technology is able to measure the gene expression levels of thousands of genes for a sample simultaneously. Gene expression data have been used in machine learning and data mining tasks and achieved promising results in areas including tumor diagnosis [2][3], gene grouping [4], gene selection [5], and dynamic modeling [6].

Before performing any machine learning and data mining tasks, a preprocessing step is always recommended to smooth, generalize, and scale the data [7]. Data scaling is particularly important for models that utilize distance measures; e.g., nearest neighbor classification and clustering. Feature scaling is also helpful to improve performance of the models in most cases. The most commonly used data scaling algorithms are *min-max mapping* and *Z-score* (also called *standardization*), and the details of these algorithms will be given in later sections. Microarray gene expression data typically have a small number of samples, so that the above two algorithms may not scale the attribute values well. Another shortcoming of these two algorithms is that they are not robust to outliers. That is, if the number of examples is small, and outliers exist in the examples, the features will be poorly scaled, and the performance of the model might be negatively affected.

In this study, we propose a data scaling algorithm which can map both original and future data into a desired interval, be suitable to tasks with small number of samples, and

X. H. Cao is with Center for Data Analytics and Biomedical Informatics, Department of Computer and Information Sciences, College of Science and Technology, Temple University, 1925 N. 12th Street, Philadelphia, PA, U.S.A xi.hang.cao@temple.edu

Z. Obradovic is with Center for Data Analytics and Biomedical Informatics, Department of Computer and Information Sciences, College of Science and Technology, Temple University, 1925 N. 12th Street, Philadelphia, PA, U.S.A zoran.obradovic@temple.edu

TABLE I: Data Scaling Algorithm Feature Comparison

Features	Minmax	Z-score	Proposed
Data scaling	✓	✓	✓
Interval mapping	✓		✓
Robust to outliers			✓

at same time, be robust to outliers. Our contribution is summarized as follow:

- Adapted from the idea of histogram equalization, we develop an algorithm that maps the original data uniformly into a desired interval.
- With no assumption on the sample distribution, the algorithm utilizes the generalized logistic functions to approximate the cumulative density functions.
- The algorithm is suitable in tasks with small number of samples (e.g., gene expression classification), and is robust to outliers.

Table I shows the comparison of features of different data scaling algorithms.

The remaining text is organized as follows: in Section II, the details of the data scaling algorithms that are currently used will be described; the details of the proposed algorithm are in Section III, followed by the experiments and results in Section IV; the summary will be in Section V.

## II. DATA SCALING

In machine learning and data mining, data scaling is also called data normalization. The objective of data scaling is to transfer or consolidate the data into forms appropriate for modeling and mining. Data scaling is a recommended step in data preprocessing, and it is always a good practice to perform data scaling before any modeling and mining. Data scaling is a necessary step for methods that utilize distance measures, such as, nearest neighbor classification and clustering. In addition, artificial Neural Network models require the input data to be normalized, so that the learning process can be more stable and be faster [8]. Two data scaling algorithms are widely used: *min-max mapping*, and *z-score*.

### A. Min-max Mapping Algorithm

In *min-max mapping*, the original data are linearly transformed. Suppose that  $min_A$  and  $max_A$  are the minimum and the maximum of attribute  $A$ . The *min-max mapping* algorithm maps a value,  $v$ , of  $A$  to  $v'$  in the interval of  $[min'_A, max'_A]$ , using the following formula

$$v' = \frac{v - min_A}{max_A - min_A} (max'_A - min'_A) + min'_A. \quad (1)$$

The advantage of *min-max mapping* is that it preserves the relationships of the original data values. However, it also has disadvantages: when the future input cases for scaling fall outside of the original data range of  $A$ , the mapped value will be out of the bounds of the interval  $[\min'_A, \max'_A]$ ; in addition, it is very sensitive to outliers.

### B. Z-score Algorithm

In the *Z-score* algorithm, the new value,  $v'$ , of an attribute,  $A$ , is scaled from the original value,  $v$ , using the formula

$$v' = \frac{v - \bar{A}}{\sigma_A}, \quad (2)$$

where  $\bar{A}$  is the mean of the original values in attribute  $A$ , and  $\sigma_A$  is the standard deviation of the original values in attribute  $A$ . After the scaling, the new values in attribute  $A$  will have value 0 as the mean, and value 1 as the standard deviation. This algorithm is less sensitive to outliers, compared to *min-max mapping*. However, it will not map the original data into an interval. When the number of examples is small, the mean and standard deviation of attribute  $A$  calculated by data may not be able to approximate the true mean and standard deviation well, so future input values will scale poorly.

## III. PROPOSED ALGORITHM

We propose a data scaling algorithm which maps the values into the open interval  $(0,1)$ , and be robust to outliers. In addition, the proposed algorithm transforms the original distribution of values in an attribute into a uniform distribution, so that data points will be separated evenly within the desired interval - any points, which are originally close, are more distinguishable after scaling.

### A. Scaling Formula

The new value  $v'$  of an original value,  $v$ , in attribute  $A$ , is computed by the following formula

$$v' = \int_{-\infty}^v p_X(x) dx, \quad (3)$$

where  $p_X(\cdot)$  is the distribution density function (*pdf*) of the values (represented by a random variable,  $X$ ) in an attribute  $A$ . Here, we assume that the value the distribution of the random variables are continuous. Formula (3) can be also written as

$$v' = P_X(v), \quad (4)$$

where  $P_X(\cdot)$  is the cumulative density function (*cdf*) of the values in an attribute  $A$ .

*Theorem 1:* The mapping in formula (3) maps the random variable,  $X$ , the values of attribute  $A$ , to a random variable  $Y$ , which is a uniform random variable in the interval  $(0,1)$ .

*Proof:* Let  $P_Y(\cdot)$  denotes the *cdf* of random variable  $Y$ ,  $P(e)$  denotes the probability of an event,  $e$ , and  $P_X^{-1}(\cdot)$

denotes the inverse of the *cdf* of random variable  $X$

$$\begin{aligned} Y &= P_X(X) \\ P_Y(y) &= P(Y < y) = P(P_X(X) < y) \\ &= P(X < P_X^{-1}(y)) = P_X(P_X^{-1}(y)) \\ &= y \\ p_Y(y) &= \frac{dP_Y(y)}{dy} = 1 \end{aligned}$$

The idea, which uses the cumulative density function (*cdf*) of data to map the values from one interval to another interval, is originated from the Histogram Equalization technique [9] in the field of Digital Image Processing. It is used to enhance the contrast of an image. The new values of an attribute after scaling by the *cdf* will span evenly in the close interval  $[0,1]$ , thus any samples originally were close with each other will become relatively more distant. This enhancement of sample separation may help to improve the classification performance. However, in practice, we do not know the distribution of the data, and thus we do not know the *cdf*.

### B. Cumulative Density Function Approximation

From the data, we do not know the exact form of the cumulative density function (*cdf*) of an attribute  $A$ ; therefore, we need to approximate the *cdf*. We can find the empirical cumulative density function (*ecdf*) of an attribute  $A$  by using the formula

$$\hat{P}_X(v) = \frac{1}{n} \sum_{i=1}^n 1_{x_i \leq v}, \quad (5)$$

where  $\hat{P}_X(v)$  is the *ecdf* at a value  $v$ ,  $n$  is the number of examples, and  $x_i$  is the value of attribute  $A$  in the  $i^{th}$  example.

Unfortunately, in most of the cases, the *ecdf* has no analytical form representation. Moreover, original data tend to be noisy, so the *ecdf* usually is very bumpy. Therefore, we propose to use a generalized logistic function (GLF) to approximate the *ecdf*. Using a Logistic Function to approximate the *cdf* of a normal distribution was proven viable and accurate [10]. In this algorithm, we do not make any assumption on the distribution of the data; therefore, we use a more general form of the Logistic function

$$L(x) = C + \frac{K - C}{(1 + Qe^{-B(x-M)})^{1/\nu}}; \quad (6)$$

because the range of an *ecdf* is  $[0,1]$ , the parameter  $C$  should equal 0, and the parameter  $K$  should equal 1. Formula (6) can be rewritten as

$$L(x) = \frac{1}{(1 + Qe^{-B(x-M)})^{1/\nu}}. \quad (7)$$

Compared to the Logistic Function used in [10], this general form of Logistic Function provides us with the flexibility to approximate a more variety of distributions. One of the notable properties of (7) is that it maps the values in the

interval  $(-\infty, \infty)$  to the interval  $(0, 1)$ . Compared to *ecdf* and *min-max mapping* (they map the values in the interval  $[\min_A, \max_A]$  to the interval  $[0, 1]$ , where  $\min_A$  and  $\max_A$  are the minimum and maximum value of attribute  $A$ , respectively), this property makes our proposed scaling algorithm robust to outliers, and guarantees that the scaled data will be in  $(0, 1)$ ; contrast to *ecdf* and *mim-max mapping*, if the future data are not in  $[\min_A, \max_A]$ , the scaled values are going to be out of the bound of  $[0, 1]$ .

In order to approximate the *ecdf*, we need to learn the parameters  $Q$ ,  $B$ ,  $M$ , and  $\nu$  from the data, so that the GLF could best fit the *ecdf*. The sum of squared differences of the GLF and the *ecdf* can be represented by

$$\eta = \sum_{i=1}^n \|L(x_i) - \hat{P}_X(x_i)\|^2, \quad (8)$$

noting that  $\eta$  is a function of the parameters of the GLF. The best set of parameters is the minimizer of  $\eta$ , so the key to find the most appropriate GLF to approximate the *ecdf* is to solve an optimization problem

$$\underset{B, M, Q, \nu}{\text{minimize}} \eta(B, M, Q, \nu). \quad (9)$$

Because (7) and (8) are differentiable, we can easily derive the derivatives of  $\eta$  with respect to the parameters

$$\begin{aligned} \frac{d\eta}{dB} &= \sum_{i=1}^n -T_1 \frac{Qe^{-B(x_i-M)}(x_i-M)}{T_2}, \\ \frac{d\eta}{dM} &= \sum_{i=1}^n T_1 \frac{BQe^{-B(x_i-M)}}{T_2}, \\ \frac{d\eta}{dB} &= \sum_{i=1}^n T_1 \frac{e^{-B(x_i-M)}}{T_2}, \\ \frac{d\eta}{dB} &= \sum_{i=1}^n -T_1 \frac{\ln(Qe^{-B(x_i-M)} + 1)}{\nu^2(Qe^{-B(x_i-M)} + 1)^{1/\nu}}, \end{aligned}$$

where

$$\begin{aligned} T_1 &= 2(\hat{P}_X(x_i) - L(x_i)) \\ T_2 &= \nu(Qe^{-B(x_i-M)} + 1)^{1/\nu+1}. \end{aligned}$$

Therefore, the local minimum of the objective function,  $\eta(B, M, Q, \nu)$ , can be solve efficiently with any of the gradient descent algorithms. In order to achieve a good local minimum (or even global minimum) of the objective function, the values of the parameters should be carefully initialized. We arrive at the following initialization of the parameters:

$$\begin{aligned} Q_0 &= 1 \\ M_0 &= \hat{P}_X^{-1}(0.5) \end{aligned} \quad (10)$$

$$B_0 = \frac{\ln(9)}{\max_A - M_0} \quad (11)$$

$$\nu_0 = \log_{10}(1 + Q_0 e^{-B_0(\min_A - M_0)}) \quad (12)$$

The reason behind (10) is that when  $Q$  and  $\nu$  are around 1, the GLF with an input valued at  $M$  should be around 0.5,

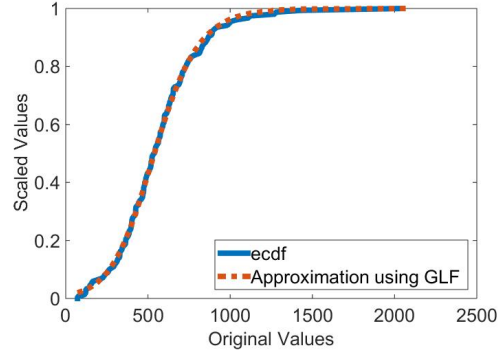


Fig. 1: An example showing the approximation of a *ecdf* using a generalized logistic function (GLF).

namely  $L(M) = \frac{1}{2^{1/\nu}} \approx 0.5$ ; therefore,  $M \approx L^{-1}(0.5) \approx \hat{P}_X^{-1}(0.5)$ , and  $\hat{P}_X^{-1}(0.5)$  can be approximated by the median of the original values in attribute  $A$ . If we assume that as the GLF at  $\max_A$  and  $\min_A$  (the maximum and minimum of the original values in attribute  $A$ ), the output values are approximately 0.9 and 0.1 respectively, we obtain the initializations in (11) and (12). With these initialization we could find a set of parameters which make the GLF fit the *ecdf* reasonably well, as shown in Fig. 1.

#### IV. EXPERIMENTS AND RESULTS

One of the datasets used to evaluate our proposed method is originally published in [11], and we download the data from the website which was made available by [12]. The data are consisted of multiple samples from 14 human cancer tissues and 12 normal tissues, and each sample was obtained by measuring the expression levels of 15,009 genes. By coupling the cancer and normal samples from the same tissue type, we extracted 4 binary classification data sets for diagnosing the four most common tumor types worldwide (*lung*, *breast*, *colon*, and *prostate*) [13]. Another dataset is from [14], and samples have measurements of 12,625 genes from the Myeloma cells of patients to diagnose whether there are bone lesions. Summaries of the data sets are in Table II.

To assess how different data scaling algorithms affect the classification performances, we use Logistic Regression and Support Vector Machine as the classification models. These two classification models have been used extensively in biological and medical researches due to their simplicity and accessibility. The program codes are implemented in

TABLE II: Summary of data used in Experiments

Dataset	#Genes	# of Samples (positive/negative)
<i>Breast</i> [12]	15,009	17/15
<i>Colon</i> [12]	15,009	15/11
<i>Lung</i> [12]	15,009	20/7
<i>Prostate</i> [12]	15,009	14/9
<i>Myeloma</i> [14]	12,625	137/36

TABLE III: Results of 5 data sets (5 binary classification tasks) using different data scaling algorithms and classification models. Raw: raw data (no data scaling); Minmax: min-max mapping; Z-score: z-score scaling; Proposed: proposed algorithm. Best performances are emphasized in bold

Dataset	Model	Raw	Minmax	Z-score	Proposed
<i>Breast</i>	Logistic Regression	0.3235	0.7882	<b>0.8353</b>	0.8118
	Support Vector Machine	0.4941	<b>0.8000</b>	0.7882	<b>0.8000</b>
<i>Colon</i>	Logistic Regression	0.5000	0.9394	0.9333	<b>0.9818</b>
	Support Vector Machine	0.8000	0.9333	0.9091	<b>0.9818</b>
<i>Lung</i>	Logistic Regression	0.4500	0.8429	0.7286	<b>0.8643</b>
	Support Vector Machine	0.2857	0.7929	0.7714	<b>0.8571</b>
<i>Prostate</i>	Logistic Regression	0.4643	0.7143	0.7381	<b>0.8016</b>
	Support Vector Machine	0.5000	0.7698	0.7460	<b>0.8571</b>
<i>Myeloma</i>	Logistic Regression	0.5000	0.7374	0.7541	<b>0.7644</b>
	Support Vector Machine	0.6281	0.7597	0.7609	<b>0.7658</b>

MATLAB. For datasets with a small number of samples, namely, *Breast*, *Colon*, *Lung*, and *Prostate*, the results are obtained using leave-one-out cross validation; for the *Myeloma* data set, 10-fold cross validation is used to obtain the result. The performances are measured by the Area Under the ROC Curve (AUC). The results are shown in Table III.

In all the dataset classification tasks, most models learned with raw data (no data scaling) have very poor performance. Models learned with scaled data have significantly better performances compared to the models learned with raw data. Models learned with the data scaled by the proposed algorithm generally achieve the best AUCs. The advantage of the proposed algorithm is more notable in the datasets that with small number of samples, such as, *colon*, *lung*, and *prostate*. In the dataset with relatively larger number of samples, such as, *Myeloma*, the AUC differences of the data scaling algorithms are relatively small.

## V. SUMMARY

In this study, we propose a data scaling algorithm to transform data to an appropriate interval for machine learning and data mining tasks. In the proposed algorithm, the values of an attribute are transformed in the (0,1) interval using the cumulative density function (*cdf*) of the attribute. Since obtaining the analytical form of the *cdf* is difficult, a generalized logistic function (GLF) is used to fit the empirical *cdf*, and the optimized GLF is used for data scaling. The proposed algorithm maps original data uniformly in the desired interval, and it is robust to outliers. Experimental results show that models learned using data scaled by the proposed algorithm generally outperform the ones using min-max mapping and z-score, which are currently the most used data scaling algorithms.

## ACKNOWLEDGMENT

This work was funded, in part, by DARPA grant [DARPA66001-11-1-4183] negotiated by SSC Pacific grant. We gratefully thank Dr. Mohamed Ghalwash for his helpful feedback on the revisions of the manuscript.

## REFERENCES

- [1] G. Piatetsky-Shapiro, T. Khabaza, and S. Ramaswamy, "Capturing best practice for microarray gene expression data analysis," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 407–415.
- [2] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.
- [3] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.
- [4] W.-H. Au, K. C. Chan, A. K. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 2, no. 2, pp. 83–101, 2005.
- [5] M. Holec, J. Kléma, F. Zelezn, and J. Tolar, "Comparative evaluation of set-level techniques in predictive classification of gene expression samples," *BMC bioinformatics*, vol. 13 Suppl 10, p. S15, 2012.
- [6] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modelling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 552–564, 2012.
- [7] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.
- [8] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*. Pearson Education Upper Saddle River, 2009, vol. 3.
- [9] R. Gonzalez and R. Woods, "Digital image processing: Pearson prentice hall," *Upper Saddle River, NJ*, 2008.
- [10] S. R. Bowling, M. T. Khasawneh, S. Kaewkuekool, and B. R. Cho, "A logistic approximation to the cumulative normal distribution," *Journal of Industrial Engineering and Management*, vol. 2, no. 1, pp. 114–127, 2009.
- [11] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences*, vol. 98, no. 26, pp. 15 149–15 154, 2001.
- [12] A. Statnikov, I. Tsamardinos, Y. Dosbayev, and C. F. Aliferis, "Gems: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data," *International journal of medical informatics*, vol. 74, no. 7, pp. 491–503, 2005.
- [13] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, "Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. international agency for research on cancer, lyon," *globocan.iarc.fr (accessed 10 August 2015)*, 2013.
- [14] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J. D. Shaughnessy Jr, "The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma," *New England Journal of Medicine*, vol. 349, no. 26, pp. 2483–2494, 2003.