The impact of scaling on Support Vector Machine in Breast Cancer Diagnosis

Elsayed Badr
Scientific Computing Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

Mustafa Abdulsalam
Scientific Computing Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

Hagar Ahmed
Scientific Computing Department, Faculty of Computers and Informatics, Benha University, Benha, Egypt

ABSTRACT
By using support vector machine (SVM) and the grid technique Badr *et al*. [1] introduced new scaling techniques on the data set Wisconsin from UCI machine learning with a total 569 rows and 33 columns. These scaling techniques overcame the standard normalization techniques. In this paper, three new scaling techniques are proposed by using SVM and the grid technique on the the data set Wisconsin from UCI machine learning with a total 569 rows and 32 columns. These scaling techniques are: (i) de Buchet for p = (∞) (ii) Lp-norm for p = (∞) (iii) Entropy . Experimental results show that SVM with new scaling techniques achieves 98.60 % , 98.42 % and 98.42 % accuracy against to the standard normalization by 96.49 %.

General Terms
Data Mining, Classification

Keywords
Machine Learning, Breast Cancer, Support vector machine, scaling techniques.

1. INTRODUCTION
Improving the accuracy of identifying the breast cancer disease is very important task. Breast cancer disease is the second most common type of cancer after lung cancer. Breast cancer is the most widespread by 12.3% of all cancer for males and females of all ages. It is the most spreading in women worldwide, accounting 25.4% of the whole cases diagnosed in 2018 [1]. Defects in breast cancer diagnosis by experts can be avoided by expert systems and artificial intelligent techniques. These expert systems can examine the medical data in shorter time and help junior physicians.

Tomlin [2] performed a computational study comparing arithmetic mean, geometric mean, equilibration, Curtis and Reid scaling technique [3], Fulkerson and Wolfe scaling technique [4], and various combinations on six test problems. The conclusion of Tomlin's comparative study was that geometric mean scaling method, optionally followed by equilibration or Curtis and Reid scaling technique are the best combined scaling techniques. Larsson [5] extended Tomlin's study by comparing entropy [5], Lpnorm [6] and de Buchet [7] scaling techniques over a dataset of 135 randomly generated LPs. Larsson concluded that the entropy scaling technique can improve the conditioning number of the constraint matrix. Elble and Sahinidis [8] expanded on Tomlin's and Larsson's studies by conducting a computational study comparing arithmetic mean, de Buchet, entropy, equilibration, geometric mean, IBM MPSX [9], Lpnorm, binormilization, and various combinations of the aforementioned scaling techniques over Netlib and Kennington set. They used four measures to evaluate each scaling technique: (i) scaling time, (ii) solution time, (iii) number of iterations for the solution of the LP, and (iv) maximum conditioning number of the constraint matrix. Elble and Sahinidis concluded that equilibration is the best scaling technique.

In a previous paper [10], the autors reviewed and compared both the CPU- and GPU-based implementations of seven scaling techniques, namely: (i) arithmetic mean, (ii) de Buchet, (iii) entropy, (iv) equilibration, (v) geometric mean, (vi) IBM MPSX, and (vii) Lp-norm scaling methods. They have performed a computational study over Netlib and Kennington set and concluded that arithmetic mean, equilibration and geometric mean are the best serial scaling techniques. In this paper a computational study is performed over a set of sparse randomly generated LPs in order to highlight the impact of scaling prior to the application of IPM, EPSA and simplex algorithms. To the best of our knowledge, this is the first paper that investigates the effect of scaling on IPM, EPSA and simplex algorithms.

The scaling techniques can improve the accuracy of classifiers. Elsayed Badr et al. [11] proposed ten efficicent scaling techniques for optimizing SVM. These scalling techniques are efficient for linear programming approach [12-20]. The scalling techniques that they applied with SVM on WDBC dataset are arithmetic mean, de Buchet for three cases (p=1, 2), equilibration, geometric mean, IBM MPSX, *Lp*-norm for three cases (p=1 or 2).

The rest of this paper is organized as follows. The algorithms that are used in the study: SVM described in Section 2. The proposed model is introduced in section 3. In Section 4, detailed descriptions of new scaling techniques, de Buchet for p=(∞), entropy and Lp-norm for p=(∞) are proposed. Experimental design which has data description, experimental setup, measure for performance evaluation and a comparative study are introduced in section 5. In Section 6 the main results and discussion are proposed. Finally, conclusions and future works are introduced in section 7.

2. PRELIMINARIES
In this section, Support vector machine (SVM), and grew wolf optimizer (GWO) are presented and discussed.

Support Vector Machine (SVM)
Support vector machine SVM developed by Vapnik [9], the support vector machine (SVM) was primarily intended for binary classification. Its main objective is to determine the optimal hyperplane f (w, x) = w ⋅ x +b separating two classes in a given dataset having input features x ∈ Rp , and labels y

$\in \{-1, +1\}$. SVM learns by solving the following constrained optimization problem:

$$\min \quad \frac{1}{p} w^T w + c\sum_{i=1}^{p} \xi_i$$

$$\textit{subject to} \quad y_i'(w.x + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \;\; i = 1,...,p$$

where $w^T w$ is the Manhattan norm, ξ is a cost function, and C is the penalty parameter (may be an arbitrary value or a selected value using hyper-parameter tuning). Its corresponding unconstrained optimization problem is the following:

$$\min \quad \frac{1}{p} w^T w + c\sum_{i=1}^{p} \max(0, 1 - y_i'(w_i x_i + b)) \qquad (1)$$

where $wx + b$ is the predictor function. The objective of Eq. 1 is known as the primal form problem of L1-SVM, with the standard hinge loss. The problem with L1-SVM is the fact that it is not differentiable[18], as opposed to its variation, the L2-SVM:

$$\min \quad \frac{1}{p} ||w||_2^2 + c\sum_{i=1}^{p} \max(0, 1 - y_i'(w_i x_i + b))^2$$

The L2-SVM is differentiable and provides more stable results than its L1 counterpart.

3. THE PROPOSED CLASSIFICATION MODEL

A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set [21] or evaluation on a held-out validation set [22]. A typical soft-margin SVM classifier equipped with an RBF kernel has at least two hyperparameters that need to be tuned for good performance on unseen data: a regularization constant C and a kernel hyperparameter γ. Both parameters are continuous, so to perform grid search, one selects a finite set of "reasonable" values for each, say: $C \in \{10,100,1000\}$ and . Grid search then trains an SVM with each pair (C, γ) in the Cartesian product of these two sets and evaluates their performance on a held-out validation set (or by internal cross-validation on the training set, in which case multiple SVMs are trained per pair). Finally, the grid search algorithm outputs the settings that achieved the highest score in the validation procedure.

4. SCALING TECHNIQUES

Here, we introduce the mathematical notations of ten scaling techniques in addition to the normalization scaling techniques with ranges [0, 1] and [-1, 1]. First of all, we introduce the following mathematical preliminaries as shown in Table 1.

The scaled matrix is expressed as *RAS*, such that $R = diag(r_1, \ldots, r_m)$ and $S = diag(s_1, \ldots, s_n)$. All scaling techniques proposed in this section apply first rows scaling and after that columns scaling. Then, the matrix after full scaling (row and column) is given by:

$$A^R = RA; \; A^{RS} = A^R S \qquad (2)$$

Table 1. Mathematical preliminaries for scaling techniques

Symbol	Description
$A\ (a_{ij})$:	m x n matrix (with m (observations) and n (attributes)).
r_i:	The scaling agent of row i
s_j:	The scaling agent of column j
R:	Diagonal matrix such that $R = diag\ (r_1, \ldots, r_m)$
S:	Diagonal matrix such that $S = diag\ (s_1, \ldots, sn)$
N_i:	$N_i = \{j : A_{ij} \neq 0\}$, such that $1 \leq i \leq m$
M_j:	$M_j = \{i : A_{ij} \neq 0\}$ such that $1 \leq j \leq n$
n_i:	The number of elements for the set N_i
m_j:	The number of elements for the set M_j
$A^R\ (a_{ij}^R)$	The scaled matrix by row R scaling agent.
$^{RS}(a_{ij}^{RS})$	The final scaled matrix.

1) **de Buchet scaling technique**: The de Buchet scaling model is based on the relative divergence and is formulated as shown in Equation (3):

$$\min_{r,s>0} \left(\sum_{i,j \in \overline{Z}} \{a_{ij} r_i s_j + 1/a_{ij} r_i s_j\}^p \right)^{1/p} \qquad (3)$$

where p is a positive integer and $\overline{Z}$ is the number of the nonzero elements of A. We focus on the case $p = \infty$.

Case $p = \infty$, Equation (3) is formulated as shown in Equation (4):

$$\min_{r,s>0} \max_{i,j \in \overline{Z}} |\log(a_{ij} r_i s_j)| \qquad (4)$$

The row scaling factors are described in Equation (5):

$$R_i = 1/\left\{ \left(\max_{j \in N_i} |a_{ij}| \right) \left(\min_{j \in N_i} |a_{ij}| \right) \right\}^{1/2} \qquad (5)$$

Similarly, the column scaling factors are presented in Equation (6):

$$s_j = 1/\left\{ \left(\max_{i \in M_j} |a_{ij}^R| \right) \left(\min_{i \in M_j} |a_{ij}^R| \right) \right\}^{1/2} \qquad (6)$$

The de Buchet for the case $p = \infty$ scaling technique corresponds to the geometric mean scaling technique that will be described later.

2) L_p-norm scaling technique: The L_p-norm scaling model is formulated as shown in Equation (7):

$$\min_{r,s>0} \sum_{i,j \in \overline{Z}} (|\log(a_{ij} r_i s_j)|^p)^{1/p} \qquad (7)$$

where p is a positive integer and $\overline{Z}$ is the cardinality number of the nonzero elements of the constraint matrix A.

We focus now our attention on the case $p = \infty$.

Case $p = \infty$, the model is equivalent to the de Buchet method (case $p = \infty$) and geometric mean scaling techniques which were proposed in [11].

3) Entropy scaling technique is equivalent to the arithmetic scaling technique that was introduced in [11].

4) Normalization scaling technique [0, 1]: Equation (8) is used for normalization scaling method with range [0, 1] such that a, a', max_k and min_k are the original value, the scaled value, the maximum value and the minimum value of feature k respectively.

$$a' = \frac{a - min_k}{max_k - min_k} \quad (8)$$

5. EXPERIMENTAL DESIGN

In this section, we introduce data description, measure for performance evaluation and the comparative study.

5.1 Data description

In this work, we have run the proposed model on the Wisconsin diagnosis Breast Cancer (WDBC) dataset that available the UCI Machine Learning Repository [23]. The dataset consists of 569 instances divided into two classes. The two classes malignant and benign have 357 and 212 cases respectively. Each record in the database has thirty-three attributes.

5.2. Experimental setup

The proposed model was developed by MATLAB. SVM, implementation was enhanced, which is originally developed by Chang and Lin [24]. Table 3 describes the experiments computing environment.

Table 2. Description of the computing environment

CPU	Intel (R) Core (TM) i5- 7200U CPU@ 2.70 GHz
RAM Size	4 GB RAM
MATLAB version	R2015a (8.5.0.197613)

Salzberg [25] introduced the k-fold CV which is used to guarantee the valid results. In this paper, k = 10.

5.3. Measure for performance evaluation

In order to test the performance of the proposed model, we use accuracy. According to the confusion matrix, accuracy is defined as follows:

$$Acc = (TruPos + TruNeg) / [TruPos + FlsPos + TruNeg + FlsNeg] \times 100\% \quad (9)$$

Where: Acc: Accuracy; TruPos: true positive; TruNeg: true negative; FlsPos: false positive qnd FlsNeg,: false negative.

5.4. Comparative study

In this study, we compare the performance of the proposed SVM using grid search technique with different scaling techniques. The best C and γ are computed by grid search. The searching space of parameters C and γ are set to C = $\{2^{-5}, 2^{-3},\ldots, 2^{15}\}$ and γ = $\{2^{-15}, 2^{-13},\cdots, 2^{1}\}$, respectively.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

Table 3 and Table 4 show a comparison among classification accuracies of SVM with normalization scaling [0, 1], de Buchet scaling for p=(∞), L_p-norm scaling for p=(∞) and entropy scaling technique. It is apparent from these tables that the average accuracy rates achieved by SVM with de Buchet scaling technique for p = ∞ (98.60 %), L_p-norm scaling technique for p = ∞ (98.42) and entropy scaling technique (98.42 %) are better than that obtained by SVM with normalization scaling techniques (96.49%)

Table 3: Accuracy for WBCD database using SVM with C and γ which were calculated by grid search technique (Without scaling and Normalization scaling [0,1]

Fold	Normalization scaling [0,1] (S1)			de Buchet p =inf (S2)		
	C	γ	Acc.%	C	γ	Acc.%
1	2^13	2^-7	100	2^5	2^-5	100
2	2^15	2^-9	98.25	2^11	2^-9	100
3	2^15	2^1	92.98	2^11	2^-9	96.49
4	2^15	2^-1	94.74	2^11	2^-9	96.49
5	2^15	2^-1	94.74	2^3	2^-11	100
6	2^15	2^1	96.49	2^11	2^-9	100
7	2^15	2^-3	98.25	2^11	2^-9	100
8	2^15	2^-13	96.49	2^11	2^-9	94.74
9	2^15	2^1	94.74	2^13	2^-11	98.25
10	2^15	2^1	98.25	2^13	2^-11	100
Avg.	**30310.4**	**0.91**	**96.49**	2871.2	0.0044	98.60
CPU Time	7.263212			15.417493		

Table 4: Accuracy for WBCD database using SVM with C and γ which were calculated by grid search technique (Normalization scaling [-1,1] and de Buchet scaling(p=1))

Fold	Entropy (S3)			Lp-norm p = inf (S4)		
	C	γ	Acc. %	C	γ	Acc. %
1	2^3	2^-7	100.00	2^11	2^-9	98.21
2	2^15	2^-9	98.25	2^9	2^-9	98.25
3	2^9	2^-5	96.49	2^11	2^-9	98.25
4	2^-1	2^-5	96.49	2^3	2^-3	98.25
5	2^9	2^-9	100.00	2^11	2^-9	98.25
6	2^5	2^-5	98.25	2^11	2^-9	96.49
7	2^7	2^-7	98.25	2^11	2^-9	100
8	2^-1	2^-3	98.25	2^11	2^-9	100
9	2^9	2^-9	100.00	2^11	2^-9	98.25
10	2^15	2^-9	98.25	2^9	2^-11	98.25
Avg.	**6724.1**	**0.024**	**98.42**	1536.8	0.014	98.42
CPU Time	12.516496			15.03757		

Table 5 summarize the results of all scaling techniques that obtained by SVM according the accuracies and CPU times. It is apparent from Table 5 that the normalization scaling technique [0, 1] overcomes all other scaling techniques according to CPU time only. On the other hand, the de Buchet

(p =inf) scaling technique outperforms all scaling techniques according to the accuracy.

Table 5: Accuracy and CPU Time for WBCD database using SVM with C and γ which were calculated by grid search technique

No.	Techniques	Accuracy%	CPU Time
S1	Normalization [0,1]	96.49	7.263212
S2	de Buchet p =inf	98.60	15.417493
S3	Entropy	98.42	12.516496
S4	Lp-norm p = inf	98.42	15.03757

7. CONCLUSION AND FUTURE WORK

In this paper, three new scaling techniques were proposed by using SVM and the grid technique on the the data set Wisconsin from UCI machine learning with a total 569 rows and 33 columns. These scaling techniques are: (i) de Buchet for p = (∞) (ii) Lp-norm for p = (∞) (iii) Entropy. Experimental results showed that SVM with new scaling techniques achieved 98.60 % , 98.42 % and 98.42 % accuracy against to the standard normalization by 96.49 %.

8. REFERENCES

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A, "Global Cancer Statistics 2018," GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, in press.

[2] Tomlin, J. A. 1975. On scaling linear programming problems. *Mathematical Programming Studies* 4, 146-166. DOI= http://dx.doi.org/10.1007/BFb0120718.

[3] Curtis, A. R. and Reid, J. K. 1972. On the automatic scaling of matrices for Gaussian elimination. *IMA Journal of Applied Mathematics* 10, 1, 118-124. DOI= http://dx.doi.org/10.1093/imamat/10.1.118

[4] Fulkerson, D. R. and Wolfe, P. 1962. An algorithm for scaling matrices. *SIAM Review* 4, 2, 142-146. DOI= http://dx.doi.org/10.1137/1004032.

[5] Larsson, T. 1993. On scaling linear programs-Some experimental results. *Optimization* 27, 4, 335-373. DOI= http://dx.doi.org/10.1080/02331939308843895

[6] Hamming, R. W. 1971. Introduction to Applied Numerical Analysis. McGraw-Hill, New York.

[7] de Buchet, J. 1966. Experiments and statistical data on the solving of large-scale linear programs. In *Proceedings of the Fourth International Conference on Operational Research*, Hertz, D. A. and Melese, J., Eds. Wiley-Interscience, New York, 3-13.

[8] Elble, J. M. and Sahinidis, N. V. 2012. Scaling linear optimization problems prior to application of the simplex method. *Computational Optimization and Applications* 52, 2, 345-371. DOI= http://dx.doi.org/10.1007/s10589-011-9420-4

[9] Benichou, M., Gauthier, J. M., Hentges, G., and Ribiere, G. 1977. The efficient solution of large-scale linear programming problems-Some algorithmic techniques and computational results. *Mathematical Programming* 13, 1, 280-322. DOI= http://dx.doi.org/10.1007/BF01584344

[10] Ploskas, N. and Samaras N. 2013. A Computational Comparison of Scaling Techniques for Linear Optimization Problems on a GPU. *Optimization Methods and Software*. Paper under review.

[11] Elsayed Badr, Mustafa Abdul Salam, Sultan Almotairi and Hagar Ahmed " From Linear Programming Approach to Metaheuristic Approach: Scaling Techniques" Complexity (submitted)

[12] Triantafyllidis, C. and Samaras, N. "Three nearly scaling-invariant versions of an exterior point algorithm for linear programming", *Optimization.* **2014**, *64*(10), 2163–2181.

[13] Ploskas, N. and Samaras, N. "A computational comparison of scaling techniques for linear optimization problems on a graphical processing unit", *International Journal of Computer Mathematics.* **2015**, *92*(2), 319–336.

[14] E. M. Badr and H. elgendy (2020) "A Hybrid water cycle - particle swarm optimization for solving the fuzzy underground water confined steady flow" Indonesian Journal of Electrical Engineering and Computer Science Vol 19, No1: 2020

[15] Elsayed M. Badr, Mahmoud I. Moussa in *Wireless Networks (*2019*)*, An upper bound of radio *k*-coloring problem and its integer linear programming model, First Online: 18 March 2019.

[16] Badr, E.;Aloufi,K.A Robot's Response Acceleration Using the Metric Dimension Problem. *Preprints* 2019, 2019110194 (doi:10.20944/preprints201911.0194.v1).

[17] E.S. Badr, K. Paparrizos, Baloukas Thanasis and G. Varkas (2006), Some computational results on the efficiency of an exterior point algorithm, in Proc. of the 18th National Conference of Hellenic Operational Research Society (HELORS), 15-17 June, Rio, Greece, pp. 1103-1115

[18] E. S. Badr, K. Paparrizos, N. Samaras, and A. Sifaleras (2005), On the Basis Inverse of the Exterior Point Simplex Algorithm, in Proc. of the 17th National Conference of Hellenic Operational Research Society (HELORS), 16-18 June, Rio, Greece, pp. 677-687.

[19] E.S. Badr, M. Moussa, K. Paparrizos, N. Samaras, and A. Sifaleras, Some computational results on MPI parallel implementation of dense simplex method, World Academy of Science, Engineering and Technology (WASET), 23, 2008,778–781.

[20] E. M. Badr and Sultan Almotiari (2019) " On a Dual Direct Cosine Simplex Type Algorithm and Its Computational Behavior" Mathematical Problems in Engineering Volume 2020, Article ID 7361092, 8 pages. https://doi.org/10.1155/2020/7361092

[21] Chin-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin (2010). A practical guide to support vector classification. Technical Report, National Taiwan University.

[22] Chicco D (December 2017). "Ten quick tips for machine learning in computational biology". BioData Mining. 10 (35): 35. doi:10.1186/s13040-017-0155-3. PMC 5721660. PMID 29234465.

[23] Vapnik, V.N. "The nature of statistical learning theory", Springer: New York, 1995.

[24] Chang, C.C. and C.J. Lin, LIBSVM: a library for support vector machines. 2001, Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

[25] Salzberg, S. L., On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1(3):317–328, 1997.

[26] Elsayed Badr, Mustafa Abdul Salam, Sultan Almotairi and Hagar Ahmed, "From Linear Programming Approach to Metaheuristic Approach: Scaling Techniques" Complexity, Hindawi, Submitted.

[27] E. M. Badr and H. elgendy (2020) "A Hybrid water cycle - particle swarm optimization for solving the fuzzy underground water confined steady flow" Indonesian Journal of Electrical Engineering and Computer Science Vol 19, No1: 2020

[28] Elsayed M. Badr, Mahmoud I. Moussa in *Wireless Networks (*2019*)*, An upper bound of radio *k*-coloring problem and its integer linear programming model, First Online: 18 March 2019.

[29] Badr, E.;Aloufi,K.A Robot's Response Acceleration Using the Metric Dimension Problem. *Preprints* 2019, 2019110194 (doi:10.20944/preprints201911.0194.v1).

[30] E.S. Badr, K. Paparrizos, Baloukas Thanasis and G. Varkas (2006), Some computational results on the efficiency of an exterior point algorithm, in Proc. of the 18th National Conference of Hellenic Operational Research Society (HELORS), 15-17 June, Rio, Greece, pp. 1103-1115

[31] E. S. Badr, K. Paparrizos, N. Samaras, and A. Sifaleras (2005), On the Basis Inverse of the Exterior Point Simplex Algorithm, in Proc. of the 17th National Conference of Hellenic Operational Research Society (HELORS), 16-18 June, Rio, Greece, pp. 677-687.

[32] E.S. Badr, M. Moussa, K. Paparrizos, N. Samaras, and A. Sifaleras, Some computational results on MPI parallel implementation of dense simplex method, World Academy of Science, Engineering and Technology (WASET), 23, 2008,778–781.

[33] E. M. Badr and Sultan Almotiari (2019) " On a Dual Direct Cosine Simplex Type Algorithm and Its Computational Behavior" Mathematical Problems in Engineering Volume 2020, Article ID 7361092, 8 pages. https://doi.org/10.1155/2020/7361092.

[34] EM Badr, MA Salam, M Ali, H Ahmed, Social Media Sentiment Analysis using Machine Learning and Optimization Techniques, International Journal of Computer Applications (0975 – 8887) Volume 178 – No. 41, August 2019.