SCHOOL OF ENGINEERING AND ENGINEERING TECHNOLOGY

FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA

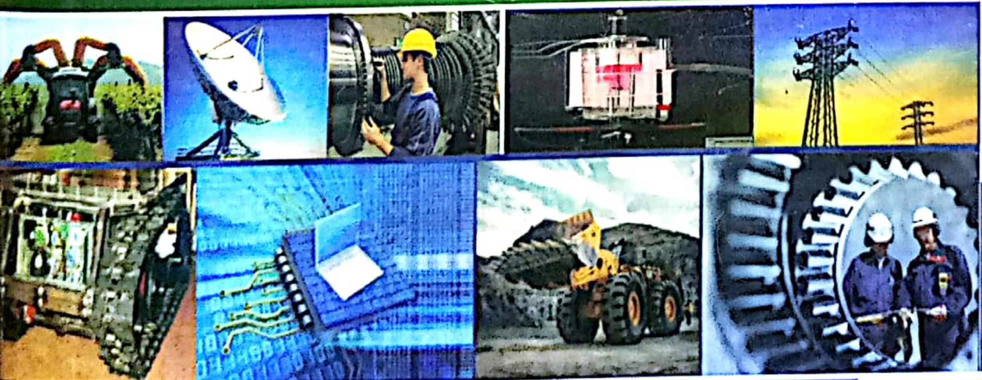
2nd INTERNATIONAL ENGINEERING CONFERENCE IEC 2017

Book of Proceedings

VOLUME 1

Theme

GREEN RESEARCH, INNOVATION & SUSTAINABLE DEVELOPMENT: A MEANS TO DIVERSIFICATION OF MONO-CULTURAL ECONOMIES



Date: 17th - 19th October, 2017

EDITED BY:

Engr. Dr. A. Nasir, Engr. Dr. S.M. Dauda, Engr. Dr. A. Onumanyi, Engr. Dr. A.S. Abdulkareem, Engr. Dr. A.G. Isah, Engr. D. A. Muriana, Engr. Dr. C. Alenoghena, Engr. H. Bello-Salau, Engr. Dr. H. Ohiza, Engr. Dr. L. Olatomiwa Engr. Dr. M. Masin, Engr. Musa Umar Engr. M. Abubakar, Engr. Engr. F. Talha, Engr. H.O. Aminulai, Engr. A. Yusuf

SCHOOL OF ENGINEERING AND ENGINEERING TECHNOLOGY
FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA

2nd INTERNATIONAL ENGINEERING CONFERENCE IEC 2017

Theme

GREEN RESEARCH, INNOVATION AND SUSTAINABLE DEVELOPMENT: A MEANS TO DIVERSIFICATION OF MONO-CULTURAL ECONOMIES

17th - 19th October, 2017

MEMBERS OF THE CONFERENCE ORGANISING COMMITTEE (COC)

Dr. A. Nasir	-	Chairman
Dr. A.S. Kovo	-	Member
Dr. M.A. Olutoye	-	Member
Dr. Mrs B.A. Adejumo	-	Member
Dr. A.S. Abdulrahman	-	Member
Dr. S.M. Dauda	-	Member
Dr. J. Tsado	-	Member
Dr. A.B. Hassan	-	Member
Prof. A.M. Aibinu	-	Member
Dr. O.M. Olaniyi	-	Member
Dr. M. Abdullahi	-	Member
Dr. M. Saidu	-	Member
Dr. A.S. Abdulkareem	-	Member/Secretary

CONTENTS

FORWARD

The School of Engineering and Engineering Technology, Federal University of Technology, Minna, organized the 1st International Engineering Conference in 2015 with tremendous attendance and success. The huge success of the 1st IEC gave the committee the strength to organize this 2nd International Engineering Conference (IEC 2017) with the theme: **Green Research, Innovation and Sustainable Development: A Means to Diversification of Mono-Cultural economies.**

The conference intends to provide opportunities for researchers, engineers, captains of industries, scientists, academics, security personnel and others who are interested in diversification of mono-cultural economies using green research and Innovation towards sustainable development to present and brainstorm on ideas and come out with a communique that will give the way forward in this respect. To arrive at this communique, the following sub-themes were carefully coined to guide the authors' submissions:

- Innovative Research, Development and Commercialization Policy.
- Electrical Power System and Electronic Engineering for National Economic Development
- Flood Disaster Management and Adaptation for Sustainable Development.
- Sustainable Water Supply, Sanitation and Water Quality Management. Sustainable Environmental Engineering and Management.
- Solid Minerals Potential Exploration for National Economic Development.
- Post-Harvest Technology and Food Security.
- Small and Medium Scale Enterprises for National Economic Development.
- Sustainable and Emerging Renewable Energy Technologies for Economic Development.
- Information and Communication Technology for Sustainable Economic and Security.
- Content Initiatives and Development.
- Structural Integrity and Geotechnical Survey for Sustainable Infrastructural Development.
- Sustainable Engineering Education and Curriculum Development.
- Engineering Entrepreneurship for National Economic Development.
- Green and Sustainable Built-Environment.
- Energy Economy, Planning and Management.

The conference editorial and Technical Board have members from the United Kingdom, Saudi Arabia, South Africa, Malaysia, Australia and Nigeria. The conference received submissions from 10 countries viz: United States of America, United Kingdom, Canada, the republic of Cyprus, Turkey, Malaysia, Botswana, South Africa, Netherlands and Nigeria. I am very happy to state that 142 papers were received and subjected to blind peer review process. Each of the paper was reviewed by two personalities who have in-depth knowledge of the subject discussed by the paper. At the end of the review process, 135 papers were accepted and recommended for presentation and publication in the conference proceedings. The conference proceedings will be indexed in Scopus.

Let me on behalf of the organizing committee seize this opportunity to thank you all for participating in the conference. Our sincere gratitude to the reviewers for finding time to do a thorough review. Thank you all and we hope to see you during the 3rd International Engineering Conference (IEC 2019)..

Engr. Dr. A. Nasir
Chairman, Conference Organizing Committee

ACKNOWLEDGEMENT

The Chairman and members of the Conference Organizing Committee (COC) of the 2nd International Engineering Conference (IEC 2017) wish to express our gratitude to the Vice Chancellor and the management of the Federal University of Technology, Minna, the Dean and all staff of the School of Engineering and Engineering Technology (SEET) for the moral and financial support towards the successful hosting of this conference. We also thank the entire staff of the university who contributed in one way or the other. We are sincerely grateful to you all.

2nd International Engineering Conference (IEC 2017)
Federal University of Technology, Minna, Nigeria

Performance Analysis of Data Normalization Methods

*Ajiboye, J. A[1], Aibinu, M. A[2]
[1]Electrical and Electronics Engineering Department, Federal University of Technology, PMB 65 Minna Niger State, Nigeria.
[2]Mechatronics Engineering Department, Federal University of Technology, PMB 65 Minna Niger State, Nigeria
*Corresponding author email: ajiboye2003@yahoo.com, +2348035532375

ABSTRACT
Statistical Data Normalization is a very important input preprocessing operation that should be done before data is fed into the training network. However, there is need for a suitable selection of normalization technique since normalization on the input has potential of varying the structure of the data and may impact on the outcome of the analysis. This paper investigates and evaluates some important statistical normalization techniques by studying thirty published papers that used wine dataset available in the UCI repository and their impact on performance accuracy. Results reveal that Min-Max normalization technique had the best performance accuracy of 95.91% on the average among all the other normalization types.

Keywords: *ANN, GA, Min-Max, Z-Score.*

1 INTRODUCTION

In an unprocessed datasets the presence of distortions, unwanted values, inadequate and omitted values and values recorded in error and insufficient sampling is inevitable. Omission of these major qualities in a data set may be as a result of human error or computer/equipment error during data entry which will definitely affect the result of data analysis.

The pre-processing of data before use is very necessary, which includes data cleaning, data preparation, data integration, data transformation and data reduction. Therefore, Data Normalization is the process of cleaning of data by inputting missing values, filtering the unwanted data, identifying and getting rid of outliers and resolving all inconsistencies. It involves rescaling of attributes. However, care must be taken during the process of data normalization so as to avoid over cleaning of data. Hence appropriate pre-processing data technique must be adopted. Although there is no generally defined rule for normalizing datasets, the choice is solely dependent on the discretion of the user (Vaishali *et al.,* 2011).

Data Normalization can be defined as transformation done on a single data input for even distribution and scaling into a range that is acceptable for the network. It has a lot of importance. Data input can be transformed into better form for the network use thereby enhancing the performance of the network. Normalization process on raw data makes the data fit for training without which it will be very slow.

Training time is being sped up when data is normalized since it involves scaling of data that has the same value range for each input thereby reducing to the barest minimal the differences within the network. For inputs that are on widely different scales, data normalization enhances modeling application as well as the quality of the data. It also corrects distortion within the network and enhances quality of images.

2 DATA NORMALIZATION TECHNIQUES

Data normalization ensures that the quality of the data is sustained prior to being fed to any learning algorithm. Several types of data normalization exist. It is important to reduce bias within the neural network for one feature to another. To achieve this, data normalization is used to scale the data in the same range of values for each input feature. There are various techniques used for data normalization such as Min-Max, Z-score, Decimal Scaling, Median Normalization, Sigmoidal Normalization, Statistical Column Normalization, Mean and Standard Deviation. These techniques are discussed here.

2.1 MIN-MAX NORMALIZATION

This technique maps the input data to a predefined range of 0 and 1 or -1 and 1. Min-Max normalization technique can be employed for preserving privacy during the mining process (Manikandan *et al.,* 2013). The Min-Max normalization method normalizes the values of the attributes of a data set according to its defined minimum and maximum values as shown in Equation (1).

$$x' = (x_{\max} - x_{\min}) X \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} + x_{\min} \tag{1}$$

Where
x_{min} is the lower bound of attribute x_i
x_{max} is the upper bound of attribute x_i
x' is the Normalized Value of attribute x_i

Min-max normalization preserves the relationships among the original data values. A problem may occur if a value of an unseen data point to be predicted is out of x_{min} and x_{max} interval.

2.2 Z-SCORE

This is also known as Zero-Mean Normalization. In this normalization method the values of an attribute x_i are normalized according to their mean and standard deviation, as shown in Equation (2). (Jayalakshmi & Santhakumaran, 2011).

$$x^{!} = \frac{(x_i - \mu_i)}{\sigma_i} \tag{2}$$

Where

μ_i is the mean and

σ_i is the standard deviation

If μ_i and σ_i are not known they can be estimated from the sample. Z-score normalization may be sensitive to small values of σ_i.

2.3 DECIMAL SCALING NORMALIZATION

In this method, the decimal point of the values of an attribute x_i is moved to its maximum absolute value as seen in Equation (3) (Luai *et al.,* 2006). The number of decimal points moved depends on the maximum absolute value of the data set.

$$x^{!} = \frac{x_i}{10^c} \tag{3}$$

Where

$x^{/}$ is the Normalized Value of attribute x_i

c is the smallest integer such that $\max(|x^{/}|)<1$

2.4. MEDIAN NORMALIZATION

This method normalizes each sample by the median of the unprocessed data inputs of all the inputs in the sample. It is a useful normalization technique that can be employed when there is a need to compute the ratio between two hybridized samples. Median is not influenced by the magnitude of extreme deviations as shown in Equation (4) (Jayalakshmi and Santhakumaran, 2011).

$$x^{!} = \frac{x_i}{median(a_i)} \tag{4}$$

2.5. SIGMOID NORMALIZATION

This normalization method is the simplest one used for most of the data normalization (Jain et al., 2005). The data value of attribute x is normalized to $x^{/}$ as shown in Equation (5).

$$x^{!} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

The sigmoid normalization function is used to scale the samples in the range of 0 and 1 or -1 to +1. There are several types of non-linear sigmoid functions available. Out of that, tan sigmoid function is a good choice to speed up the normalization process. If the parameters to be estimated from noisy data the sigmoid normalization, method is used.

2.6. STATISTICAL COLUMN NORMALIZATION:

In the statistical column normalization, each data sample is normalized with a column normalization value. (Jayalakshmi & Santhakumaran, 2011). The normalization of each column can be done by normalizing the columns to a length of one. Each sample is computed by dividing the normalized column attribute and multiplied by a small bias value as shown in Equation (6).

$$x^{!} = \frac{x_i - n(c_a)}{n(c_a)} X 0.1 \tag{6}$$

Where

$n(c_a)$ is the normalized column attribute

$x^{/}$ is the Normalized Value of attribute x_i

2.7. MEAN AND STANDARD DEVIATION NORMALIZATION

Network inputs and targets can be scaled by normalizing the mean and standard deviation of the training set, such that inputs and targets will have zero mean and unity standard deviation (Abdi *et al.,* 2010). It can be calculated as shown in Equation (7).

$$y^{!} = (x_i - x_{\min}) X \frac{y_{std}}{x_{std}} + y_{\min} \tag{7}$$

Where

x_{min} is the lower bound of attribute input x_i

x_{std} is the standard deviation of attribute x_i

$y^{/}$ is the Normalized Value of attribute y

y_{min} is the lower bound of attribute input y

y_{std} is the standard deviation of attribute y

3 REVIEWED PAPERS

Wine datasets were collected from the UCI repository (Blake & Men, 2003) for the purpose of this review. Wine data set has a total of 178 patterns and 13 features grouped into three classes. Chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars, should be sufficient to recognize the source of the wine. The analysis determined 13 quantities, including alcohol content, hue, color intensity and content of 9 chemical compounds. The number of data samples from Classes 1, 2 and 3 is 59, 71 and 48, respectively.

In this work, thirty papers were reviewed that used the wine data set from the UCI repository with the aim of analysing the normalization techniques used and the proportion of performance accuracy.

4 ANALYSIS

Table 1 contains the summary of thirty published papers that used wine dataset available in the UCI repository. The table shows the technique used by each paper and the performance accuracy. As seen in Fig. 1, out of the thirty papers reviewed, twenty four used purely ANN technique (80%), one used ANN and SVM (7%), one used ANN and Fuzzy Logic (7%), one used Clustering (7%), one used Fuzzy Logic (7%), one used GA (7%), one used KLFANN and GA (7%). This clearly reveals that ANN is the most used technique among all the other ones.

TABLE 1: SUMMARY OF REVIEWED PAPERS ON WINE DATASET

S/N	Reference	Type of Normalization	Technique Used	Accuracy
1	Jiang, *et al.* (2004)	Z-Score	ANN	94.94%
2	Doherty, *et al.* (2007)	Min-Max Normalization	ANN	94.90%
3	Alpaydin (1997)	Z-Score	ANN	94.87%
4	Bilenko, *et al.* (2004)	Median	ANN	89.40%
5	Borgelt & Kruse (2003)	Z-Score	ANN and Fuzzy Logic	92.20%
6	Orsenigo & Vercellis (2009)	Min-Max Normalization	ANN	90.10%
7	Duch (2004)	Min-Max Normalization	ANN	96%
8	Yang, *et al.* (2011)	Z-Score	ANN	86.86%
9	Hsu & Lin (2002)	Min-Max Normalization	ANN	99.44%
10	Guvenir (1998)	Min-Max Normalization	ANN	95%
11	Raymer *et al.* (2003)	Z-Score	GA	98.90%
12	Viswanath, *et al.* (2006)	Z-Score	ANN	91.03%
13	Thimm & Fiesler (1997)	Sigmoidal	ANN	90%
14	Rodriguez (2009)	Median	ANN and SVM	93%
15	Eklund (2002)	Mean and Standard Deviation	ANN	91.09%
16	Calders *et al.* (2013)	Z-Score	ANN	94%
17	Prabhu & Anbazhaga (2011)	Z-Score	Clustering	92.13%
18	Deshpande & Karypis (2002)	Min-Max Normalization	ANN	95%
19	Domingos (1996)	Min-Max Normalization	Fuzzy Logic	96.90%
20	Ali & Pazzani (1996)	Mean and Standard Deviation	ANN	93.30%
21	Li *et al.* (2005)	Min-Max Normalization	ANN	96.23%
22	Ozgur C. (2014)	Z-Score	ANN	86%
23	Cortez, *et al.* (2009a)	Modified Min-Max Normalization	ANN	95%
24	Kraipeerapun, *et al.* (2006)	Min-Max Normalization	ANN	96.53%
25	Cortez, *et al.* (2009b)	Min-Max Normalization	ANN	96.20%
26	Sharma (2014)	Min-Max Normalization	ANN	99.20%
27	Fu, *et al.* (2012)	Min-Max Normalization	ANN	98.82%
28	Swain, *et al.* (2012)	Min-Max Normalization	ANN	96.66%
29	Xiang, *et al.* (2004)	Min-Max Normalization	KFLANN and GA	90.44%
30	Chittineni & Raveendra (2012)	Min-Max Normalization	ANN	97.19%

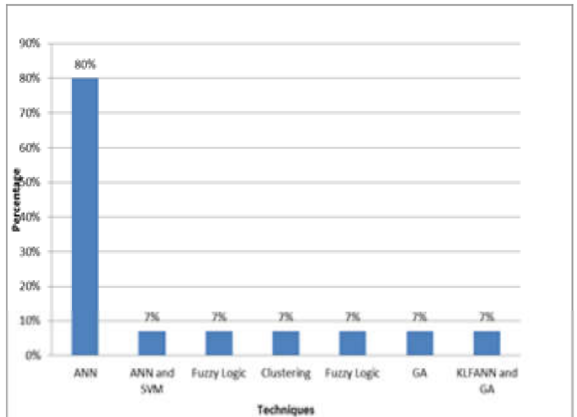


Fig. 1 Proportion of Techniques used in Reviewed Papers

Fig. 2 shows the percentages of the different normalization types used in the papers reviewed. Min-Max normalization has the highest percentage of 50% while modified min-max and sigmoid had the least percentage of 3% each.

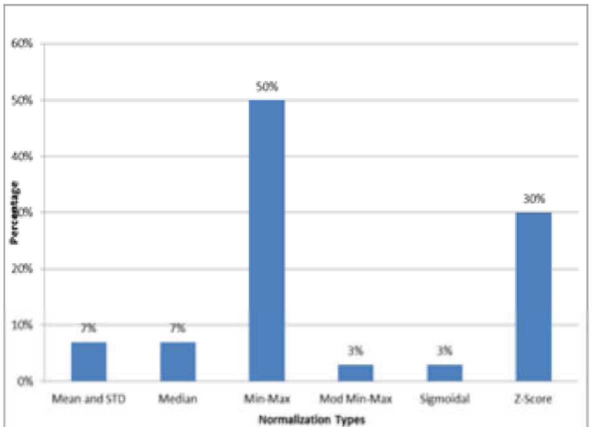


Fig. 2 Proportion of Normalization Types in the Reviewed Papers

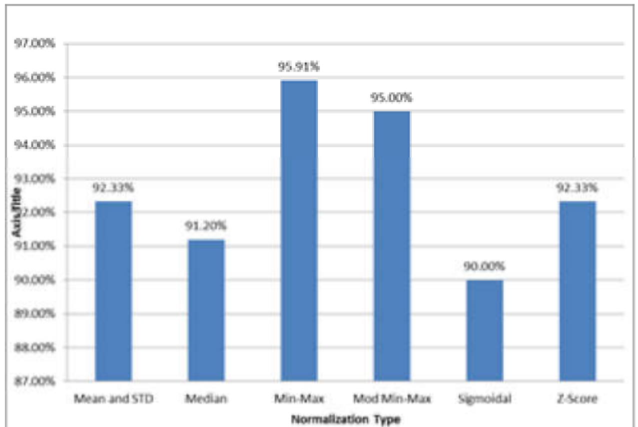


Fig. 4 Average Percentage of Normalization Types in the Reviewed Papers

Fig. 3 presents the results for average performance accuracy of the techniques used for the reviewed papers in the wine datasets. These techniques are classified into seven groups. Firstly investigation of the performance of the different techniques was done. From the figure it was observed that GA performs best with an average percentage accuracy of 98.90%. However, the highest percentage of accuracy was observed in the ANN with a value of 99.44%.

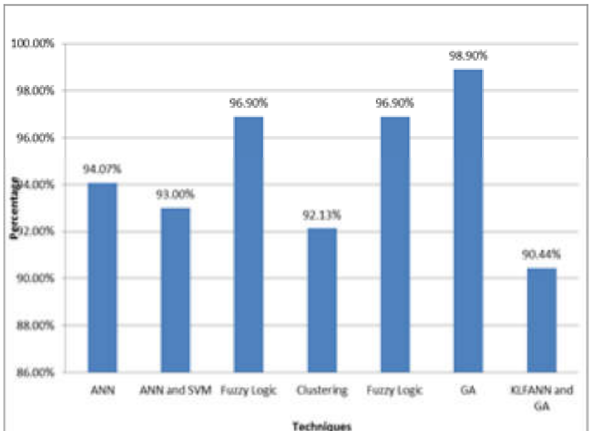


Fig. 3 Average Performance Accuracy of Techniques used in the Reviewed Papers

Fig. 4 shows the average performance accuracy for each type of normalization adopted in the papers reviewed. The Min-Max normalization has the highest average performance accuracy with a value of 95.91% and closely followed by modified Min-Max with a value of 95%. This clearly shows that the Min-Max normalization gives the best performance accuracy while the sigmoidal normalization gave the least value of 90%.

5 CONCLUSION

Appropriate choice of normalization technique is very important and impact greatly on performance accuracy. In this paper, six important data normalization techniques used in some published work were reviewed and evaluated. The effect of normalization techniques on performance accuracy for wine dataset as available in the UCI repository was studied. From the results, Min-Max normalization technique outperformed other normalization method. In conclusion, it is important to carefully select normalization method to avoid negative influence on system performance.

REFERENCE

Abdi M.N., Khemakhem M., & Ben-Abdallah H. (2010). Off-Line Text-Independent Arabic Writer Identification using Contour-Based Features. International Journal of Signal and Image Processing, *1*(1), 4-11.

Ali K. M., & Pazzani M. J. (1996). Error Reduction through Learning Multiple Descriptions. Machine Learning Journal, *24*(3), 173-202.

Alpaydin E. (1997). Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Revolution Journal, *11*(5), 115-132.

Bilenko M., Basu S., & Mooney R.J. (2004). Integrating Constraints and Metric Learning in Semi-Supervised Clustering, Proceedings of the 21st International Conference on Machine Learning (ICML), 81–88.

Blake C.L & Men C.J (2003). UCI Repository of Machine Learning Databases, http:llwww.ics.uci.edu/ mleamlMLRepository.hM. University of California, Imine, Dept. of Information and Computer Science.

Borgelt C., & Kruse R., Speeding. (2003). Fuzzy Clustering with Neural Network Techniques. Fuzzy Systems Journal, *2*(1), 852–856.

Calders T., Karim A., Kamiran F., Ali W., & Zhang X. (2013). Controlling Attribute Effect in Linear Regression. Proceedings of the 13th International Conference on Data Mining (ICDM), 71–80.

Chittineni S., Raveendra B. B.(2012). Determining Contribution of Features in Clustering Multi-Dimensional Data Using Neural Network, IJITCS, *4*(10).

Cortez P., Teixeira J., Cerdeira, A., Almeida F., Matos T., & Reis J. (2009a). Using Data Mining for Wine Quality Assessment. *5808*(1), 66–79.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T.& Reis, J. (2009b). Modeling Wine Preferences by Data Mining from Physicochemical Properties, Decision Support Systems, *47*(4), 547–553.

Deshpande M., & Karypis G. (2002). Using Conjunction of Attribute Values for Classification. Technical Report, Department of Computer Science and Engineering University of Minnesota, USA.

Doherty K. A. J., Adams, R. G., & Davey, N. (2007). Unsupervised Learning with Normalised Data and Non-Euclidean Norms. Applied Soft Computing Journal, *7*(1), 203-210.

Domingos P. (1996). Unifying Instance-Based and Rule-Based Induction. Machine Learning Journal, *24*(1), 141–168.

Duch W. (2004). Visualization of Hidden Node Activity in Neural Networks: Visualization Methods. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, 38-43.

Eklund P. W. (2002). A Performance Survey of Public Domain Supervised Machine Learning Algorithms. Technical Report, The University of Queensland Australia.

Fu J., Canqin H., Jianguo X., & Junbao Z. (2012) Pattern Classification Using an Olfactory Model with PCA Feature Selection in Electronic Noses: Study and Application, Journal of Sensors *12*(3), 2818-2830.

Guvenir H. A. (1998). A Classification Learning Algorithm Robust to Irrelevant Features. Proceedings of AIMSA '98, 281-290.

Hsu C., & Lin C. (2002). A Comparison Methods for Multi-Class Support Vector Machine. IEEE Transactions on Neural Networks, *13*(1), 415–425.

Jain A., Nandakumar K., & Ross A., (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*. *38*(12), 2270-2285.

Jayalakshmi T. and Santhakumaran A. (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, *3*(1), 89-93.

Jiang Y. & Zhou Z.H. (2004). Editing Training Data for kNN Classifiers with Neural Network Ensemble, Lecture Notes in Computer Sciience, *3173*(1), 356-361.

Kraipeerapun P., Fung C.C & Wong K.W. (2006). Multiclass classification using neural networks and interval neutrosophic sets, in Proceedings of the 5th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Venice, Italy. 123-128.

Li H., Qi F., Wang S., & Gervasi, O. (2005). A Comparison of Model Selection Methods for Multi-class Support Vector Machines, Springer-Verlag Berlin Heidelberg, 1140-1148.

Luai A., Zyad S., & Basel K. (2006). Data Mining: A Preprocessing Engine. Journal of Computer Science. *2*(9), 735-739.

Manikandan, G., Sairam, N., Sharmili S. and Venkatakrishnan S. (2013). Achieving privacy in data mining using normalization, Indian Journal of Science and Technology, *6*(1), 4268-4272.

Orsenigo C. & Vercellis C. (2009). Multicategory Classification via Discrete Support Vector Machines. Computational Management Science Journal, *6*(1), 101–114.

Ozgur C. (2014) Electrical and Electronics Engineering, Adana Science and Technology University, Adana.

Prabhu P., & Anbazhagan N. (2011). Improving the Performance of k-means Clustering for High Dimensional Data Set. International Journal on Computer Science and Engineering (IJCSE), *3*(6), 308.

Raymer M. L., Doom T. E., Kuhn L. A., & Punch W. F. (2003). Knowledge Discovery in Medical and Biological Datasets using a Hybrid Bayes Classifier/Evolutionary Algorithm, IEEE Transaction Systems, *33*(5), 802–813

Rodriguez M . (2009). M ulti-Class Boosting. Notes on AdaBoost Algorithms, Department of Computer Science, University of California, Santa Cruz.

Sharma L. (2014). Neural Network Based Classifier (Pattern recognition) for Classification of Iris Data Set. International Journal of Recent Development in Engineering and Technology, *3*(2).

Swain M., Sanjit K., Dash, Sweta D., & Ayeskanta M. (2012). An Approach for Iris Plant Classification using Neural Network, International Journal on Soft Computing, *3*(1).

Thimm G., & Fiesler E. (1997). High-Order and Multilayer Perceptron Initialization. IEEE Transactions on Neural Networks, *8*(1), 349–359.

Vaishali R. P., and Rupa G. M. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. International Journal of Computer Science *8*(1), 331-336.

Viswanath, P., Narasimba, M., & Bhatnagar, S. (2006). Partition Based Pattern Synthesis Technique with Efficient Algorithms for Nearest Neighbor Classification. Pattern Recognition Journal. *27*(14), 1714–1724.

Yang, Y., Liao, L., Meng, G., & Lee, J. (2011). A Hybrid Feature Selection Scheme for Unsupervised Learning and its Application in bearing Fault Diagnosis. Journal of Expert Systems with Applications, *38*(9), 11311–11320.

Xiang Y., Alex, T., & Leng P., (2004). Genetic Algorithm Based K-Means Fast Learning Artificial Neural Network, Nanyang Technological University.

.