Effect Z-score Normalization on Accuracy of classification of liver disease

Sondos Jameel Mukhyber[1],DhahirAbdulhadeAbdulah[2] , Amer D. Majeed[3]

[1,2]Department of Computer Science College of Science, University of Diyala
[3] College of Medicine, University of Diyala

Email [1] :scicompms11@uodiyala.edu.iq , Email [3] : amer.dmk58@gmail.com

Article History: *Do not touch during review process(xxxx)*

Abstract: Data normalization is one of the pre-processing approaches where the data is either scaled or transformed to make an equal contribution of each feature. The success of classification algorithms depends upon the quality of the data to obtain a generalized predictive model of the classification problem. The importance of data normalization for improving data quality. Therefore, this study aims to investigate the impact of z-score normalization method on accuracy of classification of liver diaseas. In this paper we apply z-score normalization on three classification algorithm Artificial Neural Network, Support Vector Machine, and K-nearest neighbour with two liver datasets. It has been observed from the results classification algorithms effected with z-score normalization.

Keywords: ANN, SVM, Knn, z- score normalization

1. Introduction

Many data mining projects are based on data sets collected for different purposes, ranging from routinely collected data to process improvement projects and data required for regulatory purposes. In some cases, a data set might be big and sufficient for extraction of knowledge. In other cases, the data set might be small and insufficient to extract meaningful knowledge and then accuracy will be low [1].

Data mining seeks to detect unrecognized associations between data items in an existing database. It is the process of extracting valid, previously unseen or unknown, comprehensible information from big databases. The growth of the size of data and number of existing databases exceeds the ability of humans to analyse this data, which creates both a need and an opportunity to extract knowledge from databases [2].

Medical databases have collected big quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data often concerned with the treatment of incomplete knowledge, with management of inconsistent pieces of information and the manipulation of different levels of representation of data [3].

Data transformation such as normalization may improve the accuracy and efficiency of mining algorithms that incloud distance measurements such as neural networks, nearest neighbor, and clustering classifier. Such methods provide best results if the data to be analysed have been normalized, that is, scaled to specific ranges such as [0.0, 1.0] [4].

2.LITERATURE REVIEW

Salama et al [5] proposed a paper for reducing the influence of normalization on data classification. The author in this paper explain that the data normalization is a pre-processing technique usually used before feature selection and classification. Complex real time pattern recognition systems use features that are generated by many various feature extraction algorithms with different kinds of sources. These features may have different dynamic ranges. Popular distance measures, for example the Euclidean distance, implicitly assign more weighting to features with big ranges than those with small ranges. Feature normalization is thus required to approximately equalize ranges of the features and make them have approximately the same effect in the computation of similarity.

Bikesh Kumar Singh et al [6] proposed a paper for investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. This paper investigates and evaluates some common feature normalization techniques and studies their impact on performance of classifier with application to breast tumor classification using ultrasound images. For evaluating the feature normalization techniques, back-propagation artificial neural network (BPANN) and support vector machine (SVM) classifier models are used. Results show that normalization of features has significant effect on the classification accuracy.

U. Rajendra Acharya et al [7] proposed a paper for automated diagnosis of glaucoma using texture and higher order spectra features. Computational decision support systems for the early detection of glaucoma can help prevent this complication. They presented a novel method for glaucoma detection using a combination of texture and higher order spectra (HOS) features from digital fundus images. Support vector machine, sequential minimal optimization, naïve Bayesian, and random-forest classifiers are used to perform supervised classification. The results explain that the texture and HOS features after z-score normalization and feature selection, and when combined with a random-forest classifier, performs best than the other classifiers and correctly identifies the glaucoma images with an accuracy of more than 91%. The impact of feature ranking and normalization is also studied to improve results. Our proposed novel features are clinically significant and can be employed to detect glaucoma accurately.

LeiliShahriyari [8] presented a research to explain the effect of normalization methods on the performance of supervised learning algorithms applied to HTSeqFPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. She compare three most common normalization methods: scaling, standardizing using z-score and vector normalization by visualizing the normalized data set and evaluating the performance of 12 supervised learning algorithms on the normalized data set. Additionally, for each of these normalization methods. Among all 12 learning algorithms and 6 different normalization techniques, the Bernoulli naïve Bayes model after standardizing files had the better performance in terms of maximizing the accuracy.

T.Jayalakshmi and Dr.A.Santhakumaran [9] proposed a research showed various normalization methods used in back propagation neural networks to enhance the reliability of the trained network. The experimental results showed that the performance of the diabetes data classification model using the neural networks was dependent on the normalization methods.

3.ROPOSED METHODOLGY

In this paper we propose a method for building predictive model for liver disease using various classification algorithms. The comparative analysis of the proposed method has been done and performance is measured using various classification metrics. The brief details of each steps involved for diseased prediction are described as follows:

A. Data Selection

Tow liver patient data sets used in this study:

- The Indian Liver Patient Dataset (ILPD) was selected from UCI Machine learning repository for this study [10]. It is a sample of the entire Indian population collected from Andhra Pradesh region. The dataset consist of 583 instances based on eleven different biological parameters. The Status value was reported based on these parameters as either Liver patient (416 cases) or not liver patient (167 cases) to represent the liver infection.
- The Iraqi liver patient dataset, we collected it from Baqubah Teaching Hospital. The dataset consist of 534 instances based on eleven different biological parameters. The Status value was diagnosed based on these parameters as either Liver patient (383 cases) or not liver patient (151 cases) to represent the liver infection.

B. Data Pre-processing

- **Handle miss value –** It refers to identifying missing values in the data and assigning the empty values with mean values. For Indian Liver Disease Patients data, Albumin and Globulin ratio has four missing values which is replaced by mean values. For Iraqi liver patient dataset, It does not contain any missing values.
- **Normalization –** is one of the data pre-processing techniques used in most of Data Mining System. An attribute of a dataset is normalized by scaling its values so that they fall within a smallspecified range, such as 0.0 to 1.0. There are various techniques of normalization are available such as z-score, max-min and decimal normalization. In this paper we used z-score normalization.

C. Training and testing

Train/Test is a method to measure the accuracy of an algorithm. It is called Train/Test because it split the data set into two sets: a training set and a testing set. In this paper we split the two data set 80% for training, and 20% for testing.

D. Classification using datamining techniques

•**SVM (Support Vector Machine) –** SVM have attracted a great deal of attention in the last decade and actively tested to various domains applications. SVMs are mostly used for learning classification, regression or ranking function. SVM are based on statistical learning theory and structural risk minimization principal and have the intent of determining the location of decision boundaries also known as hyperplane that produce the optimal separation of classes. SVM is the most robust and exact classification technique, there are many problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally costly, as solving quadratic programming methods require large matrix operations as well as time consuming numerical computations [11].

• **ANN (Artificial Neural Network) –** ANN is a classification model which is collected by interconnected nodes. It can be viewed as a circular node which is represented as an artificial neuron that reveals the output of one neuron to the input of other. The ANN model is useful in revealing the hidden relationships in the historical data, thus facilitating the prediction and forecasting of diseases of patients. ANN model is accurate enough to make significant and relevant decisions regarding data usage.

• **K-nearest neighbour –** Knn can be said as a classification, non-parametric algorithm which stores all available cases and its works is to classify new cases based on a similarity measure. It is non-parametric as it's doesn't make any assumption on the underlying data distribution. KNN uses Euclidean distance to predict the class. It works like this, a case is classified by majority vote of its neighbour, then the case is assigned to the class which is most common amongst its K nearest neighbour which is measured by measured by Euclidean distance.

E. Assessment Criteria

• Accuracy: This performance measure is calculated by performing ratio of total number of correctly diagnosed cases to the total number of cases.

Accuracy = (TP + TN) / (TP + FP + TN + FN) (1)

4. RESULTS

All the three-classification algorithm is been tested for the Iraqi liver patient dataset and the Indian liver patient dataset before apply z-score normalization, then the performance of classification algorithm as following:

Table 1. Accuracy comparison of classification algorithms before z-score normalization

S. No	The Iraqi liver patient dataset	The Indian Liver Patient Dataset
	Accuracy (%)	Accuracy (%)
ANN	74.76	76.06
SVM	80.87	80.34
Knn	70.09	72.64

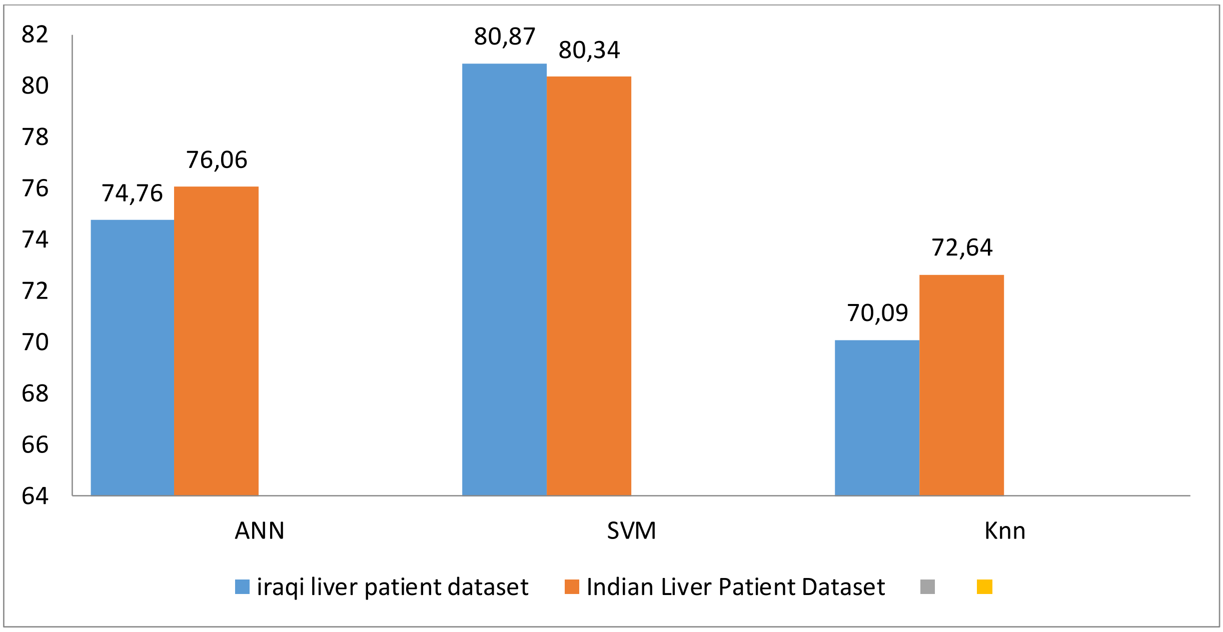

Fig.1: Graphical representation of accuracy comparison of algorithms for Iraqi liver patient dataset before apply z-score normalization

The classification algorithm above tested for the Iraqi liver patient dataset and the Indian liver patient dataset after apply z-score normalization, then accuracy of classification algorithm as following:

Table 2. Performance comparison of classification algorithms after z-score normalization

S. No	The Iraqi liver patient dataset	The Indian Liver Patient Dataset
	Accuracy (%)	Accuracy (%)
ANN	95.32	84.61
SVM	89.71	80.34
Knn	87.85	77.77

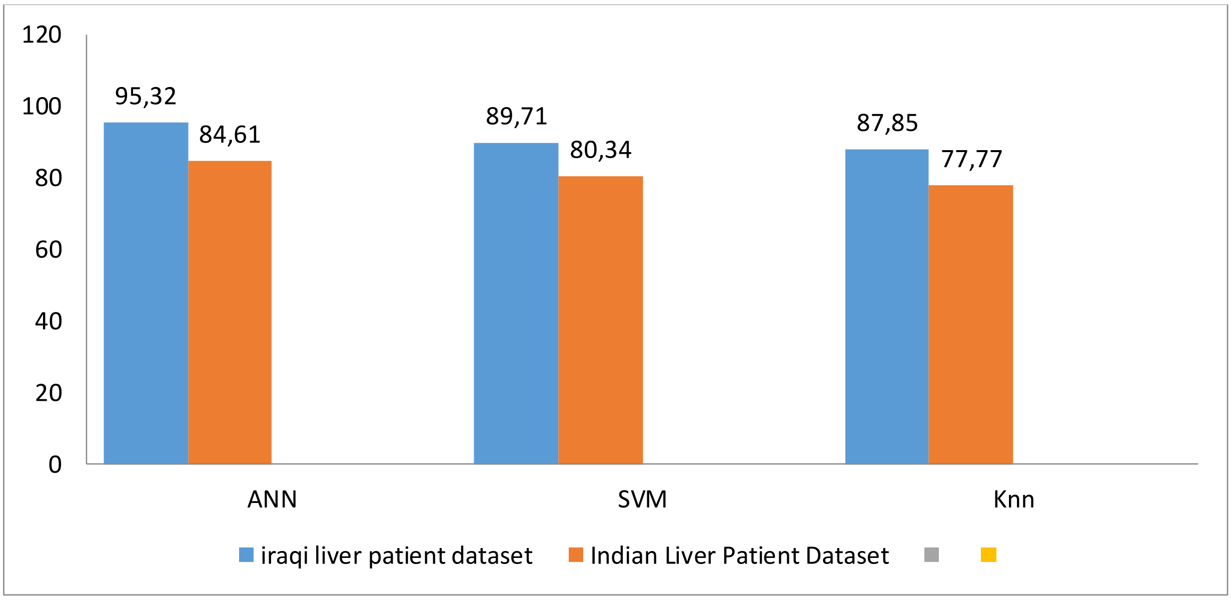

Fig.2: Graphical representation of accuracy comparison of algorithms for Iraqi liver patient dataset after apply z-score normalization

5. CONCLUSION

This paper investigates one of the data normalization approach for the improvement of classification accuracy. We have considered z-score normalization method from different research areas for the empirical analysis on two liver datasets. We apply three classification algorithms on two liver dataset before z- score normalization and compared these classification algorithms after apply z-score normalization . The results show that the algorithms effected with proposed normalization methodacrosstwo liver dataset and highest accuracy was obtained by ANN algorithm with 95.32% for Iraqi liver patient dataset and 84.61% for Indian liver patient dataset. SVM also effected and increased to 89.71% for Iraqi liver patient dataset and remained without effect for Indian liver patient dataset while knn increased to 87.85% for Iraqi liver patient dataset and 77.77% for Indian liver patient dataset.

REFERENCES

[1] Dr.Luai Al Shalabi, " Coding and Normalization: The Effect of Accuracy, Simplicity, and training time ",https://www.researchgate.net/publication/265300249,January 2006.

[2] K. Cios, W. Pedrycz, and R. Swiniarski, "Data Mining Methods for Knowledge Discovery", Kluwer Academic Publishers, 1998.

[3] Tapas Ranjan Baitharu and Subhendu Kumar Pani, "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security (CMS 2016).

[4] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, USA, 2001.

[5] Salama, M. A., Hassanien A. E. and Fahmy A. "Reducing the influence of normalization on data classification". In Proceedings of International Conference on Computer Information Systems and Industrial Management Applications, 2010, pp. 609 – 613.

[6] Bikesh Kumar Singh, Kesari Verma, and A. S. Thoke, "Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification", International Journal of Computer Applications (0975 – 8887), Volume 116 – No. 19, April 2015.

[7] U. Rajendra Acharya, Sumeet Dua, Xian Du, VinithaSree S, and Chua Kuang Chua, "Automated Diagnosis of Glaucoma Using Texture and Higher Order Spectra Features", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 15, NO. 3, MAY 2011.

[8] LeiliShahriyari, "Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeqFPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma" available on OXFOURD academic,*Briefings in Bioinformatics,* Volume 20, Issue 3, May 2019, Pages 985–994.

[9] T.Jayalakshmi, and Dr.A.Santhakumaran, "Statistical Normalization and Back Propagation for Classification", International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011.

[10] https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)#

[11] Lavesson . N, and Davidsson . P., "Generic Methods for Multi-Criteria Evaluation", in Proc. of the Siam Int. Conference on Data Mining, Atlanta, Georgia, USA: SIAM Press, 2008, pp. 541-546