International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms

M. Banu Priya[1], P. Laura Juliet[2], P.R. Tamilselvi[3]

[1]Research scholar, M.Phil. Computer Science, Vellalar College for Women, Erode12.
[2]Assistant Professor, Department of Computer Applications, Vellalar College for Women, Tamilnadu, India.
[3]Assistant Professor of Computer Science, Government Arts & Science College, Komarapalayam, India.

Abstract-*Data Mining is one of the most critical aspects of automated disease diagnosis and disease prediction. It involves data mining algorithms and techniques to analyze medical data. In recent years, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in several countries. In this thesis, liver patient datasets are investigate for building classification models in order to predict liver disease. This thesis implemented a feature model construction and comparative analysis for improving prediction accuracy of Indian liver patients in three phases. In first phase, min max normalization algorithm is applied on the original liver patient datasets collected from UCI repository. In liver dataset prediction second phase, by the use of PSO feature selection, subset (data) of liver patient dataset from whole normalized liver patient datasets is obtained which comprises only significant attributes. Third phase, classification algorithms are applied on the data set. In the fourth phase, the accuracy will be calculated using root mean Square value, root mean error value.J48 algorithm is considered as the better performance algorithm after applying PSO feature selection. Finally, the evaluation is done based on accuracy values. Thus outputs shows from proposed classification implementations indicate that J48 algorithm performances all other classification algorithm with the help of feature selection with an accuracy of 95.04%.*

Key Words: Classification, SVM, MLP, Bayesian Network, J48, Random Forest.

1. INTRODUCTION

The liver is a large, meaty organ that sits on the right side of the belly. Weighing about 3 pounds, the liver is reddish-brown in color and feels rubbery to the feel. The liver has two large sections, called the right and the left lobes. The gallbladder sits below the liver, along with parts of the pancreas and intestines. The liver and these organs behavior together to digest, absorb, and process food. The liver's main job is to strain the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver hides bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions [1].

Liver disease is any trouble of liver function that causes sickness. The liver is responsible for many dangerous functions within the body and should it become diseased or damaged, the loss of those functions can cause significant injury to the body. Liver disease is also referred to as hepatic disease. Liver disease is a large term that covers all the potential problems that cause the liver to fail to perform its designated functions. Usually, more than 75% or three quarters of liver tissue needs to be affected before a decrease in function occurs [1].

DISEASE OF LIVER

Several diseases states can disturb the liver. Some of the diseases are Wilson's disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have on the whole detrimental effects if alcohol is taken over long periods of time. Hemochromatosis can cause liver problems [1].

Common Liver Disorder

- **Fatty liver** is a revocable condition where large vacuoles of triglyceride fat acquire in liver cells via the process of limit. It can occur in people with a high level of alcohol consumption as well as in people who never had alcohol.
- **Hepatitis** (usually caused by a virus spread by excess contamination or direct contact with infected body fluids).
- **Cirrhosis** of the liver is one of the most serious liver diseases. It is an action used to indicate all forms of diseases of the liver characterized by the significant loss of cells. The liver gradually contracts in size and becomes leathery and hard. The regenerative action continues under liver cirrhosis but the progressive loss of liver cells exceeds cell replacement.
- **Liver cancer.** The risk of liver cancer is higher in those who have cirrhosis or who had valid types of viral hepatitis; but more often, the liver is the site of secondary (metastatic) cancers spread from other organs.

Data Description: Databases of 583 records/entries are taken from the ILPD (Indian Liver Patient Dataset) Dataset for the purpose of solving problem of this paper. This dataset is downloaded from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Entire ILPD dataset contains information about 583 Indian liver patients. In which 416 are liver patient records and 167 non liver patient records .The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not.

2. LITERATURE REVIEW

P.Rajeswari,G.Sophia Reena et al.,[2010]has proposed the data classification is based on liver disorder. The training dataset is developed by collecting data from UCI repository consists of 345 instances with 7 different attributes. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms .FT tree algorithms, and KStar algorithms and on the whole performance made know FT Tree algorithm when tested on liver disease datasets, time taken to run the data for result is fast when compare to other algorithm with accuracy of 97.10%Based on the experimental results the classification accuracy is found to be better using FT Tree algorithm compare to other algorithms [2].

Sa'diyah Noor Novita Alfisahrin, Teddy Mantoro et al., [2013] have proposed to identify if the patients have the liver disease based on the 10 important attributes of liver disease using a Decision Tree, Naive Bayes , and NB Tree algorithms. The result shows NB Tree algorithm has the highest accuracy; however the Naïve Bayes algorithm gives the fastest computation time.For future study, the performance of NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients.For future study, the performance of NB Tree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients [3].

S.Dhamodharan [2014] has proposed there are many liver disorders that require clinical care of the physician. They predict three major liver diseases such as liver cancer, cirrhosis, hepatitis with the help of distinct symptoms. The primary goal is to predict the class types from classes such as liver cancer, cirrhosis, hepatitis and "no diseases". In this paper Naïve Bayes and FT tree algorithm accuracy are compared and the result is obtained. The result concludes that the accuracy of Naïve Bayes algorithm is much better than the other algorithms. [4].

S. E. Seker, Y. Unal, Z. Erdem, and H. Erdinc Kocer et al.,[2014] has proposed have applied the data mining techniques, such as KNN, SVM, MLP or decision trees over a unique dataset, which is collected from 16,380 analysis results for a year. This study can be useful for further studies like reducing the number of analysis, since the prediction can be correlated and furthermore the correlation can be utilized for detecting the anomaly on the analysis [5].

A.S.AnneshKumar,Dr.C.JothiVenkateswaran et al.,[2015]has proposed describes the categorization of liver disorder through feature selection and fuzzy K-means classification. Various liver disorders also share same attribute values and it needs more effort to classify liver disorder type correctly with basic attributes. So Fuzzy based classification gives better performance in these confusing classes and achieved above 94 percentage accuracy for each type of liver disorder [6].

P.Thangarajul, R.Mehala et al., [2015] has proposed to analyze the data of liver diseases using particle swarm optimization algorithm (PSO) with K Star classification. In two aspect for classifying the existence of disease are not.The proposed algorithm enhanced the performance of accuracy when compared to existing classification algorithms. PSO-Kstar algorithm is best suitable algorithm for the classification of liver disorders as it improved the performance in prediction accuracy as discussed earlier .PSO-KStar algorithm is considered is one of the good data mining algorithm with respect to understandability, transformability and accuracy gives 100% [7].

Onwodi Gregory [2015] has proposed two real liver patient datasets were investigated for building classification models in order to predict liver diagnosis. Eleven data mining classification algorithms were applied to the datasets and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. Based on the experimental results the classification accuracy is found to be better using FT Tree algorithm compare to other algorithms., it also shows the enhanced performance according to the attributes and it gives 78.0% of Accuracy, 77.5% of Precision, 86.4% of Sensitivity and 38.2% of Specificity results respectively[8].

Dr.S.Vijayarani, Mr.S.Dhayanand et al., [2015] has proposed description of this research work is to predict liver diseases using classification algorithms. The algorithms used in this work are Naïve Bayes and support vector machine (SVM Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time. From the results, this work concludes the SVM classifier is considered as a best classification algorithm because of its highest classification accuracy values. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time from the implementation results it is observed that the SVM is a better Classifier for predict the liver diseases and comparing the execution time, the Naïve Bayes classifier needs minimum execution time [9].

Ebenezer Obaloluwa Olaniyi, Khashman Aadnan et al. [2015] has proposed back propagation neural network and radial basis function neural network are designed to diagnose these

diseases and also prevent misdiagnosis of the liver disorder patients. The algorithms were compared with the c4.5, CART, Naïve Bayes, Support Vector Machine (SVM) and concluded that the radial basis function neural network is the optimal model because it has a recognition rate of 70% which has proved more accurate and efficient than the other algorithms. [10].

Tapas RanjanBaitharua, Subhendu Kumar Panib et al., [2016] has proposed focuses on the aspect of Medical diagnosis by learning pattern through the collected data of Liver disorder to develop intelligent medical decision support systems to help the physicians. In this paper the use of several classification (J.48, SVM, Random Forest, etc) algorithms to classify these diseases and compare the effectiveness, correction rate among them. In this paper, a comparative analysis of data classification accuracy using Liver disorder data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.By analyzing the results Multilayer perceptron gives the overall best classification result with the accuracy 71.59% than other classifiers [11].

Anju Gulia, Dr. Rajan Vohra, Praveen Rani et al., [2014] has proposed to implements hybrid model construction and comparative analysis for improving prediction accuracy of liver patients in three phases. In first phase, classification algorithms are applied on the original liver patient datasets collected from UCI repository. In second phase, by the use of feature selection, a subset (data) of liver patient from whole liver patient datasets is obtained which comprises only significant attributes and then applying selected classification algorithms on obtained, significant subset of attributes. SVM algorithm is considered as the better performance algorithm, because it gives higher accuracy in respective to other classification algorithms before applying feature selection. But, Random Forest algorithm is considered as the better performance algorithm after applying feature selection. In third phase, the results of classification algorithms with and without feature selection are compared with each other. The results obtained from our experiments indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection with an accuracy of 71.8696% [12].

3. SYSTEM METHODOLOGY

NORMALIZATION

Normalization is the process of classify data into an associated table it also eliminates redundancy and increases the reliability which improves output of the query. To normalize a database, we divide the dataset into tables and establish relationships between the tables. Dataset normalization can essentially be defined as the practice of optimizing table structures. Optimization is accomplished as a result of a thorough investigation of the various pieces of data that will be stored within the database, in particular concentrating upon how this data is interrelated.

FEATURE SELECTION

PSO feature extraction model for liver dataset and applied an improve probability in many medical application such as training artificial neural networks, linear constrained function optimization, wireless network optimization, data classification, and many other areas where GA can be applied. Computation in PSO is based on a swarm of processing elements called particles in which each particle represent a candidate solution. [13].

RANDOM FOREST (RF)

Random forests is a machine learning regression method for classification that drive by constructing liver data into a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees [14]. It is unexcelled in accuracy among current algorithms. It output classification efficiently on large liver dataset. It can handle thousands of input attributers without variable deletion. It gives estimates of what variables are important in the classification. Random Forests grows many classification trees. To classify a new liver object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and says the tree "votes" for that class. The forest chooses the classification having the most votes).

SUPPORT VECTOR MACHINE (SVM)

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 [15].

J-48

J-48 classification is an algorithm used to generate a decision tree developed by Ross Quinlan. J-48 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J-48 can be used for classification, and for this reason, C4.5 is often referred to as a classifier. It induces decision trees and rules from liver datasets, which could contain categorical and numerical attributes. The rules could be used to predict categorical values of attributes from new liver records.

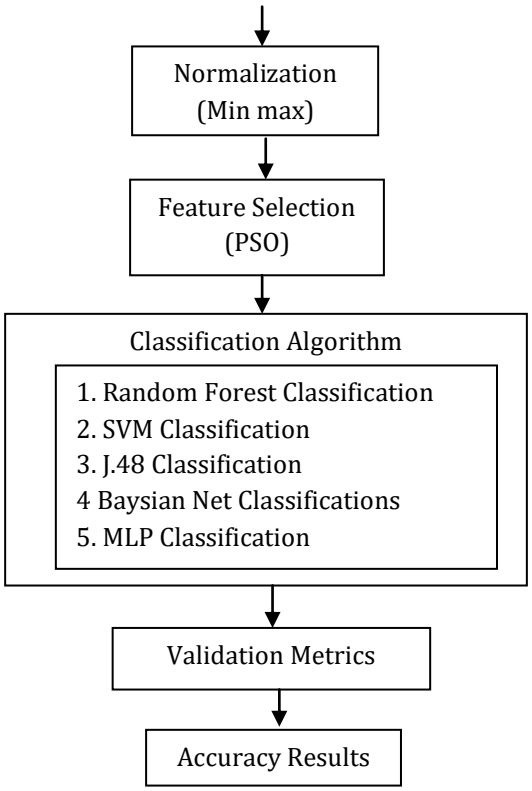
MLP (Multilayer Perceptron)

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps liver datasets of input data onto a set of appropriate outputs. An MLP classification is a multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP classification liver dataset utilizes a supervised learning technique called back propagation for training liver the network. MLP is a change of the standard linear perceptron and can distinguish data that are not linearly separable liver dataset process.

BAYESIAN NETWORKS

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several choices for data modeling. Two, a Bayesian network Classification can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data.

SYSTEM ARCHITECTURE



4. RESULTS AND DISCUSSION

PERFORMANCESMETRICS ANALYSIS

The following table 4.1 and Fig 4.1 describe a Mean absolute error analysis for PSO feature extraction model. In this table contains mean absolute Error details are shows,

Table 4.1 Mean Absolute Error

Classification Algorithm	Mean Absolute Error
J.48	0.507
MLP	0.703
SVM	0.712
Random Forest	0.604
Bayesnet	0.572

Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$MAE = 1/n \sum |y_j - y^_j|$$

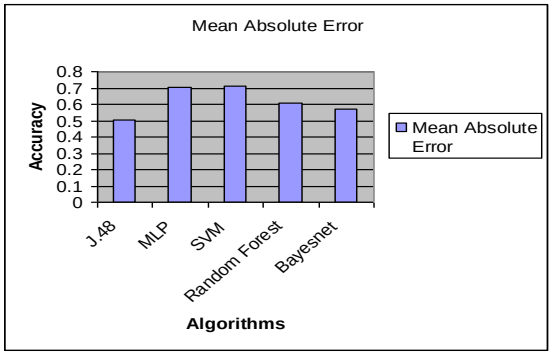


Fig 4.1 Mean Absolute Error Rate

The following table 4.2 and Fig 4.2 describe a Root Mean Square Error (RMSE) analysis for PSO feature extraction model. In this table contains Root Mean Square Error details are shows,

Table 4.2 Root Mean Square Error

Classification Algorithm	Root Mean Square Error
J.48	0.487
MLP	0.403
SVM	0.425
Random Forest	0.467
Bayesnet	0.406

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far liver dataset from the regression line data points.

RMSE is a measure of spread out these residuals is liver dataset.

$$RMSE = \sqrt{(f-o)^2}$$

f = forecasts (expected values or unknown results), o = observed values (known results).

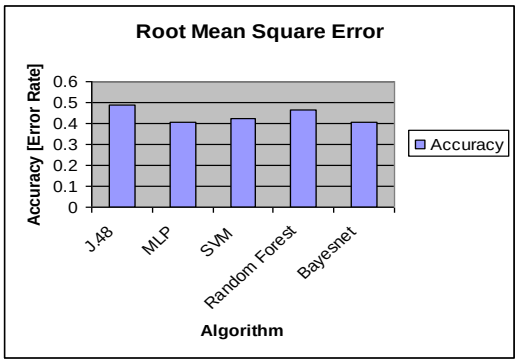


Fig 4.2 Root Mean Square Error

The following table 4.3 and Fig 4.3 describe a relative squared error analysis for PSO feature extraction model. In this table contains relative squared error details are shows,

Table 4.3 Relative Squared Error

Classification Algorithm	Relative Squared Error
J.48	73.33
MLP	69.23
SVM	71.45
Random Forest	68.44
Bayesnet	74.25

The **root relative squared error** is relative to what it would have been if a simple liver disease predictor and just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

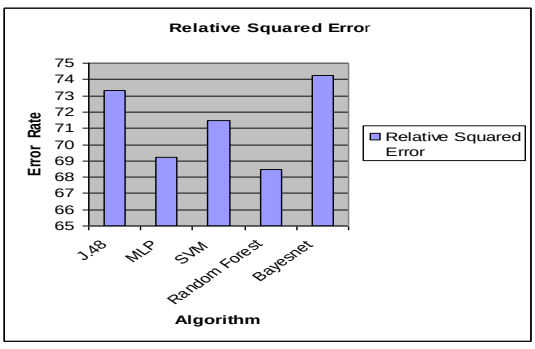


Fig 4.3 Relative Squared Error

Mathematically, the root relative squared error T_i of an individual program j is evaluated by the equation. Where $P_{(ij)}$ is the value predicted by the individual program *i* for sample case *j* (out of *n* sample cases); T_j is the target value for sample case *j*; and $\overline{T}$ is given by the formula:

$$RRSE = Tj = 1/n \sum_{j-1} T_j$$

The following table 4.4 describes an overall classification algorithm for accuracy values analysis. In this table contains existing and proposed accuracy a values shows,

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

Table 4.4 Accuracy

Classification Algorithm	Greedy Step Wise	PSO
J.48	68.77	95.04
MLP	68.26	77.54
SVM	71.35	73.44
Random Forest	70.32	80.22
Bayesnet	67.23	90.33

The following Fig4.4 describes an overall classification algorithm for accuracy values analysis. In this fig contains existing and proposed accuracy a values shows,

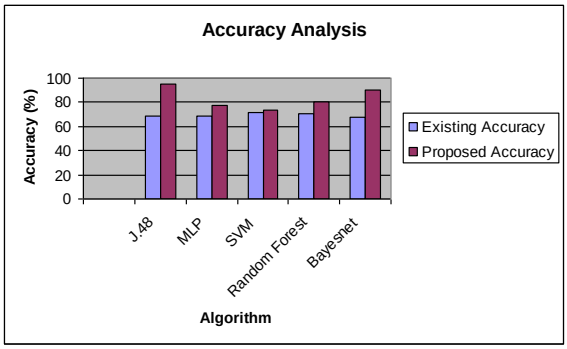


Fig 4.4 Accuracy

5. CONCLUSION AND FUTURE WORK

In this thesis the proposed system concludes that PSO feature selection methods for Indian Liver Patient Dataset. This thesis analyzed the liver disease using algorithms such as J48, MLP, SVM, Random Forest, and Bayesnet Classification. These algorithm gives various result based on PSO feature selection model .It has been seen that bayes net and J48 Classification gives better results compare to other classification algorithms.

There are many criterions for evaluating the selected feature subset, here this thesis used features such as Total bilirubin, Direct_ bilirubin, Total_protiens, Albumin, A/G ratio, SGPT, SGOT, Alkphos to evaluate the performance of different classification algorithm. In future, we have attempted to

classify different feature selection algorithms into four groups: complete search, heuristic search, meta-heuristic methods and methods that use artificial neural network.

PSO has been widely used for feature selection to improve liver classification performance. Further, a lot of work is being done using multi-objective PSO for feature selection to improve liver classification performance and to reduce number of features selected as well. Most of the existing multi-objective feature selection based on PSO algorithms use binary tournament selection to select gbest and uniform and non-uniform mutation. There is a scope to further reduce search space for better liver classification accuracy if enhanced selection and mutation procedures are being used.

The future methodology is used to analyze the liver region into separable compartments i.e. liver etc. However, the method requires further improvement mostly regarding feature selection of the liver into multiple components: renal cortex, renal column, renal medulla and renal pelvis. Apart from that, it is planned to expand the database on which the system will be tested. And also the proposed method in this thesis can be employed for detecting the heart diseases in future with the heart dataset and classification of the diseases.

REFERENCES

[1] https://www.medicinenet.com/liver_disease/article.htm

[2] P. Rajeswari ,G. Sophia Reena , Analysis of Liver Disorder Using Data Mining Algorithm,Global Journal of Computer Science and Technology,2010.

[3] Sa'sdiyah Noor Novita Alfisahrin,Teddy Mantoro, Data mining Techniques For Optimatization of Liver Disease Classification,International conference on advanced Computer Science Application and Technologies,2013.

[4] S. Dhamodharan , Liver Disease Prediction Using Bayesian Classification , National Confrence on Advanced Computing,Application&Technologies,2014

[5] S.E.Sekar ,Y.Unal, Z.Erdem,and H.Erdinc Kocer,Ensembled Correlation Between Liver Analysis Output, International Journal of Biology and Biomedical ngineering,ISSN:1998-4150

[6] A.S.Aneesh kumar,Dr.C.Jothi Venkateswaran , A novel approach for Liver disorder Classification using Data Mining Techniques ,Engineering and Scientific International Journal ,ISSN 2394-7179,ISSN 2394-7187,2015.

[7] P. Thangarajul,R.Mehala, Performance Analysis of PSO-KStar Classifier over Liver Diseases, International Journal of Advanced Research in Computer Engineering, 2015.

[8] Onwodi Gregory, Prediction of Liver Disease (Biliary Cirrhosis) Using Data Mining Technique, International Journal of EmergingTechnology&Research, ISSN (E):2347-5900, ISSN (P):2347-6079, 2015.

[9] Dr.S.Vijayarani,Mr.S.Dhayanand,Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research(IJSETR), 2015.

[10] Ebenezer Obaloluwa Olaniyi khashman Aadnan, "Liver DiseaseDiagnosisBasedon Neural Networks" , Advances in Computational Intelligence, Proceedings of the 16th International Conference on Neural Networks (NN '15), November 7-9, 2015.

[11] Tapas RanjanBaitharua, Subhendu Kumar Panib, "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset", International Conference on Computational Modeling and Security, 2016.

[12] Anju Gulia, Dr. Rajan Vohra, Praveen Rani, "Liver Patient Classification UsingIntelligent Techniques", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5110-5115.

[13] Dash M., Liu H., "Feature Selection for Classification," Intelligent Data Analysis, Elsevier, pp. 131 -156, 1997.

[14] Breiman L, Random Forests, Machine Learning, 45, 5-32, (2001)

[15] 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