British Journal of Mathematics & Computer Science
8(5): 377-394, 2015, Article no.BJMCS.2015.171
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

A Global Indicator for Measuring the Efficiency of Machine Learning Classifier Based on Multi-Criteria Approach

Hegazy Zaher[1*] and Mohamed Abdullah[2]

[1]Department of Mathematical Statistics, Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt.
[2]Department of Operations Research, Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt.

Article Information

DOI: 10.9734/BJMCS/2015/16358
Editor(s):
(1) Anonymous.
(2) Tian-Xiao He, Department of Mathematics and Computer Science, Illinois Wesleyan University, USA.
Reviewers:
(1) Hiram Ponce, Faculty Engineering, Universidad Panamericana, Mexico.
(2) Anonymous, India.
(3) Scheila de Avila e Silva, Universidade de Caxias do Sul, Brazil.
(4) Li, Xing, Department of Health Sciences Research, Mayo Clinic College, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=1034&id=6&aid=8841

Original Research Article

Received: 28 January 2015
Accepted: 08 April 2015
Published: 15 April 2015

Abstract

The main challenge that faces any researcher in the field of machine learning is determining the quality of an indicator used for measuring the efficiency of classifier techniques. This issue based on Multiple-Criteria Decision Making (MCDM) has not been tackled by any researcher until now. The previous work concerned with a single classical criterion (Accuracy Level) ignoring other important criteria in real-life. This paper presents a novel indicator for measuring the efficiency of classifier techniques. This measure is a global indicator with multi-criteria approach based on the technique for preference by similarity to the ideal solution (TOPSIS). This indicator is characterized by its ability to taking in account all previous criteria. In addition, two novel criteria are created by authors: Learning Efficiency Ratio (LER), and the CPU time efficiency. The classifiers evaluation process includes the classical classifiers: Support Vector Machines (SVM), Multi-layer perceptron (MLP), Gene Expression Programming (GEP), Single Decision Tree (STR), and the techniques that achieved the best results in literature. Inaddition, the latest classifiers: Tropical Collective Machine Learning (TCML), and Dempster-Shafer Collective Machine Learning (DSCML) using the proposed indicator. The comparison is performed using twenty-five standard datasets (benchmarks). The results supported by statistical analysis (T-test) show the efficiency and effectiveness of the proposed global

**Corresponding author: Hgsabry@yahoo.com;*

indicator for selecting the best classifier and its ability to measure the classifier efficiency based on multi-criteria. Results promise the optimistic use of the global indicator in the classifiers evaluation process for real-life problems.

Keywords: Social machine learning; multi-criteria; TOPSIS; generalization ability; classifier evaluation; global indicator.

1 Introduction

The classification process is considered one of the most important issues in the field of machine learning. This type of learning is categorized as supervised learning [1]. The traditional classification process depends on a single criterion for a classifier evaluation until now. This trend continued for many decades without any development. The existence of a global indicator used for measuring the efficiency of classifier techniques is a significant challenge. The literature review has many attempts to create an indicator used for measuring the efficiency of classifier techniques. The classifier evaluation is categorized into three categories. First category includes the comparison within the same technique for example, comparing the multi-layer perceptron with different parameters as number of hidden of layers. Second category includes the comparison among the same family of techniques for example, the multi-layer perceptron and the back-propagation network. Third category includes the comparison of different approaches as comparing the methods based on biological or mathematical sound.

The core idea behind this paper is introducing a novel global indicator with multi-criteria approach used for evaluating the efficiency of machine learning classifiers based on classical criteria and two proposed criteria by the authors. The two novels criteria: Learning Efficiency Rate (LER) and CPU time are very important in the practical life. No doubt, the learning of classifier process has three primary goals. The first goal is classifying the data with maximum accuracy. The second goal is avoiding overfitting for the classifier. The third goal is optimizing the computer resources utilization. Previous work was directed to find metrics that measure the first goal only, however there is a clear shortage of metrics that measure second, and third goal. This shortage is the primary motivator for authors to propose a global indicator based on three metrics cover the three goals of learning process. The remainder of this paper is structured as follows: Section two presents the related work that concerns with single criterion used for classifier evaluation. Section three gives a brief review of TOPSIS. Section four introduces the proposed indicator based on multi-criteria. Section five displays twenty-five datasets (benchmarks), experiments and results. Section six presents a discussion. Finally, section seven summarizes the conclusion.

2 Related Work

The classifier evaluation is a difficult issue that attracts many researchers to tackle this issue. The previous work is categorized into two major classes: the binary classification, and the multi-classifications. Each class is categorized into sub-categories. The binary classification is categorized into three sub-categories. First of them is based on the Receiver Operator Characteristic (ROC) curves. Provost et al. prove that using the accuracy as a single indicator gives misleading for the results of the classification process [2]. They suggest using the Receiver Operator Characteristic (ROC) curves in particular the large skewed data. Receiver Operator Characteristic (ROC) curve is one of the common measures used to introduce the results of binary classification. The main strength of this measure is the ability of visualization the trade-off between the two types of error I, II. In addition, allowing to the modeler to choose the suitable threshold value. However, the main weakness of ROC that can deal with binary classification only. In particular, data sets must be not highly skewed. In this case the Precision-Recall displays a more comprehensive picture for evaluation the classifier technique. Second of them is based on the alternatives of the Receiver Operator Characteristic (ROC**)** curves. Drummond and Holte suggest

evaluating the classifier using the cost curves [3,4]. This measure is considered as an alternative to ROC curves. Later, George Forman suggests a new evaluation method called "Bi-Normal Separation "[5]. He introduces a comparison for various methods of feature selection applied on benchmarks. The results achieved prove a reasonable improvement but it fail in the criteria of precision. Thirdly, the sub-category based on the Area under ROC curve (AUC). Fawcett Presents the Area under ROC curve (AUC). This criterion based on measuring the area under the curve of Receive Operating Characteristic (ROC) [6]. Also, can be named in the literature "C statistic". Where the classifier has the closer the value to 1.0 is the better classifier. This trend attracts many researchers to focus their effort for tackling this problem. Ferri et al. propose using the criterion of (AUC-ROC) for splitting process in the decision tree [7]. Cortes and Mohriprove that (AUC-ROC) is the most appropriate criterion used in the evaluating the boosting algorithm rank [8]. Later, Joachims introduces a novel method for maximizing the generalization ability of support vector machines based on (AUC-ROC) criterion [9]. Parti and flash present an algorithm called "Rule selection" to create the convex hull in the space of ROC [10]. Herschtal and Raskutti present a new technique used for optimizing the AUC-ROC within the domain of neural networks [11]. Also, Sirinivasan presents an algorithm called ILP. This algorithm is used as heuristic based on the ROC criterion [12].

The second major category of multiple classes is categorized into two sub-criteria. First of them is based on the structured loss minimization approaches. Petterson focuses on maximizing the F-measure through the phase of training in the support vector machines (SVM), and approaches of Decision-theoretic [13]. The second sub-criterion is based on F-measure. Quevedo et al. present a technique based on F-measure that maximizes the process of classifier evaluation. Confusion matrix and two types of errors defined as a matrix model shows a forecasted class to each case in test of dataset, where the actual class is known [14]. The strength of confusion matrix is introducing a simple way for comparing the frequencies of the actual class versus the predicted class. However the main weakness of the confusion matrix is inability to express the classifier evaluating in unique measure. The summary of previous work (traditional) as follows: The accuracy of the classifier that can be measured by different Criteria: specifity, recall…etc. The classical metrics are derived on single criterion. There is a clear shortage of global indicator based on multi-criteria as the speed of the classifier learning, *the generalization* ability for unseen data.

3 TOPSIS

Hwang and Yoon present the technique of preference by similarity to the ideal solution (TOPSIS) [15]. Topsis is one of the most recent techniques used in the multi-criteria to can identify the best solution from a finite set of solution alternatives. This technique ranks the alternatives based on minimizing simultaneously the Euclidean distance from the point called ideal point, and maximizing Euclidian distance from the point called nadir point. The ideal point is defined as an alternative has a maximum value according to all criteria considered. The nadir point is defined as an alternative has the worst value for all criteria considered [16]. Practically, TOPSIS has been proved its effectiveness and efficiency to tackle a large amount of problems from finite alternatives [17]. The superiority of TOPSIS in Multi-Criteria Decision Making field (MCDM) is result of the intuitive and simplicity for understanding and implementing. In addition, the benefits are generated from introducing human choice based on rationality [15].

3.1 Performance Measures

The confusion matrix is used as primary key that can derive by it the performance measures listed in Table 1, each element is explained in Table 2.

Table 1. Shows the confusion matrix

Actual class	Predicted class	
	True	False
True	TP	FN
False	FP	TN

Table 2. Shows the current criteria used for measuring the efficiency of the classifier

Criterion	Description	Formula
Accuracy	Defined as a weighted arithmetic mean of precision and inverse precision	(TP+TN) / (TP+TN+FP+FN)
Sensitivity	Defined as the percentage of actual positives that is correctly identified.	TP/ (TP+FN)
Specificity	Defined as the percentage of negatives that is correctly identified.	TN/ (TN+FP)
PPV	Defined as the percentage of true positive to total positive instants.	TP / (TP+FP)
NPV	Defined as the percentage of true negative to total negative instants.	TN/ (TN+FN)
Precision	Defined as the percentage of retrieved instances that are relevant.	TP/ (TP+FP)
Recall	Defined as the percentage of retrieved instances that are retrieved.	TP / (TP+FN)
F- measure	Defined as the harmonic mean of recall and precision	2 (Precision. Recall)/ (Precision+ Recall)

Such True positives (TP) are defined as the number of correctly labeled of a positive class to total positive class. False positives (FP) are defined as the number of incorrectly labeled of a positive class to total positive class. True negatives (TN) are defined as the number of correctly labeled of a negative class to total negative class. False negatives (FN) are defined as of incorrectly labeled of a negative class to total negative class. It is noticed that the sensitivity equals recall, and the precision equals PPV. The global indicator uses one from the two criteria. Both Recall and PPV will be excluded from the indicators list.

4 The Proposed Global Indicator

The existence of global indicator with multi-criteria is a significant challenge that faces any researcher in the field of machine learning. This challenge is the primary motivator for this paper. The global indicator is composed of three major criteria: the first criterion is classical metrics accuracy that used in the literature review of machine learning. This metric is composed of weighted mean for sub-criteria listed in the Table 2. These weights of sub-criteria are calculated by multi-criteria approach (TOPSIS) based on two dimensions: VC-dimension, and information gain.VC-Dimension is defined as the VC-dimension of a function set F (VCdim (F)) is the cardinality of the largest dataset that can be shattered by F [18]. Where the information gain is defined as reduction in entropy caused by partitioning the set of examples *S* for attribute *A*. This measure is calculated by the following equation:

$$Gain(S,A) \equiv Entropy(S) - \sum_{V \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_V) \quad (1)$$

Where *Values (A)* is the set of all possible values for attribute *A* and (*Sv)* is the subset of *S* for which attribute *A* has value *v* [19].

Where learning of classifier process has three primary goals: classifying the data with maximum accuracy, avoiding overfitting for the classifier which means the efficiency of learning not memorizing, and maximizing the computer resources utilization [20]. The great risk that faces the modeler is overfitting problem. Over fitting is defined as performing very well on the training data however fails to generalize well to unseen data [21,22]. This problem despite its importance is not

considered until now. This problem appears in the medical diagnosis such fail the classifier for determining the patient has cancer or not. Also, extends to image processing and military process. What is the importance of any classifier perform well on the training data where fails to classify the unseen data. It is clear when the overfitting indicator increases this means a shortage of learning process efficiency. From that, authors create an indicator measures the efficiency of learning process. This paper proposes a new indicator for avoiding overfitting or called Learning Efficiency Ratio.

$$LER = \frac{\text{classification error of testing process}}{\text{classification error of testing process)}} \times 100 \quad (2)$$

Computer resources utilization is measured by CPU time for learning process. There are many real-life problems require the speed of classification process as military defense for terrorist attacks which require the classifier performs his task with minimum time. Same importance appears in the medical diagnosis, stock purchasing decisions for trader in market of money stock. This problem is not considered until now through evaluating the Classifier. Also, this paper proposes a new indicator for Computer resources utilization indicator.

CPU indicator= (Minimum CPU/Actual CPU time)x 100 (3)

CPU time= CPU time of training +CPU time of testing (4)

Minimum CPU time= the minimum CPU time of classifier achieved (5)

Actual CPU time=The CPU time of evaluated classifier (6)

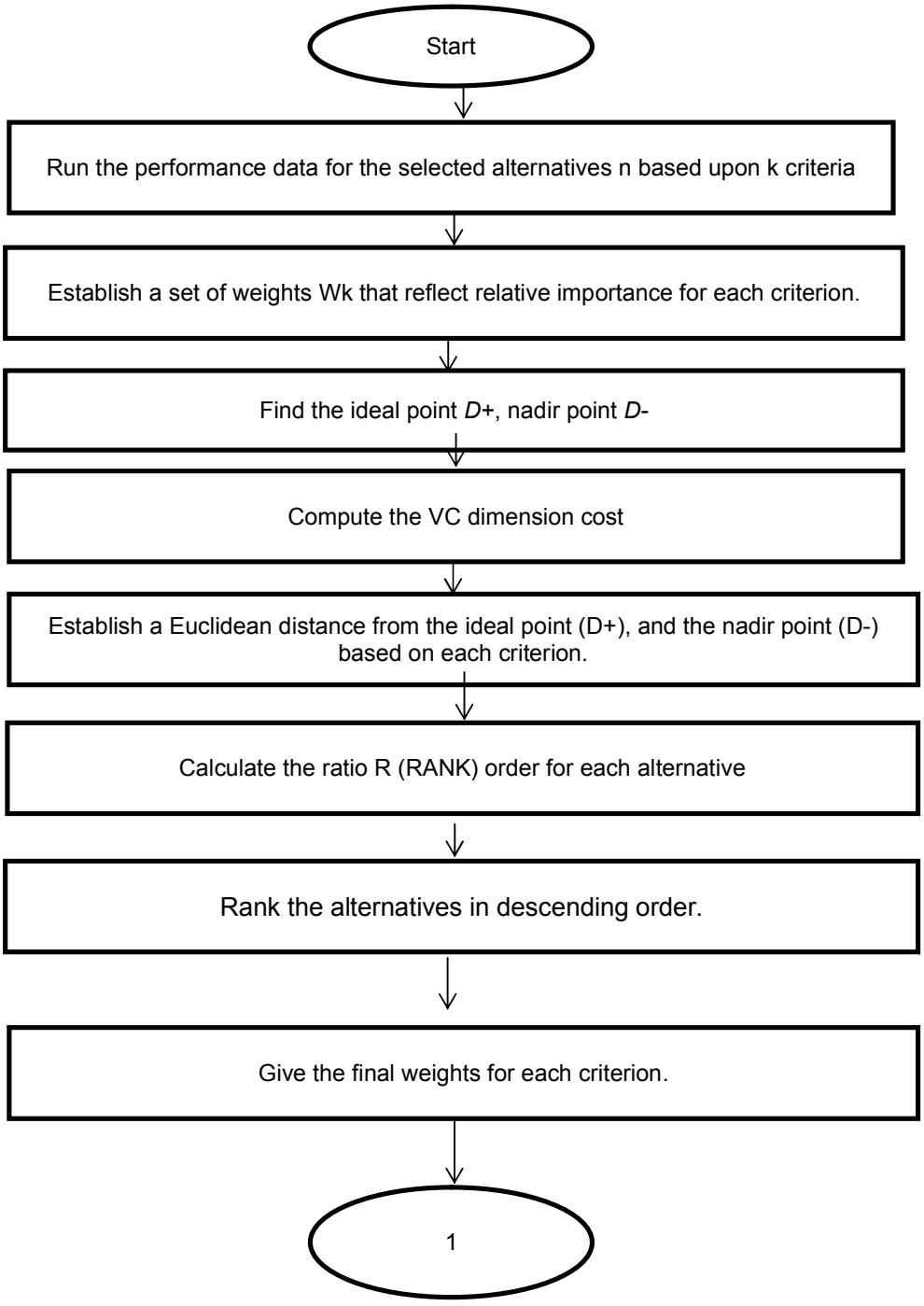
The paper presents a global indicator based on the three indicators of accuracy, avoid overfitting, and resources utilization indicator. Each indicator is percentage of one hundred so the range is unified by normalization. Therefore, it is suitable to compute the geometric mean. The equations of TOPSIS are illustrated in following steps.

1- Run the obtained data for n metrics (Accuracy, Sensitivity, Specifity, NPV, Precision, and F-measures over two criteria (VC dimension, information gain).
2- Mapping the raw measures X_{ij} to standardized measures S_{ij+}
3- Compute the relative importance W_k for each criterion.
4- Identify the ideal choice extreme performance on VC dimension, information gain $^{S+}$
5- Identify the nadir choice(reverse extreme performance on VC dimension, information gain)$^{S-}$
6- Compute R the Euclidean distance over each criterion to ideal (D+), and nadir (D-) such
7- $R = \frac{D^-}{(D^- + D^+)}$
8- Rank order of metrics by maximizing the ratio in step 7.
9- Give the final weight of metrics

The output of TOPSIS is used in the proposed global indicator illustrated in the Fig. 1.

5 Experiments and Results

This section presents twenty-five datasets (benchmarks) used in the experiments. The results are achieved by DTREG software package.


Start
Run the performance data for the selected alternatives n based upon k criteria
Establish a set of weights Wk that reflect relative importance for each criterion.
Find the ideal point D+, nadir point D-
Compute the VC dimension cost
Establish a Euclidean distance from the ideal point (D+), and the nadir point (D-) based on each criterion.
Calculate the ratio R (RANK) order for each alternative
Rank the alternatives in descending order.
Give the final weights for each criterion.
1

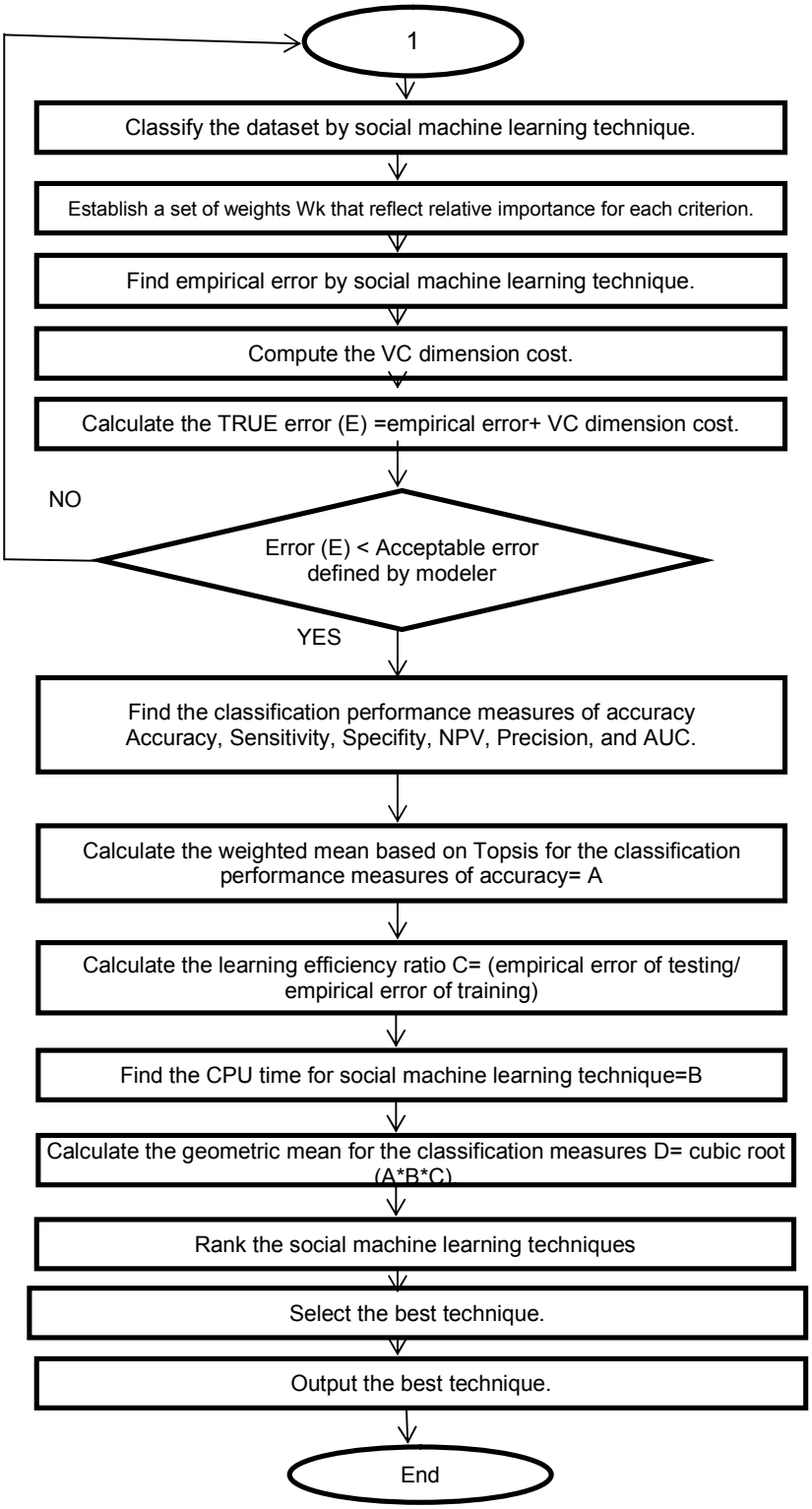


Fig. 1. The proposed global indicator is shown in the following flowchart

5.1 Datasets

Datasets presents classification benchmark problems used in the experiments. Data of benchmarks and best results including the references are illustrated in Table 3.

5.2 Simulation Procedure

The simulation is based on standard of machine leaning community that divide the data set, 50% training and 30% validating 20% testing were produced. These experiments produced: 20 times repeated; the foldcross validation is 10.

5.3 Vovel Benchmark

This subsection is composed of three stages. The first stage displays a comparison between classical classifiers (included the best classifier in the literature review) and latest classifiers [23,24] based on a single criterion (accuracy level)on Vovel benchmark. Second stage presents the rank of different single criterion based on TOPSIS. Third stage introduces a comparison between classical classifiers (included the best classifier in the literature review) and latest classifiers based on a proposed global indicator.

Table 4 illustrates the comparison of classifier techniques based on a single criterion (the accuracy of classifier). According to results of Table 4, STR has the best results among the individual machine learning that is closed to the best method in the literature (Cart-DB). TCML dominates others through the achieved accuracy level 0.931313131 compared with best results obtained in the literature review 0.90000000. However, DSCML has the superior performance of classifier is 0.943900. The results of DSCML dominate TCML, where the two latest classifiers DSCML and TCML are based on social learning. But DSCML has additional merit that is ability to tackle the uncertainty problem that TCML cannot.

Table 5 presents the relative importance of each criterion for classical six metrics obtained by TOPSIS. The highest rank is the F-measures 0.186289 and the lowest rank is NPV 0.123456. These weights are the input for the third stage.

Table 6 shows the weighted mean for the classical criteria, in addition the CPU efficiency indicator, and learning efficiency indicator. Based on multi-criteria the global indicator ranks the best method in the literature (Cart-DB) by 0.842262 that is highest rank among the classical classifiers except the latest classifiers TCML and DSCML. However, for general comparison the global indicator proves the dominance of DSCML that achieved global indicator 0.908158compared with other classifiers [23,24].

5.4 Telugu Vovel Benchmark

This subsection is composed of three stages. The first stage displays a comparison between classical classifiers (included the best classifier in the literature review and latest classifiers [23,24] based on a single criterion (accuracy level)on Telugu Vovel benchmark. Second stage presents the rank of different single criterion based on TOPSIS. Third stage introduces a comparison between classical classifiers (included the best classifier in the literature review) and latest classifiers based on a proposed global indicator.

Table 3. Displays the twenty-five classification benchmarks with best results achieved in the literature review used in testing the efficiency of global indicator for determining the best classifier

N	Data set	No of vectors	No. of attributes	The best method in the literature review	Accuracy level	References
1	Appendicitis	106	8	PVM (logical rules)	89.6	Weiss, Kapouleas
2	Wisconsin breast cancer	699	9	NB + kernel est	97.5	WD, WEKA,
3	Breast Cancer (Ljubljana data)	286	9	MLP+backprop	71.50%	Weiss, Kapouleas
4	Hepatitis	155	19	Weighted 9-NN	92.9±	Karol Grudziński
5	Statlog version of Cleveland Heart disease	303	13	Lin SVM 2D QCP	85.9±5.5	MG, 10xCV
6	Cleveland heart disease.	303	13	IncNet+ transformations	90	Norbert Jankowski
7	Diabetes.	786	8	Logdisc	77.7	Statlog
8	Hypothyroid	7200	21	C-MLP2LN rules+ASA	99.36	Rafał/Krzysztof/Grzegorz
9	Hepatobiliary disorders	536	4	IB2-IB4	44.6	WEKA
10	Landsat Satellite image dataset	6435	39	MLP+SCG	91	Michie, D.J. Spiegelhalter
11	Ionosphere	351	34	3-NN + simplex	98.7	Our own weighted kNN
12	Sonar: Mines vs Rocks	208	60	1-NN, 5D from MDS, Euclid, std	97.1	our, GM (WD)
13	Vovel	462	10	CART-DB, 10xCV	90	Shang, Breiman
14	Telugu Vovel	871	3	3-NN, Manhattan	87.8±4.0	Kosice
15	Wine data	178	13	kNN, Manhattan, k=1	98.7	GM-WD, std data
16	DNA-Primate splice-junction gene sequences	3190	3	RBF, 720 nodes	98.5	kNN GM - GhostMiner
17	Credit management	15000	7	Discrim	96.7	Statlog
18	Australian credit dataset	690	14	Cal5	86.9	Statlog
19	4 x 4 digit dataset	18000	16	Discrim	78.6	Statlog
20	Karhunen-Loeve digits	18000	40	Discrim	92.5	Statlog
21	Vehicle dataset	846	18	Discrim	78.4	Statlog
22	Letters	20000	16	ALLOC80	93.6	Statlog
23	Chromosome dataset	40000	16	Discrim	89.3	Statlog
24	Satellite image (Sat Image)	6435	36	K-NN	90.6	Statlog
25	Image segmentation	2310	11	Discrim	89.4	Statlog

These datasets used in the experiments are generated from UCI machine learning database. Available: http://duch- links.wikispaces.com/Classification results

Table 7 illustrates the comparison of classifier techniques based on a single criterion only the accuracy of classifier. According to results of Table 7, SVM has the best results among the classifier technique that is closed to best method in the literature (3 NN-Manhattan)-Kosice.

TCML dominates others through the achieved accuracy level0.908151better than best results obtained in the literature review 0.878000, DSCML has the superior performance of classifier is 0.912700. The results of DSCML dominate TCML, however the two latest classifiers DSCML, and TCML are based on social learning. But DSCML has additional merit that is ability to tackle the uncertainty problem that TCML cannot

Table 8 present the relative importance of each criterion of classical six metrics obtained by TOPSIS. The highest rank is the NPV 0.191356801and the lowest rank is Precision 0.125456895. These weights are the input for the third stage.

Table 4. Displays accuracy level for classifier techniques, the best method in the literature and DSCML on Vovel

Machine learning technique	Accuracy level
SVM	0.895050
MLP	0.893939
GEP	0.885858
STR	0.896969
The best method in the literature(Cart-DB)- shang-breiman	0.900000
TCML	0.931313
DSCML	0.943900

The best results available: http://duch- links.wikispaces.com/Classification results

Table 5. Displays weights by TOPSIS for individual machine learning techniques on Vovel

Criterion	Weights by TOPSIS
Accuracy	0.181234
Sensitivity	0.170100
Specifity	0.161234
NPV	0.123456
Precision	0.178654
F-measures	0.186289

Table 6. Displays global indicator plus different criterion for machine learning techniques on Vovel

Criterion	Weighted criterion for SVM	Weighted criterion for MLP	Weighted criterion for GEP	Weighted criterion For STR	The best technique in the previous literature review (Cart-DB)	Weighted criterion TCML	Weighted criterion DSCML
Accuracy	0.164026	0.162012	0.160548	0.162561	0.165765	0.168786	0.169856
Sensitivity	0.164594	0.167217	0.165833	0.164928	0.166556	0.167886	0.175443
Specifity	0.080618	0.034550	0.041731	0.059637	0.045678	0.038696	0.045562
NPV	0.038087	0.028490	0.037722	0.040160	0.034567	0.035273	0.039864
Precision	0.172871	0.168031	0.166393	0.169593	0.168435	0.171493	0.183454
F-measures	0.180260	0.179084	0.177468	0.178713	0.179098	0.181308	0.198651
Weighted mean (based on TOPSIS)=	0.800458	0.739387	0.749697	0.775592	0.760103	0.763443	0.812832
Learning Efficiency Rate	0.821335	0.801666	0.812345	0.824456	0.895666	0.931234	0.954335
CPU efficiency	0.812334	0.776544	0.674134	0.798765	0.877654	0.943344	0.965567
Global indicator	0.811331	0.772106	0.700146	0.799356	0.842262	0.875324	0.908158

Table 7. Displays Accuracy level for individual machine learning techniques and the best method in the literature on Telugu Vovel

Machine learning technique	Accuracy level
SVM	0.876337
MLP	0.865189
GEP	0.848300
STR	0.866486
The best method in the literature (3 NN-Manhattan)-Kosice	0.878000
TCML	0.908151
DSCML	0.912700

The best results available: http://duch- links.wikispaces.com/Classification results

Table 8. Displays weights by TOPSIS for individual machine learning techniques on Telugu Vovel

Criterion	Weights by TOPSIS
Accuracy	0.167543
Sensitivity	0.157654
Specifity	0.189543
NPV	0.191356
Precision	0.125456
F-measures	0.168765

Table 9 shows the weighted mean for classical criteria, in addition the CPU efficiency indicator, and learning efficiency indicator. Based on multi-criteria the global indicator ranks the best method in the literature (3 NN-Manhattan)-Kosiceby 0.856899 that is highest rank among the classical classifiers except the latest classifiers TCML and DSCML. However, for general comparison the global indicator proves the dominance of DSCML that achieved global indicator 0.929930 compared with other classifiers [23,24].

5.5 Cleveland Heart Disease Benchmark

This subsection is composed of three stages. The first stage displays a comparison between classical classifiers (included the best classifier in the literature review), and latest classifiers [23,24] based on a single criterion (accuracy level)on Cleveland heart disease benchmark. Second stage presents the rank of different single criterion based on TOPSIS. Third stage introduces a comparison between classical classifiers (included the best classifier in the literature review), and latest classifiers based on a proposed global indicator.

Table 10 illustrates the comparison of classifier techniques based on a single criterion only the accuracy of classifier. According to results of Table 10, SVM is closed to the best results among the individual machine learning that near the best method in the literature Inc net Transformation. The dominance of TCML that achieved accuracy level 0.927392better than best results obtained in the literature review 0.900000. DSCML has the superior performance of classifier is 0.934000. The results of DSCML dominate TCML, however the two latest classifiers DSCML, and TCML are based on social learning. But DSCML has additional merit that is ability to tackle the uncertainty problem that cannot TCML.

Table 11 presents the relative importance of each criterion of classical six metrics obtained by TOPSIS. The highest rank is the accuracy 0.220234 and the lowest rank is Sensitivity 0.109877. These weights are the input for the third stage.

Table 12 shows the weighted criteria for classical six criteria, in addition the CPU efficiency indicator, and learning efficiency indicator. Based on multi-criteria the global indicator ranks the best method in the literature Inc net Transformation by0.828771 that is highest rank among the classical classifiers excluding latest classifiers TCML and DSCML. However, for general comparison the global indicator proves the dominance of DSCML that achieved global indicator 0.924893compared with other classifiers [23,24].

Table 9. Displays global indicator plus different criteria for machine learning techniques on Telugu V

Criterion	Weighted criterion For SVM	Weighted criterion For MLP	Weighted criterion For GEP	Weighted criterion For STR	The best technique in the previous literature 3 NN-Manhattan	Weighted criterion TCML	Weighted criterion DSCML
Accuracy	0.148499	0.148307	0.147153	0.143498	0.150123	0.152154	0.165567
Sensitivity	0.153161	0.153360	0.148986	0.147517	0.144976	0.151874	0.176554
Specifity	0.069499	0.052373	0.074463	0.085028	0.081323	0.063181	0.097554
NPV	0.050721	0.060886	0.102679	0.080571	0.110454	0.108819	0.112457
Precision	0.129406	0.116885	0.114644	0.115678	0.114324	0.116583	0.123445
F- measures	0.162274	0.160627	0.156602	0.156750	0.158765	0.159651	0.165786
Weighted mean (based on TOPSIS)=	0.713563	0.692440	0.744527	0.729042	0.759960	0.752262	0.841363
Learning efficiency	0.821335	0.801666	0.782345	0.824456	0.922331	0.951234	0.966651
CPU efficiency	0.812334	0.776544	0.771345	0.798765	0.897654	0.963344	0.988778
Global indicator	0.780830	0.759024	0.765906	0.783023	0.856899	0.883376	0.929930

Table 10. Displays accuracy level for individual machine learning techniques and the best method in the literature and TCML on Cleveland heart disease

Machine learning technique	Accuracy level
SVM	0.897290
MLP	0.891089
GEP	0.874587
STR	0.894389
The best method in the literature (Inc net Transformation)-norbet	0.900000
TCML	0.927392
DSCML	0.934000

The best results available: http://duch- links.wikispaces.com/Classification results

Table 11. Displays weights by TOPSIS for individual machine learning techniques on Cleveland heart disease

Criterion	Weights by TOPSIS
Accuracy	0.220234
Sensitivity	0.109877
Specifity	0.134568
NPV	0.186678
Precision	0.169876
F-measures	0.178911

Table 12. Displays global indicator plus different criterion for machine learning techniques and TCML on Cleveland heart disease

Criterion	Weighted criterion for SVM	Weighted criterion for MLP	Weighted criterion for GEP	Weighted criterion for STR	The best technique in the literature IncNet transformation	Weighted criterion TCML	Weighted criterion DSCML
Accuracy	0.148499	0.148307	0.147153	0.143498	0.149123	0.152154	0.165456
Sensitivity	0.153161	0.153360	0.148586	0.147510	0.148765	0.151874	0.165765
Specifity	0.069499	0.052373	0.074463	0.085028	0.076544	0.063181	0.076545
NPV	0.05072	0.060886	0.102679	0.080571	0.098765	0.108810	0.123454
Precision	0.119406	0.116885	0.114644	0.115678	0.132233	0.116583	0.132456
F-measures	0.162274	0.160627	0.156602	0.156750	0.143245	0.159651	0.165755
Weighted mean (based on TOPSIS)=	0.703563	0.692440	0.744129	0.729038	0.748677	0.752255	0.829434
Learning efficiency	0.821335	0.801666	0.782345	0.824456	0.876544	0.951234	0.976678
CPU efficiency	0.812334	0.776544	0.771345	0.798765	0.867433	0.963344	0.976655
Global indicator	0.777176	0.755406	0.759942	0.783031	0.828771	0.883374	0.924893

5.6 Diabetes Benchmark

This subsection is composed of three stages. The first stage displays a comparison between classical classifiers (included the best classifier in the literature review), and latest classifiers [23,24] based on a single criterion (accuracy level) on Cleveland heart disease benchmark. Second stage presents rank the different single criterion based on TOPSIS. Third stage introduces a comparison between classical classifiers (included the best classifier in the literature review) and latest classifiers based on a proposed global indicator.

Table 13 illustrates the comparison of classifier techniques based on a single criterion only the accuracy of classifier. According to results of Table 13, STR has the best results among the individual machine learning that is closed to the best method in the literature logistic discrimination. The dominance of TCML that achieved accuracy level 0.788511better than best results obtained in the literature review 0.777000, also better than individual machines learning. But the DSCML has the superior performance of classifier is 0.808000. The results of DSCML dominate TCML, however the two latest classifiers DSCML, and TCML are based on social learning. But DSCML has additional merit that is ability to tackle the uncertainty problem that cannot TCML.

Table 13. Displays Accuracy level for individual machine learning techniques and the best method in the literature and TCML on Diabetes

Machine learning technique	Accuracy level
SVM	0.772456
MLP	0.762402
GEP	0.771540
STR	0.748041
The best method in the literature (logistic discrimination LOG DISC)-Statlog	0.777000
TCML	0.788511
DSCML	0.808000

The best results available: http://duch- links.wikispaces.com/Classification results

Table 14 presents the relative importance of each criterion of classical six metrics obtained by TOPSIS. The highest rank is the accuracy 0.21098765 and the lowest rank is Specifity 0.13133456. These weights are the input for the third stage.

Table 15 shows the weighted mean for classical criteria, in addition the CPU efficiency indicator, and learning efficiency indicator. Based on multi-criteria the global indicator ranks the best method in the literature logistic discrimination by 0.831468 that is highest rank among the classical classifiers excluding latest classifiers TCML and DSCML. However, for general comparison the global indicator proves the dominance of DSCML that achieved global indicator 0.925919 compared with other classifiers [23,24].

Table 14. Displays weights by TOPSIS for individual machine learning techniques and TCML on Diabetes

Criterion	Weights by TOPSIS
Accuracy	0.21098765
Sensitivity	0.17234567
Specifity	0.13133456
NPV	0.17435678
Precision	0.13987665
F- measures	0.17289543

Table 15. Displays global indicator plus different criterion for machine learning techniques and DSCML

Criterion	Weighted criterion For SVM	Weighted criterion For MLP	Weighted criterion For GEP	Weighted criterion For STR	the best technique in the literature LOGDISC	Weighted criterion TCML	Weighted criterion DSCML
Accuracy	0.148499	0.148307	0.147153	0.143498	0.146555	0.152154	0.165935
Sensitivity	0.153161	0.153360	0.148586	0.147510	0.150876	0.151874	0.165444
Specifity	0.069499	0.072373	0.074463	0.085028	0.076578	0.063181	0.086543
NPV	0.080721	0.060886	0.112679	0.090571	0.128455	0.108810	0.112223
Precision	0.119406	0.116885	0.114644	0.115678	0.112312	0.126583	0.123453
F- measures	0.162274	0.160627	0.156602	0.156750	0.141234	0.159651	0.167854
Weighted mean (based on TOPSIS)=	0.733560	0.712438	0.754127	0.739035	0.756010	0.762253	0.831452
Learning efficiency	0.821335	0.801666	0.782345	0.824456	0.876543	0.951234	0.976556
CPU efficiency	0.812334	0.776544	0.771345	0.798765	0.867433	0.963344	0.977654
Global indicator	0.788067	0.762609	0.769184	0.786593	0.831468	0.887270	0.925919

5.7 Test the Efficiency of Global Indicator to Select the Best Classifier

This subsection presents an extensive test based on twenty-five familiar benchmarks to prove the ability of global indicator to select the best classifier. This test is supported by statistical test (T-test) with confidence level 0.9999.

Table 16 shows the evaluation of classical classifiers (included the best classifier in the literature review), and the latest classifiers TCML and DSCML based on the proposed global indicator. This evaluation is used twenty-five familiar benchmarks to detect the ability of global indicator for determining the best classifier from a classifier space. The global indicator presents an evaluation to every classifier as percentages of one hundred. This percentage displays the efficiency of the selected classifier based on multi-criteria that explained in section 4.

The global indicator is tested on Appendicitis benchmark to evaluate different classifiers: SVM, MLP, GEP, STR, and the technique that has best results in literature **(**logical rules). In addition, the latest classifiers TCML [23], DSCML [24]. The global indicator gives evaluation 0.786656for SVM, 0.766775 for MLP, 0.765454 for GEP, 0.813636 for STR, 0.854356 for the technique has best results in the literature review PVM (logical rules).

Table 16. Displays the comparison between classical classifiers including the best classifier in the literature review with the latest classifiers TCML and DSCML

No	Dataset type	SVM	MLP	GEP	STR	The technique has best results in literature	TCML	DSCML	P-value
1	Appendicitis	0.786656	0.766775	0.765454	0.813636	0.854356	0.864346	0.876544	0.005***
2	Wisconsin breast cancer	0.877678	0.882767	0.877889	0.897798	0.898765	0.897776	0.976586	0.007***
3	Ljubljana dataset	0.812883	0.833454	0.843456	0.845667	0.876663	0.901838	0.912838	0.009***
4	Hepatitis	0.878987	0.856785	0.879858	0.864447	0.886689	0.890239	0.931009	0.004***
5	Statlog Cleveland Heart	0.887789	0.813873	0.834165	0.865432	0.908987	0.912389	0.940081	0.005***
6	Cleveland heart disease.	0.777176	0.755406	0.759942	0.783031	0.828771	0.883374	0.924893	0.007***
7	Diabetes.	0.788067	0.762609	0.769184	0.786593	0.831468	0.887270	0.925919	0.006***
8	Hypothyroid	0.665675	0.687643	0.698764	0.712867	0.734562	0.776857	0.786542	0.007***
9	Hepatobiliary disorders	0.677664	0.687994	0.702765	0.720977	0.736664	0.787765	0.798787	0.003***
10	Landsat Satellite image	0.876562	0.727638	0.876727	0.822276	0.897762	0.901827	0.911233	0.005***
11	Ionosphere	0.821773	0.798663	0.776365	0.799873	0.813763	0.836536	0.865543	0.007***
12	Sonar: Mines vs Rocks	0.776738	0.778787	0.798765	0.801873	0.811238	0.837729	0.876544	0.007***
13	Vovel	0.811331	0.772106	0.700146	0.799356	0.842262	0.875324	0.908158	0.004***
14	Telugu Vovel	0.780830	0.759024	0.765906	0.783023	0.856899	0.883376	0.929930	0.005***
15	Wine data	0.778783	0.786598	0.798989	0.724577	0.812675	0.825436	0.945365	0.008***
16	DNA-Primate splice-	0.887687	0.898778	0.898678	0.882201	0.910289	0.928769	0.936789	0.009***
17	Credit management	0.767578	0.776868	0.789866	0.812458	0.832457	0.834678	0.850023	0.003***
18	Australian credit dataset	0.787789	0.821334	0.856467	0.786579	0.879076	0.898672	0.909765	0.004***
19	4 x 4 digit dataset	0.777734	0.758689	0.778658	0.813459	0.823547	0.865788	0.886754	0.006***
20	Karhunen-Loeve digits	0.675754	0.676009	0.710097	0.723568	0.745346	0.786789	0.798676	0.006***
21	Vehicle dataset	0.712334	0.689865	0.676579	0.698765	0.767678	0.797665	0.821344	0.009***
22	Letters	0.778876	0.780125	0.798744	0.712461	0.813467	0.823467	0.832567	0.008***
23	Chromosome dataset	0.88789	0.898799	0.876891	0.867898	0.904535	0.912358	0.933245	0.006***
24	Satellite image (SatImage)	0.808891	0.812863	0.832789	0.846537	0.873712	0.886579	0.891234	0.007***
25	Image segmentation	0.813031	0.832561	0.845356	0.856765	0.893457	0.908123	0.912345	0.006***

With deep analysis of the archived results, the global indicator detects the classifier PVM (logical rules) has best results. For more details, the ability of the global indicator detects the efficiency of the latest classifiers TCML, and DSCML.

The global indicator is tested on the Vovel benchmark to evaluate different classifiers: SVM, MLP, GEP, STR, best results in literature logical rules. In addition, the latest classifiers TCML [23] DSCML [24]. The global indicator gives evaluation 0.777176for SVM, 0.755406for MLP, 0.759942for GEP, 0.783031for STR, 0.834170for the technique has best results in the literature review PVM (logical rules). 0.883374 for TCML and 0.922193 for DSCML.

The results show that the global indicator detects the classifier Logistic discrimination with best results. For more details, the ability of the global indicator detects the efficiency of the latest classifiers TCML [23] DSCML [24]. These results are clear in the twenty five familiar benchmarks of classification generated from site illustrated above [25].

6 Discussion

The experiments performed illustrate and show the significant benefits that generated from the proposed indicator. Also, the experiments clear the risk that faces the modeler within absence the global indictor. Such, the modeler applies a single criterion for measuring the efficiency of the classifier will mislead him to select the best classifier. This fact supported by the results achieved from the experiments on twenty-five datasets. Without loss of generality if the classifier select based on single criterion as Specifity, precision, and F-measure, the modeler will not select the best. The solution of this problem presented by the proposed indicator based on multi-criteria that are taking in account other criteria as CPU efficiency, and the Learning Efficiency Ratio (LER) these criteria proposed by the authors. The proposed global indictor selects the DSCML as the best classifier based on multi-criteria. The other benchmarks reflect the same results.

Statistical tests are performed to ensure the achieved results. The last column in the Table 16 shows the T-test for comparing the DSCML with the technique has the best results in the literature review.

The results achieved indicate to the superior performance of DSCML classifier compared than the best classifiers in the literature review based on statistical tests shown in the last column in Table 16. This superior due to the following reasons:

- The technique is based on the concept of social learning that merges multiple classifiers simultaneously to optimize the classification process.
- The ability of DSCML to tackle the classification process under uncertainty that fail any classifier to overcome it.

The quality of the global indicator for these reasons:

1. The global indicator is based on multiple criteria, not a single criterion as they existed in the traditional classifier evaluation.
2. Taking into account the learning efficiency (avoiding the overfitting problem).
3. Taking into account the CPU efficiency that its importance clearly appears in the military process, financial process, or medical process that requires the speed of solution.

7 Conclusion

This paper presents a novel indicator for tackling an important issue that faces researchers in the area of classification. This paper suggests a global indicator that concerns all dimensions

presented in previous literature review. In addition, two criteria (that not be considered before) the LER and CPU time efficiency. A comparison between classical classifier techniques and latest classifier techniques is done to evaluate the efficiency of them. The results achieved from a comparison on twenty-five datasets prove the efficiency of the global indicator to detect the best classifier, and support using it as effective tool for evaluating the classifier techniques for large scale of classification problems.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and regression tress. Belmont, CA: Wadsworth; 1984.

[2] Provost FJ, Fawcett T, kohavi R. The case against accuracy estimation for comparing induction algorithm. Proceedings of the 15th International Conference on Machine Learning. 1998;445-453.

[3] Drummond C, Holte R. Explicitly representing expected cost: an alternative to ROC representation. Processdings of knowledge Discovery and Datamining. 2000;198-207.

[4] Drummond C, Holte R. What ROC curve cannot do and what can. ROCAI. 2004;19-26.

[5] Forman G. An extensive empirical study of feature selection metrics for text classification. Special issue on variable and feature selection. Journal of Machine Learning Research. 2003;3(3):1289-1305.

[6] Fawcet T. Roc graphs: Notes and practical consideration for data mining reseachers (Technical report). HP Laboratories, Palo Alto; 2003.

[7] Ferri C, Flach P, Henrandez –Orallo J. Learning decsion tree using area under the ROC curve. Processdings of the 22 nd International Conference on Machine Learning. 2002; 139-146.

[8] Cortes C, Mohri M. AUC optimization vs. error rate minimization. Proceeding neural information processing system s 15(NIPS). MIT Press; 2003.

[9] Joachims T. A support vector machines for multi-variate performance measures. Processdings of the 22 nd International Conference on Machine Learning ACM Press; 2005.

[10] Prati R, Flach P. A roc convex hull rule learning algorithm. Processdings of the ECML/PKDD Workshop on Advances in Inductive Rule Learning. 2004;144-153.

[11] Herschtal A, Raskutti B. Optimizing area under the ROC curveusing gradient descent. Processdings of the 21 st International Conference on Machine Learning. New York, NY, USA: ACM Press. 2004;49.

[12] Srinivasan A. The aleph mannual version 4; 2003. Acceessed 13 January 2015. Avialable: http//web .comlab.ox.ac.uk/ oucl/research/areas/ machlearn/Aleph

[13] Petterson J, Caetano TS. Revese multi-label learning. In Advances in Neural Information Processing Systems. 2010;24:1912-1920.

[14] Quevedo J, Luaces O, Bahamonde A. Multilabel classifier with a probabilistic thresholding strategy. Pattern Recognition; 2012;(45).

[15] Hwang CL, Yoon K. Multiple attribute decision making: Methods and Applications. Springer. Berlin. Heidelberg. New York; 1981.

[16] Jee DH, Kang JK. A method for optimal material selection aided with decision making theory. Materials and Design. 2000;21(3):199-206.

[17] Yong D. Plant location selection based on fuzzy TOPSIS. International Journal of the Advanced Manufacturing Technology. 2006;28(7-8):839-844.

[18] Alessandro Moschitti. Vapnik-Chervonenkis (VC) Dimension; 2011. Acceessed 19 January 2015. Avialable: http://www.cedar.buffalo.edu/~srihari/CSE574/Chap16/Chap16.1-InformationGain.pdf

[19] Chapter 16.1. information gain. 2010. Acceessed 29 January 2015.
Avialable: http://disi.unitn.it/moschitti/ML2010-11/VC-dim.pdf

[20] Defossez A, Bach F. Averaged least-mean-square: bias-variance trade-offs and optimal sampling distributions. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS); 2015.

[21] Lacoste-Julien S, Lindsten F, Bach F. Sequential kernel herding: frank-wolfe optimization for particle filtering. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS); 2015.

[22] Bach F. Breaking the curse of dimensionality with convex neural networks. Technical Report, HAL-01098505; 2014.

[23] Zaher H, Abdullah M, Ragaa N. A social learning approach for minimizing true risk of collective machine learning. International Journal of Advanced Research in Computer Science and Software Engineering. 2013;(3):1172-1179.

[24] Zaher H, Abdullah M, Ragaa N. Social learning under uncertainty based on dempster-shafer approach for minimizing true error of machine learning. Journal of Advanced Mathematics and Computer Science. Science Domain. Manuscript number. 2015;7(4):280-292.

[25] Experiments are generated from UCI machine learning database.
Available: http://www.ics.uci.edu/˜mlearn/MLRepository.html.
Available: http://duch- links.wikispaces.com/Classification result

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=1034&id=6&aid=8841