International Journal of Engineering & Technology, 7 (4.39) (2018) 980-984

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Design of Cost Sensitive Classifiers for E-Learning Data Sets Tuning with Cost Ratio

Mr.C.S.Sasikumar[1]*, Dr.A.Kumaravel[2]

[1]Research Scholar, Department of CSE, Bharath Institute of Higher Education and Research, India
[2]Professor and Dean, School of Computing, Bharath Institute of Higher Education and Research, India
**Corresponding author E-mail: sasi_kumin@yahoo.com*

Abstract

An examination on costly classifiers impacting essential choices utilizing expectations is a vital research field for the information mining analysts. Notwithstanding, the determination of parameters for such classifiers assumes an imperative job in getting more exactness and less expense in the basic setting. Following this rule, a measurement dependent on a levelheaded number, dictated by the proportion between the quantities of false positives to false negatives in assessing the classifiers is considered. Cost delicate models too the cost dazzle models are normally both acknowledged by their execution through least blunder or most extreme exactness. Thus in setting of understudies points of interest from East London locale and from Yorkshire district should be connected with more significant measures to locate the correct minimal effort esteems. In this paper, we analyze the cost touchy classifiers and measure their execution by shifting the parameter (False Positive and False Negative). We recognize distinctive examples of conduct of these classifiers for various scope of qualities. Add up to cost for four unique reaches are investigated independently and the exhibitions in the two-distinctive setting of understudy detail from East London district and from Yorkshire locale are considered. Add up to cost of understudy subtle elements from East London area happens to be more than expense of Yorkshire district while tuning the parameters. These discoveries can bolster the choices of diagnosing which methods for instruction is more important to choose with foruming or non-foruming with more certainty.

Keywords: *cost sensitive classification, learning, data mining, prediction*

1. Introduction

Many classifiers expect with the intention of the wrong classification costs (false negative and false positive expense) are same. In many genuine applications, this presumption may false. The parameters and conditions for cost estimations are reconsidered with the accompanying phrasing. The cost administering values are arranged in the cost framework which has indistinguishable structure from disarray grid as appeared in the table 1.

Table 1: Template for Cost Matrix based on confusion matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	C_{11}	C_{12}
	Negative	C_{21}	C_{22}

Right when misclassification costs are known, the best estimation for surveying classifier execution is indicate cost. Indicate cost is the fundamental appraisal metric used in this paper and is in like manner used to survey all of the three expense touchy learning techniques. The recipe for aggregate expense is appeared in condition.

Total Cost = (FN × CFN) + (FP × CFP) where CFN is cost of false negative and CFP is cost of false positive qualities denoted by C_{21}, C_{12} respectively.

Recall and precision are expressed by the ratios TP/(TP+FN) and TP/(TP+FP) respectively.

The F-measure is characterized as a symphonious mean of exactness (P) and recall(R): i.e F =2PR/(P + R).

Sensitivity = TP / (TP + FN)
Specificity= TN / (FP + TN)

The probability proportion for a positive outcome is = sensitivity /(1- specificity).

The probability proportion for a negative outcome is = (1-sensitivity)/ specificity.

The contributions for the order calculations are the cost frameworks of differing proportions of false negative to false positive. We separate the yield from the perplexity grid. There are two difficulties as for the preparation of cost delicate classifier. The misclassification costs assume a pivotal job in the development of a cost touchy learning model for accomplishing expected grouping results. On the off chance that C (I, j), where i,j take esteems either 1 or 2, be the expense of anticipating an occurrence having a place with class I when in truth it has a place with class j, at that point we are keen on C(1,2)/C(2,1) or the opposite of this. Our primary goal is to discover adequate proportion as it changes enormously crosswise over various settings.

Table 2: Algorithm components based on Ratio formats

Ratio Pattern (#FP: #FN)	Uniform	Non-Uniform (Relatively Prime)
Normal	CSTMC-U	CSTMC-NU
Reverse	CSTMC-RU	CSTMC-NRU

A metaclassifier that makes its base classifier cost-delicate. Two methods can be used to introduce cost-affectability: reweighting planning cases as demonstrated by the total cost delegated to each class; or anticipating the class with slightest expected misclassification cost (instead of the no doubt class). Execution can routinely be upgraded by using a Bagged classifier to improve the probability evaluations of the base classifiers (J48 Decision Tree, Hoeffding Tree, LMT (Logical Model Tree), REPTree). The paper is created as seeks after: territory 2 contains related work, fragment 3 deals with dataset used and segment 4 oversees technique used, Section 5 with test outcomes and portion 6 closes.

2. Literature Review

Research in data mining [1][2] proves that EDM mainly deals with
"1. Anticipating understudies' future learning conduct by making understudy models that join such point by point data as understudies' information, meta-insight, inspiration, and frames of mind.
2. Finding or enhancing space models that describe the substance to be educated and ideal instructional successions.
3. Concentrate the impacts of various types of instructive help that can be given by learning programming; and
4. Progressing logical information about learning and students through building computational models that consolidate models of the understudy, the product's teaching method and the area.
To achieve these objectives, instructive information mining research utilizes specialized strategies like expectation, grouping, relationship mining, displaying etc[3]"
In view of EDM survey, Romero and Ventura[4] demonstrated that " future EDM explore center around the accompanying perspectives: - coordinate EDM devices with e-learning frameworks - institutionalize information and models - make EDM devices simpler for teachers and non-master clients - alter customary digging calculations for instructive setting". In [5], a precise survey on learning investigation is furnished and different information mining calculations with money saving advantage examination are looked into in [6-9].

3. Methods & Materials

3.1 Cost-Sensitive Learning (CSL)

Most classifiers expect that the misclassification costs (false negative and false positive cost) are the comparable. In most authentic applications, this assumption may false. Another point of reference is illness assurance: misclassifying a dangerous development is generously more veritable than the false alert since the patients could lose their life in perspective of a late end and treatment [24]. The parameters and conditions for cost checks are refreshed with the going with expressing.
The cost administering values are classified in the cost network which has indistinguishable structure from perplexity lattice as appeared in the accompanying table.

Table 3: Template for Cost Matrix based on confusion matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	C_{11}	C_{12}
	Negative	C_{21}	C_{22}

Exactly when misclassification costs are known, the best estimation for evaluating classifier execution is mean expense. Signify cost is the fundamental appraisal metric used in this paper and is in like manner used to evaluate every one of the three cost-tricky learning techniques. The recipe for total cost is showed up in condition.
Total Cost = (FN × CFN) + (FP × CFP) where CFN is cost of false negative and CFP is cost of false positive values denoted by C_{21} , C_{12} respectively.
Recall and precision are expressed by the ratios TP/(TP+FN) and TP/(TP+FP) respectively.
The F-measure is characterized as a symphonious mean of exactness (P) and recall(R): i.e F =2PR/(P + R).
Sensitivity = TP / (TP + FN)
Specificity= TN / (FP + TN)
The probability proportion for a positive outcome is = sensitivity /(1- specificity).
The probability proportion for a negative outcome is = (1-sensitivity)/ specificity.
The accompanying segment portrays the calculations to create false negatives and false positives which anticipated that would be in the base dimension. The contributions for these calculations are the cost lattices of changing proportions of false negative to false positive. We separate the yield from the perplexity lattice. There are two difficulties as for the preparation of cost touchy classifier. The misclassification costs assume a vital job in the development of a cost touchy learning model for accomplishing expected grouping results. In the event that C(i, j), where i,j take esteems either 1 or 2, be the expense of anticipating an occasion having a place with class I when in truth it has a place with class j, at that point we are keen on C(1,2)/C(2,1) or the reverse of this. Our fundamental target is to discover adequate proportion as it shifts incredibly crosswise over various settings.

Table 4: Algorithm components based on Ratio formats

Ratio Pattern (#FP:#FN)	**Uniform**	**Non Uniform (Relatively Prime)**
Normal	CSTMC-U	CSTMC-NU
Reverse	CSTMC-RU	CSTMC-NRU

3.2 Meta Classifier

3.2.1 Cost Sensitive Classifier

Name
Cost Sensitive Classifier
Synopsis
A metaclassifier that makes its base classifier cost-fragile. Two methods can be used to exhibit cost-affectability: reweighting getting ready events as demonstrated by the total cost distributed to each class; or foreseeing the class with minimum expected misclassification cost (rather than the more then likely class). Execution can frequently be upgraded by using a Bagged classifier to improve the probability assessments of the base classifier.

3.3 Base Classifiers

3.3.1 J48 Decision Tree

Depiction
Class for producing a pruned or unpruned C4.5 choice tree.

3.3.2 HoeffdingTree

A hypothetically engaging component of Hoeffding Trees not shared by other gradual choice tree students is that it has sound certifications of execution. Utilizing the Hoeffding bound one can demonstrate that its yield is asymptotically about indistinguishable to that of a non-gradual student utilizing endlessly numerous models.

3.3.3 LMT (Logical Model Tree)

Classifier for building 'strategic model trees', which are arrangement trees with calculated relapse capacities at the leaves. The calculation can manage twofold and multi-class target factors, numeric and ostensible characteristics and missing qualities.

3.3.4 REPTree

Description

Fast decision tree learner. Constructs a choice/relapse tree utilizing data gain/change and prunes it utilizing decreased mistake pruning (with backfitting). Just sorts esteems for numeric characteristics once. Missing qualities are managed by part the relating cases into pieces.

4. Dataset Description

These Datasets are about student details from the region of East London and from the region of Yorkshire and the following attributes and its description which are in the tables before pre-processing.
A Hoeffding tree (VFDT) is a steady, whenever choice tree acceptance calculation that is fit for gaining from monstrous information streams, expecting that the dispersion creating models does not change after some time. Hoeffding trees misuse the way that a little example can regularly be sufficient to pick an ideal part characteristic. This thought is bolstered scientifically by the Hoeffding bound, which measures the quantity of perceptions (for our situation, precedents) expected to assess a few insights inside an endorsed accuracy (for our situation, the integrity of a trait).

5. Methodology Proposed

Cost touchy learning is tied in with contrasting the False Negative (FN) values and False Positive (FP) values. The datasets which I'm utilizing here is an ongoing information. This dataset comprises of 15 qualities and it is separated to 9 characteristics (counting 1 class) after pre-preparing process to be specific id_student, date_registred, sexual orientation, date, sum_click, activity_type, date summit, center, weight. The Class is an ostensible class characteristic which is having two kind of qualities foruming (forumng) and n-foruming (nforumng).
After the pre-processes, we need to select the meta classifier. Data classification is a process that helps for efficient prediction using the data set. In data classification, there are many meta classifiers are present and we use particularly a classifier and that is CostSensitiveClassifier. Inside these meta classifiers, there are numerous base classifiers namely J48, Hoeffding Tree, LMT, REPTree. After selecting each base classifier, we need to choose the cost matrix 2:2 matrix (1.0,1.0,1.0,1.0). In that Matrix, we have to fix and change the False Positive (FP) values and False Negative (FN) values. Here we need to find the Cost Sensitive value for CSTMC - U, CSTMC - RU using the formulae f(x) = ((FP*CFP) +(FN*CFN)). After finding the values compare the values and find the less cost value. The various steps in proposed methodology is shown in figure 1.

○ In Weka Application, data pre-processes is initialized, that is uploading the entire dataset into the application. The steps are

- Choose the dataset **East London Region** with17 Attributes and 83732 Instances&

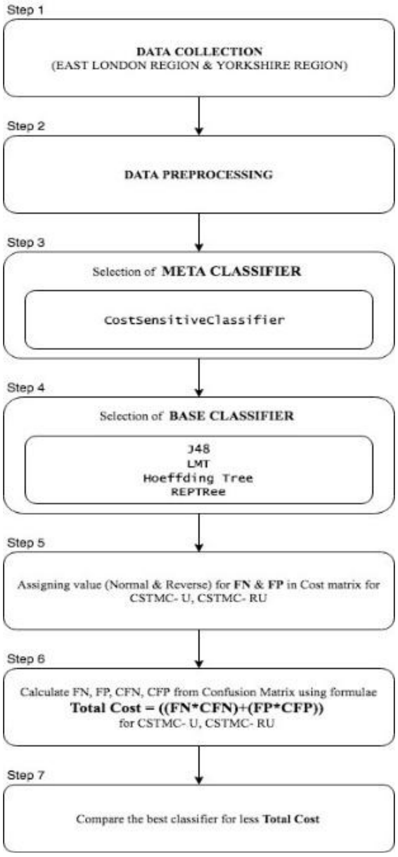
Step 1
DATA COLLECTION
(EAST LONDON REGION & YORKSHIRE REGION)
Step 2
DATA PREPROCESSING
Step 3
Selection of META CLASSIFIER
CostSensitiveClassifier
Step 4
Selection of BASE CLASSIFIER
J48
LMT
Hoeffding Tree
REPTRee
Step 5
Assigning value (Normal & Reverse) for FN & FP in Cost matrix for CSTMC- U, CSTMC- RU
Step 6
Calculate FN, FP, CFN, CFP from Confusion Matrix using formulae
Total Cost = ((FN*CFN)+(FP*CFP))
for CSTMC- U, CSTMC- RU
Step 7
Compare the best classifier for less Total Cost

Figure 1: Proposed methodology

6. Experiment Results

We consider the implementation of base tree classifiers for cost sensitive meta learners in Weka platform. The top such classifiers are J48, LMT, Hoeffding Tree, REPTree. For this purpose, we adopt the data from clinical records for East London Region & Yorkshire Region students detail. Table3 shows the total cost.

Table 6: Total cost for Variation of false positive and false negative in Cost sensitive Meta Classifiers for East London Region dataset using CSTMC-U and CSTMC-RU

Total Cost for East London Region Normal					Total Cost for East London Region Reverse				
(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**	(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**
1	2611	2813	3069	2596	1	2611	2813	3069	2596
2	3909	4018	4836	3800	2	2917	3462	3093	3040
3	4451	4872	5263	4527	3	2982	3630	3181	3161
4	4678	5206	5275	4927	4	3038	3487	3359	3120
5	4789	5653	5285	5075	5	3074	3401	3297	3069
6	4878	5813	5859	4440	6	3074	3192	3254	3075
7	4934	6203	5296	5231	7	3074	3261	3327	3046
8	5017	6112	5378	5286	8	3074	3114	3367	3069
9	5111	6296	5375	5300	9	3074	3121	3241	3067
10	5165	6178	5395	5356	10	3074	3099	3318	3079

Dataset is being ordered utilizing meta classifier Cost Sensitive Classifier. Under meta classifier we have chosen four base classifiers in particular J48, Logical Model Tree (LMT), Hoeffding Tree, REPTree. In the wake of choosing the base classifiers, we have to pick the cost network esteems False Negative and False Positive to discover which is creating the less cost esteem
The esteem which is settled in the cost grid must be in two kind, they are typical shape and turn around form.After characterization, perplexity framework esteem (FN and FP) needs to determined with CFP and CFN utilizing the formulae Total Cost = (FN ×

CFN) + (FP × CFP) where CFN is cost of false negative what's more, CFP is cost of false positive qualities meant by C21, C12respectively. Fig 2 and 3 demonstrates the aggregate expense for East London area and its turn around.

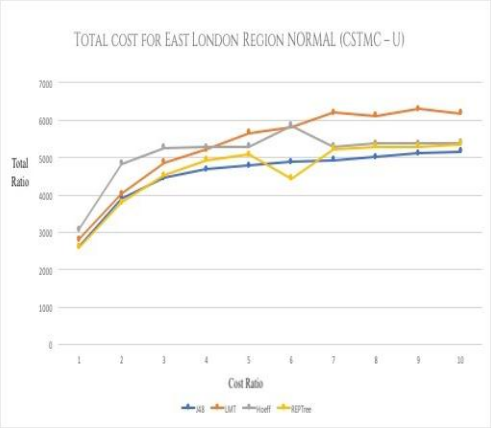

Figure 2: Total cost for East London region

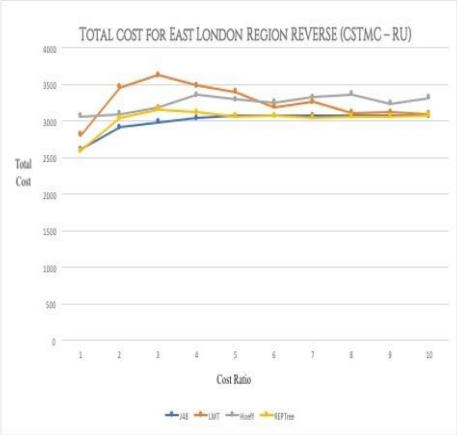

Figure 3: Total cost for East London region(reverse)

Dataset is being grouped utilizing meta classifier CostSensitiveClassifier. Under meta classifier we have chosen four base classifiers in particular J48, Logical Model Tree (LMT), Hoeffding Tree, REPTree

After choosing the base classifiers, we have to pick the cost framework esteems False Negative and False Positive to discover which is creating the less cost esteem.
The esteem which is settled in the cost framework must be in two kind, they are ordinary shape and switch shape.

After arrangement, disarray network esteem (FN and FP) needs to determined with CFP &CFN utilizing the formulae Total Cost = (FN × CFN) + (FP × CFP) where CFN is cost of false negative and CFP is cost of false positive qualities signified by C21, C12 separately.
With the aggregate cost got from the formulae the esteem must be thought about and the less aggregate cost will be picked.
Table 4 demonstrates the aggregate expense for Yorkshire area and fig. 4 and 5 demonstrates the graphical portrayal of aggregate expense for Yorkshire area and its turn around.

Table 7: Total cost for Variation of false positive and false negative in Cost sensitive Meta Classifiers for Yorkshire Region dataset Using CSTMC-U and CSTMC-RU

Total Cost for Yorkshire Region Normal					Total Cost for Yorkshire Region Reverse				
(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**	(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**
1	2435	2556	2741	1413	1	2435	2556	2741	1413
2	3250	3551	3926	3258	2	2767	3226	2813	2804
3	3734	4209	4225	3733	3	2783	3426	2825	2841
4	3911	4409	4216	3947	4	2815	3311	2840	2830
5	4056	4754	4246	4134	5	2796	3069	2852	2802
6	4245	4956	4583	4242	6	2796	3003	2838	2785
7	4288	5164	4335	4228	7	2796	2951	2842	2798
8	4274	5282	4366	4269	8	2796	2883	2844	2807
9	4291	5050	4370	4283	9	2796	2825	2852	2828
10	4208	5060	4325	4295	10	2796	2840	2853	2832

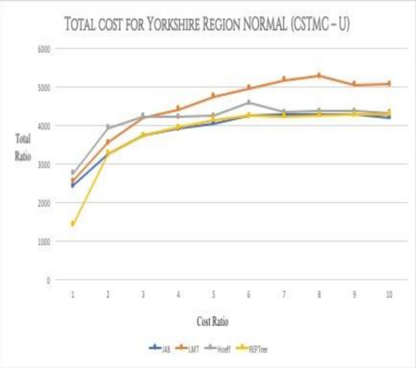

Figure 4: Total cost for Yorkshire region

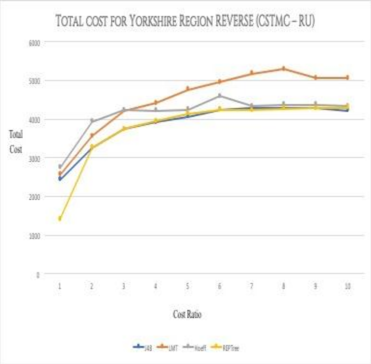

Figure 5: Total cost for Yorkshire region(reverse)

Table8: Total cost for variation of false positive and false negative in Cost sensitive Meta Classifiers for East London & Yorkshire Region dataset using CSTMC-U and CSTMC-RU

Total Cost for East London & Yorkshire Region Normal					**Total Cost for East London & Yorkshire Region Reverse**				
(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**	(Cost Ratio)	**J48**	**LMT**	**Hoeff**	**REPTree**
1	4993	5391	5647	5082	1	4993	5391	5647	5082
2	7317	7660	8507	7352	2	5794	6587	906	5915
3	8064	8907	9265	8338	3	5913	6819	5905	5948
4	8606	9725	9492	8926	4	5869	6754	5921	5986
5	8875	10242	9544	9105	5	5869	6547	5916	5974
6	9050	10907	9559	9251	6	5869	6213	5931	5950
7	9181	11367	9593	9486	7	5869	6378	5953	5912
8	9306	11228	9591	9725	8	5869	6303	5923	5936
9	9464	11655	9597	9705	9	5869	6120	5923	5926
10	9479	11916	9619	9669	10	5869	6028	5910	5901

With the total cost derived from the formulae the value has to be compared and the less total cost will be chosen.

Table 5 demonstrates the variety for the two districts and fig 6 and 7 demonstrates the graphical portrayal of the equivalent and its turn around. Add up to cost is expanding more for false negative than that of false positive in the both East London area and Yorkshire district understudy subtle elements exploratory setups. In addition, the size of aggregate expense is more than that in the East London Region than the Yorkshire Region. We present the outcomes for the four sections of the primary calculation as appeared in the above diagrams. These certainties

are introduced as patterns and conduct in each portion of ρ values additionally the needs of impacts of these fragments likewise exhibited for probability proportion.

We have thought about all the table and found certain qualities in CSTMC-U, CSTMC-RU. In the wake of investigating the table qualities J48 and REPTree are the bases classifiers creating minimum cost delicate qualities. Among these two base classifiers J48 tree is the slightest touchy esteem creating classifier.

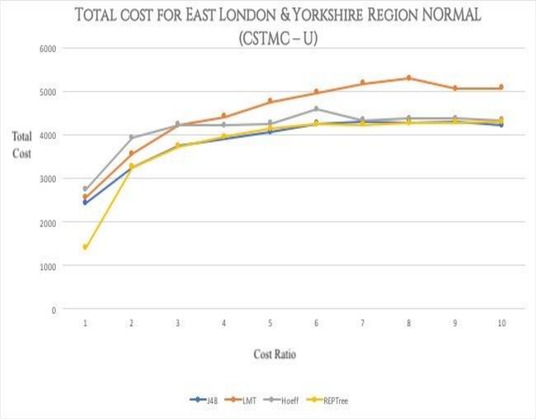

Figure 6: Total cost for both regions

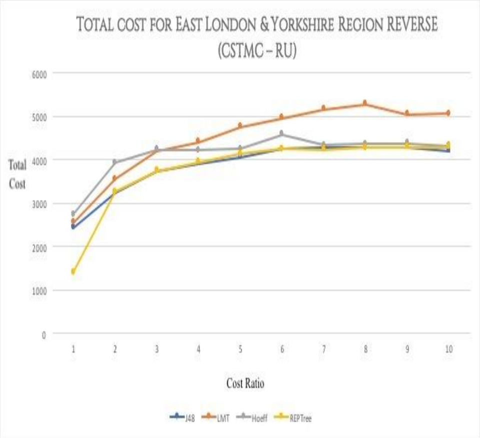

Figure 7: Total cost for both regions (reverse)

7. Conclusion

The cost delicate models for East London district and Yorkshire locale understudy points of interest informational collections are built and demonstrated the conduct for four scopes of cost proportion ρ. In this cost touchy process unmistakably demonstrates the requirement for isolated principles in basic leadership diversely relying upon the cost proportion in various setting like the datasets we utilized. With these two datasets, East London area and Yorkshire locale we have discovered two different ways of instruction framework foruming and non-foruming and we characterized which method for framework is reacting better positive reaction. More over the greatness of aggregate expense is more in Yorkshire district than East London locale. The future work can be stretched out with the examination for different sorts of cost delicate meta classifiers to quantify the blunder cost as talked about in this paper.

References

[1] Baker, R. S. J. d. 2011. "Data Mining for Education." In International Encyclopedia of Education, 3rded., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.

[2] Baker, R. S. J. D., and K. Yacef. 2009. "The State of Educational Data Mining in 2009: A Review and Future Visions." Journal of Educational Data Mining 1 (1): 3–17.

[3] Hamilton, L., R. Halverson, S. Jackson, E. Mandinach, J. Supovitz, and J. Wayman. 2009. UsingStudent Achievement Data to Support Instructional Decision Making (NCEE 2009-4067).Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center forEducation Evaluation and Regional Assistance.

[4] Romero, C.,&Ventura,S.(2010),Educational data mining: A review of the state of the art,IEEE Transactions on systems man and Cybernetics Part C.Applications and review, 40(6),601-618.

[5] ZacharoulaPapamitsiou,&Anastasios A. Economides. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, *17*(4), 49-64. Retrieved from http://www.jstor.org/stable/jeductechsoci.17.4.49

[6] Cios, K.J., Pedrycz W., Swiniarski, R.W. & Kurgan, L.A. (2007), Data Mining: A Knowledge Discovery Approach, Springer, New York.

[7] Klosgen, W. &Zytkow, J. (2002), Handbook of data mining and knowledge discovery, Oxford University Press, New York.

[8] Quinlan, J.R. (1993), C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco.

[9] Witten, I.H. & Frank E. (2000), Data Mining – Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco.