INVESTIGATION A NEW APPROACH TO DETECT AND TRACK FRAUD IN VIRTUAL LEARNING ENVIRONMENTS BY USING CHAYD MODEL

Sayyed Jalalaldin Gharibi Karyak[1] and Sayyed Hedayat Tarighinejad[2]

[1]Technical and Vocational University Yasooj, Iran
[2]Student of Computer Sciences in Islamic Azad University of Yasuj, Iran

ABSTRACT

Virtual University is the environment that with utilizes the appropriate multimedia tools and having good communication infrastructure is a provider of e-learning services, so that usually does not have require to physical location as a traditional university and students are able in any place and at any time be willing to use a lot of services provided, such as e-courses or electronic tests. There are many solutions in order to identify and detect fraud in the online environment that use of these methods can be identified took place fraud, but still is great importance of avoid discussion and fraud detection in virtual university. In this research, we aimed are to investigate a new approach to detect and track fraud in virtual learning environments by using decision tree. (Chayd model). The results showed that the accuracy of the model is 84.54% which is indicative of high performance and high precision in predicting fraud from the teachers, students and hackers.

KEYWORDS

E-learning, fraud, Chayd, data mining, decision tree

1. INTRODUCTION

Internet and virtual space are a wonderful source of awareness and knowledge. With the major changes that have occurred as a result of development of new technologies in various fields of human life, as well as education is not exempt from this rule change. virtual space with resources and great facilities that puts its users in the field of science and knowledge and with features such as lack of time and space, interactivity and puts a lot of capabilities in the field of education training and education to provide its users and in addition to these capabilities, also there are challenges such as fraud. Virtual University is the environment that with utilizes the appropriate multimedia tools and having good communication infrastructure is a provider of e-learning services, so that usually does not have require to physical location as a traditional university and students are able in any place and at any time be willing to use a lot of services provided, such as e-courses or electronic tests. Sometimes this system, in addition to traditional universities and sometimes separately, it pays to provide educational services. This system with utilizes of capabilities and facilities offered by Internet network and multimedia tools and technologies with the aim of raising the level of culture to prevent the removal of material resources, as well as the scientific capital of the country creating to a large extent, enhance the scientific level of society and possibility of the widespread distribution of knowledge, expertise and abilities in the university. Fraud, as any illicit behavior which may be by those who have regulate, manage and scoring and the evaluation is conducted by a fraudulently. Fraud in the evaluation of university education is a very serious crime which may ultimately lead to the deportation of the University. Because of the relatively new of e- learning and educational technology on web based, especially

DOI : 10.5121/ijmnct.2017.7203

in Iran, which is also the subject of a new study of this issue is very vital. Also main reason is that the use of e-learning shows a relatively high growth in recent years and this is caused that, before the creation of large problems deal to issue of fraud virtual education and thus can be, it can be controlled or at least have a plan to control it.

2. THEORETICAL FOUNDATIONS VIRTUAL EDUCATION

It refers to any type of learning that is done in a way other than traditional face to face. The contents of the courses may be transferred via the internet or the use of video, active images and interactive two-way. (Bedford et al., 2009). It also to providing educational content and teaching experiences to learners that they can in anywhere the world have take advantage of this kind of training, namely be done personal training by using computer technology, information and communication. A virtual training system provides a virtual environment to surround can learn the content without the presence of the instructor. Naturally, in this system should be considered in all aspects of environmental and effective spiritual components in teaching.

3 . CHALLENGES IN A VIRTUAL CLASSROOM

1. How to presence and absence.
2. Method of entry of students into the network.
3. Rules Compilation for registration.
4. Paying any tuition.
5. Does the evaluation of learning should be a stage, or final?
6. Method of conducting such evaluations.
7. How to prevent fraud and identify cheaters. (Trenholm, 2005).

4. DATA MINING

Data mining is a concept that has been developed and defined to analyze large amounts of data. Researchers such Hand, Manila and Esmet have provided a good definition of data mining: "Often, the analysis of massive amounts of collection of observed data to find a collection of transparent communication and summarizing data in different ways, as that concept and applied for people. (Hand et al., 2001). . By definition, it is clear that used data mining to analyze massive amounts of data to find the necessary information which is not understandable directly. Data mining is a very interesting scientific application, for example is used in diagnosing diseases, financial and non-financial fraud detection and analysis of weather data (Li et al., 2004). The main motivation for the use of data mining comes from the value that can be achieved by implementing successful in business and in most processes to reduces costs and we're saving money. (Lukawiecki, 2008).

5. RESEARCH HISTORY

Etter, Cramer and Finn (2007) were compared the two distinguishes groups to assess their perceptions about the discovery of the fraud and information technology. A group of students from the university were affiliated with a particular religion or another group of students from other region research institute. Sample for university students related to a particular religion are 237 students and are considered 202 students for a regional research institute. The results show that students affiliated to a particular religion compared to students from regional research institutes are less inclined to fraud.

Mirza and Staples (2010) conducted research on the use of webcams, as an innovative way to reduce or eliminate fraud. They were in an online course for nursing students. Students need to buy a webcam to participate in the exam training, rather, gained skills before the actual exam. After completing the test, was done a survey of 44 students, while they were under to consistently surveillance. Students reported that the possibility of less fraud is sad and disappointment feeling but only 19 people thinks this webcam can help prevent cheating.Sottile and Watson (2010) conducted a survey of 635 students to find that, more students are fraud online or face-to-face in classrooms. Over 30% of students admitted that they had been fraud in both environments. However the difference between traditional and online fraud in terms of statistical, is not less than 0.05 percent. The students admitted that they had more than 4 times as likely to fraud in an online class compared to when the test was face to face classes. (Watson and Sottile, 2010), (Kidwell and Kent 2008).

Mac Cabe and Trevino (1993) in order to determine whether students in Australia, most of done the fraud in the online classroom or when their education in face to face classes. Surveys sent randomly to 1500 students. Just returned 459 usable survey which was 210 surveys consisting of face-to-face students and 248 online students. The results of this study suggest that online students had less fraud than face to face students.

6. Research Methodology Clustering In Data Mining

Clustering is a data mining technique which makes significant clusters and widely used that they have similar characteristics and uses of automated techniques which are different classification techniques. (Zhou and Kapoor, 2011). Using clustering techniques, we can do the same cluster in a cluster and identify them with meaningful labels, so after that we detected fraud, we determined that this type of fraud is related to which of those students, faculty and hacker. (Etter, Cramer and Finn, 2007).

7. Modeling

At this stage, they were selected and used different modeling techniques and set their parameters with the optimal value, so as an example, there are methods, models and different techniques to build models, therefore, is often required rollback methods to the data production stage. The most important issue in the construction of the model is that, this is an iterative process, so we have to identify fraud; we need to build the best model to detect fraud. We learn in search of a suitable model can guide us to go back and dosome Changes in the data used and even improve speech issue. (Zhou and Kapoor, 2011).

8. Decision Tree Algorithms

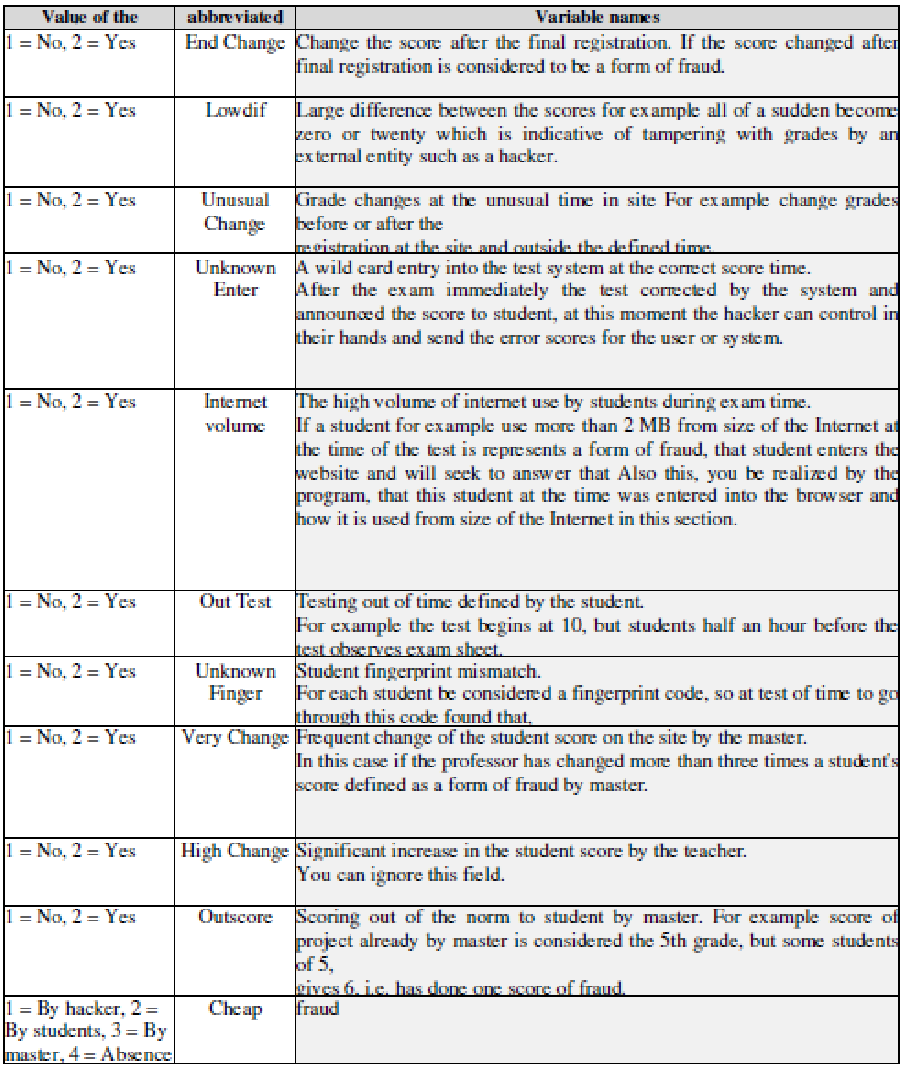
Decision tree based on a series of decision rules are used for prediction. In this study, we want to observed the result of these projections, so to do this act, it is very efficient to use of decision trees. In the decision tree the more important traits are transferred to the high decision tree nodes as well as decision trees puts aside minor characters. This allows, before entering other data mining techniques such as neural networks, we have a good view of the importance of features and can select the input algorithm consciously. (Etter, Cramer and Finn, 2007).

9.The Introduction Of Data And Variables

To evaluate the proposed idea, there is a need to use a set of training data sample that this set can include several important fields. For example, exam date, exam raw score, the date of appeal, score of appeals, the scores of students during the semester including the research, projects, midterms, also a former student information such as average scores of previous terms, the difference between the scores, the high volume of Internet usage during the test, a significant increase in the student scores and the things of this kind. These items can be received by a data set is ready via the web or to be taken from an educational institution or online University. Data used in this study is related to the data of 325 students which is extracted from the data warehouse, University of California Irvine, that variable "fraud" as the response variable (can be predicted) which has four parameters for detection fraud as follows:

1. By hacker
2. By students
3. By master
4. Absence of fraud and other variables which can be used as an explanatory variable (predictor).

Table 1: Variable names and their abbreviated titles



Value of the	abbreviated	Variable names
1 = No, 2 = Yes	End Change	Change the score after the final registration. If the score changed after final registration is considered to be a form of fraud.
1 = No, 2 = Yes	Lowdif	Large difference between the scores for example all of a sudden become zero or twenty which is indicative of tampering with grades by an external entity such as a hacker.
1 = No, 2 = Yes	Unusual Change	Grade changes at the unusual time in site For example change grades before or after the registration at the site and outside the defined time.
1 = No, 2 = Yes	Unknown Enter	A wild card entry into the test system at the correct score time. After the exam immediately the test corrected by the system and announced the score to student, at this moment the hacker can control in their hands and send the error scores for the user or system.
1 = No, 2 = Yes	Internet volume	The high volume of internet use by students during exam time. If a student for example use more than 2 MB from size of the Internet at the time of the test is represents a form of fraud, that student enters the website and will seek to answer that Also this, you be realized by the program, that this student at the time was entered into the browser and how it is used from size of the Internet in this section.
1 = No, 2 = Yes	Out Test	Testing out of time defined by the student. For example the test begins at 10, but students half an hour before the test observes exam sheet.
1 = No, 2 = Yes	Unknown Finger	Student fingerprint mismatch. For each student be considered a fingerprint code, so at test of time to go through this code found that.
1 = No, 2 = Yes	Very Change	Frequent change of the student score on the site by the master. In this case if the professor has changed more than three times a student's score defined as a form of fraud by master.
1 = No, 2 = Yes	High Change	Significant increase in the student score by the teacher. You can ignore this field.
1 = No, 2 = Yes	Outscore	Scoring out of the norm to student by master. For example score of project already by master is considered the 5th grade, but some students of 5, gives 6, i.e. has done one score of fraud.
1 = By hacker, 2 = By students, 3 = By master, 4 = Absence	Cheap	fraud

According to the results the Cheat variable giving a value of 1 to 4 and among them, a value of 1 is lowest frequency with 76 records and the value 2 which has the highest rate with 87 records.

Also, this variable does not have any missing values. This variable has a nominal and plays a role in the response variable (can be predicted) in the model. Likewise, other variables can be carried out statistical description.

10. Results Of Chayd Model

The Chayd model implementation process shows in Rapid Miner software in Figure 1.

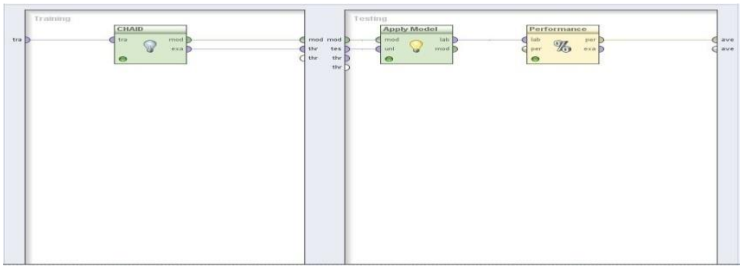


Figure 1: The process of implementing the Chayd model

After implementation Chayd model is created decision tree as shown in Figure 2.

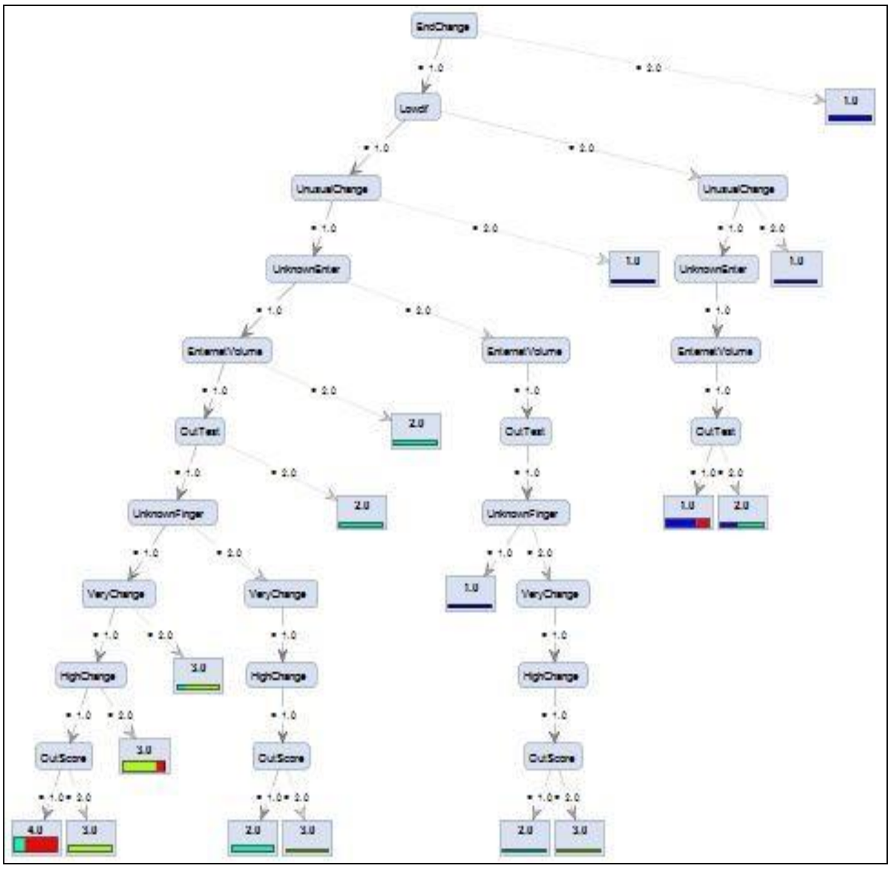
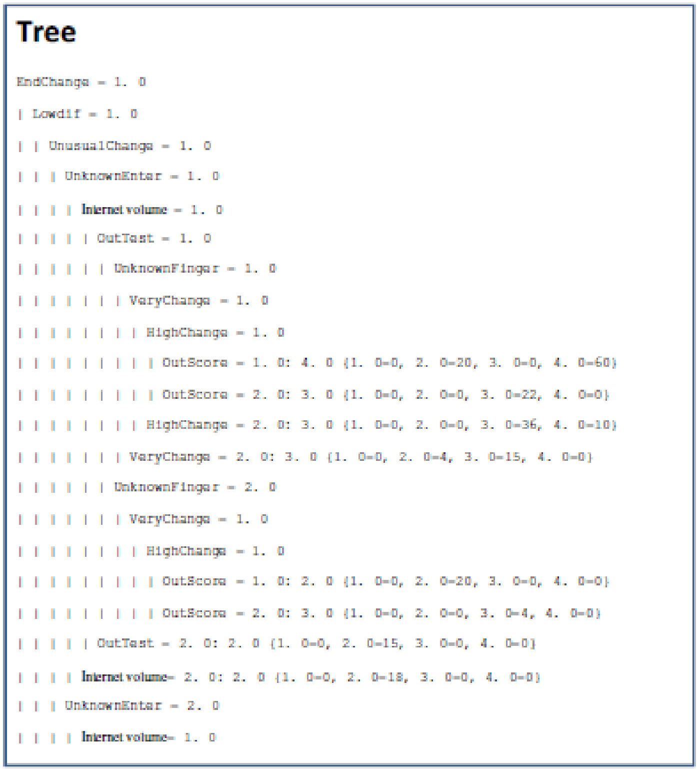


Figure 2: Decision Tree of resulting from Chayd model.

Showing created decision tree in Figure 2, shown as a rule in Figure 3.



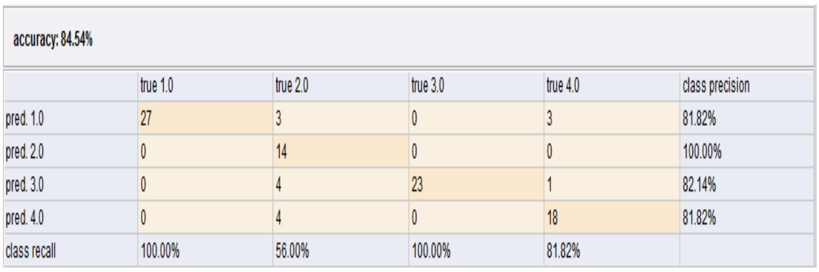
Tree

```
EndChange = 1. 0
| Lowdif = 1. 0
| | UnusualChange = 1. 0
| | | UnknownEnter = 1. 0
| | | | Internet volume = 1. 0
| | | | | OutTest = 1. 0
| | | | | | UnknownFinger = 1. 0
| | | | | | | VeryChange = 1. 0
| | | | | | | | HighChange = 1. 0
| | | | | | | | | OutScore = 1. 0: 4. 0 {1. 0=0, 2. 0=20, 3. 0=0, 4. 0=60}
| | | | | | | | | OutScore = 2. 0: 3. 0 {1. 0=0, 2. 0=0, 3. 0=22, 4. 0=0}
| | | | | | | | HighChange = 2. 0: 3. 0 {1. 0=0, 2. 0=0, 3. 0=36, 4. 0=10}
| | | | | | | VeryChange = 2. 0: 3. 0 {1. 0=0, 2. 0=4, 3. 0=15, 4. 0=0}
| | | | | | UnknownFinger = 2. 0
| | | | | | | VeryChange = 1. 0
| | | | | | | | HighChange = 1. 0
| | | | | | | | | OutScore = 1. 0: 2. 0 {1. 0=0, 2. 0=20, 3. 0=0, 4. 0=0}
| | | | | | | | | OutScore = 2. 0: 3. 0 {1. 0=0, 2. 0=0, 3. 0=4, 4. 0=0}
| | | | | OutTest = 2. 0: 2. 0 {1. 0=0, 2. 0=15, 3. 0=0, 4. 0=0}
| | | | Internet volume= 2. 0: 2. 0 {1. 0=0, 2. 0=18, 3. 0=0, 4. 0=0}
| | | UnknownEnter = 2. 0
| | | | Internet volume= 1. 0
```

Figure 3: Showing decision tree as a rule by resulting from implementation Chayd model

Finally, Chayd model by using the created tree doing a prediction operation for each of the records related to the test data set. The results of this prediction as shown in Table 2.

Table 2: Results of fraud variables predicted based on Chayd model



accuracy: 84.54%

	true 1.0	true 2.0	true 3.0	true 4.0	class precision
pred. 1.0	27	3	0	3	81.82%
pred. 2.0	0	14	0	0	100.00%
pred. 3.0	0	4	23	1	82.14%
pred. 4.0	0	4	0	18	81.82%
class recall	100.00%	56.00%	100.00%	81.82%	

According to the results of Table 2, we observe that Chayd model with 84.54 percent accuracy is done classification procedure for test data set. In other words, this model for 84.54% of the records contained in the test data set is correctly diagnosed the type of fraud. For the records that the type of its fraud is a type of hacker fraud (value 1) the model has done the procedure to predict with 100% accuracy, to records that the type of its fraud is a type of students fraud (value 2) the model has done the procedure to predict with 56% accuracy, to records that the type of its fraud is a type of Master fraud (value 3) the model has done the procedure to predict with 100% accuracy and for records that do not fraud (value 4) the model the model has done the procedure to predict with 81.82% accuracy.

Finally, can the results obtained from the Chayd model shown in Table 3.

Table 3: Results of the Chayd model

Absence of fraud	Master fraud	Students fraud	Hacker fraud	Model accuracy	Model
82.82%	100%	56%	100%	84.54%	Chayd

11. CONCLUSIONS AND RECOMMENDATIONS

Our goal in this research is to find out and choose the features between the all features which be reduced dimension of gene database, so that from losing important data. There are many and varied ways to reduce property, according to the research field and the characteristics of each method have chosen one of them or a combination of them in different cases. Data mining is a technique that can help to reduce these characteristics. The data used in this study included 325 samples which are discussed in the first place to describe the data and obtain an overview of the data. In this section presented a report from number of contained records in each class. Through this report, we found that class 1, devotes to its more records than any other class. Also we found that are not lost none of our data. But we have some outlier's data. In the third phase, we enter to data prepare or in other words, data preprocessing which was very time consuming. At first, outlier data with this strategy, we replaced with the closest data to it. Then, through the Chayd model, we calculate the accuracy of the model. The results showed that the Chayd model has a very high accuracy (84.54%) to detect fraud. Were excluded non-critical features and due to the variable our goal was to have four values 1, 2, 3, and 4 as follows:

1 - Detection of fraud made by the hackers.
2 - Detection of fraud made by the students.
3 - Detection of fraud made by the teachers.
4 - Detection of the Absence of the fraud.

From the available data considered 70% to build models and the remaining of 30%, as data to test the model that in the end, discussed to data modeling, after reduction feature by Chayd model. Based assessments and analysis, we have done our resarch, we managed that to recognize fraud taken by hackers, students and teachers by using existing Chayd model and data mining.

REFERENCES

[1] Bedford, W., Gregg, J., & Clinton, S. (2009). Implementing technology to prevent online cheating: A case study at a small southern university (SSRU). Merlot Journal of Online Learning and Teaching, 5(2).

[2] Etter, S Cramer, J. J. & Finn, S. (2007) "Origins of Academic Dishonesty: Ethical Orientations and Personality Factors Associated with Attitudes about Cheating with Information Technology", Journal of Research on Technology in Education, Vol 39, No. 2, pp 133-155.

[3] Lorenzetti, J. P. (2006) "Proctoring Assessments: Benefits & Challenges", Distance Education Report pp 5-6.

[4] Mirza, N. & Staples, E. (2010) "Webcam as a New Invigilation Method: Students' Comfort and Potential for Cheating", Journal of Nursing Education, Vol 49, No. 2, pp 116-119.

[5] Trenholm, Sven. "A review of cheating in fully asynchronous online courses: A math or fact-based course perspective." Journal of educational technology systems 35. 3 (2007): 281-300.

[6] Trenholm, S. (2005). Current issues in teaching mathematics online. Presentation given at CIT (Conference on Instructional Technologies), Binghamton, New York, May 26, 2005.

[7] Watson G. Sottile, J. (2010) "Cheating in the Digital age: Do Students Cheat More in Online Courses?" Online Journal of Distance Learning Administration, Vol 13, No 1

[8] Zhou W. , G. Kapoor, Detecting Evolutionary Financial Statement Fraud, Decision Support Systems,Vol.50(3),2011,pp.250-576.Changes in the data used and even improve speech issue. (Zhou and Kapoor, 2011).