ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 5, July 2012

Intrusion Detection and Classification Using Improved ID3 Algorithm of Data Mining

Sandeep Kumar
Dept. of Computer Engineering
Netaji Subhas Institute of Technology
New Delhi, India

Prof. Satbir Jain
Dept. of Computer Engineering
Netaji Subhas Institute of Technology
New Delhi, India

***Abstract*—Intrusion detection technology exists a lot of problems, such as low performance, low intelligent level, high false alarm rate, high false negative rate and so on. There is a need to develop some robust decision tree in order to produce effective decision rules from the attacked data. In this paper, ID3 decision tree classification method is used to build an effective decision tree for intrusion detection, then convert the decision tree into rules and save them into the knowledge base of intrusion detection system. These rules are used to judge whether the new network behavior is normal or abnormal. Experiments show that: the detection accuracy rate of intrusion detection algorithm based on ID3 decision tree is over 97%, and the process of constructing rules is easy to understand, so it is an effective method for intrusion detection. This paper introduces the use of ID3 algorithm of decision tree and we use Havrda and Charvat Entropy instead of Shannon Entropy. This decision tree evaluates less false positive and true negative alarm rates compare to existing algorithms. This Decision Tree helps in taking the better decision to analyze the data.**

***IndexTerms*—ID3 algorithm; intrusion detection; data mining; decision tree**

I. INTRODUCTION

As advances in computer network technology expand for communications and commerce in recent times, the rate of intrusions increase more than double every year. Intrusion detection is the process of identifying actions that attempt to compromise the confidentiality, integrity or availability of computers or networks. The use of data mining algorithms for detecting intrusions is now considered to build efficient and adaptive intrusion detection systems (IDS) that detect unauthorized activities of a computer system or network. IDS was first introduced by James P. Anderson in 1980 [1], and later in 1986, Dr. Dorothy Denning proposed several models for IDS based on statistics, Markov chains, time-series, etc [2]. Anomaly based intrusion detection using data mining algorithms such as decision tree (DT), naïve Bayesian classifier (NB), neural network (NN), support vector machine (SVM), k-nearest neighbors (KNN), fuzzy logic model, and genetic algorithm have been widely used by researchers to improve the performance of IDS [3]-[8]. However, today's commercially available IDS are signature based. Signature based IDS performs pattern matching techniques to match an attack pattern corresponding to known attack patterns in the database and produces very low false positives (FP), but it requires regular updates of rules or signatures and not capable of detecting unknown attacks. On the other hand, anomaly based IDS builds models of normal behavior and automatically detects anomalous behaviors. Anomaly detection techniques identify new types of intrusions as deviations from normal usage [9], but the drawback of these techniques is the rate of false positives (FP). The use of data mining algorithms for anomaly based IDS are to include an intelligent agent in the system that can detect the known and unknown attacks or intrusions.

Intrusion detection systems (IDS) gather and analyze information from a variety of systems and network sources for signs of intrusions. IDS can be host-based or network based systems. Host-based IDS located in servers to examine the internal interfaces and network-based IDS monitor the network traffics for detecting intrusions. Network-based IDS performs packet logging, real-time traffic analysis of IP network, and tries to discover if an intruder is attempting to break into the network. The major functions performed by IDS are[21]: (1) monitoring users and systems activity, (2) auditing system configuration, (3) assessing the data files, (4) recognizing known attacks, (5) identifying abnormal activities, (6) managing audit data, (7) highlighting normal activities, (8) correcting system configuration errors, and (9) stores information about intruders. A variety of IDS have been employed for protecting computers and networks in last decades, but still there some issues that should be consider in the current IDS like low detection accuracy, unbalanced detection rates for different types of attacks, and high false positives. In this paper, we proposed a new decision tree based learning algorithm for classifying different types of network attacks, which improves the detection rates (DR) and reduces false positives (FP) using KDD99 benchmark network intrusion detection dataset in comparison with other existing methods.

II. OVERVIEW OF DECISION TREE TECHNOLOGY

Decision tree technology is a common and fast classification method. Its construction process is top-down, divide-and-rule. Essentially it is a greedy algorithm. Starting from root node, for each non-leaf node, firstly choose an attribute to test the sample set; Secondly divide training sample set into several sub-sample sets according to testing results ,each sub-sample set constitutes a new leaf node; Thirdly repeat the above division

process, until having reached specific end conditions. In the process of constructing decision tree, selecting testing attribute and how to divide sample set are very crucial. Different decision tree algorithm uses different technology. In practice, because the size of training sample set is usually large, the branches and layers of generated tree are also more. In addition, abnormity and noise existed in training sample set will also cause some abnormal branches, so we need to prune decision tree. One of the greatest advantages of decision tree classification algorithm is that: It does not require users to know a lot of background knowledge in the learning process. As long as training samples can be expressed as the form of attribute-conclusion, you can use this algorithm to study. But decision tree technology also has a lot of deficiency, such as: When there are too many categories, classification accuracy is significantly reduced; It is difficult to find rules based on the combination of several variables. At present, there are a lot of decision algorithms, such as: ID3, SLIQ, CART, CHAID and so on. But ID3 algorithm [10] is the most representative and widely used. It is proposed by Quinlan in 1993.

III. PROPOSED APPROACH

The decision tree (DT) is very powerful and popular data mining algorithm for decision-making and classification problems. It has been using in many real life applications like medical diagnosis, radar signal classification, weather prediction, credit approval, and fraud detection etc. DT can be constructed from large volume of dataset with many attributes, because the tree size is independent of the dataset size. A decision tree has three main components: nodes, leaves, and edges. Each node is labeled with an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to possible values of the attribute. An edge connects either two nodes or a node and a leaf. Leaves are labeled with a decision value for categorization of the data. To make a decision using a decision Tree, start at the root node and follow the tree down the branches until a leaf node representing the class is reached. Each decision tree represents a rule set, which categorizes data according to the attributes of dataset. The decision tree building algorithms may initially build the tree and then prune it for more effective classification. With pruning technique, portions of the tree may be removed or combined to reduce the overall size of the tree. The time and space complexity of constructing a decision tree depends on the size of the data set, the number of attributes in the data set, and the shape of the resulting tree. Decision trees are used to classify data with common attributes. The ID3 algorithm builds decision tree using information theory, which choose splitting attributes from a data set with the highest information gain [11]. The amount of information associated with an attribute value is related to the probability of occurrence. The concept used to quantify information is called entropy, which is used to measure the amount of randomness from a data set. When all data in a set belong to a single class, there is no uncertainty, and then the entropy is zero. The objective of decision tree classification is to iteratively partition the given data set into subsets where all elements in each final subset belong to the same class. The entropy calculation is shown in equation 1. Given probabilities p1, p2,..,ps for different classes in the data set

$$Entropy: H(p1,p2,...ps) = \sum_{i=1}^{n}(pi\ log(1/pi)) \quad (1)$$

Given a data set, D, H(D) finds the amount of entropy in class based subsets of the data set. When that subset is split into s new subsets S = {D1, D2,…,Ds} using some attribute, we can again look at the entropy of those subsets. A subset of data set is completely ordered and does not need any further split if all examples in it belong to the same class. The ID3 algorithm calculates the information gain of a split by using equation 2 and chooses that split which provides maximum information gain.

$$Gain\ (D,S) = H(D) - \sum_{i=1}^{s} p(Di)H(Di) \quad (2)$$

Here, Shannon Entropy has been used in ID3 algorithm to calculate the Information Gain contained by data, which helps to make Decision Tree. However, the results obtained from Shannon Entropy, are rather complex, have more numbers of node and leaf and Decision Rules. Thus it makes the decision making process time consuming. Therefore, to minimize these problems, new algorithm has been proposed by modifying ID3 algorithm using Havrda and Charvat Entropy instead of Shannon Entropy

A . Proposed Algorithm using Havrda And Charvat Entropy

Classification is a mapping of the database to the set of classes. Each tuple in the database is assigned to exactly one class. The classes that exist for a classification problem are indeed equivalence classes. In actuality, the problem usually is implemented in two phases:
1) Create a specific model by evaluating the training data. This step takes the training data as input and gives the output as the definition of the developed model. The developed model classifies the training data as accurate as possible.
2) Apply the established model in step 1 by classifying tuples from the target database.

The ID3 algorithm works by recursively applying the procedure above to each of the subsets produced until "pure" nodes are found a pure node contains elements of only one class or until there are no attributes left to consider. It can be stated in pseudo code, as is shown in Figure 1.

```
function ID3 (I, 0, T) {
/* I is the set of input attributes
* O is the output attribute
* T is a set of training data
*
* function ID3 returns a decision tree
*/
if (T is empty) {
```

return a single node with the value "Failure";
}
if (all records in T have the same value for O) {
return a single node with that value;
}
if (I is empty) {
return a single node with the value of the most frequent value of O in T;
/* Note: some elements in this node will be incorrectly classified */
}
/* now handle the case where we can't return a single node */
compute the information gain for each attribute in I relative to T;
let X be the attribute with largest Gain(X, T) of the attributes in I;
let {x_j| j=1,2, .., m} be the values of X;
let {T_j| j=1,2, .., m} be the subsets of T when T is partitioned according the value of X;
return a tree with the root node labeled X and arcs labeled x_1, x_2, .., x_m, where the arcs go to the trees ID3(I-{X}, O, T_1), ID3(I-{X}, O, T_2), .., ID3(I-{X}, O, T_m);
}

Figure 1: The ID3 algorithm

B. Definition and Role of Havrda and Charvat Entropy

So far, we have only considered Shannon's entropy. However, many measures of entropies have been introduced in the literature to generalize Shannon's entropy, e.g. Renyi's entropy [12], Kapur's entropy [13],and Havrda-Charvat's structural entropy [14]. We are particularly interested in the Havrda-Charvat's structural entropy for reasons that will be clear later. The structural entropy is defined as entropy measure by formula shown under :

$$H(X) = (2^{1-\propto} - 1)^{-1} (\sum_x p^{\propto}(x) - 1)$$

Where $\alpha > 0$ and $\alpha \neq 1$.
This formula calculates Entropy. To avoid deduced solution in decision tree making process, Havrda and Charvat entropy based ID3 algorithm is proposed which gives good solution in reasonable time

IV. EXPERIMENTAL ANALYSIS

A. Intrusion Detection Dataset

DARPA in concert with Lincoln Laboratory at MIT launched the DARPA 1998 dataset for evaluating IDS. The refined version of DARPA dataset which contains only network data (i.e. tcpdump data) is termed as KDD dataset. The Third International Knowledge Discovery and Data Mining Tools Competition were held in colligation with KDD-99 dataset. KDD training dataset consists in single connection vectors where each single connection vectors consists of 41 features and is marked as either normal or an attack, with exactly one particular attack type [15]. The Experimental data comes from KDD CUP 1999 dataset [16]. It is test set widely used in Intrusion detection field. It includes about 4.9 million simulative attack records and 22 types of attack. Because the entire data set is to large. In KDD99 dataset, each example represents attribute values of a class in the network data flow, and each class is labeled either normal or attack. The classes in KDD99 dataset categorized into five main classes (one normal class and four main intrusion classes: probe, DOS, U2R, and R2L).In KDD99 dataset these four attack classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that tabulated in Table I.

TABLE I : Different Types Of Attacks in KDD99 DATA SET

4 Main Attack Classes	22 Attack Classes
Denial of Service (DoS)	back, land, neptune, pod, smurt, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Probing	ipsweep, nmap, portsweep, satan

1) Normal connections are generated by simulated daily user behavior such as downloading files, visiting web pages.

2) Denial of Service (DoS) : is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. DoS attacks are classified based on the services that an attacker renders unavailable to legitimate users like apache2, land, mail bomb, back, etc.

3) Remote to User (R2L) occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine. which include sendmail and Xlock.

4) User to Root (U2R) is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system. Most common exploits of U2R attacks are regular buffer overflows, load-module, Fd-format and Ffb-config.

5) Probing (Probe) is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

There are 41 input attributes in KDD99 dataset for each network connection that have either discrete or continuous values and divided into three groups. The first group of attributes is the basic features of network connection, which include the duration, prototype, service, number of bytes from

source IP addresses or from destination IP addresses, and some flags in TCP connections. The second group of attributes in KDD99 is composed of the content features of network connections and the third group is composed of the statistical features that are computed either by a time window or a window of certain kind of connections.

Table II : KDD Cup 99 Dataset 41 Features

#	Feature Name	#	Feature Name
1	Duration	22	is_guest_login
2	protocol_type	23	Count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	Urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_ rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_ port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound cmds	41	dst_host_srv_rerror_ rate
21	is_hot_login		

B. Experimental & Result Analysis

This section describes the experimental results and performance evaluation of the proposed system. The proposed system is implemented in MATLAB and the performance of the system is evaluated By Accuracy and Error of classification. For experimental evaluation, we have taken KDD cup 99 dataset [16], which is mostly used for evaluating the performance of the intrusion detection system. For evaluating the performance, it is very difficult to execute the proposed system on the KDD cup 99 dataset since it is a large scale. Here, we have used subset of 10% of KDD Cup 99 dataset for training and testing.

The input to the proposed system is KDD Cup 1999 dataset, which is divided into two subsets such as, training dataset and testing dataset. At first, the training dataset is classified into five subsets so that, four types of attacks DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root), Probe) and normal data are separated.

We have divided the behavior of user into two classes namely attack and normal, where the behavior of user is the collection of different attacks belonging to the five classes as explained in table1. The aim of our Decision Tree experiment is to differentiate between normal and attack behavior of user. In our experiments normal data are classified and all attacks are classified. The training dataset contains normal data as well as four types of attacks, which are given to the proposed system for identifying the suitable attributes. Then, using the Decision Tree learning strategy, the system generates definite and indefinite rules and finally, Tree generated from the definite rules. In the testing phase, the testing dataset is given to the proposed system, which classifies the input as a normal or attack. The obtained result is then used to compute overall accuracy of the proposed system. The overall accuracy of the proposed system is computed based on the correct classified instance are normally used to estimate the rare class prediction.

Our experiments show the detection rate of correct classified instance, False Positive Rate and Error Rate is misclassified instance by using 14 features of KDD Dataset to Decision Tree algorithm. The accuracy of proposed system is 97.74 % and Error Rate is 2.81%.

Following fundamental formulas are used to estimate the performance of the system: Detection rate (DR) and Error Rate (ER).

$$\text{Detection rate} = \frac{\text{Total no.of correct classified instance} * 100}{\text{Total no.of instance}}$$

$$\text{Error rate} = \frac{\text{Total no.of misclassified instances} * 100}{\text{Total no.of instance}}$$

Fig 2. Shows the detection rate of different dataset. Proposed method has 98 % accuracy.

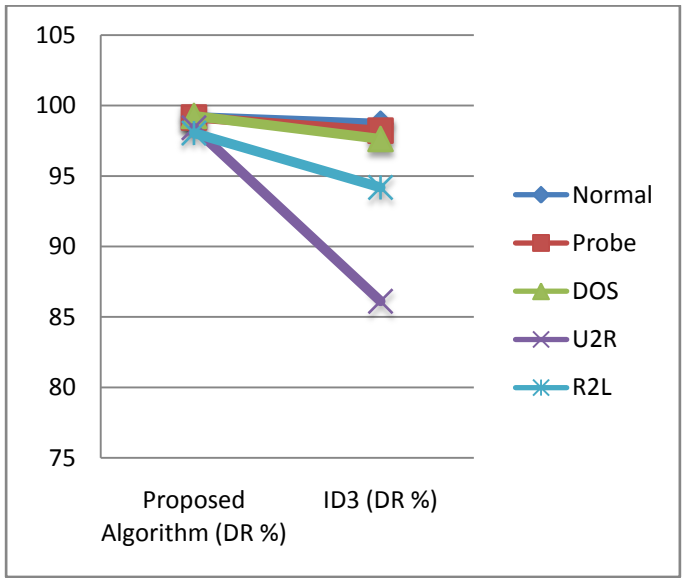

Fig 2.Analysis of Detection Rate

For a good IDS FPR and Error rate should be low Fig 3, 4. shows that the FPR and Error rate of proposed algorithm is lower as compare to ID3. Proposed algorithm based on decision tree, can not only help people understand intrusion rules, but also can greatly improve the accuracy of intrusion

detection algorithm. At the same time, the decision tree can be converted into rules, then add them into the knowledge base of intrusion detection system. These rules are used to judge whether the new network behavior is normal or abnormal.

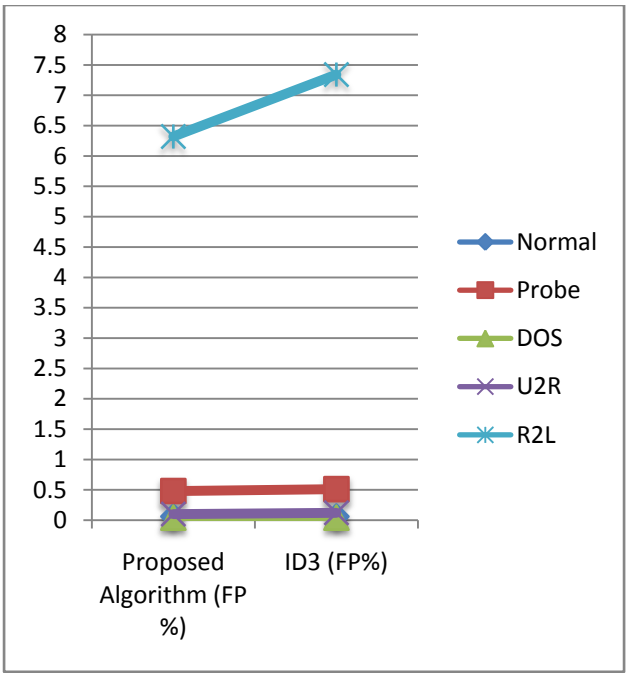

Fig 3.Analysis of False Positive Rate

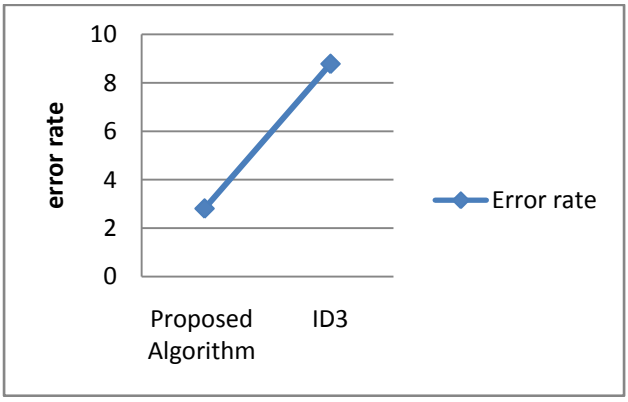

Fig 4. Analysis of Error Rate

IV. Conclusion

In this paper, we propose an intrusion detection algorithm based on ID3 decision tree. In the process of constructing intrusion rules, information gain ratio is used with Harvard and charvat entropy in place of shanon entropy . The experiment result shows that: The intrusion detection algorithm based on ID3decision tree is feasible and effective, and has a high accuracy rate. How to detect intrusion behavior and new intrusion behavior as much as possible will be the focus of next work.

References

[1] James P. Anderson, "Computer security threat monitoring and surveillance," Technical Report 98-17, James P. Anderson Co., Fort Washington, Pennsylvania, USA, April 1980.

[2] Dorothy E. Denning, "An intrusion detection model," IEEE Transaction on Software Engineering, SE-13(2), 1987, pp. 222-232.

[3] Barbara, Daniel, Couto, Julia, Jajodia, Sushil, Popyack, Leonard, Wu, and Ningning, "ADAM: Detecting intrusion by data mining," IEEE Workshop on Information Assurance and Security, West Point, New York, June 5-6, 2001.

[4] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. Decision trees in intrusion detection systems," In Proc. of 2004ACM Symposium on Applied Computing, 2004, pp. 420-424.

[5] Mukkamala S., Janoski G., and Sung A.H., "Intrusion detection using neural networks and support vector machines," In Proc. of the IEEE international Joint Conference on Neural Networks, 2002, pp.1702-1707.

[6] J. Luo, and S.M. Bridges, "Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection," International Journal of Intelligent Systems, John Wiley & Sons, vol. 15, no. 8, 2000, pp. 687-703.

[7] YU Yan, and Huang Hao, "An ensemble approach to intrusion Detection based on improved multi-objective genetic algorithm," Journal of Software, vol. 18, no. 6, June 2007, pp. 1369-1378.

[8] Shon T., Seo J., and Moon J., "SVM approach with a genetic algorithm for network intrusion detection," In Proc. of 20th International Symposium on Computer and Information Sciences (ISCIS 2005), Berlin: Springer-Verlag, 2005, pp. 224-233.

[9] Lazarevic, A., Ertoz, L., Kumar, V., Ozgur,. A., Srivastava, and J., "A comparative study of anomaly detection schemes in network intrusion detection," In Proc. of the SIAM Conference on Data Mining, 2003.

[10] J.R.Quinlan, "C4.5:Programs for Machine Learning[J],"NewYork: Morgan Kaufman,1993

[11] J. R. Quinlan, "Induction of Decision Tree," Machine Learning Vol. 1, pp. 81-106, 1986.

[12] A. Renyi. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, volume 1, pages 547{561. University of California Press, 1961.

[13] J. N. Kapur. Generalised entropy of order and type. The Mathematics Seminar, 4:78{94, 1967.

[14] J. Havrda and F. Charvat. Quantication method of classication processes: Concept of structural-entropy. Kybernetika, 3:30{35, 1967.

[15] R. Shanmugavadivu, Dr. N. Nagrajan "Network Intrusion Detection System Using fuzzy Logic" Indian Journal of computer Science and Engineering.

[16] http://kdd.ics.uci.edu/database/kddcup99/kddcup99.ml.

[17] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. decision trees in intrusion detection systems," In Proc. of 2004 ACM Symposium on Applied Computing, 2004, pp. 420-424.

[18] M.L. Shyu, S.C. Chen, K. Sarinnapakorn, L. Chang, "A novel anomaly detection scheme based on principal component classifier," in Proc. Of the IEEE Foundations and New Directions of Data Mining Workshop, Melbourne, FL, USA, 2003, pp. 172-179.

[19] T. Chen, B. C. Vemuri, A. Rangarajan, S. J. Eisenschenk, Group-Wise Point-Set Registration Using a Novel CDF-Based Havrda-Charvát Divergence.

[20] Mohammadreza Ektefa, Sara Memar, Fatimah sidi , Lilly Suriani Affendey" Intrusion Detection using Data Mining Technique"IEEE 2010.

[21] Dewan Md. Farid, Nouria Harbi, Emna Bahri, Mohammad ZahidurRahman, Chowdhury Mofizur Rahman "Attacks Classification in Adaptive Intrusion Detection using Decision Tree" 2010.