Journal of Mathematical Modelling and Applied Computing

Volume No. 12

Issue No. 3

September - December 2024

ENRICHED PUBLICATIONS PVT.LTD

JE - 18,Gupta Colony, Khirki Extn,
Malviya Nagar, New Delhi - 110017.
E- Mail: info@enrichedpublication.com
Phone :- +91-8877340707

Journal of Mathematical Modelling and Applied Computing

Aims and Scope

The Journal of Mathematical Modelling and Applied Computing is an Indian research journal, which publishes top-level original and review papers, short communications and proceedings on Interdisciplinary Integrative Forum on Modelling, Simulation and Scientific Computing in Engineering, Physical, Chemical Biological, Medical, Environmental, Social, Economic and Other Systems using Applied Mathematics and Computational Sciences and Technology.

Journal of Mathematical Modelling and Applied Computing

Journal of Mathematical Modelling and Applied Computing

(Volume No. 12, Issue No. 3, September - December 2024)

Contents

Id3 Algorithm in Data Mining Applied to Diabetes Database

Dr. R. Jamuna
Professor, Department of Computer Science S. R. College,
Bharathidasan university, Trichy.

ABSTRACT

The past decade has seen a flurry of promising breakthroughs in data mining predictions. Many of these developments hold the potential to prevent dreadful diseases to improve the quality of life. The advances in medical science and technology have corresponded to the use of computer algorithms as an intermediary between the medical researchers and technocrats. Diabetes Mellitus is a major killer disease of mankind today. Data mining techniques can be used to highlight the significant factors causing such a disorder. Even though total cure is not possible for this pancreatic disorder the complications can be avoided by awareness using data mining algorithms. In this paper, eight major factors playing significant role in the Pima Indian population are analyzed. Real time data is taken from the dataset of National Institute of Diabetes and Digestive and Kidney Diseases. The data is subjected to an analysis by logistic regression method using spss 7.5 statistical software, to isolate the most significant factors. Then the significant factors are further applied to decision tree technique called the Iterative Dichotomiser-3 algorithm which leads to significant conclusions. Conglomeration of data mining techniques and medical research can lead to life saving conclusions useful for the physicians.

Keywords: BMI, Diabetes, decision tree, logistic regression, plasma.

INTRODUCTION

Diabetes Mellitus is a major killer disease of mankind today. Data mining techniques can be used to highlight the significant factors causing such a disorder. Even though total cure is not possible for this pancreatic disorder the complications can be avoided by awareness about the factors playing major role in the cause of this disorder using data mining algorithms. In this paper, eight major factors, Prg (No. of times pregnant), Plasma (Plasma glucose concentration in Salvia), BP (Diastolic blood pressure), Thick (Forceps skin fold thickness), Insulin (Two hours serum insulin), Body (Body Mass Index; weight/height), Pedigree (Diabetes pedigree function), Age (in years), Response (1: Diabetic 0: Non-Diabetic), playing significant role in the Pima Indian population are analyzed. Real time data is taken from the large dataset of http://www.niddk.nih.gov/ which is the home page for the National Institute of Diabetes and Digestive and Kidney Diseases. First the data is sampled by eliminating any record which has a zero value for any field from the total real time data base. Next the data is subjected to an analysis by Logistic regression method by using spss 7.5 statistical software to show the most significant factors among the eight factors taken.

Then the significant factors are applied to a Iterative Dichotomiser-3 algorithm which generates Decision Trees using Shannon Entropy for further investigations. Decision tree technique called the ID3 algorithm of data mining leads to significant conclusions about this diabetes disorder which poses to be the greatest threat to mankind in the coming era. Conglomeration of data mining techniques and medical data base research can lead to life saving conclusions for the physicians at critical times to save the mankind.

1. METHODSOFBUILDINGDECISIONTREESINDATAMINING

In data mining, a decision tree is a predictive model; that is, a mapping of observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification tree or reduction tree. In these tree structures, leaves represent [1] classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or decision trees. In decision theory and decision analysis, a decision tree is a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It can be used to create a plan to reach a goal. Decision trees are constructed in order to help with making decisions [2]. A decision tree is a special form of tree structure and a descriptive means for calculating conditional probabilities.

Decision tree learning is a common method used in data mining. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents a possible value of target variable given the values of the variables represented by the path from the root. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate. In data mining [3], trees can be described also as the synergy of mathematical and computing techniques that aids on the description, categorization and generalization of a given set of data. Data comes in records of the form:

$(X_i, y) = (x_1, x_2, x_3 ... x_k, y)$

The dependent variable, y, is the variable that we are trying to understand, classify or generalize. The other variables x_1, x_2, x_3 etc. are the variables that will help us for predictions.

1.1. Decision Trees in Data mining [10]

- An internal node is atest onan attribute.
- A branch represents an outcome of the test.
- A leaf node represents a class label.
- At each node, one attribute is chosen to split training examples into distinct classes.
- A new case is classified by following a matching path to a leaf node.

1.2. Types of building Decision Trees [4]

- Top-down tree construction
- At start, all training examples are at the root.
- Partition the examples recursively by choosing one attribute each time like age, Pdf etc…
- Bottom-up tree pruning
- Remove sub trees or branches, in a bottom-up manner, to improve the estimated accuracy on new cases.

1.3. Choosing the Splitting Attribute

- At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A Goodness function is used for this purpose.
- Typical goodness functions:
- Information Gain (Id3)
- Information Gain Ratio
- Gini Index

Take all unused attributes and count their entropy concerning test samples.

- Choose attribute for which entropy is minimum.
- Make node containing that attribute.

2. PRINCIPLESUSEDINTHEANALYSISOFLARGEDATASETS

2.1. Information Entropy by Claude Shannon

In information theory, the Shannon entropy [7] or information entropy is a measure of the uncertainty associated with a random variable. It can be interpreted as the average shortest message length, in bits, that can be sent to communicate the true value of the random variable to a recipient. This represents a fundamental mathematical limit on the best possible lossless data compression of any communication: the shortest average number of bits that can be sent to communicate one message out of all the possibilities is the Shannon entropy. [8]

Information is measured as follows:-

- Given a probability distribution, the information required to predict an event is the distribution's entropy.
- Entropy gives the information required in bits.

Formula for computing the entropy:

ShannonEntropy $(p_1, p_2, ..., p_n) = -p_1 \log p_1 - p_2 \log p_2 \ ... - p_n \log p_n$

2.2 Definition of Information Entropy

The information entropy of a discrete random variable X, that can take the range of possible values {x1... xn} is defined to be,

$$H(X)=E(I(X))=\sum_{i=1}^{n} p(x_i)\log_2(1/p(x_i))$$

$$=\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

I(X) is the information content or self-information of X, which is itself a random variable; and $p(x_i) = P(X = x_i)$ is the probability mass function of X.

Introduced by Claude Shannon in 1948, ID3 (Iterative Dichotomiser-3) is an algorithm used to generate a decision tree. However, it does not always produce the smallest tree, and is therefore a heuristic. Occam's razor is formalized using the concept of information entropy as:

$$H(X)=-\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

2.2.1 Information Gain

- Uses Shannon Entropy
- IG calculates effective change in entropy after making a decision based on the value of an attribute.
- For decision trees, it's ideal to base decisions on the attribute that provides the largest change in entropy, the attribute with the highest gain.

2.3. Introduction to ID3 Algorithm for Diabetes Database

Id3 begins by choosing a random subset of the training instances. This subset is called the window. The procedure builds a decision tree that correctly classifies all instances in the window. The tree is then tested on the training instances outside the window. If all the instances are classified correctly then the procedure halts. Otherwise it adds some of the instances incorrectly classified to the window and repeats the process. This iterative strategy is empirically more efficient than considering all instances at once. In building a decision tree ID3 selects the feature which minimizes the entropy function given below and thus best discriminates among the training instances. Data have been collected from about 768 Indian Origin females who were tested for the presence of diabetes mellitus of which 268 were found to be positive. Sample of 336 records are selected deleting the record sets with zero values.

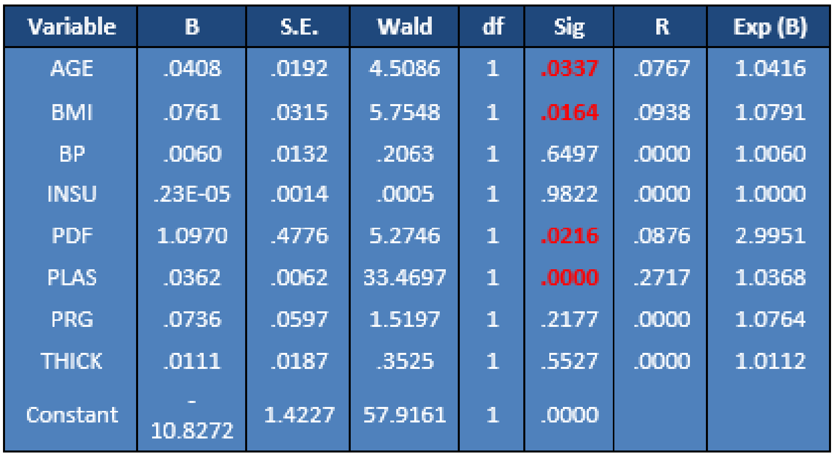
3. RESULTS OF SOFTWARE BASED ANALYSIS OF DATASET.

3.1. Logistic Regression Outputs from SPSS7.5

Logistic regression method was applied to bring out the significance factors like age, obesity, etc. in the cause of the diabetes disorder, in Pima Indian diabetes database using SPSS 7.5 software. These factors are fuzzified to form a sample decision tree by ID3 algorithm.

Table 1: Logistic Regression Outputs from SPSS 7.5 [11]

------------- Variables in the Equation --------------



Variable	B	S.E.	Wald	df	Sig	R	Exp (B)
AGE	.0408	.0192	4.5086	1	.0337	.0767	1.0416
BMI	.0761	.0315	5.7548	1	.0164	.0938	1.0791
BP	.0060	.0132	.2063	1	.6497	.0000	1.0060
INSU	.23E-05	.0014	.0005	1	.9822	.0000	1.0000
PDF	1.0970	.4776	5.2746	1	.0216	.0876	2.9951
PLAS	.0362	.0062	33.4697	1	.0000	.2717	1.0368
PRG	.0736	.0597	1.5197	1	.2177	.0000	1.0764
THICK	.0111	.0187	.3525	1	.5527	.0000	1.0112
Constant	-10.8272	1.4227	57.9161	1	.0000		

1. Results of Logistic Regression SPSS 7.5 version.

Total number of cases :	336
(Un-weighted)	
Number of selected cases :	336
Number of unselected cases :	0

Dependent Variable Encoding:

Original value	Internal value
0.00	0
1.00	1

Dependent Variable : Response

Beginning Block Number 0.

Initial Log Likelihood Function -2 Log Likelihood 426.33781

* Constant is included in the model.

```
Beginning Block Number 1.
Method: Enter variable(s) Entered on Step Number
1.      Age          BMI          Bp
        Insulin      PDF          Plasma
        Prg          Thick

Estimation terminated at iteration number 4 because Log Likelihood decreased by less than .01 percent.
2.      Log          Likelihood   288.920
        Goodness of Fit          351.718
        Cox & Snell - R^2        .336
        Nagelkerke - R^2         .467
        Chi-Square   df     Significance
Model   137.417      8      .0000
Block   137.417      8      .0000
Step    137.417      8      .0000

Classification Table for Response
The Cut Value is .50
Predicted
.00     1.00    Percent Correct
        0       I       1
Observed        +-------+      +
.00     0       I 200  I25  I   88.89%
                +-------+----------+
1.00    1       I      45 I   66  I   59.46%
                +---------+--------+
                        Overall      79.17%
```

From the observations of Table [1.1], we find that the following factors playing significant role in the cause of diabetes.

3.2. Analysis of Logistic Regression Results

- Age: sig = .0337 so 97% confidence level.
- Body Mass Index: sig = .0164 = .02 so 98% confidence level.
- PDF (Diabetes Pedigree Function): sig=.0216 so 98% confidence level. Implication of Hereditary Nature in the disease.

- Plasma (Glucose Concentration in Saliva): sig = .0000 100% confidence level as shown in Fig[1.1]

Table 2: Sample Dataset from Pima Indian diabetes database [5]

PATIENT	AGE	BMI	PLASMA	PDF	DIABETIC/ NOT
P1	YOUNG	HIGH	MEDIUM	LOW	NO
P2	YOUNG	HIGH	LOW	MEDIUM	NO
P3	YOUNG	NORMAL	MEDIUM	HIGH	YES
P4	MIDDLE	HIGH	HIGH	HIGH	YES
P5	OLD	HIGH	MEDIUM	HIGH	YES
P6	MIDDLE	HIGH	LOW	HIGH	NO
P7	OLD	NORMAL	HIGH	HIGH	NO
P8	OLD	HIGH	HIGH	HIGH	NO
P9	MIDDLE	NORMAL	LOW	HIGH	YES
P10	OLD	HIGH	MEDIUM	HIGH	YES
P11	YOUNG	HIGH	LOW	LOW	NO
P12	YOUNG	HIGH	MEDIUM	MEDIUM	NO
P13	MIDDLE	HIGH	HIGH	HIGH	NO
P14	YOUNG	NORMAL	MEDIUM	MEDIUM	NO
P15	MIDDLE	HIGH	HIGH	HIGH	NO

Table 3: Sample Fuzzified Range of Dataset. [5]

FUZZIFICATIONOF OF DATA SET-RANGE			
Age youth:21-38 middle:39 -81 large:>81	BMI: small:0-33 medium:34-52 large:>52	Plasma Glucose: small:128-158 medium:159-197 large:>197	Pdf: small:0.12-0.54 medium:0.54-2.33 large:>2.33

3.3 The ID3 Algorithm Applied to Diabetes Database [6]

1. Select a random subset W (called the "window") from the training set. Build a decision tree for the current window. Select the best feature which minimizes the entropy function H:

$H = \Sigma -p_i \log p_i$(optimal values are available and the optimum entropy may be found by discrete probabilistic methods)

Where pi is the probability associated with i^{th} class. The entropy is calculated for each value. The sum of the entropy is calculated for each value. The sum of the entropy weighted by the probability of each value is the entropy for the feature. Categorize training instances into subsets by this feature. Repeat this process recursively until each subset contains instances of one kind (class) or some statistical criterion is satisfied.

2. Scan the entire training set for exceptions to the decision tree.

3. If exceptions are found, insert some of them into W and repeat from Step 2. The insertion may be done either by replacing some of the existing instances in the window or by augmenting it with the new exceptions. In practice a statistical criterion can be applied to stop the tree from growing as long as most of the instances are classified correctly. Fig [1.2]

3.3.1. ID3 ALGORITHM [9]

- Establish Classification Attribute as in Table [1.2].
- Compute Classification Entropy.
- For each attribute in R, calculate Information Gain using classification attribute.
- Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
- Remove Node Attribute, creating reduced table RS.
- Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced.

3.3.2 Establishing a Target Classification

Is the test patient diabetic?

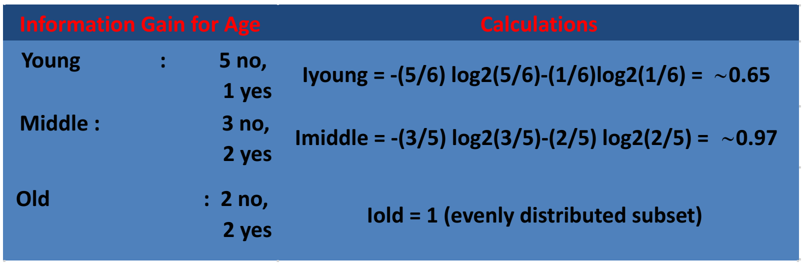
- 5/15 yes, 10/15 no
- Calculating for the Classification Entropy

I_E= -(5/15)$\log_2$(5/15)-(10/15)$\log_2$(10/15) = ~0.918

3.3.3 Example – Information Gain for Age

- Age: 6 Young, 5 Middle, 4 Old
- 3 values for the attribute age, so we need 3 entropy calculations.

Table 4. Information Gain for Age



Information Gain for Age		Calculations
Young :	5 no, 1 yes	Iyoung = -(5/6) log2(5/6)-(1/6)log2(1/6) = ~0.65
Middle :	3 no, 2 yes	Imiddle = -(3/5) log2(3/5)-(2/5) log2(2/5) = ~0.97
Old :	2 no, 2 yes	Iold = 1 (evenly distributed subset)

$$IG_{Age} = IE(S) - \left[(6/15)*I_{young} + (5/15)*I_{middle} + (4/15)*I_{old}\right]$$

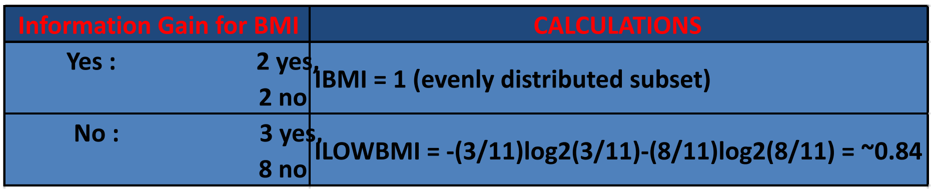
$$IG_{Age} = 0.918 - 0.85 = 0.068$$

We must calculate Information Gain of remaining attributes to determine the root node.

3.3.4 Example - Information Gain for BMI

- 4 yes, 11 no
- 2 values for the attribute BMI, so we need 2 entropy calculations.

Table 5. Information Gain for BMI



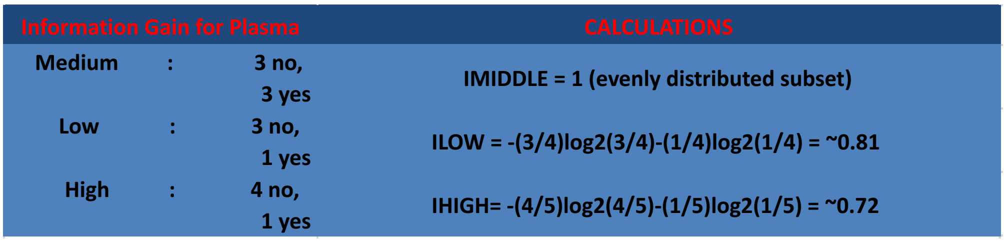
Information Gain for BMI	CALCULATIONS
Yes : 2 yes, 2 no	IBMI = 1 (evenly distributed subset)
No : 3 yes, 8 no	ILOWBMI = -(3/11)log2(3/11)-(8/11)log2(8/11) = ~0.84

$$IG_{BMI} = IE(S) - [(4/15*)I_{BMI} + (11/15)*I_{LowBMI}]$$
$$IG_{BMI} = 0.918 - 0.886 = 0.032$$

3.3.4 Example – Information Gain for Plasma

- 6 Middle, 4 Low, 5 High
- 3 values for attribute Plasma, so we need 3 entropy calculations

Table 6. Information Gain for Plasma



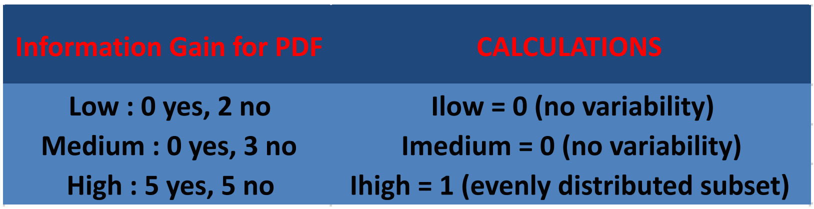
Information Gain for Plasma	CALCULATIONS
Medium : 3 no, 3 yes	IMIDDLE = 1 (evenly distributed subset)
Low : 3 no, 1 yes	ILOW = -(3/4)log2(3/4)-(1/4)log2(1/4) = ~0.81
High : 4 no, 1 yes	IHIGH= -(4/5)log2(4/5)-(1/5)log2(1/5) = ~0.72

$$IG_{Plasma} = IE(S) - [(6/15)*I_{Medium} + (4/15)*I_{Low} + (5/15)*I_{High}]$$
$$IG_{Plasma} = 0.918 - 0.856 = 0.062$$

3.3.4 Example – Information Gain for PDF

- PDF: 2 Low, 3 Medium, 10 High
- 3 values for attribute PDF, so we need 3 entropy calculations.

Table 7. Information Gain for PDF



Information Gain for PDF	CALCULATIONS
Low : 0 yes, 2 no	Ilow = 0 (no variability)
Medium : 0 yes, 3 no	Imedium = 0 (no variability)
High : 5 yes, 5 no	Ihigh = 1 (evenly distributed subset)

We can omit calculations for Low and Medium since thay always end up with not-diabetic category.

$$IG_{PDF} = IE(S) - \left[\lfloor (10/\ 15)^{*}\ I_{high} \rfloor\right]$$
$$IG_{PDF} = 0.918 - 0.667 = 0.248$$

3.3.6 Choosing the Root Node

Table 8. Finding Maximum Gain Factor

Finding Maximum Gain Factor	values
IGage	0.068
IGBMI	0.032
IGPlasma	0.062
IGPDF	0.248

Our best pick is PDF, and we can immediately predict the patient is not diabetic when PDF is Low or Medium. It also indicates that diabetes is a hereditary disease since PDF indicates the diabetes pedigree function from genes.

Fig-1 ROOT OF DECISION TREE IS THE PATIENT DIABETIC?

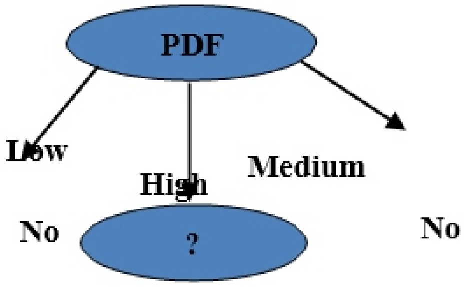


Table 1.3: Example – After Root Node Creation

Since we selected the PDF attribute for our Root Node, it is removed from the table for future calculations.

Table 9.Table of sample dataset after first iteration

PATIENT	AGE	BMI	PLASMA	PDF	DIABETIC/ NOT
P3	YOUNG	NORMAL	MEDIUM	HIGH	YES
P4	MIDDLE	HIGH	HIGH	HIGH	YES
P5	OLD	HIGH	MEDIUM	HIGH	YES
P6	MIDDLE	HIGH	LOW	HIGH	NO
P7	OLD	NORMAL	HIGH	HIGH	NO
P8	OLD	HIGH	HIGH	HIGH	NO
P9	MIDDLE	NORMAL	LOW	HIGH	YES
P10	OLD	HIGH	MEDIUM	HIGH	YES
P13	MIDDLE	HIGH	HIGH	HIGH	NO
P15	MIDDLE	HIGH	HIGH	HIGH	NO

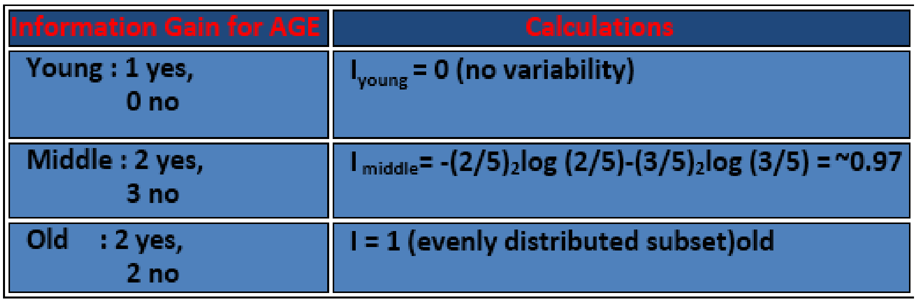
4. ITERATION-II

Calculating for Entropy IE (PDF) we get 1, since we have 5 yes and 5 no.

4.1. Example – Information Gain for AGE

- Age: 1 Young, 5 Middle, 4 Old
- 3 values for attribute age, so we need 3 entropy calculations.

Table 10. – Information Gain for AGE



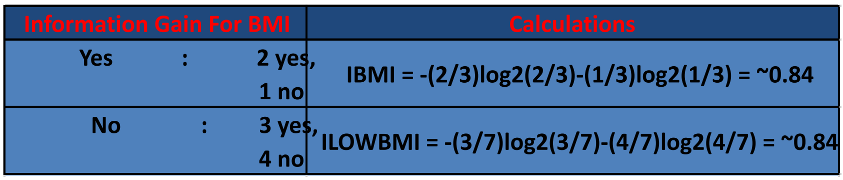
Information Gain for AGE	Calculations
Young : 1 yes, 0 no	I_{young} = 0 (no variability)
Middle : 2 yes, 3 no	I_{middle}= -(2/5)$_2$log (2/5)-(3/5)$_2$log (3/5) = ~0.97
Old : 2 yes, 2 no	I = 1 (evenly distributed subset)old

$$IG_{age} = IE(S_{pdf}) - [(5/10)*I_{high} + (4/10)*I_{low}]$$
$$IG_{age} = 1-0.885=0.115$$

4.2. Example – Information Gain for BMI

- BMI: 3 yes, 7 no
- 2 values for attribute BMI, so we need 2 entropy calculations.

Table 11. Information Gain for BMI



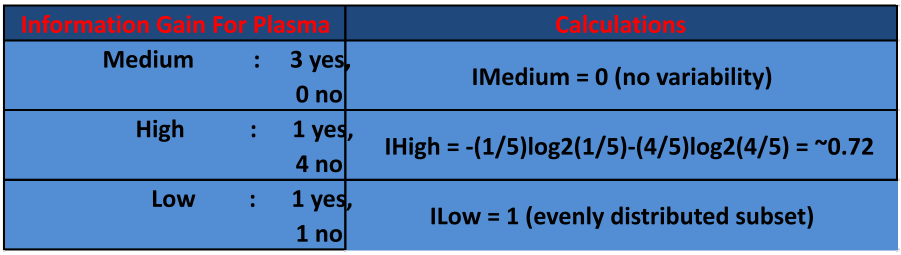
Information Gain For BMI	Calculations
Yes : 2 yes, 1 no	IBMI = -(2/3)log2(2/3)-(1/3)log2(1/3) = ~0.84
No : 3 yes, 4 no	ILOWBMI = -(3/7)log2(3/7)-(4/7)log2(4/7) = ~0.84

$$IG_{BMI} = IE(S_{PDF}) - \lfloor (3/10)*I_{BMI} + (7/10)*I_{LowBMI} \rfloor$$
$$IG_{BMI}=1-0.965=0.035$$

4.3. Example - Information Gain for Plasma

- Plasma: 3 Medium, 5 High, 2 Low
- 3 values for attribute weight, so we need 3 entropy calculations.

Table 12. Information Gain for Plasma



Information Gain For Plasma	Calculations
Medium : 3 yes, 0 no	IMedium = 0 (no variability)
High : 1 yes, 4 no	IHigh = -(1/5)log2(1/5)-(4/5)log2(4/5) = ~0.72
Low : 1 yes, 1 no	ILow = 1 (evenly distributed subset)

$$IG_{Age} = IE(S_{PDF}) - [(5/10)*I_{High} + (2/10)*I_{Low}]$$
$$IG_{Age}=1 -0.561= 0.439$$

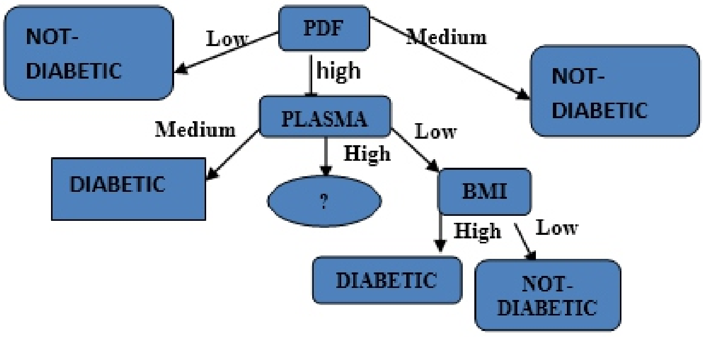
EXAMPLE - CHOOSING THE LEVEL 2 NODE

Table 13. Finding Maximum Gain Factor

Finding Maximum Gain Factor	Calculations
IGAge	0.115
IGBMI	0.035
IGPlasma	0.439

- Plasma has the highest gain, and is thus the best choice.

Fig. 2: Example – Decision Tree: 1



Since there are only two items for BMI where Plasma =low and Plasma=medium the result is inconsistent.

Table 14: Example – Updated Table

AGE	BMI	DIABETIC/NOT
MIDDLE	HIGH	YES
OLD	NORMAL	NO
OLD	HIGH	NO
MIDDLE	HIGH	NO
MIDDLE	HIGH	NO

5. RESULTS AND DISCUSSIONS

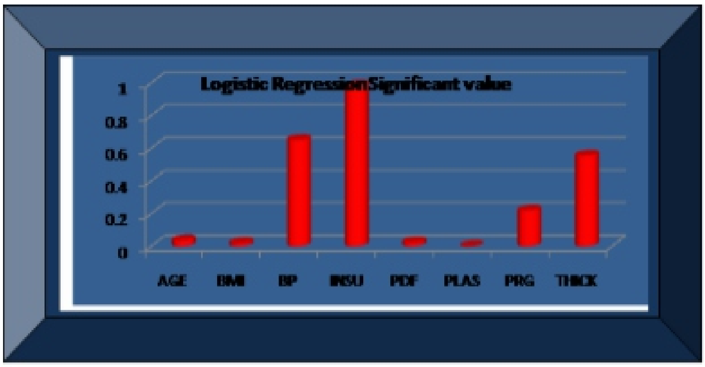
- All patients with large BMI in the Fig [2] are diabetic.
- All patients with large age factor (OLD) in the Table [14] are not diabetic even with large BMI.
- Obesity or age alone cannot indicate the sure possibility of getting the disorder since hereditary factors and phenotypic factors like lifestyle, stress and habits play a role.
- Due to inconsistent patterns in the data, there is no way to proceed since middle age patients may be diabetic or non diabetic depending on other factors, so iterations are terminated.
- ID3 attempts to make the shortest decision tree out of a set of learning data, shortest is not always the best classification.
- Requires learning data to have completely consistent patterns with no uncertainty.

6. CONCLUSIONS

Data mining method using logistic regression implies that Age, Obesity, PDF and Plasma level are to be taken care of for the onset of diabetes mellitus. Pdf factor is the diabetes pedigree function value which shows the hereditary nature of the disorder which signifies that candidates with diabetes disorder in ancestors should take necessary monitoring of glucose levels periodically. ID3 algorithm applied to the sample database gives the decision tree prediction with major factors influencing diabetes. Pdf at first level of decision and Plasma glucose concentration in saliva at the next level and, the body mass index at the third level of decision have minimum entropy. Their significant role in the cause of diabetes is implied by the decision tree. Similarly larger decision tree can be drawn to show the significance of all factors which are responsible for the cause of diabetes by mining the total dataset. The paper on a small scale tries to bring out the dominant factors alone by applying Iterative Dichotomiser ID3 algorithm of data mining. As our mankind has a great threat of this pancreatic disorder more in the coming era the sample data is chosen from diabetes database. The same idea can be applied to any disease database on a large sampling to bring out more useful diagnostic findings before complications affect the human population.

TABLES AND FIGURES

Figure 3: Significant Factors from Logistic Regression output



REFERENCES

1. *Ankerst, M., Elsen, C., Ester, M. and Kriegel, H.P. Visual classification: An interactive approach to decision tree construction. In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99), San Diego, CA, Aug. 1999, pp. 392-396.*
2. *Almuallim H., an Efficient Algorithm for Optimal Pruning of Decision Trees. Artificial Intelligence, 1996, 83(2): 347-362.*
3. *Brodley C.E. and Utgoff. P.E., Multivariate decision trees, Machine Learning, 1995, 19: 45-77.*
4. *Quinlan, J. R (1985). Induction of Decision Trees, Machine Learning 1: 81-106, 1986.*
5. *http://www.niddk.nih.gov/ Home page for the National Institute of Diabetes and Digestive and Kidney Diseases.*
6. *Navathe Elmasri, (2007). Fundamentals of Database Systems (5th Edition), 975-977.*
7. *Shannon, Claude E. Prediction and Entropy of Printed English. (Retrieved 04/23/2010). http://languagelog.ldc.upenn.edu/myl/Shannon1.*
8. *Shannon, C.E. (1948). A mathematical theory of Shannon communication, Bell System Technical Journal 27: 379-423 and 623-656. http://cm.bell labs.com/ cm/ms/what/ Shannon day/paper.html.*
9. *Ross, Peter (10/30/2000). Rule Induction: Ross Quinlan's ID3 Algorithm (Retrieved 04/23/2010). http://www.dcs.napier.ac.uk/~peter/vldb/dm/node11.html.*
10. *Quinlan, J.R., Simplifying decision trees, International Journal of Man machine Studies, 1987, num. 27, pp. 221-234.*
11. *R.Jamuna, K.Meena. Data mining by Logistic Regression Techniques in Pima Indian Diabetes Database. Bio-Science Research Bulletin, Vol. 22 (No. 2) July- December 2006.*

Control Charts for Waiting Time using Method of Weighted Variance and Power Transformation for (M/M/S) : (∞ : FCFS) Model

Dr. Mrs. M. V. Khaparde[*], **Dr. Mrs. S. D. Dhabe**[**]

[*] Professor, Department of Statistics, R.T.M. Nagpur University, Nagpur (Maharashtra.State) India

[**] Associate Professor, Sydenham Institute of Management Studies, Research And Entrepreneurship Education(SIMSREE), B road, Churchgate , Mumbai 20

ABSTRACT

In this paper to monitor the waiting time of the (M/M/S) : (∞ : FCFS) queuing model , control chart for the random waiting time is constructed using method of weighted variance and Nelson's power transformation. The performance measure average run length for these charts is obtained and compared.

Keywords : False alarm rate, Type II error, Average queue length, Average run length, average queue length, average waiting time , Weibull distribution

1. INTRODUCTION

Various types of control charts for the random queue length N and waiting time i.e. Ws for the (M/M/1):(∞/FCFS) queuing model are constructed by Khaparde M.V. and Dhabe S.D. In this paper control charts for random waiting time for (M/M/s) : (∞ / FCFS) queuing model are constructed .

2. (M/M/S) : (∞/FCFS) queuing model

Notations

Let P_n denote steady state probability of having exactly n customers in the system

λ = mean arrival rate , μ = mean service rate per busy server

s = number of parallel servers , ρ = Traffic intensity $\lambda/s\mu$

W_s = waiting time per customer in the system

W_Q = waiting time per customer in the queue

f (W_s) = density function of waiting time of the customer in the system.

f(W_Q)= density function of waiting time of the customer in the queue

Multichannel queuing theory deals with the condition in which there are several service stations in parallel and each element in the waiting line can be served by more than one station. Each service facility is prepared to deliver the same type of service. The new arrival selects one station without any external pressure. When a waiting line is formed, a single line usually breaks down into shorter lines in front of each service station. The arrival rate l and service rate m are mean values from Poisson distribution and exponential distribution respectively. Service discipline is first come first serve and customers are taken from a single queue i.e. any empty channel is filled by the next customer in line.

When $n < s$, there is no queue because all arrivals are being serviced, and the rate of servicing will be nm as only n channels are busy, each at the rate of m. When $n = s$, all channels will be working and when $n > s$, there will be $(n - s)$ persons in the queue and rate of service will be sm as all the s channels are busy.

3. CONSTRUCTION OF CONTROL CHARTS

For any queuing system, average queue length and average waiting time are the main observable characteristics. Customers want to have waiting time in the system as minimum as possible i.e.queue length should be small. Haim Shore (1999) has made pioneering attempt of extending the application of statistical process control to queuing systems. He obtained control limits for the random queue length N for (M/M/s) queuing model. These control limits are explicitly expressed in terms of mean, standard deviation and skewness of the distribution of r.v.N. This control chart monitors the stability of the queuing system in terms of N. If an out of control signal is generated it will indicate a change in the parameter arrival rate or service rate which determine N.

To monitor the waiting time of the customers in the queuing system ,the following control charts for random waiting time are constructed .

3.1 The following two control charts for WQ are constructed .

i) Control chart sW^{1}_{Q}

I) Control chart sW^{2}_{Q}

- This is simple Shewhart control chart and - This chart is constructed using method of weighted variance.

3.2 The following three control charts for r.v.Ws are constructed which are referred to as sW_{s1} , sW_{s}^{2} and sW_{s}^{3}

I) Control chart sW_{s1-} This is simple Shewhart control chart

ii) Control chart sW_s^2- This chart is constructed using method of weighted variance.

iii) Control chart sW_s^3-This chart is constructed using Nelson's transformation.

4. Control charts for r.v.WQ for (M/M/s : ∞/FCFS) model

In this section ,Shewhart control chart for WQ for (M/M/s :   /FCFS) model is constructed

Waiting time distribution of WQ

Construction of control limits for WQ , needs the distribution of WQ , its expectation and variance. The r.v.WQ denote the waiting time of customer in the queue. Assuming that the queue discipline is FCFS, from queuing theory ,the distribution f(x) of WQ is given by

$$f(x)dx = 1-\frac{(\lambda/\mu)^s}{(s-1)!\left(s-\frac{\lambda}{\mu}\right)}p_0 = 1-\frac{(\lambda/\mu)^s}{s!(1-\rho)}p_0, \quad x=0$$

$$= \frac{(\lambda/\mu)^s}{(s-1)!}\mu e^{-(s\mu-\lambda)x} p_0, \quad x>0 \ldots\ldots\ldots\ldots\ldots\ldots 4.1$$

4.1Moments of WQ

First two raw moments of WQ are obtained

$$\mu_1^1 = E[W_Q] = [W_Q=0]P[W_Q=0] + \int_\epsilon^\infty xf(x)dx, \quad \text{as } \epsilon \to 0$$

Now

$$\int_\epsilon^\infty xf(x)dx = \int_\epsilon^\infty \frac{(\lambda/\mu)^s p_0\mu}{(s-1)!}\left\{xe^{-(s\mu-\lambda)x}\right\}dx \to \left\{\frac{(\lambda/\mu)^s \mu p_0}{(s-1)!}\right\}\frac{1}{(s\mu-\lambda)^2}$$

$$= \frac{(\lambda/\mu)^s}{(s\mu)(s!)(1-\rho)^2}p_0 \quad \frac{\square p_s}{s\mu(1-\rho)^2}$$

$$E[W_Q] = \frac{P[N\geq s]}{s\mu(1-\rho)_Q} = \frac{(\lambda/\mu)^s}{(s-1)!}p_0\mu\frac{1}{(s\mu-\lambda)2} \ldots\ldots\ldots\ldots\ldots\ldots 4.1.1$$

$$\mu_1^2 = E[W_Q^2] = \int_\epsilon^\infty x^2 f(x)dx$$

on simplification

$$E[W_Q^2] = \frac{(\lambda/\mu)^s}{(s-1)!}p_0\mu\frac{2}{(s\mu-\lambda)^3}$$

Let σ^2 denote variance of W_Q

$$\sigma^2 = V[W_Q] = \mu_2^1 - (\mu_1^1)^2$$

$$\sigma^2 = V[W_Q] = E[W_Q^2] - \{E[W_Q]\}^2$$

$$= \frac{p_0\mu\,(\lambda/\mu)^s}{(s-1)!\,(s\mu-\lambda)^2}\left[\frac{2}{(s\mu-\lambda)} - \frac{(\lambda/\mu)^s p_0\mu}{(s-1)!(s\mu-\lambda)^2}\right] \ldots\ldots\ldots\ldots\ldots 4.1.2$$

5. Control chart $sW_Q^{\ 1}$

Knowing E[WQ] and V[WQ] ,the 3 sigma control limits for WQ are given by

$$\begin{aligned} UCL &= E[W_Q] + 3\sqrt{V(W_Q)} \\ CL &= E[W_Q] \qquad \text{.......... .... 5.1} \\ LCL &= E[W_Q] - 3\sqrt{V(W_Q)} \end{aligned}$$

False alarm rate

Let α_u denote type I error probability generated in the upper tail or false alarm rate which is given by α_u = P [WQ > UCL]

where UCL is obtained from 5.1 and P[W_Q > UCL] is obtained using expression 4.1

6. Control chart $sW_Q^{\ 2}$

Control chart for the r.v. WQ using method of weighted variance

In order to obtain control limits for W_Q using this method, the probability P_{WQ} defined as follows ,is needed

$$P_{W_Q} = P\left|W_Q \le E(W_Q)\right]$$

$$= \int_0^{E[W_Q]} f(x)dx \qquad = \int_0^{E[W_Q]} \frac{(\lambda/\mu)^s p_0 \mu}{(s-1)!} e^{-(s\mu-\lambda)x} dx$$

Solving the integral and substituting for E[WQ],

$$P_{W_Q} = \frac{\mu p_0 (\lambda/\mu)^s}{(s-1)!(s\mu-\lambda)} \left\{ 1 - e^{\frac{-\mu p_0 (\lambda/\mu)}{(s\mu-\lambda)}} \right\} \quad 6.1$$

If the underlying population is symmetric then P_{WQ} = 0.5 and the chart for weighted variance reduce to Shewhart chart. However, if the underlying population is skewed to the right then P_{WQ} > 0.5 and the distance of UCL from the Center Line (CL) is larger than that of LCL similarly if the underlying population is skewed to the left then P_{WQ} < 0.5 and the distance of the LCL from the CL is larger than that of UCL.

6.1 Control limits using method of weighted variance

The 3 sigma control limits using method of weighted variance are given by:

$$\begin{aligned} UCL &= E[W_Q] + 3\sqrt{V(W_Q)}.\sqrt{2P_{W_Q}} \\ CL &= E[W_Q] \qquad \text{...........................6.1.1} \\ LCL &= E[W_Q] - 3\sqrt{V(W_Q)}.\sqrt{2(1-P_{W_Q})} \end{aligned}$$

Where $E[W_Q]$, $V[W_Q]$ and PW_Q are obtained using 4.1.1, 4.1.2 and 6.1

6.2 False alarm rate (FAR)

Let α_u denote type I error probability generated in the upper tail or false alarm rate which is given by $\alpha_u = P[W_Q > UCL]$

where UCL is obtained from 6.1.1 and the corresponding probability can be obtained from expression 4.1

7. Control charts for the r.v.Ws for (M/M/S : ∞/FCFS) model

The distribution of r.v.WQ and its moments are obtained in section 4. In this section ,control limits for the r.v.Ws using Shewhart method and method of weighted variance are to be constructed. In order to obtain control limits for the r.v.Ws, the expressions for E[Ws] and V[Ws] are required.

Let Ws denote the waiting time of the customer in the system.

$$\therefore W_s = W_Q + \frac{1}{\mu} \quad \text{.......... ....7.1}$$

where WQ is the waiting time of the customer in the queue and (1/μ) is the service rate of individual channel.

$$E[W_s] = E\left[W_Q + \frac{1}{\mu}\right] = E[W_Q] + \frac{1}{\mu} \quad \text{... 7.2}$$

and

$$V[W_s] = V\left[W_Q + \frac{1}{\mu}\right] = V[W_Q] \quad \text{... 7.3}$$

where, $E(W_Q)$ and $V(W_Q)$ are obtained from 4.1.1 and 4.1.2.

sW_s^1

7.1 Control chart s

Control limits for Ws using Shewhart method

The 3 sigma control limits for r.v.Ws are

$$UCL = E[W_s] + 3\sqrt{V(W_s)}$$

$$CL = E[W_s]$$

$$LCL = E[W_s] - 3\sqrt{V(W_s)}$$

where $E[W_s]$ and $V[W_s]$ are obtained using expressions 7.2 and 7.3.

7.2 False alarm rate

It is of interest to know the probability that the time of waiting in the line plus the service time exceeds time t. This probability is denoted by P [Ws > t]. This probability is given by

$$P[W_s > t] = e^{-ut}\left[1 + \frac{W}{s}\left(\frac{1-e^{-\mu s t\left(1-\frac{\lambda}{\mu s}-\frac{1}{s}\right)}}{1-\frac{\lambda}{\mu s}-\frac{1}{s}}\right)\right] \text{.......... ....7.2.1}$$

Where W is the probability that a customer has to wait in line, which is the sum of all probabilities that all service facilities are being used or that s or more customers are in line.

$$W = \frac{P_0}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu s}\right)^n \quad W = \left(\frac{\lambda}{\mu}\right)^s \frac{P_0}{s!\left(1-\frac{\lambda}{\mu s}\right)}$$

Let α_u denote false alarm rate given by $\alpha_u = P[Ws > UCL]$,replacing t by UCL in expression

5.3.1 the expression for false alarm rate α_u is obtained and is given by

$$\alpha_u = P[W_S > UCL] = e^{-\mu UCL}\left[1 + \frac{W}{S}\left(\frac{1-e^{-\mu UCL\left(1-\frac{\lambda}{\mu s}-\frac{1}{s}\right)}}{1-\frac{\lambda}{\mu s}-\frac{1}{s}}\right)\right] \text{.......... ....7.2.2}$$

7.3 Numerical analysis of Control chart sW_s^1

In order to study effect of ρ on control limits, one set of values of λ, μ and s is selected. For this set of values of ρ, p_0, LCL, UCL, α_u and ARL are obtained and are displayed in table 1 From this table, it is observed that keeping λ and μ fixed, if s is increased, the value of α_u increases which results in the corresponding decrease in the values of ARL. We also observe that for some combination of λ, μ and s, α_u turns out to be 0. This will mean that in a queuing system with those particular combinations of λ, μ and s, there are no chances of system going out of control, which means that system is performing very well.

Table 1 : Lower and Upper Control limits and the associated values of α_u for r.v. Ws for M/M/S queue using sW 1 chart with L = 3

Sr. No				p_o	variance					α_u	
	λ	μ	s			ρ	LCL	CL	UCL		ARL
1	20	15	2	0.2	0.0078	0.666667	0	0.12	0.3853	0	--
2	20	15	3	0.2542	0.0005	0.444444	0.005	0.0738	0.1427	0.0025	400
3	20	15	4	0.2621	0.000006	0.333333	0.0441	0.0679	0.0917	0.0339	30
4	20	15	5	0.2633	0.000008	0.266667	0.0582	0.0668	0.0755	0.0715	14
5	20	15	6	0.2635	0.000001	0.222222	0.0635	0.0667	0.0698	0.0936	11
6	10	15	2	0.5	0.0007	0.3333	0	0.075	0.1579	0.0026	385
7	10	15	3	0.5121	0.00005	0.222222	0	0.0675	0.0892	0.0508	20
8	10	15	4	0.5133	0.00004	0.166667	0.0607	0.0667	0.0728	0.0951	11
9	10	15	5	0.5134	0	0.133333	0.065	0.0666	0.0683	0.1131	9
10	100	35	3	0.0111	0.03968	0.952381	0	0.2108	0.8084	0	-
11	100	35	4	0.0464	0.00043	0.714286	0	0.0398	0.1025	0	-
12	100	35	5	0.0546	0.00006	0.571429	0.0071	0.0312	0.0553	0.0016	625
13	100	35	6	0.0567	0.000013	0.47619	0.0185	0.0293	0.0401	0.0175	57
14	100	35	7	0.0572	0.000003	0.408163	0.0237	0.0287	0.0337	0.0428	23
15	16	15	2	0.5333	0.000308	0.533333	0	0.0931	0.2597	0.000003	3333333
16	16	15	3	0.339	0.00024	0.355556	0.0238	0.0703	0.1167	0.0121	82
17	16	10	2	0.1111	0.057284	0.8	0	0.2777	0.9958	0	-
18	16	10	3	0.1871	0.002411	0.533333	0	0.1195	0.2668	0.0002	5000
19	16	10	4	0.1992	0.000301	0.4	0.0517	0.1037	0.1557	0.0165	61
20	12	6	3	0.1111	0.019204	0.666667	0	0.2407	0.6564	0	-

8. Control chart sW_s^2

Control limits for Ws using method of weighted variance

To obtain control limits, the method of weighted variance needs the probability PWS, where PWS is given by

$$\begin{aligned} P_{W_s} &= P[W_s \le E(W_s)] \\ &= 1 - P[W_s > E(W_s)] \square \quad 8.1 \end{aligned}$$

This probability can be obtained using7.2.1. The control limits of PWs using method of weighted variance are given by :

$$\begin{aligned} UCL &= E[W_s] + 3\sqrt{V(W_s)}\sqrt{2P_{W_s}} \\ CL &= E[W_s] \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 8.2 \\ LCL &= E[W_s] - 3\sqrt{V(W_s)}\sqrt{2(1-P_{W_s})} \end{aligned}$$

8.1. False alarm rate

Let au denote the false alarm rate which is given by au = P [Ws > UCL] Where UCL is obtained using 8.2 and P[Ws > UCL] is obtained using expression 7.2.1

8.2 Numerical analysis

In order to study effect of ρ on control limits. the same set of values of λ, μ and s as in charts sW_s^1 are selected. For this set UCL, CL and LCL, PWs, P0, au and ARL are obtained and are displayed in table 2. From this table it is observed that if we keep λ and μ fixed and s is increased then au increases and consequently the associated ARL decreases very rapidly.

If ARL of this chart is compared with the ARL of charts sW_s^1 then the increase in values of ARL can be noticed. This means that the performance of this schart is better than performance of control chart sW_s^1 . This improvement in ARL is due to the

presence of factor PWs in control limits which takes into account skewness of the underlying distribution of Ws for that particular combination of λ, μ and s.

Table 2 Lower and Upper Control limits and the associated values of au for r.v. Ws for M/M/ssqueue using method of weighted variance sW_s^2 with L = 3

Sr. No	λ	μ	s	p_o	Pws	ρ	LCL	CL	UCL	α_u	ARL
1	20	15	2	0.2	0.9975	0.666667	0.1	0.12	0.496	0	-
2	20	15	3	0.254	0.9411	0.444444	0.05	0.074	0.168	0.0008	1250
3	20	15	4	0.262	0.913	0.333333	0.06	0.068	0.1	0.0245	41
4	20	15	5	0.263	0.902	0.266667	0.06	0.067	0.078	0.0642	16
5	20	15	6	0.264	0.8956	0.222222	0.07	0.067	0.071	0.0903	11
6	10	15	2	0.5	0.9186	0.333333	0.04	0.075	0.187	0.0008	1250
7	10	15	3	0.512	0.888	0.222222	0.06	0.068	0.096	0.0392	26
8	10	15	4	0.513	0.883	0.166667	0.06	0.067	0.075	0.0889	11
9	10	15	5	0.513	0.8804	0.133333	0.07	0.067	0.069	0.1111	9
10	100	35	3	0.011	1.0000.	0.952381	0.21	0.211	1.056	0	-
11	100	35	4	0.046	0.9954	0.714286	0.04	0.04	0.128	0	-
12	100	35	5	0.055	0.9672	0.571429	0.03	0.031	0.065	0.00052	1923
13	100	35	6	0.057	0.9449	0.47619	0.03	0.029	0.044	0.0115	18
14	100	35	7	0.057	0.9306	0.408163	0.03	0.029	0.036	0.036	28
15	16	15	2	0.533	0.9777	0.533333	0.06	0.093	0.326	0	-
16	16	15	3	0.339	0.9176	0.355556	0.05	0.07	0.133	0.0062	161
17	16	10	2	0.111	0.9999	0.8	0.28	0.278	1.293	0	0
18	16	10	3	0.187	0.9651	0.533333	0.08	0.12	0.324	0.000032	31250
19	16	10	4	0.199	0.9277	0.4	0.08	0.104	0.175	0.0098	102
20	12	6	3	0.111	0.9927	0.666667	0.19	0.241	0.827	0	-

9 Control Chart sW_s^3

Nelson's control chart for Ws for M/M/s model

Khaparde M. V. and Dhabe S. D. have constructed control chart using power transformation for r.v.Ws for (M/M/1: ∞/FCFS) model.Like (M/M/1 : ∞/FCFS) model, in (M/M/s: ∞/FCFS) model also the distribution of Ws is exponential. But this exponential distribution is a special case of Weibull distribution.

$$W\left(\frac{1}{5\mu-\lambda},1\right)$$

Using the transformation $Y = (W)^{\frac{1}{3.6}} = W^{0.277}$,Y transforms to Weibull

$$W\left[\left(\frac{1}{5\mu-\lambda}\right)^{0.2777},3.6\right]$$

and thus follows approximate normal distribution. The mean of Y is given by

$$E(Y) = \left(\frac{1}{5\mu-\lambda}\right)^{0.2777}\Gamma\left(1+\frac{1}{3.6}\right) = (0.901)\left(\frac{1}{5\mu-\lambda}\right)^{0.2777} \text{.......... ....} 9.1$$

This expression is used to set the center line of the control chart for Y. The standard deviation of Y is given by.

$$\sqrt{V(Y)} = \left(\frac{1}{5\mu-\lambda}\right)^{0.2777}\sqrt{\Gamma\left(1+\frac{2}{3.6}\right)-\left\{\Gamma\left(1+\frac{1}{3.6}\right)\right\}^2}$$

$$= \left(\frac{1}{5\mu-\lambda}\right)^{0.2777}(0.278)\text{.................. ....}9.2$$

9.1 Control limits for Ws using Nelson Chart

Using the above approximation the control limits for Ws are given by

$$UCL = E[W_s] + L\sqrt{V(W_s)}$$

$$= (0.901)\left(\frac{1}{5\mu-\lambda}\right)^{0.2777} + L\left(\frac{1}{5\mu-\lambda}\right)^{0.2777}(0.278)\text{.................. ...}9.1.1$$

$$CL = (0.901)\left(\frac{1}{5\mu-\lambda}\right)^{0.2777}.$$

$$LCL = (0.901)\left(\frac{1}{5\mu-\lambda}\right)^{0.2777} + L\left(\frac{1}{5\mu-\lambda}\right)^{0.2777}(0.278)\text{.................. ...}9.1.2$$

where L is the distance of control limits from the center line. Taking L = 3, we get 3 sigma control limits.

9.2 Derivation and definition of α

Let a be the probability of type I error then $\alpha = \alpha_u + \alpha l$

where au and al are the risk probabilities generated in the upper and lower tail respectively and are defined as

$$\alpha_u = P[Ws > UCL] \; ; \; al = P[Ws < LCL]$$

The distribution of Ws is exponential. But after using transformation the distribution of Ws is not complete symmetrical about its mean. Therefore the probability of Ws exceeding upper control limit is obtained from C.D.F. of Weibull distribution.

$$\therefore \alpha_u = P[Ws > UCL] = 1 - P[Ws < UCL] = 1 - F[UCL]$$

where F(.) is the distribution function of 2 parameter Weibull distribution W(η, ν).

$$F[UCL] = 1 - \exp\left\{ -\left(\frac{UCL}{\eta} \right)^{\nu} \right\} \square 9.2.1$$

but $\quad \nu \quad = 3.6$

$$= 1 - \exp\left\{ -\left(\frac{UCL}{\eta} \right)^{3.6} \right\}$$

$$\alpha_l \; = P[W_s \leq LCL] = \exp\left\{ \left(-\frac{LCL}{\eta} \right)^{3.6} \right\} \text{.......... ....9.2.2.}$$

where $\eta = \left(\frac{1}{s\mu - \lambda} \right)^{0.2777}$

9.3 Numerical analysis We have selected same set of values of λ, μ and s as that of control chart sW_s^1 and sW_s^2, thecontrol limits αu and αl are obtained. These are given in table 3 .

For this chart αu = 0.000732 and αl = 0.000059 for all values of λ, μ and s.

If we are interested in detecting the shift in upper as well as lower direction then ARL is given by

$$ARL = \frac{1}{\alpha_u + \alpha_l} = \frac{1}{(0.000732) + (0.000059)} = \frac{1}{0.000791} = 1264.2225$$

This ARL remains same for all values of λ, μ and s. The main difference that can be observed in the ARL of this chart and the ARL of earlier two charts is that, in the chart sW_s^1 and sW_s^2, if we keep λ, μ fixed then ARL decreases with increase in value of s (the number of servers) but in this chart it remains same.

Using Nelson's chart if we want to detect the shift in the upward direction only then we have to consider value of αu only and in that case

$$ARL = \frac{1}{\alpha_u} = \frac{1}{0.000732}$$
$$= 1366.1202$$
$$\cong 1366$$

10 Conclusion- The comparison of the above three control charts for waiting time on the basis of ARL reveals that since AR L is highest for the third control chart sW_s^3 where Nelson transformation is used ,Therefore it is the best chart.

Table 3 Lower and Upper Control limits and the associated values of αu and αl for r.v. Ws for M/M/s queue using Nelson transformation sW_s^3 , with L = 3

Sr. No.	λ	µ	s	ρ	UCL	CL	LCL	α_u	α_l
1	20	15	2	0.667	0.9154	0.475362	0.04	0.0007	0.000059
2	20	15	3	0.444	0.7097	0.368567	0.03	0.0007	0.000059
3	20	15	4	0.333	0.6229	0.323469	0.02	0.0007	0.000059
4	20	15	5	0.267	0.5702	0.296092	0.02	0.0007	0.000059
5	20	15	6	0.222	0.5332	0.276912	0.02	0.0007	0.000059
6	20	15	7	0.19	0.5052	0.262377	0.02	0.0007	0.000059
7	10	15	3	0.222	0.6464	0.335689	0.02	0.0007	0.000059
8	10	15	4	0.167	0.5855	0.304033	0.02	0.0007	0.000059
9	10	15	5	0.133	0.5443	0.282669	0.02	0.0007	0.000059
10	100	35	3	0.952	1.1097	0.576263	0.04	0.0007	0.000059
11	100	35	4	0.714	0.6229	0.323469	0.02	0.0007	0.000059
12	100	35	5	0.571	0.5231	0.271657	0.02	0.0007	0.000059
13	100	35	6	0.476	0.4703	0.244247	0.02	0.0007	0.000059
14	100	35	7	0.408	0.4356	0.226211	0.02	0.0007	0.000059
15	16	15	2	0.533	0.8337	0.432957	0.03	0.0007	0.000059
16	16	15	3	0.356	0.6811	0.353685	0.03	0.0007	0.000059
17	16	10	2	0.8	1.1806	0.613102	0.05	0.0007	0.000059
18	16	10	3	0.533	0.8337	0.432957	0.03	0.0007	0.000059
19	16	10	4	0.4	0.7178	0.372769	0.03	0.0007	0.000059
20	12	6	3	0.667	1.0549	0.547813	0.04	0.0007	0.000059

REFERENCES

1 Bai D.S. and Choi I.S. (1995) X and R control charts for skewed populations. Journal of Quality Technology. Vol. 27, No. 2, 120-131.

2 Dhabe S.D. & Khaparde M.V.(2011) Control charts for random queue length for (M/M/1) : ((∞ / FCFS) queuing model using skewness and power transformation Bulletin of Pure and Applied Sciences vol 30 E (Math & Stat) , Issue (No. 1) , pp 71-83

3 Grant E.L. and Leavenworth R.S. Statistical Quality Control. Sixth edition. McGraw Hill International editions.

4 Khaparde M.V & Dhabe S.D. (2010) Control charts for random queue length N for (M/M/1) : (∞ / FCFS) queuing model International Journal of Agricultural and applied Sciences ,Vol 06 (No 1), pp 319-334 .

5 Khaparde M.V & Dhabe S.D.(2011) Control charts for random waiting time using power transformation for (M/M/1) : (∞ / FCFS) model International Journal of Mathematical Sciences and Engineering Applications (IJMSEA) ,Pune Vol 5 , No VI , pp 121- 137.

7 McCool J.I. and Motley T.J. (1998) Control charts applicable when fraction Non-conforming is small. Journal of quality technology. Vol. 30, No. 3. 240-247.

8 Mitra A. Fundamentals of Quality Control and improvement (Second edition). Pearson Education (Singapore) Pte. Ltd.

9 Montgomery D.C. Introduction to Statistical Quality Control (Fourth edition) John Wiley and Sons.

10 Shore H. (2000) General Control Charts for attributes. IIE Transactions. 32, 1149-1160.

11 Taha H.A. Operations Research : An introduction Seventh edition. Prentice Hall of India Private Limited.

12 Wagner H.M. Principles of operations Research with applications to managerial decisions Eastern economy edition.

13 Winston W.L. (Second edition) Operation Research applications and algorithm. Dusbury Press Boston.

Hub-and-Spoke MPLS Layer 3 VPN (L3VPN) Topology

Akshay[1], Pooja Ahlawat[2]

[1]M.Tech. Student, Department of Computer Science & Engineering, R.N. College of Engineering & Management, Maharshi Dayanand University, Rohtak, Haryana, India

[2]Assistant Professor, Department of Computer Science & Engineering, R.N. College of Engineering & Management, Maharshi Dayanand University, Rohtak, Haryana, India

ABSTRACT

There are two kinds of MPLS L3VPN topologies: Mull-Mesh and Hub-and Spoke. This paper gives the theoretical understanding of Hub-and-Spoke MPLS Layer 3 VPN Topology and its implementation using GNS3 simulator. This paper will also discuss its merit and demerits and comparison with Full-Mesh.

Keywords: MPLS, VPN, GNS3, Hub-and-spoke

1. INTRODUCTION TO HUB-AND-SPOKE MPLS L3VPN TOPOLOGY

The most commonly encountered topology is a hub-and- spoke topology, where a number of remote offices (spokes) are connected to a central site (hub. The remote offices usually can exchange data (there are no explicit security restrictions on inter-office traffic), but the amount of data exchanged between them is negligible. The hub-and-spoke topology is used typically in organizations with strict hierarchical structures, for example, banks, governments, retail stores, international organizations with small in-country offices, and so on.

Often, customers do not want their sites to have full interconnectivity. This means they do not want or need the sites to be fully meshed. A typical scenario involves one main site at a company with many remote sites. The remote sites or spokes need connectivity to the main or hub site, but they do not need to communicate between them directly. Perhaps the connectivity is possible but not wanted for security reasons. This scenario is commonly referred to as the hub-and-spoke scenario. It can also be achieved across MPLS VPN, but care must be taken.

2. IMPLEMENTATION OF HUB-AND-SPOKE MPLS L3VPN

To implement the Hub-and-spoke MPLS L3VPN Topology, the following is needed:

(a) The spoke sites can communicate only with the hub site.

(b) Spoke-to-spoke traffic needs to be sent to the hub site first.

To achieve this, the following things are needed:

(a) Two different RTs

(b) Different RDs

When hub-and-spoke connectivity is required, two different RTs are required. One RT is used to identify routes (re)advertised from the hub site, and another is required to identify routes advertised from the spoke sites.

The PE router connected to the hub site CE router imports routes advertised from the spoke sites, and the PE routers connected to the spoke site CE routers import routes (re)advertised from the hub. Crucially, spoke sites do not import routes directly from other spoke sites.

In addition, it is important to note that the PE router connected to the hub site CE router requires two VRFs for the VPN, whereas the PE routers connected to the spoke site CE routers requires only one VRF for the VPN.

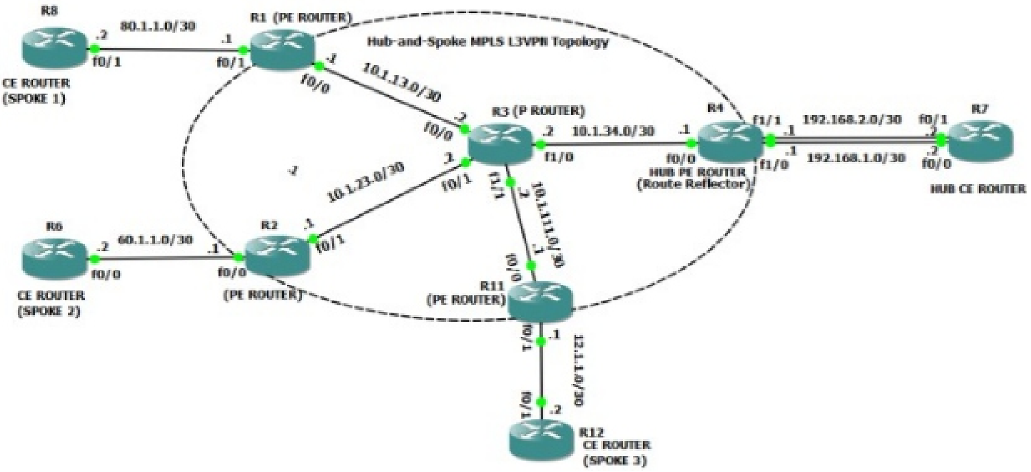


Figure 1: Hub-and-Spoke MPLS L3VPN Topology implemented using GNS3

Here, there are four 7600 series Routers running as a PE routers. Out of these four PE routers, one PE router, R4 acts as a Route reflector for other PE routers, namely R1, R2 and R11. All these PE routers, R1, R2, R4, and R11 have LDP neighbourship with P router R3.

The advantage of configuring R4 router as a RR (Route Reflector) is that all PE routers (R1, R2, and R11) have exactly only and only one MP-BGP session with router R4, thus avoiding one MP-BGP session for each PE router. The VPNv4 routes learnt from one PE router will be reflected to other PE routers by a Route Reflector (R4 Router).

Table 1: Address table

PE/P/CE	ROUTER	INTERFACE	IP ADDRESS	SUBNET
PE	R1	F0/0	10.1.13.1	255.255.255.252
		F0/1	80.1.1.1	255.255.255.252
	R2	F0/0	10.1.34.1	255.255.255.252
		F1/0	192.168.1.1	255.255.255.252
		F1/1	192.168.2.1	255.255.255.252
	R4	F0/0	10.1.34.4	255.255.255.252
		F0/1	10.1.1.4	255.255.255.252
	R11	F0/0	10.1.111.1	255.255.255.252
		F0/1	12.1.1.1	255.255.255.252
P	R3	F0/0	10.1.13.2	255.255.255.252
		F0/1	10.1.23.2	255.255.255.252
		F1/0	10.1.34.2	255.255.255.252
	R6	F1/1	10.1.111.2	255.255.255.252
		F0/0	80.1.1.22	255.255.255.252
CE	R7	F0/0	192.168.1.2	255.255.255.252
		F0/1	192.168.2.2	255.255.255.252
	R8	F0/1	80.1.1.2	255.255.255.252
	R12	F0/1	12.1.1.2	255.255.255.252

Here, in this implemented Hub-and-Spoke MPLS L3VPN topology, the R4 is a hub PE router, having direct point-to- point connectivity (direct connectivity) with Hub CE router. The R4 (Hub) PE router has two connections (Fast Ethernet connections) with Hub CE router.

On R4 Router, two VRFs are created, one VRF is for importing routes from all Spokes (CE Routers), and other VRF is for exporting routes to all spokes.

With Hub-and-Spoke, the manageability of customer location is better. Whenever, a new spoke site is provisioned, we make route-target entries into these two VRFs on the Hub PE (R4) router.

Here, in this implemented Hub-and-Spoke topology, BGP is used as a PE-CE routing protocol between R4 (Hub) Router and R7 (Hub) CE Router, OSPF is used as a PE-CE routing protocol in between rest of PE-CE connectivity.

Whenever, any CE pings other location IP, it always goes through Hub CE (R7) router.

For example, R8 router does a trace route and it goes through R7 as:

```
R8# traceroute 12.1.1.2
```

```
Type escape sequence to abort. Tracing the route to 12.1.1.2
```

```
1 80.1.1.1 80 msec 44 msec 56 msec
2 10.1.13.2 [MPLS: Labels 19/32 Exp 0] 140 msec 156 msec
152 msec
3 192.168.2.1 [MPLS: Label 32 Exp 0] 216 msec 168 msec
136 msec
4 192.168.2.2 136 msec 168 msec 120 msec
5 192.168.1.1 188 msec 176 msec 124 msec
6 10.1.34.2 [MPLS: Labels 18/26 Exp 0] 336 msec 280 msec
260 msec
7 12.1.1.1 396 msec 276 msec 296 msec
8 12.1.1.2 268 msec * 268 msec R8#
```

The above entry, in red colour, is the IP address of HUB CE Router (R7).

3. HUB-AND-SPOKE MPLS L3VPN BENEFITS

The following are the benefits of Hub-and-spoke MPLS Benefits:

(a) It is very easy to add a new site/router, as no changes to the existing spoke or hub routers are required.

(b) Reduces the hub router configuration size and complexity.

(c) Scales the network through scaling of the network at specific hub point.

(d) Hub-and-spoke topology is much economical than other MPLS topologies.

4. DISADVANTAGES OF HUB-AND-SPOKE MPLS L3VPN (LAYER 3 VPN) TOPOLOGY

(a) Route distribution between a set of VRFs in a VPN with Hub-and-spoke connectivity is a little more complicated than that required for full-mesh connectivity (topology) [1].

(b) SP implementations of hub-and-spoke MPLS VPNs can force spoke site traffic to route through a centralized hub site to reach other spoke sites. Creating a hub-and-spoke topology adds a level of complexity to the service.

5. CONCLUSION

This paper concludes that Hub-and-spoke MPLS L3Topology is very beneficial when certain central site services for a particular VPN, such as Internet access, Firewalls, server farms, and so on, are housed within hub site. Or it may be because this particular VPN customer requires that all connectivity between its sites be through the central site.

The Hub-and-Spoke topology is considered where cost is a factor.

The above factors prohibit customer to deploy/provision Full- Mesh topology.

But the Full-Mesh VPN Topology is still in use.

The Hub-and-Spoke topology has other side too. The Hub- and-Spoke topology adds another level of hierarchy which become more complex than Full-Mesh.

REFERENCES

[1] Mark Lewis, "Comparing, Designing and Deploying VPNs", Cisco Press

Public Cloud Security Challenges & Solution

Surabhi Shukla
Maharana Pratap College of Technology, Gwalior (M.P), Rajiv Gandhi Proudyogiki Vishwavidyalay, Bhopal (M.P.), India

ABSTRACT

Cloud computing is a set of IT services that are provided to a customer over a network on a leased basis and with the ability to scale up or down their service requirements. Usually cloud computing services are delivered by a third party provider who owns the infrastructure. It advantages to mention but a few include scalability, resilience, flexibility, efficiency and outsourcing non-core activities. Cloud computing offers an innovative business model for organizations to adopt IT services without upfront investment. Despite the potential gains achieved from the cloud computing, the organizations are slow in accepting it due to security issues and challenges associated with it. Security is one of the major issues which hamper the growth of cloud. The idea of handing over important data to another company is worrisome; such that the consumers need to be vigilant in understanding the risks of data breaches in this new environment. This paper introduces a detailed analysis of the cloud computing security issues and challenges focusing on the cloud computing types and the service delivery types.

Keywords: CSP, TPA, VPN, data traffic, SaaS, PaaS, IaaS

1. INTRODUCTION

Cloud computing has risen as a new computing paradigm that brings unparalleled flexibility and access to shared and scalable computing resources. The increasing demand for data processing and storage in this digital world is leading a significant growth of data centers. Cloud computing encompasses activities such as the use of social networking sites and other forms of interpersonal computing; however, most of the time cloud computing is concerned with accessing online software applications, data storage and processing power. Cloud computing is a way to increase the capacity or add capabilities dynamically without investing in new infrastructure, training new personnel, or licensing new software. It extends Information Technology's (IT) existing capabilities [1]. In the last few years, cloud computing has grown from being a promising business concept to one of the fast growing segments of the IT industry. But as more and more information on individuals and companies are placed in the cloud, concerns are beginning to grow about just how safe an environment it is. Despite of all the hype surrounding the cloud, customers are still reluctant to deploy their business in the cloud. Security issues in cloud computing has played a major role in slowing down its acceptance, in fact security ranked first as the greatest challenge issue of cloud computing.

2. CLOUD COMPUTING PARADIGM

Cloud-computing data centers offer information technology resources as services. The hardware systems and software systems represent the resources the data center provides as Infrastructure as a Service (IaaS) and Platform as a Service (PaaS), respectively. Applications, such as web search, social networking, computation, etc., offered by cloud-computing data centers are hosted as Software as a Service (SaaS). These applications run on virtualized IT resources, namely, virtual machines, provided by IaaS and PaaS. Based on the request, the cloud service providers provision resources such as different types of VMs to the requests.

In cloud computing, the available service models are:

- **Infrastructure as a Service (IaaS) -** Provides the consumer with the capability to provision processing, storage, networks, and other fundamental computing resources, and allows the consumer to deploy and run arbitrary software, which can include operating systems and applications. The consumer has control over operating systems, storage, deployed applications, and possibly limited control of select networking components [3].
- **Platform as a Service (PaaS) -** Provides the consumer with the capability to deploy onto the cloud infrastructure, consumer created or acquired applications, produced using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.
- **Software as a Service (SaaS) -** Provides the consumer with the capability to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices, through a thin client interface, such as a web browser (e.g. web-based e-mail). The consumer does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings[2,6].

Four deployment models have been identified for cloud architecture solutions, described below:

- **Private cloud -** The cloud infrastructure is operated for a private organization. It may be managed by the organization or a third-party, and may exist on premise or off premise.
- **Community cloud -** The cloud infrastructure is shared by several organizations and supports a specific community that has communal concerns (e.g., mission, security requirements, policy, and

compliance considerations). It may be managed by the organizations or a third party, and may exist on premise or off premise.

- **Public cloud -** The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.
- **Hybrid cloud -** The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology, that enables data and application portability (e.g., cloud bursting for load-balancing between clouds) [2,6].

Cloud computing is viewed as one of the most promising technologies in computing today, inherently able to address a number of issues. A number of key characteristics of cloud computing have been identified [3]:

- **On-demand self-service:** A consumer is able to provision resources as needed without the need for human interaction.
- **Broad access:** Capabilities of a Cloud are accessed through standardized mechanisms and protocols.
- **Resource Pooling:** The Cloud provider's resources are pooled into a shared resource which is allocated to consumers on demand.
- **Rapid elasticity:** Resources can be quickly provisioned and released to allow consumers to scale out and in as required.
- **Measured service:** Cloud systems automatically measure a consumer's use of resources allowing usage to be monitored controlled and reported [2, 6].

3. PUBLIC CLOUD ROLE IN CLOUD COMPUTING

When used to extend existing data center footprints, public cloud can deliver big benefits for data backup and scalability. With companies such as Amazon Web Services, Google, Microsoft and Rackspace offering the ability to create virtual machines in the cloud to support and replace physical servers, cloud virtualization services are being integrated into data center infrastructures. But knowing which features to consider and which vendors to compare can be a daunting task. In most cases, established organizations with IT resources on-premises should not dispose of existing servers and move everything to the cloud. It would be a waste of money, unless the local resources were scheduled to be retired. Even then, there may be some workloads, such as latency-sensitive ones, that should run locally. Similarly, you might not want to put all of your domain controllers on a public cloud. Of course, this does not mean companies with established IT infrastructures cannot benefit from the cloud. The best approach is often to treat the cloud as an extension to the organization's existing IT footprint. In this feature, we look at the benefits of using public cloud versus an on- premises data center.

- **Workload scaling -** There are multiple ways a data center benefits from being extended to the cloud, and one involves workload scaling. There may be times your organization needs to ramp up a production workload beyond what the local data center can comfortably handle. For example, consider the way insurance companies operate. For most of the year, insurance companies typically consume a predictable level of IT resources. However, there may be open enrollment periods that occurs throughout the year. These enrollment periods are especially busy for insurance companies. As a result, existing servers may not be able to handle open enrollment workloads. Rather than buy new servers to accommodate temporary spikes in demand, the companies could use public cloud. If the company's enrollment applications are Web-based, it would be relatively easy for cloud-based Web servers to accommodate the seasonal demand. Once open enrollment is over, the cloud-based Web servers can be decommissioned.
- **Business continuity-** Another advantage of cloud-based VMs is protecting businesses in case of equipment failures or physical disasters. To protect against data center failures, some organizations build geographic clusters that span multiple data centers. Then, if a natural disaster destroys an organization's primary data centers, mission- critical workloads fail over to secondary data centers. Building geo-clusters, however, is expensive and complicated. Another way to use public cloud for business continuity is through VM replication. Not every provider or server virtualization platform supports replication, but some cloud and hypervisor combinations allow duplicate VMs to be created in the cloud and kept in sync with on- premises Vms.

Public cloud boasts a number of enterprise benefits, but it isn't perfect. Enterprises should be aware of unpredictable cost structures and other drawbacks. Public cloud services offer enterprises several advantages. They allow for flexible and affordable virtual machine deployments and can boost an organization's data backup and workload-scaling capabilities. However, public cloud isn't without its drawbacks.

- **Multi-tenant environment-** One of public cloud's biggest disadvantages is its multi-tenant environment. The host server running your virtual machine (VM) likely is hosting other companies' VMs. Because of this, public cloud providers don't give you access to the hypervisor, so you can't install host-level utilities, such as antivirus software or backup agents. This also means you can't join a hypervisor to an existing domain or cluster. There are also security implications, as well as potential downtime from cloud or WAN failure. In addition, public cloud providers own the hardware and control the underlying software, so they can make low-level changes at will.

- **Unpredictable costs-**Another disadvantage of running VMs in the cloud is that costs can be wildly unpredictable. Public cloud providers are not known for using simple billing models. Typically, you are billed based on the resources you consume. This includes storage resources, but also CPUs, memory and storage I/O. Resource consumption may be billed differently at different times of the day, and not all activity is treated equally. There are cloud providers that differentiate between various types of CPU functions, billing those functions at different rates. Because public cloud providers use complicated billing formulas, it can be difficult to estimate the cost of running cloud workloads. They can vary each month based on how heavily the workloads are used [5].
- **Backups become complicated-** Another disadvantage is how public clouds can complicate your backup processes. If you have mission-critical VMs running in the cloud, you need a way to back them up. While most cloud service providers perform their own backups, they don't necessarily offer restoration services for customers. This can be complex because most of the off-the-shelf backup products support data backup to the cloud, but not from the cloud. A cloud data backup increases the consumption of storage I/O, network I/O and WAN bandwidth, which may also increase costs [5].

4. CLOUD SECURITY PROBLEM

IT departments are struggling with inadequate tools for protecting data traveling both inside and outside their enterprises. They lack strong network segmentation controls, one of the main security shortcomings that played a role in many recent breaches. a.) The old tools used for data traffic security are clearly inadequate and are now routinely defeated or by-passed by hackers. b.)Several glaring deficiencies in how networks applications are protected today are now increasingly apparent. c.) The continued in flux of mobile devices and personal devices is creating more security challenges. d.) Extension of sensitive applications to points outside the enterprise perimeter, including across the Internet, is creating new challenges.

- **Fractured Security of Data Traffic-** Protection of data traffic from end-to-end was one of the biggest security challenges. Fragmentation of controls over data traffic security that includes a mishmash of VPNs, application- layer and network-layer encryption in use by typical enterprises. IT managers reported that they need to use two or more forms of encryption to secure data traffic in their enterprises. More than a third must contend with three or more forms of encryption for securing their data in motion.

- **Network Segmentation Shortcomings-** Digging deeper, IT professionals, want to use data traffic encryption to provide stronger network segmentation for sensitive applications. But they report being unable to deploy encryption for this purpose because of encryption management issues and device performance issues, among others. We're all familiar with the classic security architecture designed to comply with basic data protection and compliance requirements: "crunchy" on the outside, with a strong, firewalled perimeter, but "soft" on the inside, with internal networks largely trusted. But an emerging best practice for data protection is to encrypt a sensitive application's data traffic regardless of where it is. This is driven by the practical realities of modern security gaps and exploits. It is no longer a safe assumption that the firewall perimeter will always keep the bad guys out. In fact, many security consultants and pen-testers advise IT managers to assume that a breach has already occurred and malware is already present in the formerly trusted zone of the enterprise network.

Similarly, it's no longer safe to assume that a sensitive application will not be shared outside the enterprise perimeter. Enterprise applications of all sorts are now routinely extended to external parties, such as partners or suppliers or employees on the move or in home offices. While it has been a longstanding practice to encrypt traffic when it traverses an external, untrusted network, it is not always easy to police this requirement in the era of BYOD and widespread remote access. Even when an application itself may not contain overly sensitive data, a compromised application of any sort can create an opening for attackers to gain a foothold in the enterprise's more sensitive areas. An increasingly common solution, especially among compliance- oriented IT shops, is to use strong encryption on all sensitive data in motion, even on internal networks.

Almost half of those who want to encrypt but can't say it is because management of the various forms of encryption is too difficult. A little more than a third of the respondents cited the reduced performance of firewalls and network devices when they are used to encrypt traffic. The third main reason cited by those who want to encrypt but can't is the complexity of encrypting at a consistent level across a multi- vendor network environment. Different vendors implement encryption at varying levels, via different standards, or even at varying network or application layers, further contributing to the management fragmentation issues discussed earlier.

IT professionals asked how encryption keys are managed in their environments. Despite widespread availability of advanced key management systems, respondents said that they continue to manage keys manually at each network hop, firewall or VPN node. These management and device performance challenges are forcing enterprises into a dangerous trade-off.

- **Muddled Mobility-** The survey findings about management fragmentation and fractured security controls carry over into the enterprise mobile device management area. The survey found that two thirds of companies are now permitting employees to use their own devices to access corporate applications and enterprise data. But how that traffic to and from the mobile devices is protected varies widely.

The enterprises allowing BYOD indicated that they do not require data traffic to be encrypted to the personal devices, regardless of which networks the traffic will traverse. Of those that do require encryption, the most common modes of data traffic protection were application-embedded HTTPS, enterprise-controlled VPNs and security supplementing a Mobile Device Management System. But none of these was used exclusively by an overwhelming majority of managers, further underscoring the fragmented nature of security controls.

These findings also indicate a shift in how VPNs are viewed and deployed. Traditionally, a VPN served the purpose of connecting a trusted device to a trusted network. Now, as networks are increasingly untrusted, devices are BYOD, and network security controls are increasingly fragmented and soloed, VPNs and data encryption policies focus on connecting users to applications regardless of which device or network is being used.

- **Application Conundrum -** Similarly, a majority of enterprises permits employees to access corporate applications across untrusted networks such as the Internet. Of those permitting this access, a majority reported using encryption embedded in the applications themselves and separate VPN and encryption technologies controlled directly by the enterprise, with at least a third utilizing both. This is a stark illustration of the fragmented nature of data traffic protection. Because they have no single point of control and encryption policy management, enterprises are very often forced to rely on the embedded encryption supplied by an application developer. How strong is thatencryption? Is it consistent with the encryption policies used by the enterprise for other applications elsewhere? What open source components did the application developer use for that function and are the patches up to date? Are key management controls consistent with the enterprise's policies?

- **Identity and access management -** In case of public cloud, the more companies depend on their I-A-M policy which comes under the SLA between the two parties and share Q- o-S among them.

- **Authorization-** It gives the policy to manage the security concern of an individual user. Permission to the given or privileged user to access the system application; what resources should be given to the user for an individual task?
- **Identity provisioning-** Identity of user is the information which it provides at the time of registration. The same information will be seen every time till the account of user is generated. At the time of registration, user has to fulfill some standards which are necessary for the account generation and will further help in improvement of data security. Security concerns are so high that different level of identity checking is required. In the absence of this security checking who will take the responsibility of data.
- **Management of personal data-** The data which is related to user is the personal data. Data can be of any type either it can be official document or some personal stuff; but all the data which user think is important for it, is its personal data. Hardware requirement depends upon the quantity of data which is to be accessed or how the accessibility of data can be increased, so that user don't feel any inconvenience. Variety of cloud provider is there to access the data stored within the cloud data storage.
- **Key management-** policy deals with the keys used for encrypting or decrypting the document. Till date the organizations are confused what to do with the keys. The keys should be traverse manually or it should follow some sensitive way to reach the desired goal.
- **Encryption-** data in transit/rest/memory all can be encrypted but what will be the well-defined policy for this.

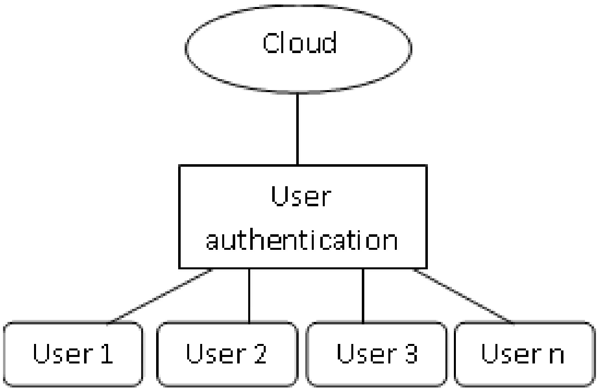


Figure (a)

- **Authentication -** high assurance security operation should be used. This may include login management interfaces, key creation, access to multiple-user accounts, firewall configuration, remote access, etc. can two factor authentication be helpful to manage critical components.

5. STRATEGY

Security remains a major concern for moving data to the cloud. Although data encryption provides protection, decisions need to be made regarding when, where and how to encrypt data heading to cloud. The proliferation of cloud technology certainly hasn't hurt the security industry. As more people climb aboard the cloud bandwagon, data security ranks at the top of every adopter's list, regardless of the platform. However, many IT pros place less of a burden on security because of increased throughput and stronger encryption standards.

While advances in security and cloud technology are robust, security pros should be careful when moving data to the cloud and pay attention to when, where and how cloud-bound data is encrypted. There are a few ways to encrypt cloud- bound data, depending on your cloud stage: before, during or after the move to cloud.

- Data encryption before taking the cloud plunge-It seems obvious to encrypt data before moving it to the cloud. But the data that must be encrypted before a move to the cloud is data at rest. The encryption of data in transit -- while extremely important -- may not suffice in every circumstance. For example, the HeartBleed vulnerability took many security pros by surprise because HTTPS/SSL was previously considered rock solid. Admittedly, HeartBleed was more of Apache Web server vulnerability than HTTPS, but many cloud providers' management interfaces reside on similar servers. However, data encrypted before it reaches the Internet is in a better position to defend against HeartBleed. The HeartBleed vulnerability focused on stealing login credentials rather than data is actual data, but access to unencrypttrivial once login information has been compromised. Accessing data that was encrypted prior to an HTTPS login is a different matter entirely.

- Encrypting data during a move to the cloud-Encrypting data in transit to the cloud is vital for security and its importance cannot be overstated. Furthermore, encrypting in-transit communications is becoming so popular that a reversal of the current trend seems highly unlikely. Many times -- though not always -- cloud data encryption in transit requires trust in the vendor destination or third-party technology. The cloud vendor or third party must be equally dedicated to security; solely relying on the encryption in-transit is risky business [4].

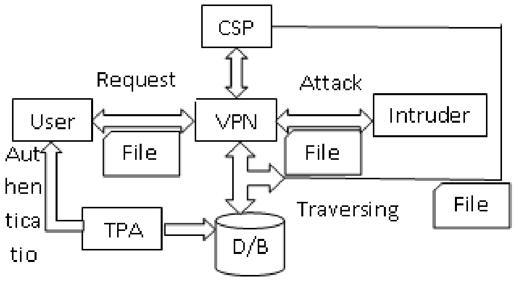


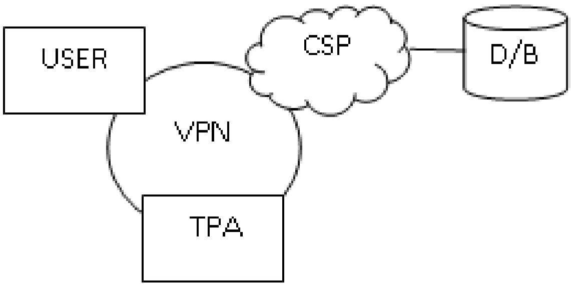
Figure (b)

- **Data encryption comes full circle after the cloud -** Data encryption following a move to the cloud brings the issue of data at rest full circle. At this point, the cloud provider is responsible for data encryption. However, several issues arise when enterprises rely on the cloud provider for data security -- including the ownership. Several cloud providers, such as Amazon Web Services and Google Cloud, have solid security mechanisms replete with encrypted files, SSL login for management and disaster recovery. But, if the data resides on AWS servers Google cloud, who owns the data and encryption keys? It only takes one lawsuit against a cloud provider to expose proprietary data -- encrypted or otherwise -- during legal proceedings. Admins need to have alternatives to relying on cloud providers for security in the event of data compromise. Whatever method is chosen, power brokers within each organization should make tolerable levels of risk clear.

We have kept all these above mentioned three major points in our mind. Data to be encrypted is a necessary task which has to be performed and we will do the same. But meanwhile we have to focus on that only encryption isn't the solution. We have to match this encryption with some more features to make the document more reliable.

Security concern should be start with the initial stage of authentication. Whatever data-file user want to access from the database server which is a cloud based server, the user has to confirm that it's a legal entity or not. With some security features we will check that the user is loyal, if user found loyal then the file can be easily accessed but if it found that the user isn't loyal then it can never view the file[4].

There are three entities user, CSP, TPA together they all generate a VPN among them. We have to focus on how the data travels among all these. If user desire for a particular file it will ask CSP for the individual file, because CSP is the mediator between the original database server. The request when reaches to the CSP,it will immediately forward this to TPA so that TPA can verify the actual user and then further process can be done. This is the core task of TPA to verify the user. After the verification the task of TPA will be finished. We can't blame TPA for any data breach because this isn't the TPA's task. It's only there for user verification.



Figure©

After the user verification the further process is then transferred to CSP, then CSP will check how and what files to be delivered to user. CSP will take the individual file and give it back to user.VPN will be the network in which the transfer of files takes place.

Implementing data encryption controls in a cloud environment can be quite challenging for organizations using those services. Ed Moyle discusses the first two steps to implementing data security in the cloud.

When it comes to data security in the cloud, two things are true: First, more and more sensitive information is going there, and second, traditional data protection controls like encryption of data at rest are unlikely to be applied once it's there. As a proof point of that statement, consider the recent 2013 Global Trends in Cloud Encryption research published in April 2014. According to the survey, 53% of organizations have transferred sensitive or confidential data to the cloud. Yet only 39% of data in software (SaaS) applications is encrypted, and that number lowers to 26% when it comes to platform (PaaS) and infrastructure as a service (IaaS) deployments.

It probably goes without saying that there are some good reasons why this is the case. First of all, data protection controls in a cloud context can be challenging to implement architecturally. Recall that in the cloud, the underlying portions of the stack below that which the customer uses are deliberately opaque; the customer (by design) cedes direct control over these layers to the cloud service provider (CSP) from a management standpoint. So, unless the CSP specifically provides data protection features (note that some do), a customer's ability to implement technical data protection controls without additional engineering might be constrained.

Secondly, from a process standpoint, these controls can be challenging to pull off. There might be legitimate reasons why a CSP requires access to enterprise data -- for example, to debug application functionality or in the case that they provide monitoring services to your company. This means that the logistics of who will hold the encryption keys, as well as whether, how and even if the CSP can get access to them for legitimate business purposes requires discussion, planning and well-thought-out procedures established in advance.

Not everyone will be in a position to implement cloud data encryption controls right away, but even so, taking a few steps now let's organizations analyze both the security benefits they'd get and where those benefits are most needed. It will also help companies understand the level of difficulty if and when they do implement, and, if done strategically, can actually make the implementation process that much easier when a business does decide to pull the trigger.

- Laying the groundwork-The first of these steps is data classification and service inventorying. This sounds like "eat your vegetables" advice at some level, but the reason why it's important is the sheer volume of cloud services that even a modest-sized organization will have in play. Most organizations have dozens of cloud services in active use; in fact, if you include SaaS applications -- both sanctioned ones and consumer-oriented services employees may use with or without your say-so -- cloud services might number into the hundreds or even thousands.

Not all of those applications will process sensitive or confidential information, so not everyone will require encryption of data. Distinguishing which applications are appropriate to apply encryption to from those where it is not is the crux of this first step. Essentially, the goal is to identify and record -- in as granular detail as possible -- where data an enterprise might want to encrypt resides in the cloud. For some situations (i.e., SaaS), "as granular as it gets" might be that the data is held at a certain CSP. For others (i.e., IaaS), it could be that you get down to the level of a certain virtual device or storage container. The point is, your organization should know which applications and environments process the data you care about, versus those which do not, and you should be able to construct a rough idea of where you'd need to apply controls.

If this sounds like a tall order, it can be. Start with a manageable subset and build on it. There are tools that assist in this regard. In a private cloud context or one where your business has a fairly extensive relationship with an IaaS provider, virtualization-aware tools can assist in the inventorying of specific hypervisors to help determine what's running where. SaaS discovery tools exist as well, but in a pinch some service-level information can be gleaned from examining user traffic .To automate the task of keeping track of specific systems and applications, can provide an assist as well to record services and usage as they're identified. The point is, organizations need to establish which data is in which environments so that they can prioritize their efforts.

- **Evaluating specific usage -** After data classification and service inventorying, the next step is where the "rubber meets the road." It's here where your organization must evaluate the specific usage, make the determination about whether it will encrypt, and decide how it will implement encryption. Note that depending on the cloud computing model or service your organization is using, it may need to select different tools to affect this. For example, if you have a high-sensitivity SaaS application and you want to encrypt data within it, affecting this is very different from encrypting a database within a PaaS or encrypting volume storage in an IaaS.

With an IaaS use case, for example, since you have access to the underlying OS on virtual images within that environment, you might choose to implement a tool that operates at the file-system level. In fact, Microsoft and most Linux distributions natively support encrypted file systems that may be viable options. There are dozens of commercial products that support this as well. For a PaaS, your choices might be more limited; you may need support from the developers actively working in the PaaS to author code that leverages CSP APIs or that leverage specific APIs in the application environment they're working in. And, of course, in a SaaS context, since the entirety of the application stack is managed by the CSP, you may find yourself looking to reverse- proxying tools or a specific SaaS-integrated product to accomplish that.

The point is the tools vary and these differences should be noted and planned around; if your organization needs to purchase multiple tools to do this, it will need to plan its budget accordingly. Using the inventory and data classification evaluation that you've already done can help prioritize which approach is most valuable and/or urgent.

6. CONCLUSION

In light of these data traffic security problems, it's no wonder that network security improvements rank among the enterprises' IT project priorities for 2015, according to the survey findings. More than half indicated that network security improvements are planned for 2015 and nearly a quarter named network security as a top IT priority for the enterprise in the coming year. In total, two-thirds of enterprises report that they are budgeting such projects. If enterprises are studying and learning from the recent parade of data breaches, then we can safely predict that several initiatives will be included in these network security projects:

1. **Proactive security:** More enterprises will establish proactive security and stronger network segmentation by encrypting sensitive data traffic over all networks.
2. **Encryption consolidation:** Reducing the number of forms of encryption and consolidating encryption control will make protecting traffic simpler and reduce the possibilities of gaps in the end-to-end data path.
3. **Simpler policy management:** Because all networks are essentially untrusted today, it makes less and less sense to focus solely on network-based VPNs that connect a device to a network. Instead, encryption policies are now focusing on connecting an authorized user to the applications they want to access and then applying the required encryption profile. The policy should be applied regardless of which devices or networks are involved, which in turn enables more consistent, enforceable and auditable encryption policies.

In the end, the IT security community already has benefited greatly from the lessons learned by the surge in hack attacks. IT security now has the attention of senior management and budget decision-makers. In the long run, this heightened prioritization and investment can only improve the overall effectiveness of security controls and allow them to evolve to meet the changing needs of users and applications in the modern enterprise.

REFERENCES

[1] "Cloud Computing Security Issues and Challenges"; International Journal of Computer Networks (IJCN), Volume (3): Issue (5): 2011; Kuyoro S. O., Ibikunle F. & Awodele O.

[2] "Observing the Clouds : A Survey and taxonomy of Cloud Monitoring"; Ward and Barker Journal of Cloud Computing:Advances, Systems and Applications (2014) 3:24

[3] "Addressing cloud computing security issues";Dimitrios Zissis , Dimitrios Lekkas; Future Generation Computer Systems 28 (2012) 583–592

[4] www.techgig.com

[5] www.techtarget.com

[6] "Exploring Cloud Computing for Naïve";Reema Ajmera & Rudra Gautam; IJCSNS International Journal of Computer Science and Network Security, VOL.14 No.12, December 2014 62

[7] NIST cloud definition, version 15 http://csrc.nist.gov/groups/SNS/cloudcomputing/.

Author Profile

Surabhi Shukla holds a B.E. in Computer Science, from RGPV and is currently pursuing M.E. in Computer Science at the same university RGPV. She has been involved with Infosysworld, as a business analyst for 1 year. Her interest area is Cloud Computing and database security. She is trying harder to secure database in cloud.

A Survey on DNA Based Cryptography

Manisha[1], Pooja Ahlawat[2]

[1]M.Tech Student, R. N. College of Engineering & Management, Maharshi Dayanand University, Rohtak, Haryana, India

[2]Associate Professor, Department of Computer Science and Engineering, R. N. College of Engineering & Management, Maharshi Dayanand University, Rohtak, Haryana, India

ABSTRACT

Two level security needs the hiding of data into a cover medium that cover medium is also inserted into another cover media. Both cover media may be same or different depending upon features and applications. The paper study the DNA, operation on the DNA. Then the paper focus on DNA based security i.e. cryptography and Steganography by using DNA. This paper discusses DNA cryptography and the difference between the traditional and the DNA cryptography. This paper also brief the various work done in the field of the DNA cryptography.

Keywords: DNA, cryptography, DNA cryptography, DES, PCR

1. INTRODUCTION

Cryptography is the science and art of secret writing [1][2]. It studies some mathematical techniques and provides mechanisms necessary to provide aspects related to information security like confidentiality, data integrity, entity authentication, and data origin authentication [2].

Symmetric algorithms are cryptosystems that either a secret key will be shared for both encryption and decryption [1][5]. The algorithms of symmetric cryptosystems are very strong against possible attacks, but mainly weakness of symmetric cryptosystems is brute-forcing the secret key. This characteristic creates the biggest critical act in any cryptosystem that uses symmetric algorithms which is distribution of the shared secret between the two parties like DES algorithms. Asymmetric algorithms use different values the clear difference between the traditional and DNA based cryptography that is specified in the table 1.

Table 1: Comparison of traditional and DNA cryptography [4]

Characteristics	Traditional Cryptography	DNA cryptography
Security	Less	More
Time	Minutes to hours	Hours to days
Storage capacity	In MB	In TB
Dependency	On Implementation environment	On environmental conditions

2. TECHNOLOGY USED IN DNA COMPUTATION

for encryption and decryption and do not need to share secret between two parties. Each party only has to keep a secret of its own. The earliest foundation of asymmetric algorithms known as public key cryptosystems comes from key exchange problem of symmetric algorithms. In 1976, Whitfield Diffie and Martin Hellman proposed a method were the sender and receiver do not have to share a secret. That was the first work on hybrid cryptosystem [1][2].

DNA cryptography, a new branch of cryptography utilizes DNA as an informational and computational carrier with the aid of molecular techniques. It is relatively a new field which emerged after the disclosure of computational ability of DNA [5]. DNA cryptography gains attention due to the vast storage capacity of DNA, which is the basic computational tool of this field. One gram of DNA is known to store about 108 tera-bytes. This surpasses the storage capacity of any electrical, optical or magnetic storage medium [5], [6]. Traditional cryptographic systems have long legacy and are built on a strong mathematical and theoretical basis. Traditional security systems like RSA, DES or NTRU are also found in real time operations. So, an important perception needs to be developed that the DNA cryptography is not to negate the tradition, but to create a bridge between existing and new technology. The power of DNA computing will strengthen the existing security system by opening up a new possibility of a hybrid cryptographic system. This needs Today, various techniques are used to carry out DNA computation. Researchers use these techniques for performing the operations on informative DNA molecules. Some of these technologies are as follows:

Gel electrophoresis: It is a phenomenon used to separate the DNA fragments according to their length. A gel of polyacrylamide or agarose is prepared. The negatively charged DNA molecules are placed in the wells which are situated at one side of this gel. On the application of an electric current to the gel, the negatively charged DNA molecules will start moving towards the positive pole, where the shorter molecules travel faster than the larger ones. Hence, a separation between them can be detected easily [14].

Polymerase Chain Reaction (PCR): As it is difficult to manipulate the small amount of DNA, an amplification process is carried out. PCR has very high amplification efficiency, hence, this technology is used to amplify and quantify the DNA. In DNA amplification using PCR, required DNA segments are cloned into vectors. For PCR amplification two things are required, a primer and a DNA template. DNA template is a single-stranded DNA sequence which contains the segment which is to be amplified and primer is a complement sequence of that segment. A primer is annealed with the DNA template. After that, DNA polymerase enzyme initiates DNA synthesis process by successively adding the nucleotides to 3' end of the primer, until the desired DNA strand is obtained. Primer always extends in the direction 5' to 3' only. The desired DNA strand starts with the primer and is always complementary to the DNA template. The whole PCR process can be divided into two steps:

- Designing the two primers and loading them separately, one at the beginning and another at the end of target DNA.
- Matching the primers with their complement sequences in template DNA [15].

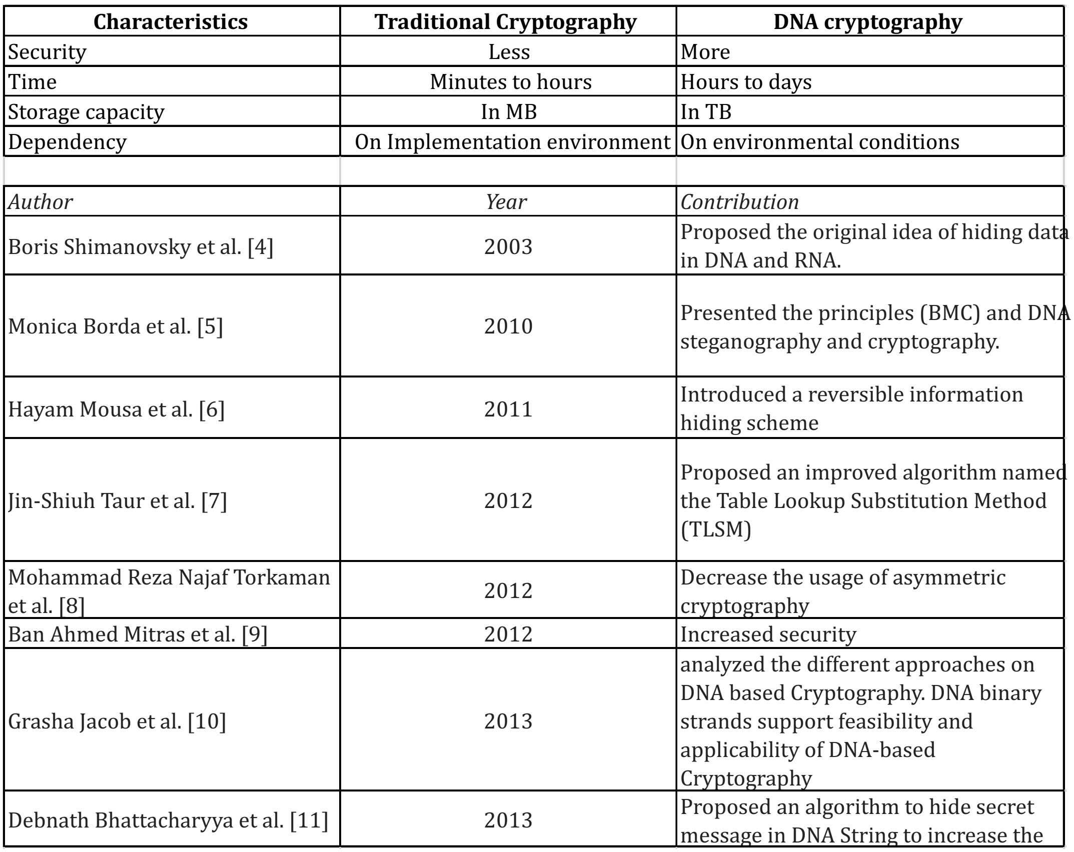
DNA Chip technology: With the help of DNA chip, a vast amount of genome-sequencing data can be manipulated [16]. It is used to find the expression of several genes in parallel. DNA chips stores data in the form of DNA sequences. In DNA chips, a huge number of spots are embedded on solid surface, generally a glass slide. Each and every spot of a chip consists of different type and number of probes. Probes are small single-stranded DNA sequences have the ability to bind with their complementary DNA sequences. Binded DNA sequences are labelled fluorescently which are observed under laser dye. Depending upon the ratio of binding between probe and DNA of each spot, data is calculated by electronic means [17].

3. RELATED WORK

DNA chromosomes indexing. Hayam Mousa et al. [6] introduced a reversible information hiding scheme for DNA sequence based on reversible contrast mapping. The scheme uses two words of the sequence with the reversible contrast mapping to achieve reversibility. Jin-Shiuh Taur et al. [7] proposed an improved algorithm named the Table Lookup Substitution Method (TLSM) to enhance the performance of an existing data hiding method called the substitution method. Moreover, a general form of the TLSM is discussed, which includes the original method as a special case.

Boris Shimanovsky et al. [4] (2003) proposed the original idea of hiding data in DNA and RNA. The first is a simple technique that hides data in non-coding DNA such as non- transcribed and non-translated regions as well as non-genetic DNA such as DNA computing solutions. The second technique can be used to place data in active coding segments without changing the resulting amino acid sequence. Monica Borda et al. [5] (2010) presented the principles of bio molecular computation (BMC) and several algorithms for DNA (deoxyribonucleic acid) steganography and cryptography: One- Time-Pad (OTP), DNA XOR OTP and Mohammad Reza Najaf Torkaman et al. [8] proposed to decrease the usage of asymmetric cryptography and introduced a novel cryptographic-steganography protocol. The main advantage of proposed cryptography protocol was using innovative DNA steganography techniques to conceal secret session key which is transferred among sender and receiver throughout unsecured channel. Ban Ahmed Mitras et al. [9] discussed a reference DNA sequence has been shared between sender and receiver. Not only this DNA reference sequence can be retrieved from EBI or NCBI databases but it can also be simply selected from any database. Therefore, by considering any sort of database, there are 163 million targets to select it. Virtually, guessing the correct DNA sequence by attacker is unachievable. Grasha Jacob et al. [10] (2013) analyzed the different approaches on DNA based Cryptography. They said that DNA binary strands support feasibility and applicability of DNA-based Cryptography. The security and the performance of the DNA based cryptographic algorithms are satisfactory for multi-level security applications of today's network. Debnath Bhattacharyya et al. [11] (2013) proposed an algorithm to hide secret message in DNA String to increase the security during transmission of data. In this paper, we propose a new Binary Coded DNA rules towards Data Hiding in DNA. K. These works are also explained in the following table i.e. table 2:

Table 2: Related Work



Characteristics	**Traditional Cryptography**	**DNA cryptography**
Security	Less	More
Time	Minutes to hours	Hours to days
Storage capacity	In MB	In TB
Dependency	On Implementation environment	On environmental conditions
Author	*Year*	*Contribution*
Boris Shimanovsky et al. [4]	2003	Proposed the original idea of hiding data in DNA and RNA.
Monica Borda et al. [5]	2010	Presented the principles (BMC) and DNA steganography and cryptography.
Hayam Mousa et al. [6]	2011	Introduced a reversible information hiding scheme
Jin-Shiuh Taur et al. [7]	2012	Proposed an improved algorithm named the Table Lookup Substitution Method (TLSM)
Mohammad Reza Najaf Torkaman et al. [8]	2012	Decrease the usage of asymmetric cryptography
Ban Ahmed Mitras et al. [9]	2012	Increased security
Grasha Jacob et al. [10]	2013	analyzed the different approaches on DNA based Cryptography. DNA binary strands support feasibility and applicability of DNA-based Cryptography
Debnath Bhattacharyya et al. [11]	2013	Proposed an algorithm to hide secret message in DNA String to increase the

4. DRAWBACK OF EXISTING WORK

The unintended user gets to know that data is hided in the particular DNA then the extraction of data is possible in the DNA based Steganography. It is due to the fact the data is in plain form in the DNA. While the cryptography makes the data in encrypted form but the data is visible to the unintended user. To make the process more robust and secure the data must hided must be cascaded by cryptography and the Steganography.

5. CONCLUSION

This paper discusses the structure of the DNA along with the DNA cryptography. This paper also briefs the work done in the area of the DNA cryptography. The difference between the traditional and DNA cryptography clears the importance of the DNA cryptography. The drawback of the previous work defines the open area of research in the field of DNA cryptography. In future an algorithm can be designed for DNA based cascaded Steganography and cryptography.

REFERENCES

[1] Torkaman M.R.N., Nikfard P., Kazazi N.S., Abbasy M.R., and Tabatabaiee S.F.: Improving Hybrid Cryptosystems with DNA Steganography. E. Ariwa and E. El-Qawasmeh (Eds.): DEIS 2011, CCIS 194, pp. 42– 52, 2011

[2] Alia, M.A., Yahya, A.: Public–Key Steganography Based on Matching Method. European Journal of Scientific Research, 223–231 (2010)

[3] Kumar, S., Wollinger, T.: Fundamentals of Symmetric Cryptography. Embedded Security in Cars, 125–143 (2006)

[4] Shimanovsky, B., Feng, J., & Potkonjak, M. (2003, January). Hiding data in DNA. In Information Hiding (pp. 373-386). Springer Berlin Heidelberg.

[5] Borda, M., & Tornea, O. (2010, June). DNA Secret Writing Techniques. In IEEE conferences.

[6] Mousa, H., Moustafa, K., Abdel-Wahed, W., & Hadhoud, M. M. (2011). Data hiding based on contrast mapping using DNA medium. Int. Arab J. Inf. Technol., Volume- 8 Issue (2), pp 147-154.

[7] Taur, J. S., Lin, H. Y., Lee, H. L., & Tao, C. W. (2012). Data Hiding In DNA Sequences Based On Table LookUp Substitution. International Journal of Innovative Computing, Information and Control, Volume 8 Issue (10).

[8] Torkaman, M. R. N., Kazazi, N. S., & Rouddini, A. (2012). Innovative approach to improve hybrid cryptography by using DNA steganography. International Journal of New Computer Architectures and their Applications (IJNCAA), Volume-2 Issue (1), pp. 224- 235.

[9] Mitras, B. A., & Aboo, A. K. (2012). Proposed Steganography Approach Using Dna Properties. International Journal of Information Technology and Business Management, Volume-14 Issue 1.

[10] Jacob, G., & Murugan, A. (2013). DNA based Cryptography: An Overview and Analysis. International Journal of Emerging Sciences, Volume 3 Issue (1), pp.36-27.

[11] Bhattacharyya, D., & Bandyopadhyay, S. K. (2013) Hiding Secret Data in DNA Sequence. International Journal of Scientific &Engineering Research Volume 4.

[12] Mitras, B. A., & Aboo, A. K. (2012). Proposed Steganography Approach Using Dna Properties. International Journal of Information Technology and Business Management, Volume-14 Issue 1.

[13] Yamuna, M., Dangi, M. K., & Singh, K. (2013). Encryption of a Binary String Using DNA Sequence. International Journal of Computer Science, Volume 2, Issue (02).

[14] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell "Molecular Cell Biology", 5th ed. New York: W. H. Freeman and Co. 2003.

[15] G. Cui, L. Qin, Y. Wang, and X.Zhang, "An encryption scheme using DNA technology," in IEEE 3rd International conference on Bio-Inspired Computing: Theories and Applications (BICTA08), Adelaid, SA, Australia, pp. 37–42, 2008

[16] P. Gwynne and G. Heebner, "Technologies in DNA chips and microarrays: I," Science, vol. 4 May, p. 949, 2001.

*[17] T. Tsukahara and H. Nagasawa, "Probe-on-carriers for oligonucleotide microarrays (DNA chips)," Science and Techn*ology of Advanced Materials,Elsevier Science, vol. 5, pp. 359–362, 2004.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

-

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.

The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.