International Conference on Recent Advances in Engineering Science and Management
PHD Chamber of Commerce and Industry, New Delhi
30 August 2015, www.conferenceworld.in, www.icraesm.org
ISBN: 978-81-931039-5-1

SNR IMPROVEMENT FOR EVOKED POTENTIAL ESTIMATION USING WAVELET TRANSFORM TECHNIQUES

Shailesh M L[1], Dr.Anand Jatti[2], Madhushree K S[3]

[1]Research Scholar, Pacific University, Udaipur, Rajasthan, Member of IEEE, (India)

[2]Professor Dept. of Electronics and Instrumentation, R.V.College of Engineering, Bangalore, Karnataka, (India)

[3]UG Student, Dept of Electronics & Communication Engineering, VTU, Belgaum, Karnataka, (India)

ABSTRACT

Digital Signal Processing uses mathematical analysis and algorithms to extract information hidden in signals derived from sensors[1]. The Biomedical Signal contaminated by noise and artifcats. The problem of estimating one signal from another is one of the most important in signal processing[9]. In many applications, the desired signal is not available or observable directly. Instead the observable signal is a degraded version of the original signal. The signal estimation problem[12] is to recover in the best way possible , the desired signal from its degraded replica. In this case, the desired signal may be corrupted by strong additive noise, such as weak evoked brain potentials measured against the strong background of on going EEG (ElectroEncephalogram) Signals.

Wavelet transform *technique of estimation improves the SNR by a large amount in almost one sweep of EP. The two different wavelet transforms such as Daubechies wavelet transform and Bi – Orthogonal wavelet transform have been used to improve the SNR.*

SNR comparison is made with the conventional ensemble averaging technique, where this technique requires more number of sweeps to improve the SNR. Comparison is made to understand the best Daubechies wavelet transform and Bi – Orthogonal wavelet transform for estimating the EP signal. In this paper, Visual Evoked Potential signals have been considered for the analysis.

Keywords: Bi – Orthogonal, Daubechies, Evoked Potential, Ensemble Averaging, SNR.

I. INTRODUCTION

The brain is the most complex structure in the well known universe. The brain dominate many highly specialized component parts each associated with specific functionalities, i.e., memory and vision. While these parts work united, each part is amenable for a specific function. To analyze the functional status of the brain such as in anesthesia, hypoxia sleep (lack of oxygen) and in certain nervous diseases, i.e., epilepsy, the brain's recordable neuro electric signals, called electroencephalogram (EEG), are processed and analyzed The brain electrical activity, that occurs in connection with an external stimulus (auditory, visual or somatosensory), is called **Evoked Potential** (EP). If the analysis is relevant to a cognitive activity, the response signal is frequently

ISBN: 978-81-931039-5-1

called as either event-related-potential (ERP) or cognitive EP in a wide range of cognitive paradigms. EPs are important diagnostic tools in investigation of physiological and psychological situation of subjects[2]. In general, EPs or ERPs are not recognizable by visual inspection since they are buried in spontaneous Electroencephalogram (EEG) with signal-to-noise ratio (SNR) as low as -5dB considering stimulus-unrelated background EEG as the noise in the measurements. The split up of the EP (the signal) and the ongoing EEG (the noise) in the measurements have been very attractive points in this paper. This needs use of powerful Bio – Medical Digital Signal Processing tools and several methods have been proposed for this purpose.

1.1 Visual Evoked Potential (VEP)

Evoked potentials (EP) constitute a relatively new method of clinical neurophysiology allowing functional evaluation of the neural system. Such non-invasive techniques give information about the functional state of different tracts within the central nervous system, specifically when the[3] clinical signs and the results of neuro imaging methods are either non informative or non-definable. Evoked potentials are very much useful in the detection of subclinical dysfunction.

The first recognition of visual evoked potentials (VEP) coincides with the discovery of electroencephalography. It was observed earlier the electrical activity of the brain is altered when an intensive light stimulus is applied. However - since these potentials are of very low amplitude widespread use of the method was made possible only by the introduction of computerized averaging techniques.

1.2 Recording of Visual Evoked Potentials (VEPs)

- VEPs are recorded from the occipital region of the scalp (visual cortex) with reference at the vertex
- The most common stimulation modalities are pattern reversal (about 2reversals per second) and flashing (about 5…7 flashes per second)
- It lasts up to 300 ms (and beyond)
- The VEP amplitude is up to 20 µV
- The maximun100 stimuli enough for averaging
- The spectral contents or frequency range 1…300 Hz[4]

1.3 Earlier Methods

During the analysis of real biomedical signals it can almost always be seen noise that distorts the signal. The presence of interference is associated with the specific acquisition of these signals. For example in the case of bioelectric signals, disturbances may come from the hardware retrieves those signals, the power line or the bioelectric activity of body cells. The bioelectric signals, which are widely used in most fields of biomedicine, are generated by nerve cells or muscle cells. The electric field propagates through the tissue and can be acquired from the body surface, eliminating the potential need to invade the bio medical system. However, using surface electrodes results in high amplitude of noise and the noise should be suppressed to extract a priori desired information.

There are many approaches to the noise reduction problem while preserving the variability of the desired signal morphology. One of the possible methods of noise attenuation is low-pass filtering such as arithmetic mean. The classical band-pass filtering is very simple method but also very ineffective because the frequency

ISBN: 978-81-931039-5-1

characteristics of signal and noise significantly overlap. The methods of noise attenuation are ensemble averaging techniques and based on transforming the input space of signal and creating a new space with the help of wavelet transform. In the case of repeatable biomedical signals, another possible method of noise attenuation is the synchronized averaging. The method assumes that the biomedical signal is quasi-cyclic and the noise is additive, independent and with zero mean.

Weighted averaging techniques gave the best signal-to-noise ratios when compare with ensemble averaging technique. By considering the statistical parameters like standard deviation should be considered for changing weight of the each sample. Two EP samples have been considered for the analysis. Weighted averages of brain evoked potentials (EP's) are obtained by weighting each single EP sweep prior to averaging. These weights are shown to maximize the signal-to-noise ratio (SNR) of the resulting average if they satisfy a generalized eigenvalue problem involving the correlation matrices of the underlying signal and noise components.

A parametric method of identification of event-related (or evoked) potentials on a single-trial basis through an AR, MA and ARMA algorithm is implemented. The basic estimation of the information contained in the single trial is taken from an average carried out on a sufficient number of trials, while the noise sources are EEG. The simulations as well as the experimental results confirm the capability of the model of drastically improving the S/N (signal-to-noise) ratio in each single trial and satisfactorily identifying the contributions of signal and noise to the overall recording.

Adaptive filters have been widely utilized in applications that include channel equalization, echo cancellation, radar, linear prediction, spectral analysis and system identification [9]. Here the discussion is about the use of adaptive filters for adaptive noise cancellation for Evoked Potentials. The implementation of the Wiener theory for adaptive noise cancellation requires infinite filter weights to minimize the output error. To make the Wiener solution realizable, a finite number of filter weights must be used. That is, adaptive filters must assume the Wiener filter as an FIR filter.

Wavelet analysis represents the next logical step: a windowing technique with variable-sized regions. Wavelet analysis allows the use of long time intervals where one[10] [11] want more precise low frequency information, and shorter regions where one want high frequency information.

In this paper ensemble averaging technique and wavelet transform techniques have been implemented for improving the output SNR values.

II. METHODS

2.1 Ensemble Averaging Technique

Signal averaging is a technique for separating a repetitive signal from noise without introducing signal distortion. Ensemble signal averaging sums a set of time epochs of the signal together with the super imposed random noise. If the epochs are properly aligned, [8] the signal waveforms directly sum together. On the other hand, the uncorrelated noise averages out time. Thus, the signal – to – Noise **(SNR)** is improved.

Signal averaging is based on the following characteristics of the signal and the noise.

1. The signal waveform must be repetitive (although it does not have to be periodic).
2. The noise must be random and uncorrelated with the signal. In this application random means that the noise is not periodic and that it can only be described (e.g. by its mean and variance).
3. The temporal position of each signal waveform must be accurately known.

ISBN: 978-81-931039-5-1

In this method SNR is improved as more number of sweeps is considered for averaging. The relation below represents that SNR improvement factor.

This can be proven mathematically as follows

The input waveform $f(t)$ has a signal portion $S(t)$ and a noise portion $N(t)$. Then

$$f(t) = S(t) + N(t) \quad (1)$$

Let $f(t)$ be sampled every T seconds. The value of any sample point in the time epoch ($i = 1, 2, \ldots, n$) is the sum of the noise component and the signal component.

$$f(iT) = S(iT) + N(iT) \quad (2)$$

Each sample point is stored in memory. The value stored in memory location i after m repetitions is

$$\sum_{k=1}^{m} f(iT) = \sum_{k=1}^{m} s(iT) + \sum_{k=1}^{m} N(iT) \quad (3)$$

The signal component for sample point i is the same at each repetition if the signal is stable and the sweeps are aligned together perfectly. Then

$$\sum_{k=1}^{m} S(iT) = m\, S(iT) \quad (4)$$

The assumptions for this development are that the signal and noise are uncorrelated and that the noise is random with a mean of zero. After many repetitions, $N(iT)$ has an rms value of σn.

$$\sum_{k=1}^{m} N(iT) = \sqrt{(m\, \sigma n^2)} = \sqrt{(m)}\, \sigma n \quad (5)$$

Taking the ratio of Eqs. (4) and (5) gives the SNR after m repetitions as

$$SNRm = \sqrt{m}\, SNR \quad (6)$$

Thus, signal averaging improves the SNR by a factor of m

$$SNRm = sqrt(m)*SNR \quad (7)$$

Where **m** is number of sweeps

2.1.1 Algorithm for Ensemble averaging Technique

1. Take the different ensemble data and store it in different arrays
2. Add the first position values of all the arrays and store it in first position of another array, likewise all the position values are to be added and stored.
3. Calculate the average by dividing it by number of sweeps.
4. For SNR plot calculate the output SNR for each sweep and store it in an array, finally plot the SNR array elements.

2.2 Wavelet Transform Technique

Wavelet transforms have evoked considerable interest in the signal processing community. They have found applications in several areas such as speech coding, edge detection, data compression, extraction of parameters for recognition and diagnostics etc. since wavelets provide a way to represent a signal on various degrees of resolution, they are convenient tool for analysis of data and manipulation of data. Wavelet transform already

ISBN: 978-81-931039-5-1

discussed in the early part of this paper. Next we will see algorithm for EP estimation using Wavelet Transform [5].

2.2.1 Daubechies Wavelets Approximation

Non-linear approximation is obtained by thresholding low amplitude wavelet coefficients.

This defines the best M-terms approximation fM of f:

$fM=\sum|\langle f,\psi j,n\rangle|>T\langle f,\psi j,n\rangle\psi j,n.$

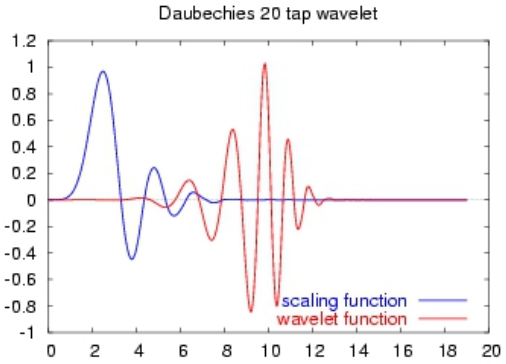


Fig. 2.1 The Shape of a Wavelet

A wavelet coefficient is an inner product $dj[n]=\langle f,\psi j,n\rangle$ with a wavelet atom $\psi j,n$.

A wavelet atom $\psi j0,n0$ can be computed by applying the inverse wavelet transform to coefficients $\{dj[n]\}j,n$ such that

dj[n] = 1 if j=j0 and n=n0,

0 otherwise.

2.2.2 Biorthogonal Wavelet

A **Biorthogonal wavelet** is a wavelet where the associated wavelet transform is invertible but not necessarily orthogonal. Designing Biorthogonal wavelets allows more degrees of freedom than orthogonal wavelets. One additional degree of freedom is the possibility to construct symmetric wavelet functions.

In the Biorthogonal case, there are two scaling functions, which may generate different multiresolutional analyses, and accordingly two different wavelet functions $\psi, \tilde{\psi}$. So the numbers *M* and *N* of coefficients in the scaling sequences $a, \tilde{a}$ may differ[6]. The scaling sequences must satisfy the following biorthogonality condition

$$\sum_{n\in\mathbb{Z}} a_n \tilde{a}_{n+2m} = 2\cdot\delta_{m,0}$$

.

Then the wavelet sequences can be determined as

$$b_n = (-1)^n \tilde{a}_{M-1-n} \qquad (n=0,\ldots,N-1)$$

$$\tilde{b}_n = (-1)^n a_{M-1-n} \qquad (n=0,\ldots,N-1)$$

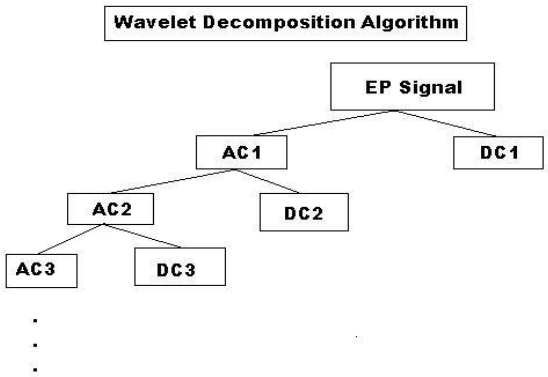
2.3 Algorithm for Wavelet Transform Technique

1. Decompose the signal by applying the discrete wavelet transform[7] on the signal and is shown in Fig.2.1
2. Remove the high frequency signal i.e. detailed coefficients and retain the low frequency components i.e.

ISBN: 978-81-931039-5-1

approximation coefficients[7][8].

3. Reconstruct the EP signal by applying inverse wavelet transform of the decomposed signal and is shown in Fig.2.2
4. Make all detailed coefficients to zero, while applying inverse wavelet transform.
5. Calculate the output SNR for different sweeps.



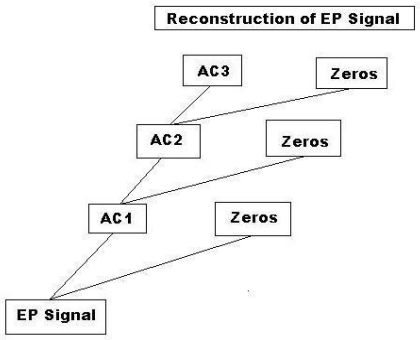
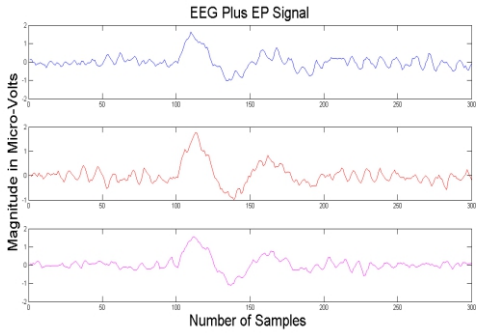


Fig. 2.2 Decomposition of EP Signal **Fig. 2.3 Reconstruction of EP Signal**

III. RESULTS



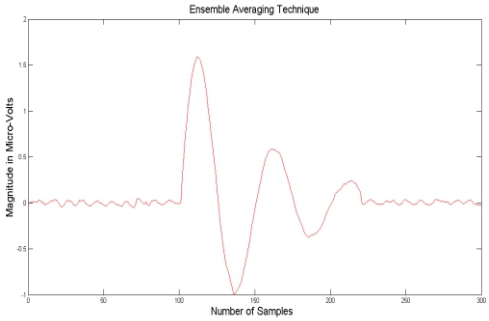
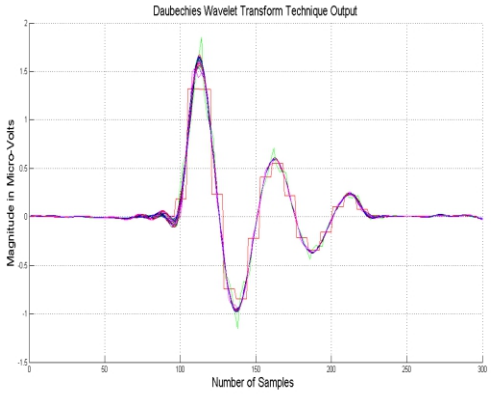


Fig. 3.1 EEG plus EP Signal **Fig 3.2 Ensemble Averaging Output**



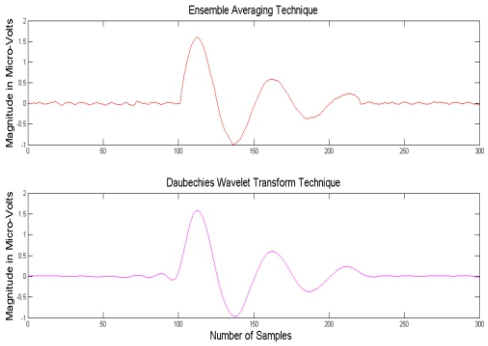
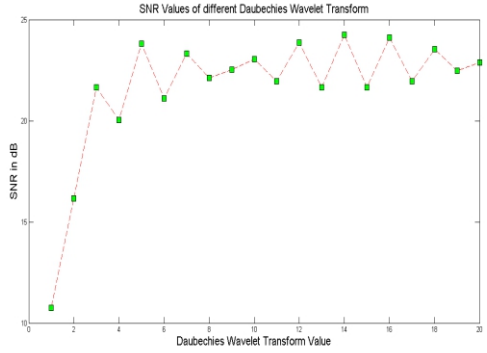


Fig. 3.3 Daubechies Wavelet Transform Output **Fig.3.4 Ensemble Averaging and Daubechies Wavelet Transform Output**

ISBN: 978-81-931039-5-1



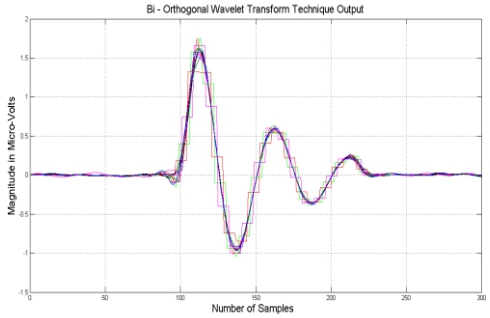
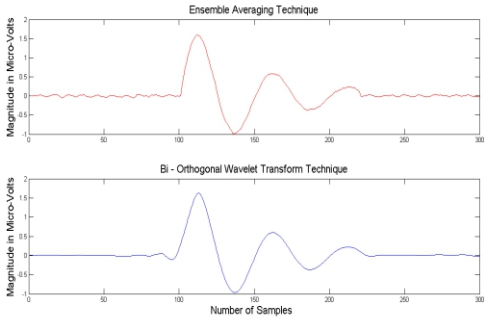


Fig. 3.5 Output SNRs vs Daubechies Values **Fig. 3.6 Bi – orthogonal Wavelet Transform Output**



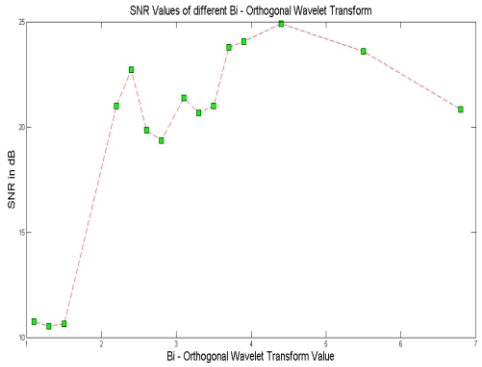


Fig.3.7 Ensemble Averaging and Bi – orthogonal Wavelet Transform Output **Fig. 3.8 Output SNRs vs Bi – orthogonal Values**

Table 3.1 Ensemble Average Technique of SNR Table for EP

Data 1	Number of Sweeps	Ensemble Averaging Technique SNR in dB
#1	10	17.1dB
#2	20	18.63 dB
#3	40	19.48 dB
#4	60	21.28 dB

ISBN: 978-81-931039-5-1

Table 3.2 Daubechies Wavelet Transform Technique of SNR Table for EP

Daubechies Value	SNR in dB
db1	10.7477
db2	16.1588
db3	21.6576
db4	20.0363
db5	23.8056
db6	21.1091
db7	23.3163
db8	22.1091
db9	22.5339
db10	23.0505
db11	21.9618
db12	23.8590
db13	21.6615
db14	24.2534
db15	21.6659
db16	24.1004
db17	21.9520
db18	23.5450
db19	22.4692
db20	22.8903

Table 3.2 Daubechies Wavelet Transform Technique of SNR Table for EP

Bi – Orthogonal Values	SNR in dB
Bior 1.1	10.7477
Bior 1.3	10.5446
Bior 1.5	10.6502
Bior 2.2	20.9838
Bior 2.4	22.7090
Bior 2.6	19.8606
Bior 2.8	19.3507
Bior 3.1	21.3890
Bior 3.3	20.6582
Bior 3.5	21.0086
Bior 3.7	23.7916
Bior 3.9	24.0510
Bior 4.4	24.9055
Bior 5.5	23.5833
Bior 6.8	20.8404

ISBN: 978-81-931039-5-1

3.1 Interpretation of Results

In this paper, simulated data's have been taken for the analysis and is shown in Fig.3.1 Fig 3.1 shows three different sweeps of data taken at sweep no.1, sweep no.16 and sweep no.60 respectively. One sweep contains 300 samples, 60 such sweeps of data have been taken for the analysis. Only three sweeps of data have shown in Fig 3.1. The simulated signal contains EP and EEG signal and both signals are added to form the contaminated signal. For the Ensemble averaging technique, 60 such sweeps of data have been considered for obtaining the output and also to calculate SNR values. Table 3.1 shows SNR values for different number of sweeps for obtaining the output.

Fig 3.2 shows the output waveform of Ensemble averaging technique. In this figure, it narrates about the repetitive signals are almost averaged to highlight EP signal. The strength of the noise signal which is an EEG (back ground signal) reduces as more number of sweeps is considered. Table 3.1 shows that as more number of sweeps is considered, SNR improves by a factor of almost square root of number of sweeps.

In wavelet transform technique of estimating EP signal, only one sweep has been considered for obtaining the output. In this paper, there are two different wavelet transforms have been used. Each wavelet transform has its own features, but most suitable for denoising or estimation of signals in noisy environment. Three levels of decomposition is processed for each wavelet transform. Hard Thresholding is used for each decomposed signal, since EP signal is low frequency signal and background signal is an EEG signal, which is an high frequency signal. In this method smooth curve is obtained since high frequency components are removed in the process. Fig.3.4 corresponding results obtained by the algorithm. The different Daubechies wavelets are used in the algorithm and the corresponding SNR values are tabulated in the table 3.2. Table 3.2 shows the different Daubechies wavelets and the corresponding SNR values are tabulated for the EP Data From the tables highest SNR values are obtained and corresponding Daubechies wavelet is highlighted. This signifies the maximum SNR is obtained if the corresponding Daubechies wavelet is used. The wavelet transform is most useful in decomposing and reconstructing the any signal, can also be data compression algorithms. Fig. 3.3 shows output waveform of different Daubechies values used in obtaining the output waveforms. Some output waveform are degraded version, but most output waveforms are approximated to the desired EP signal. A different color of output waveforms have been plotted and are shown in Fig.3.3. Fig.3.4 shows the output waveform of Ensemble averaging technique and Daubechies wavelet transform. In this figure, it is observed that a smooth curve is obtained from initial part of the signal to the end part of the signal in wavelet transform, in Ensemble averaging technique, some noise is present at initial part and end part of the signal. Fig 3.5 shows the output SNR plots of Daubechies wavelet transform technique. In this figure, different Daubechies wavelet transform values have been plotted along the X – axis and output SNR values along Y – axis. It is observed that maximum SNR is obtained for Daubechies wavelet transform value equal to db14 and is highlighted in table 3.2 as well.

In Fig 3.7 Bi – Orthogonal wavelet transform output waveform is shown and as mentioned earlier this waveform is compared with Ensemble averaging technique output waveform. Different output waveforms for different Bi – Orthogonal wavelet transform functions are shown in Fig 3.6. Here, also some of the Bi – Orthogonal wavelet transform functions will not give the desired output. It is evident that Bior 4.4 function will give the better SNR in comparison with the other functions and is shown in Fig 3.8. Table 3.3 shows that tabulated output SNR

ISBN: 978-81-931039-5-1

values with that of Bi – orthogonal function values

IV. CONCLUSION

Various estimation methods were studied for EP signals denoising. The signals were estimated using Wavelet method. It is known that signals with higher SNR and low MSE are less noisy signals. By looking at the various evaluation parameters like MSE,SNR calculated by different methods it is concluded that wavelet method gave the best denoising result with its multiresolutional capacities. Wavelet transform analyses the signals in both time and frequency domain and also signals with low noise amplitudes can be removed from the signals by selecting the best wavelet to decompose the signal and reconstruct the signal also improves the SNR.

In the ensemble signal averaging technique, it improves the SNR by a factor of $\sqrt{m}$. where m is the number of sweeps**.** The main disadvantage of this method is that more number of sweeps of data is required to improve the SNR and which is practically difficult for the subject to receive more number of stimulus and respond equally.

Wavelet-based signal processing has become common place in the signal processing community over the past few years. One of the most important applications of wavelets is removal of noise from biomedical signals and is called de-noising or estimation which is accomplished by thresholding wavelet coefficients in order to separate signal from noise. A biomedical signal is a non-stationary signal whose frequency changes overtime and for the analysis of these signals Wavelet transform is used. Wavelet transform has been a very novel method for the analysis and processing of non-stationary signals such as bio-medical signals in which both time and frequency information is required. The algorithm for estimating EP Signal based on wavelet transform shows the potential of the wavelet transform, especially for processing time-varying biomedical signals. The power of wavelet transform lies in its multi scale information analysis which can characterize a signal very well. In this paper Daubechies wavelet and Bi – orthogonal wavelet transform improves SNR in the results obtained and is more suitable for EEG and EP signal estimation.

V. ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my mentor, guide, supervisor and advisor **Dr. Anand Jatti,** Ph.D (VTU), Associate Professor, Department of Electronics and Instrumentation, R.V.College of Engineering, Bangalore for the continuous support extended during the preparation of this paper. I thank him for all his courtesies, patience, motivation, enthusiasm and immense knowledge extended to me. His guidance helped me throughout the process and producing of this paper

I express heartfelt gratitude to **Ms. Sujitha Vasu,** Sri Venkateshwara International Educational Trust, Bangalore, who enhanced my confidence level from residue level to peak level, without whom my paper would have never seen the day of Light.

REFERENCES

Journals

[1]. M P Wachowiak, G S Rash, P M Quesada, A H Desoky in IEEE Transactions on Biomedical Engineering(2000).Wavelet-based noise removal for biomechanical signals: a comparative study.

ISBN: 978-81-931039-5-1

[2]. Adinarayana Reddy, P. Chandra Sekhar Reddy,G. Hemalatha, T. Jaya Chandra Prasad, "Removal of Artifacts in Multi-channel Visual Evoked Potentials", International Journal of Modern Engineering Research (IJMER) Vol.1, Issue,2, pp-413-417 ISSN:2249-6645.

[3]. Sammaiah, A.; Narsimha, B.; Suresh, E.; Reddy, M.S.; On the performance of wavelet transform improving Eye blink detections for Brain Computer Interface(2011). Emerging Trends in Electrical and Computer Technology (ICETECT),2011..

[3]. L. Hua, Z.G. Zhang, Y.S. Hung, K.D.K. Luk, G.D. Iannetti, Y. Hua, Single-trial detection of somatosensory evoked potentials by probabilistic independent component analysis and wavelet filtering

[4]. Priyanka Khatwani, Archana Tiwari, "A survey on different noise removal techniques of EEG signals", International Journal of Advanced research in Computer and Communication Engineering Vol.2, Issue 2, February 2013.

[5]. Jeena Joy, Salice Peter, Neetha John, "Denoising Using Soft Thresholdng", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol.2, Issue 3, March 2013

Conferences:

[6]. P.O.Ranta-aho, A.Karjalainen,A.S.Koistinen, J.Kaipio, J.Partanen, "Comparison of amplitude estimates in the single trial estimation of evoked potentials", Proceedings of the 22nd Annual EMBS International Conference, July 23-28, 2000, Chicago IL.

[7]. Conor McCooey, Dinesh Kant Kumar, and Irena Cosic, "Decomposition of Evoked Potentials using Peak detection and the discrete Wavelet Transform", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual conference Shanghai, China, September 1-4, 2005.

Text Books:

[8]. Willis.J.Tomkins Bio Medical Digital Signal Processing, PHI Publications

[9]. C.Britton Rorabaugh DSP Primer, John Wiley

[10]. Stephane Mallat A wavelet tour of signal processing, PHI Publications

[11] .Time frequency and wavelets in Biomedical signal processing by Metin Akay (IEEE express)

[12] .Sophacles J. Orfanidis Optimum Signal Processing, John Wiley.

International Conference on Recent Advances in Engineering Science and Management
PHD Chamber of Commerce and Industry, New Delhi
30 August 2015, www.conferenceworld.in, www.icraesm.org
ISBN: 978-81-931039-5-1

A FAST CLUSTERING-BASED FEATURE SUBSET SELECTION ALGORITHM

Akshay S. Agrawal[1], Prof. Sachin Bojewar[2]

[1]*P.G. Scholar, Department of Computer Engg., ARMIET, Sapgaon, (India)*

[2]*Associate Professor, VIT, Wadala.*

ABSTRACT

The paper aims at proposing the fast clustering algorithm for eliminating irrelevant and redundant data. Feature selection is applied to reduce the number of features in many applications where data has hundreds or thousands of features. Existing feature selection methods mainly focus on finding relevant features. In this paper, we show that feature relevance alone is insufficient for efficient feature selection of high-dimensional data. We define feature redundancy and propose to perform explicit redundancy analysis in feature selection. A new hypothesis is introduced that dissociate relevance analysis and redundancy analysis. A clustering based method for relevance and redundancy analysis for feature selection is developed and searching based on the selected features will be performed. While the efficiency concerns the time required to find a subset of features, the effectiveness determines the quality of the subset of features. A fast clustering-based feature selection algorithm, FAST, has been selected to be used in the proposed paper. The clustering-based strategy has a higher probability of producing a subset of useful as well as independent features. To ensure the efficiency of FAST, efficient minimum-spanning tree clustering method has been adopted. When compared with FCBF, ReliefF, with respect to the classifier, namely, the tree-based C4.5, FAST not only produces smaller subsets of features but also improves the performances by reducing the time complexity.

Keyterms: Clustering, Feature subset selection, Minimum Spanning Tree, T-Relevance, F-Correlation.

I. INTRODUCTION

Data mining uses a variety of techniques to identify lump of information or decision-making knowledge in bodies of data, and extracting them in such a manner that they can be directly use in the areas such as decision support, estimation prediction and forecasting. The data is often huge, but as it is important to have large amount of data because low value data cannot be of direct use; it is the hidden information in the data that is useful. Data mine tools have to infer a model from the database, and in the case of supervised learning this requires the user to define one or more classes. The database contains various attributes that denote a class of tuple and these are known as predicted attributes. Whereas the remaining attributes present in the data sets are called as predicting attributes. A combination of values of these predicted attributes and predicting attributes defines a class. While learning classification rules the system has to find the rules that predict the class from the predicting attributes so initially the user has to define conditions for each class, the data mine system then constructs descriptions for the classes. Basically the system should given a case or tuple with certain known attribute values so that it is able to predict what class this case belongs to, once classes are defined the system should infer rules that govern

ISBN: 978-81-931039-5-1

the classification therefore the system should be able to find the description of each class [2]. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm is basically evaluated from the efficiency and effectiveness points of view. The time required to find a subset of features is concerned with the efficiency while the effectiveness is related to the quality of the subset of features. Some feature subset selection algorithms can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can remove the irrelevant while taking care of the redundant features. A Fast clustering-based feature selection algorithm (FAST) is proposed which is based on above criterion handling redundancy and irrelevancy. [1] The Minimum Spanning tree (Kruskal's algorithm) is constructed from the F-Correlation value which is used to find correlation between any pair of features. Kruskal's algorithm is a greedy algorithm in graph theory that finds a minimum spanning tree for a connected weighted graph. It finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized.

II. EXISTING SYSTEM

Feature subset selection generally focused on searching relevant features while neglecting the redundant features. A good example of such feature selection is Relief, which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.[9] But, Relief is ineffective in removing redundant features as the two predictive but highly correlated features are likely to be highly weighted. Relief-F [6] is an extension of the traditional Relief. This method enables working with noisy and incomplete data sets and to deal with multi-class problems, but is still ineffective in identifying redundant features. However, along with irrelevant features, redundant features also do affect the speed and accuracy of all the probable learning algorithms, and thus should are also important to be eliminated. FCBF is a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis. Different from these algorithms, our proposed FAST algorithm employs clustering based method to choose features.

There are different approaches available to perform learning. The wrapper methods make use of predictive accuracy of a predetermined learning algorithm to determine the effectiveness of the selected subsets.[7] The accuracy of the learning algorithms [1] is usually high. The however the generality of the selected features is limited and the computational complexity is very large. Thus the wrapper methods are computationally expensive and tend to over fit on small feature training sets. Wrapper uses a search algorithm for searching through the space of possible features and evaluates individual subset by running a model on the subset. The filter methods [3] are independent of the learning algorithms, and also have good generality. Computational complexity is low, but the accuracy of such learning algorithms is not guaranteed. The hybrid method used in our approach is a combination of filter and wrapper methods, filter method reduces search space of computation that will be considered by the subsequent wrapper.

III. PROPOSED SYSTEM

ISBN: 978-81-931039-5-1

The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes Therefore; symmetric uncertainty is chosen as the measure of correlation between either two features or a feature and the target concept.[8]

The **symmetric uncertainty (SU)** is defined as follows,

$$SU(X,Y) = \frac{2 \times Gain\left(\frac{X}{Y}\right)}{H(X) + H(Y)}$$

Where, $H(X)$ is the entropy of a discrete random variable X. Let (x) be the prior probabilities for all values of X, then (X) is defined by

$$H(X) = -\sum_{x \in X} p(x) \log 2 p(x)$$

Gain (X | Y) determines the amount by which the entropy of Y decreases. It is given by,

$$Gain\ (X|Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

Where H(X | Y) is the conditional entropy and is calculated as,

$$H\left(\frac{X}{Y}\right) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x) \log_2 p(x)$$

Where, X is a Feature and Y is a Class.

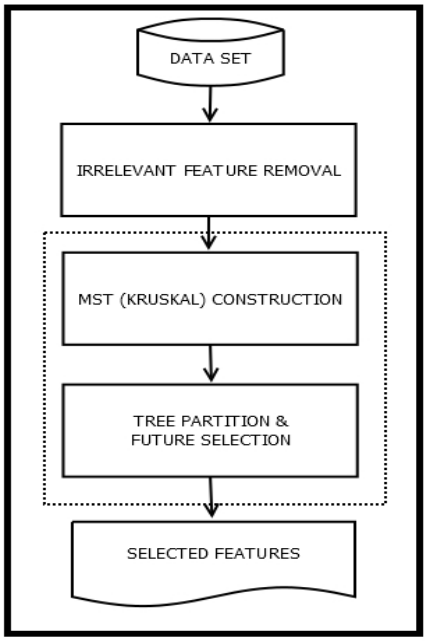


Fig. 3.1: Feature Subset Selection Process.

Given that (X, Y)be the symmetric uncertainty of variables X and Y, the relevance T-Relevance between a feature and the target concept C, the correlation F- Correlation between a pair of features, the feature redundancy F-Redundancy and the representative feature R- feature of a feature cluster can be defined as follows.

T-Relevance - The relevance between the feature $F_i \in F$ and the target concept is referred to as the T-Relevance of F_i and C, and denoted by SU (F_i,C). If SU (F_i,C) is greater than a predetermined threshold θ,

Symmetric Uncertainty of each Feature is greater than the T-Relevance threshold(θ) is checked.

ISBN: 978-81-931039-5-1

$SU(X, Y) > \theta$ then X is submittedin Feature set S

Where, 'S' is a set of Relevant Features

we say that F_i is a strong T-Relevance feature.

F-Correlation - The correlation between any pair of features and F_j (F_i, $Fj \in F \wedge i \neq j$) is called the F-Correlation of F_i and, and denoted by SU (F_i, F_j).

F-Redundancy - Let S = {F1, F2, F_i, F_k<|F|} be a cluster of features.

If $\exists$ $F_j \in S$, $(F_j) \geq$ SU(F_i, C) $\wedge$ SU (F_i, F_j) >SU (F_i, C) is always corrected for each $Fi \in S (i \neq j)$, then F_i are redundant features with respect to the given F_j (i.e. each F_i is a F-Redundancy).

R-Feature - A feature $F_i \in S$ ={F1, F2, ..., F_k } (k < |F |) is a representative feature of the cluster S (i.e. F_i is a R-Feature) if and only if, F_i = argmax$F_j \in S$ SU (F_j, C).

This means the feature, which has the strongest T Relevance, can act as an R-Feature (Most relevant Feature) for all the features in the cluster.

1) Irrelevant features have no/weak correlation with target concept;

2) Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster. [4]

IV. MST CONSTRUCTION

With the F-Correlation value computed, the Minimum Spanning tree is constructed. Kruskal"s algorithm is used which forms MST effectively. Kruskal's algorithm is a greedy algorithm in graph theory that finds a minimum spanning tree for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

Minimum spanning tree using Kruskal's algorithm is constructed and then a threshold value and step size is set. Those edges from the MST, whose lengths are greater than the threshold value are removed. The ratio between the intra-cluster distance and inter-cluster distance is calculated and the ratio as well as the threshold is recorded. The threshold value is updated by incrementing the step size. Every time the new (updated) threshold value is obtained, the above procedure is repeated. When the threshold value is maximum and as such no MST edges can be removed the above procedure is stopped. In such situation, all the data points belong to a single cluster. Finally the minimum value of the recorded ratio is obtained and the clusters are formed corresponding to the stored threshold value.

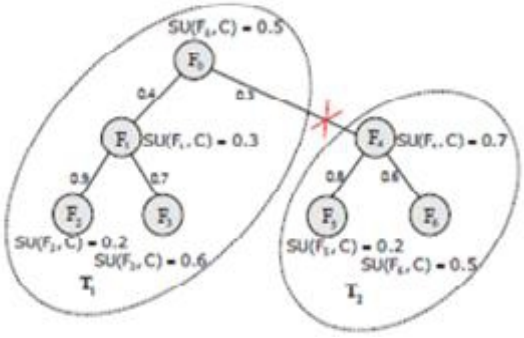


Fig. 3.2: Clustering with MST Construction.

ISBN: 978-81-931039-5-1

1. Create a forest F (a set of trees), where each vertex in the graph is a separate tree.

2. Create a set S containing all the edges in the graph.

3. While S is nonempty and F is not yet spanning.

Remove an edge with minimum weight from S. If that edge connects two different trees, then add it to the forest, combining two trees into a single tree, otherwise discard that edge. At the termination of the algorithm, the forest forms a minimum spanning forest of the graph. If the graph is connected, the forest has a single component and forms a minimum spanning tree.[1]

V. PROPOSED ALGORITHM

Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method.

Algorithm:

Inputs: D (F1, F2 … Fm, C) (High Dimensional Dataset).

Output: S-Selected feature subset for searching. [1]

Part 1: Removing irrelevant features:

The features whose SU (Fi,C) values are greater than a predefined threshold(Ø) comprise the target relevant feature subset. Consider feature input dataset (F).

F´= { F_1´,F_2´,…F_k´} (k<=M)

1. for i = 1 to m do

2. T-Relevance = SU (Fi, C)

3. if T-Relevance > θ then

4. S = S ∪ { };

Part 2: Removing redundant features:

The F-correlation SU (Fi´,Fj´) value for each pair of features.

5. G = NULL; //G is a complete graph

6. for each pair of features {F'_i, F'} ⊂ S do

7. F-Correlation = SU (F'_i , F'_j)

8. $F'i$ and/or $F'j$ to with F-Correlation as the weight of the corresponding edge;

9. MinSpanTree = Kruskal's (G); //Using Kruskal's algorithm to generate minimum spanning tree.

Part 3 : Feature selection.

10. Forest = minSpanTree

11. for each edge $E_{i,j}$ ∈ Forest do

12. if SU (F'_i , F'_j) < SU(F'_i ,C) ∧ SU(F'_i , F'_j)< SU(F'_j , C) then

13. Forest = Forest − E_{ij}

14. S = ϕ

15. for each tree T_i ∈ Forest do

16. F_R^j = argmax F_k∈ SU(F'_k , C)

17. S = S ∪ { F_R^j};

ISBN: 978-81-931039-5-1

18. Return S.

The algorithm can be expected to be divided into 3 major parts:

The first part is concerned with removal of irrelevant features;

The second part is used for removing the redundant features and The final part of the algorithm is concerned with feature selection based on the value of the Forest. [1]

5.1 Working

A. First Step:

The data set 'D' with 'm' features F= (F1,F2,…Fm) and class 'C', 'I' compute the T-Relevance '**SU**' (Fi, C) value for every feature ($1 \leq i \leq m$).

B. Second step:

Here the first step is to calculate the F-Correlation '**SU**' (F'_i , F'_j) value for each pair of features F'_i and F'_j Then, seeing features F'_i and F'_j as vertices and '**SU**' (F'_i , F'_j) the edge between vertices F'_i and F'_j a weighted complete graph G= (V,E) is constructed which is an undirected graph. The complete graph reflects the correlations among the target-relevant features. [3]

C. Third step:

Here, unnecessary edges can be removed. Each tree $T_j \in$ Forest shows a cluster that is denoted as V (T_j), which is the vertex set of T_j . For each cluster V(T_j), select a representative feature whose T-Relevance SU(F_jR, C) is the highest. All F_jR (j = 1...|*Forest*|) consist of the final feature subset $\cup F_jR$.

A clustering tree depending on the domain that the admin selects while uploading the file is created. Proposed system then stores the file in the cluster by using the minimum spanning tree method (MST). While in the searching domain; user passes the query and the results are generated in the required format. i.e. either image result, text result or a file result along with the time complexity. FAST algorithm reduces the run time complexity as compared to the other available Algorithms. It removes the redundant features by calculating the Correlations among the various features. F-correlation is calculated as SU (Fi, Fj).

A threshold value ( ) is defined to calculate the relevance among the selected features. If any feature exceeds a particular threshold value then that feature is treated as irrelevant.

F´= { F1´,F2´,…Fk´} (k<=M) [1]

VI. ADVANTAGES

Table 5.1.: Advantages and Disadvantages [5]

SR. NO.	Techniques (or) Algorithms	Advantages	Disadvantages
1.	FAST Algorithm	Improve the performance of the classifiers. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.	--
2.	Consistency Measure	Fast, Remove noisy and irrelevant data.	Unable to handle large volumes of data.

ISBN: 978-81-931039-5-1

3.	Wrapper Approach	Accuracy is high.	Computational complexity is large.
4.	Filter Approach	Suitable for very large features.	Accuracy is not guaranteed.
5.	Agglomerative linkage algorithm	Reduce Complexity.	Decrease the Quality when dimensionality becomes high.
6.	INTERACT Algorithm	Improve Accuracy.	Only deal with irrelevant data.
7.	Distributional clustering	Higher classification accuracy.	Difficult to evaluation.
8.	Relief Algorithm	Improve efficiency and Reduce Cost.	Powerless to detect redundancy.
9.	Grid based method	Jobs can automatically restart if a failure occurs.	You may need to have a fast interconnect between compute resources.
10.	Model based method	Clusters can be characterized by a small number of parameters.	Need large data sets. Hard to estimate the number of clusters.

VII. RESULT

In the proposed system data set of heart diseases [10] possessing high dimensional features containing 75 categorical, integer and real attributes have been used to eliminate the irrelevant and redundant features by selecting any one feature 'num' from the data set and to form a cluster.

Before proceeding with the actual implementation the files having dataset and features are being uploaded and the specific feature on which the clustering is to be done is inserted ('num' is the feature on which the clustering is done is selected in the proposed algorithm).

Step 1: Removal of irrelevant features:

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. T-Relevance is calculated using Symmetric Uncertainty (***SU)*** where each attribute/feature (***Fi)*** is checked with the class ***(C).***

$$T\text{-}Relevance = SU(Fi, C)$$

$$if\ T\text{-}Relevance > \theta\ then\ S = S \cup \{\};$$

$$SU(X,Y) = \frac{2 \times Gain\left(\frac{X}{Y}\right)}{H(X) + H(Y)}$$

The relevance between the feature $F_i \in$ □ and the target concept C is referred to as the T-Relevance of F_i and C, and denoted by (F_i, C). If (F_i, C) is greater than a predetermined threshold θ, we say that F_i is a strong T-Relevance feature.

ISBN: 978-81-931039-5-1

Feature Classification After T-Relevance Calculation

Selected Features (Relevant Features)

Feature Name	T-Relevance
lvx4	0.18065372510763003
rcadist	0.14693577298945426
trestbps	0.11474615071159641
ekgday	0.11317483303799542
exang	0.21921885052628132
lvf	0.09504551808814005
id	0.3429926963481016
painexer	0.18410458787218822
cathef	0.14678905923184288
rldv5	0.11906272127411502
rldv5e	0.11021037890983608
cday	0.09768578647365952
slope	0.2283922385278686

If $S(F_i, C)$ is lesser than a predetermined threshold θ, we say that F_i is a not an T-Relevance feature.

Unselected Features (Irrelevant Features)

Feature Name	T-Relevance
cigs	0.012643257627983873
restckm	0.01
pncaden	0.01
sex	0.06404418345447156
lvx2	0.022223746796010673
ca	0.018494752794235625
lvx3	0.07013514394150012
lvx1	0.01
restef	0.01
cmo	0.06126498338275048
thal	0.04687983246503802
fbs	0.04137806322972103
cyr	0.030307656998330282
ekgo	0.06365890362169223
junk	0.02764689227286641
met	0.08206595203372598

Step 2: Removal of redundant data:

Redundant features are assembled in a cluster and a representative feature can be taken out of the cluster.

Let S = {F1, F2, F_i, F_k<|F|} be a cluster of features.

If $\exists F_j \in S, (F_j) \geq SU(F_i, C) \wedge SU(F_i, F_j) > SU(F_i, C)$ is always corrected for each $Fi \in S (i \neq j)$, then F_i are redundant features with respect to the given F_j (i.e. each F_i is a F-Redundancy).

Redundant data are removed using MST.

ISBN: 978-81-931039-5-1

SR.NO	To	From	Value
1	0	20	0.10856213286124775
2	0	24	0.18291607205933239
3	0	23	0.11621050434818578
4	0	22	0.12023262230267498
5	0	21	0.11795440407281713
6	0	16	0.13098452774297195
7	0	17	0.12295684002842386
8	0	14	0.11803376127593215
9	0	15	0.12748708469310402
10	0	18	0.11981814178224007
11	0	19	0.1444157417217047
12	0	29	0.10295094687709841
13	0	28	0.1200011826719238
14	0	11	0.1357659142612397
15	0	27	0.11662254638836325

MST BY PRIMS ALGORITHM

In the proposed system the redundant values are removed using both Kruskal's and Prim's. The complete graph reflects the correlations among the target-relevant features.

SR.NO	To	From	Value
1	0	20	0.10856213286124775
2	0	24	0.18291607205933239
3	0	23	0.11621050434818578
4	0	22	0.12023262230267498
5	0	21	0.11795440407281713
6	0	16	0.13098452774297195
7	0	17	0.12295684002842386
8	0	14	0.11803376127593215
9	0	15	0.12748708469310402
10	0	18	0.11981814178224007
11	0	19	0.1444157417217047
12	0	29	0.10295094687709841
13	0	28	0.1200011826719238
14	0	11	0.1357659142612397
15	0	27	0.11662254638836325

MST BY KRUSKAL'S ALGORITHM

Step 3: Feature Selection

Here, unnecessary edges can be removed. Each tree $T_j \in Forest$ shows a cluster that is denoted as $V(T_j)$, which is the vertex set of T_j . For each cluster $V(T_j)$, select a representative feature whose *T-Relevance* $SU(F_jR, C)$ is the highest. All F_jR (j = 1...|*Forest*|) consist of the final feature subset ∪ F_jR .

ISBN: 978-81-931039-5-1

Clusters Created

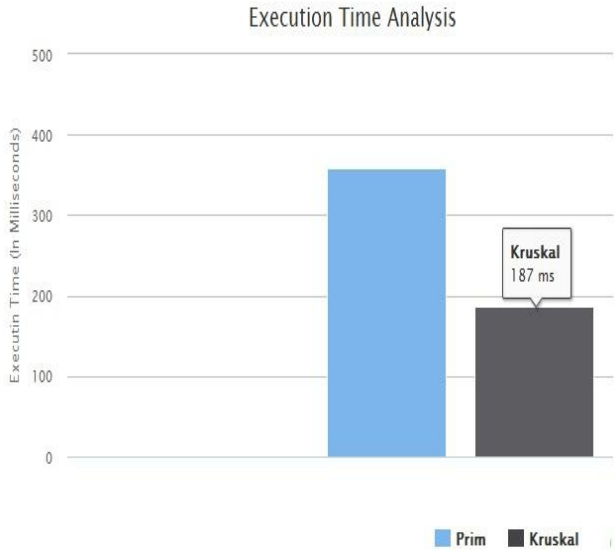
Cluster No.	Features	T-Relevance
1	[lvx4, orn1, lvf, dummy, thalrest, rdv5, cday, rdv5e, age]	[0.18865372510763003, 0.14273315749592783, 0.09504551808814005, 0.11474615071159641, 0.14226010269907755, 0.11906272127411502, 0.09768578647365952, 0.11021037890983608, 0.11467123996208545]
2	[cxmain]	[0.27570057346005 5]
3	[relrest]	[0.13248729522896688]
4	[thaldur]	[0.11821416110177616]
5	[id]	[0.34299269634810 16]
6	[exang]	[0.21921885052628132]
7	[lmt]	[0.30859851315704645]
8	[painexer]	[0.18410458787218822]
9	[tpeakbps]	[0.11643551580215977]
10	[trestbps]	[0.11474615071159641]
11	[rcadist]	[0.14883577298945426]
12	[rcaprox]	[0.29655370060133 25]
13	[cp]	[0.17516707858670133]
14	[slope]	[0.22839223852786 86]
15	[laddist]	[0.20697355597443 47]

VIII. ANALYSIS

The major amount of work for Algorithm involves the computation of SU values for T-Relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity (m) in terms of the number of features m. Assuming $(1 \leq k \leq m)$ features are selected as relevant ones in the first part, when $k = 1$, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity. In proposed system the analysis of Prim's and Kruskal's algorithm is done and the best method is being selected based on the time complexity to generate and select the efficient features by eliminating redundant and irrelevant data.

Based on the analysis, Kruskal's algorithm is consider to be the efficient algorithm as compared to Prim's as the time complexity in Kruskal's is less than of Prim's.

ISBN: 978-81-931039-5-1



IX. CONCLUSION

In this paper, we have proposed a clustering algorithm, FAST for high dimensional data. The algorithm includes (i) irrelevant features removal (ii) construction of a minimum spanning tree (MST) from, and (iii) partitioning the MST and selecting the representative features. Feature subset selection should be able to recognize and remove as much of the unrelated and redundant information. In the proposed algorithm, a cluster will be used to develop a MST for faster searching of relevant data from high dimensional data. Each cluster will be treated as a single feature and thus volume of data to be processed is drastically reduced. FAST algorithm will obtain the best proportion of selected features, the best runtime, and the best classification accuracy after eliminating redundant and irrelevant data.

Overall the system will be effective in generating more relevant and accurate features which can provide faster results.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:1 YEAR 2013.

[2] Karthikeyan.P, High Dimensional Data Clustering Using Fast Cluster Based Feature Selection , Int. Journal of Engineering Research and Applications, March 2014, pp.65-71.

[3] B.Swarna Kumari, M.Doorvasulu Naidu, Feature Subset Selection Algorithm for Elevated Dimensional Data By using Fast Cluster, In International Journal Of Engineering And Computer Science Volume 3 Issue Page No. 7102-7105, 7 July, 2014.

[4] Sumeet Pate, E.V. Ramana, A Search Engine Based On Fast Clustering Algorithm for High Dimensional Data, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 3, Issue 10, October 2014, ISSN: 2278 – 909X.

[5] Comparative study of various clustering techniques with FAST, International Journal of Computer Science and Mobile Computing, Volume 3, Issue 10, October 2014, ISSN: 2320-088X.

ISBN: 978-81-931039-5-1

[6] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge, 1988.

[7] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[8] Fayyad U. and Irani K., Multi-interval discretization of continuous-valued attributes for classification learning, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp 1022-1027,

[9] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.

[10] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

International Conference on Recent Advances in Engineering Science and Management
PHD Chamber of Commerce and Industry, New Delhi
30 August 2015, www.conferenceworld.in, www.icraesm.org
ISBN: 978-81-931039-5-1

REMOTE MONITORING OF NUCLEAR POWER PLANTS WITH THE INTEGRATION OF IOT AND CLOUD COMPUTING

S.Lavanya[1], Dr. S. Prakasam[2]

[1]Research Scholar, [2]Asst.Professor and Head, Dept of CSA, SCSVMV University, Kancheepuram, (India)

ABSTRACT

Sensors play a vital role in monitoring Nuclear power plants. As the technology advances, lots of critical unstructured data are collected by sensors everyday. Sensors are used to record and monitor more than 800,000 daily measurements. Day by day large amounts of data are generated by these sensors. They require complex processing at every level. The risk factor involved becomes clearly evident in situations of disaster like tsunami or earthquakes. The most important data should be checked regularly and any irregularities beyond a threshold should be regulated. But when utilized properly these data could be useful and could save us from a huge loss in terms of resource or human lives. Integration of cloud architecture helps to improve the performance of Wireless Sensor Networks.

Keywords: Cloud,Internet-of-Things, Nuclear Power Plants, Reliability, Sensor Networks

I. INTRODUCTION

Wireless Sensor Networks (WSN) is now-days widely used in Nuclear Power Plants (NPP). Around the clock, the current states of the plants are supervised for emission of radioactive particles to air and water. The main operating parameters in a Nuclear power plants are pressure, temperature, water levels in circuit and so on. The data produced by Industrial Internet is quite mission- critical in nature as opposed to the data collected from social media. Therefore the techniques that are used to collect, store, analyze and optimize such mission-critical data requires higher levels of reliability, performance, scalability and adaptability. While the data collected is used for a single application, current industry demands the need for a system framework that can work on a multiple application intelligently rather than a single application. This requires specialized platforms, more analytical tools and data models to work with Industrial data. Nuclear power plants are reliable energy generation sources and continue to work for long terms without shutdown. Therefore management of radioactive waste is a challenging task which could be addressed by technology.

ISBN: 978-81-931039-5-1

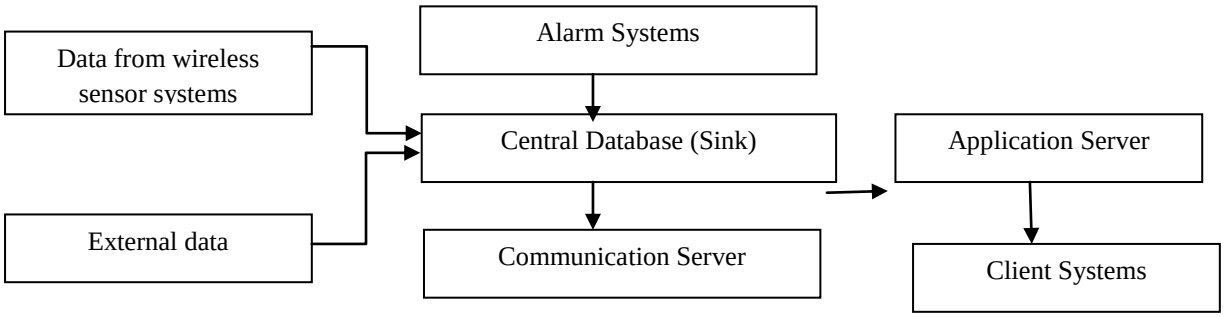


Fig. 1 Technical Components Used in Nuclear Power Stations

II. LITERATURE SURVEY

From authors [1-3], it is evident that Wireless Sensor Networks (WSN) are best suited for monitoring various parameters in Nuclear Power Plants (NPP). A number of techniques have been proposed to reduce energy and improve the reliability of the system. A good number of routing protocols are suggested that improves reliability [4].Three hop reliability models though cost effective offers a good solution for reliability [5]. But Integration of WSN with cloud helps in easy management of remotely connected sensors and helps in increasing the reliability, scalability and performance of the given network systems [6, 7]

III. EXISTING SYSTEM

Usually a network of sensors is created to collect the data sensed by the sensors from where they are sent to the base stations. Wireless Sensor Networks combined with suitable Energy harvesting techniques are used nowadays in sensor networks. Various reliability models can be applied thereby reliable data collected by these sensors are stored in the Power Plant's application server and are monitored for any irregularities. Sometimes these data may be required by other agencies for some monitoring purpose. [6]

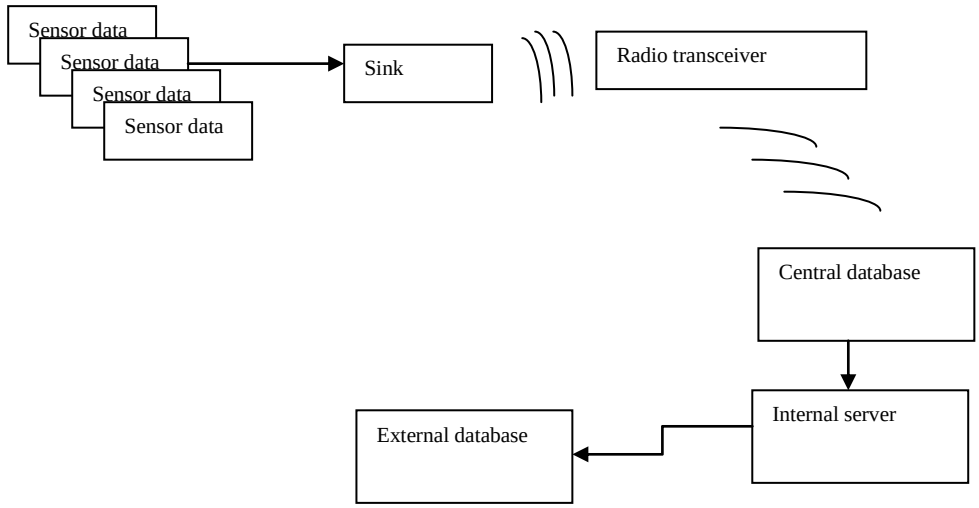


Fig2. Flow of Data in a Tradition Wireless Sensor Networks

Now-days the need for localized storage and centralized database management system is quickly changing. With the advancement of the new technology, there is a reduction in the maintenance of hosting individual networks.

ISBN: 978-81-931039-5-1

People now-days prefer easy to use web services namely cloud computing. The benefit of integration the results of WSN to cloud would be real time access of information, scalability and reduction in the risk factor [7]

IV. PROPOSED SOLUTION USING CLOUD ARCHITECTURE

Cloud computing is a term that delivers service through Internet. Basically it a front-end to access the Internet-of-Things. The combination of IOT, cloud computing and big data can enable sensing and processing of sensed data more powerful. The sensed data can be more intelligently used by much application. The benefits of using cloud and IOT are On-demand self service, Network access by many smart devices, resource pooling, elasticity and measured service. IBM Internet of Things Foundation, are providing developers with the ability to quickly and easily extend an Internet-connected device such as a sensor or controller into the cloud, build an application alongside the device to collect the data and send real-time insights back to the developer's business. At the same time, developers can quickly build mobile apps that act as remote controls to connected devices [9]

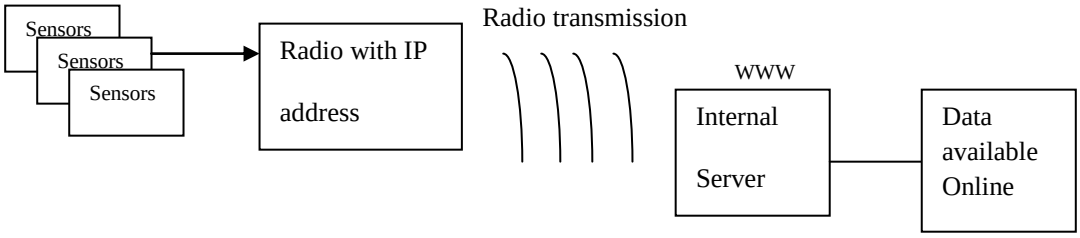


Fig 3. Flow of Data When Integration of WSN with the Cloud

Integration of Big data, IoT and Cloud seems to offer a wonderful solution. Huge volume of unstructured data is produced by sensors everyday in nuclear power plants. When proper analytics tool is used then those data can be structured which could be useful in several ways. Nuclear Regulatory commission normally defines emergency planning zones to avoid radioactive exposures. Data generated by the sensors could be stored on a cloud. Not all data needs to be stored. Few critical data can be stored in the data centre of the power plants. IPv6's huge address space has led to the development of IOT technology. The idea behind IOT is that sensors can be placed anywhere to create networks that connects devices that gather data. But every device in the IOT needs an unique address. The more important fact is that what kind of data is needed how to store and analyze them and for how long they are needed is a key requirement. To simplify, the data must be structured to satisfy the human needs. The main benefits are efficient data analysis, enhanced reporting and notifications in emergency situations that would not be possible otherwise.[10,11]

Sensors in remote locations serves as a collection points

Data is transmitted to the cloud via cellular data

ISBN: 978-81-931039-5-1

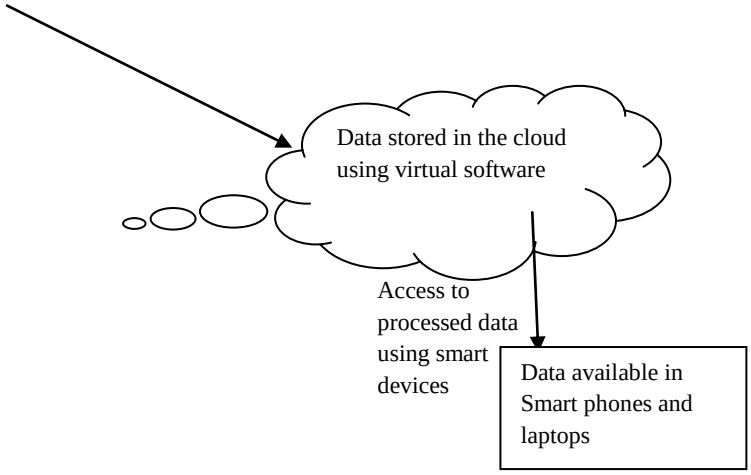


Fig. 4 Remote Access of Data Using Cloud Architecture

Pros	Cons
Data backup in regular scheduled intervals.	Security challenges may be handled properly as we have to rely on third party service providers.
No separate data management process	High bandwidth is necessary to avoid latency.
Real time data analysis and integration of data with other web services.	Trained experts are needed for integration with the cloud.
Latest updates are already available in the cloud.	It may be costly for some applications.

Table1. Merits and Demerits of Integrating Cloud with WSN

V. IMPLEMENTATION

Sensors are distributed in the sensing zone and clouds can act as a virtual sink that collects the sensed data. Open.Sen.se is an open source IOT API application that helps to transmit the sensed data through HPPT protocol to the Internet. Integration of Sensors in a heterogeneous network is quite complex and with the availability of XML templates the readings of sensors are easily convertible and stored in cloud. But this requires the use of web services and standardized language namely WSDL (Web Service Description Language).

VI. CONCLUSION

As the Internet continues to expand and become available wirelessly to more locations, users can expect to see additional web-based services offered that make the acquisition of remote data online a simple solution, providing real-time access to critical data for businesses and organizations in many industries. Moreover the integration of cloud with sensors enables the sensor data to be stored and accessed in a cost effective manner so that the needed data can be accessed anytime, anywhere in a timely manner.

ISBN: 978-81-931039-5-1

REFERENCES

[1] L. Li*a, Q. Wangb, A. Barib, C. Dengc, D. Chenc, J. Jiangb, Q. Alexandera, and B. Sura , Field Test of Wireless Sensor Network in the Nuclear Environment, Atomic Energy of Canada Limited, Chalk River, ON K0J 1J0, Canada.

[2] H.M. Hashemian, 2011, Nuclear Power Plant Instrumentation and Control -In Nuclear Power – Control, Reliability and Human Factors, InTech, Chapter 3, pp. 49–66, Available from URL: 1051/InTech uclear_power_plant_instrumentation_and_control.pdf

[3] A. Kadri, R.K. Rao and J. Jiang, 2009, "Low-Power Chirp Spread SpectrumSignals for Wireless Communication within Nuclear Power Plants", Nuclear Technology, 166(2), pp. 156–169.

[4] Muhammad Adeel Mahmood and Winston Seah, "Reliability in Wireless Sensor Networks: Survey and Challenges Ahead" ,Published in Elsevier February 8, 2012.

[5] Praful P. Maktedar , Vivek S. Deshpande , J. B. Helonde, V.M. Wadha, Performance Analysis of Reliability in Wireless Sensor Network, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278 - 3075, Volume 2 , Issue 4 March 2013.

[6] Joseph V. Cordaro, Davis Shull, Mark Farrar, and George Reeve, Ultra Secure High Reliability Wireless Radiation Monitoring System, IEEE 2012.

[7] Peng Zhang, Zheng Yan and Hanlin Sun, A Novel Architecture Based on Cloud Computing for Wireless Sensor Network**,** Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013).

[8] Dipankar Mishra and Ratnesh Kumar Gupta, Application of Cloud Computing in Hazardous Mechanical Industries, International Journal of Advancements in Research & Technology, Volume 2, Issue4, April - 2013.

[9] Rajeev Piyare and Seong Ro Lee, Towards Internet Of Things (Iots):Integration Of Wireless Sensor Network To Cloud Services for Data Collection and Sharing, International Journal of Computer Networks & Communications (IJCNC) Vol.5, No.5, September 2013.

[10] Bill Chamberlin, Iot (Internet Of Things) Will Go Nowhere Without Cloud Computing and Big Data Analytics, IBM Center for Applied Insight, Nov 2014.

[11] Atif Alamri, Wasai Shadab Ansari, Mohammad Mehedi Hassan, M. Shamim Hossain, Abdulhameed Alelaiwi, and M. Anwar Hossain, A Survey on Sensor-Cloud: Architecture, Applications, and Approaches, International Journal of Distributed Sensor Networks , Volume 2013, Article ID 917923, 18 pages, http://dx.doi.org/10.1155/2013/917923.

International Conference on Recent Advances in Engineering Science and Management
PHD Chamber of Commerce and Industry, New Delhi
30 August 2015, www.conferenceworld.in, www.icraesm.org
ISBN: 978-81-931039-5-1

SURVEY ON CLASSIFICATION OF INCOMPLETE DATA HANDLING TECHNIQUES

M.Kowsalya[1], Dr.C.Yamini[2]

[1]Research Scholar, [2]Associate Professor, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore, (India)

ABSTRACT

Data is often incomplete. Classification with incomplete data is a new subject. This study proposes a classification for incomplete survey data. The task of classification with incomplete data is a complex phenomena and its performance depends upon the method selected for handling the missing data. Missing data occur in datasets when no data value is stored for an attribute / feature in the dataset. This paper provides a brief overview to the problem of incomplete data handling techniques and discusses the various methods used with classification and missing data. It proposes a various techniques of classification use of incomplete data.

Keywords: Classification, Incomplete Data, Missing Values.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a technology with great potential that predict future trends and behaviors, and it can generate results which come out to be significant and which cannot actually predict future behavior and cannot be reproduced on a new sample of data and allow small use. It is allowing businesses to make proactive, knowledge-driven decisions. It is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing and data visualization. An incomplete data may severely affect the quality of learned patterns and the performance of algorithms. As a result, how to properly handle incomplete data is an important and challenging problem in the practice of machine learning and data mining.

There are many approaches to handling incomplete data as far as classification is concerned, from simply removing samples or features with missing values to completing the original data set by filling in specific values. In this paper, focused on the classification of incomplete data. However, the deletion of samples or features may result in the loss of useful information especially when a large portion of samples or features have missing values. In this paper, a novel scheme is developed for conducting classification on incomplete data with applying various techniques.

Pattern classification was developed starting from the 1960s. It progressed to a great extent in parallel with the growth of research on knowledge-based systems and artificial neural networks. Increasing computational resources, while enabling faster processing of huge data sets, have also facilitated the research on pattern classification, providing new developments of methodology and applications. This interdisciplinary field has been successfully applied in several scientific areas such as computer science, engineering, statistics, biology,

ISBN: 978-81-931039-5-1

and medicine, among others. These applications include biometrics (personal identification based on several physical attributes such as fingerprints and iris), medical diagnosis (CAD, computer aided diagnosis), financial index prediction, and industrial automation (fault detection in industrial process).

The complete goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

The final part of this paper is organized as follows. An overview of the previous research on incomplete data is given in Section II. Section III introduces the details of the proposed method for handling incomplete data. Section IV provides a conclusion with future research directions.

II. AN OVERVIEW OF INCOMPLETE DATA

Missing values are a common occurrence, and need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in.

Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. In incomplete datasets, some values are missing for one or more features. When choosing the right techniques for dealing with this issue, it is necessary to have a good understanding of different reasons that lead to incomplete data. Efficient treatment of missing values requires a complete understanding behind it.

2.1 Types of Incomplete Data

Little and Rubin [10] define a list of missing mechanisms, which are widely accepted by the community. There are three mechanisms under which missing data can occur:

1) Missing completely at random (MCAR): MCAR is the probability that an observation (Xi) is missing, is unrelated to the value of Xi or to the value of any other variables and the reason for missing is completely random. Typical examples of MCAR are when a tube containing a blood sample of a study subject is broken by accident (such that the blood parameters cannot be measured) or when a questionnaire of a study subject is accidentally lost [5]. This situation is rare in real world and is usually discussed in statistical theory.
2) Missing at random (MAR): MAR is the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. An example of this is accidentally or deliberately skipping an answer on a questionnaire by the participant. This mechanism is common in practice and is generally considered as the default type of missing data.
3) Not missing at random (NMAR). If the probability that an observation is missing depends on information that is not observed, this type of missing data is called NMAR. For example, high incomers may be more reluctant to provide their income information [5]. This situation is relatively complicated and there is no universal solution.

ISBN: 978-81-931039-5-1

2.2 Review of Work on Incomplete Data

In the past decades, significant efforts have been devoted to this area from the point of view of statistical theory, machine learning and so on. Various methods for handling incomplete data have been introduced and these methods can be summarized as follows: samples or features deletion, missing values imputation and learning with missing data.

2.2.1 Samples or Features Deletion

Samples or features with missing values are simply removed from the dataset. This method is easy to implement and usually performs well when the missing rate is low. However, it is obvious that it may ignore some potentially valuable information and create bias in the dataset.

2.2.2 Imputation of Missing Values

Most studies on incomplete data focus on imputation. Imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population. The imputation of missing data of a feature is to generate values drawn from an estimate of the distribution of this variable. Common imputation schemes include completing missing data with specific values such as the unconditional mean or the conditional mean (if one has an estimate for the distribution of missing features given the observed features).

2.2.3 Learning with Missing Data

Some classifiers can be customized in order to handle incomplete data directly, such as Artificial Neural Network (ANN), C4.5 decision trees, Bayesian Networks (BN) , Rough sets and Logistic regression algorithm. Generally speaking, different approaches suit different datasets, which should be selected according to the property of the dataset at hand as well as the requirement on algorithm complexity and efficiency.

2.3 Statistical Framework of Incomplete Data

The statistical framework of incomplete or missing data is present based on Little and Rubin (1987).. In this framework, the dataset is denoted as X have N items (x1, x2, ..., xN), which is composed of two components, namely, observed components (xo) and missing component (xm). The framework considers a random process for both data generation and missing data mechanism with joint probability distribution as given in Equation (1).

(1). $P(X, R|\theta, \phi) = P(X|\theta) P(R|X, \phi)$ (1)

Where θ is data generation process and ϕ for missing data mechanism. The notion of missing data mechanism can be formalized using a missing data indicator matrix R.

III. METHODOLOGY

Data mining process as a five step procedure. The first step declares the selecting or segmenting the data according to some criteria e.g. all people who own vehicle, in this way subsets of the data can be determined.

The second step is preprocessing. This is the data cleaning stage where certain information is removed which is judged unnecessary and may slow down queries. In this step, storage of unnecessary values (Example : gender details of a patient when studying pregnancy), out-of-range values (Example : Salary 100), missing values, and data values which in general lead to misleading errors, are identified and attempts to correct these problematic data are made. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0.

ISBN: 978-81-931039-5-1

The third step transforms the cleaned data to a format which is readily usable and navigable by the data mining techniques.

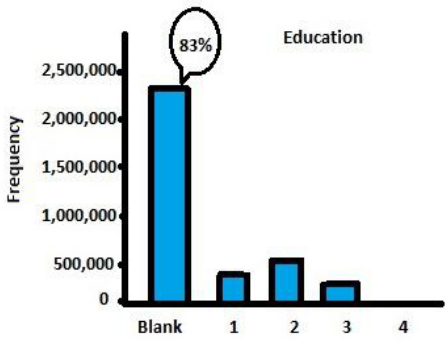
The fourth stage is concerned with using data mining techniques for the extraction of patterns from the transformed dataset. The discovered knowledge is then interpreted and evaluated for human decision-making in the last step.

Incomplete data problems usually occur in areas such as social sciences, bank or shop surveys and medical research. Suppose a set of diagnostic data of patients and normal people among different hospitals in different areas has been collected. The dataset is likely to be incomplete due to several reasons.

Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values.

In general, pattern classification with missing data concerns two different problems, handling missing values and pattern classification. Most of the approaches in the literature can be grouped into four different types depending on how both problems are solved,

- Deletion of incomplete cases and classifier design using only the complete data portion.
- Imputation or estimation of missing data and learning of the classification problem using the edited set, i.e.,
- Complete data portion and incomplete patterns with imputed values.
- Use of model-based procedures, where the data distribution is modeled by means of some procedures, e.g., by expectation–maximization (EM) algorithm.
- Use of machine learning procedures, where missing values are incorporated to the classifier.



In the two-first types of approaches, the two problems, handling missing values (data deletion and imputation, in each case, respectively) and pattern classification, are solved separately; in contrast, the third type of approach model the probability density function (PDF) of the input data (complete and incomplete cases), which is used to classify using the Bayes decision theory. Finally, in the last kind of approaches, the classifier has been designed for handling incomplete input data without a previous estimation of missing data. Our main goal is to show the most representative and useful procedures for handling missing data in classification problems, with a special emphasis on solutions based on machine learning. Due to space constraints, are not able to provide a complete and detailed study of proposed solutions for incomplete data classification. Thus, this review paper provides a wide and general overview of the state-of the- art in this field. The remainder of this paper is structured as follows.

ISBN: 978-81-931039-5-1

This work reviews the most important missing data techniques in pattern classification, trying to highlight their advantages and disadvantages. An excellent reference for missing data is the book written by little and Rubin [16], which gives an accurate mathematical and statistical background in this field.

Missing value replacement policies:

i. Ignore the records with missing values.
ii. Replace them with a global constant.
iii. Fill the missing value based on domain knowledge.
iv. Replace them with a mean or frequent value.
v. Use modeling techniques such as nearest neibours, Bay's rule, and Decision tree.

IV. TECHNIQUES USED FOR MISSING VALUES

Dealing with missing values means to find an approach that can fill them and maintain (or approximate) as closely as possible the original distribution of the data. In this section, the various methods used are discussed.

4.1 Place of Implementation during Mining

Generally, the methods that deal with missing values can be implemented at two stages [7]. They are,

(1) Before mining (Pre-replacing methods) and

(2) During mining (Embedded methods).

Pre-replacing methods replace missing values before the data mining process, while embedded methods deal with missing values during or along with the data mining process. Pre-replacing methods are either statistics based or machine-based.

4.2 Machine Learning Classification Methods with Missing Values

The pattern of missing values is an important characteristic that plays a vital role in the performance of a classifier. The problem of classification with missing data generally involves two steps. They are

(i) Handling missing values and

(ii) Classification.

Table I: Comparative Evaluation Pre-Replacing Methods to Deal with Missing Values

Method	Computation Cost	Attributes
Mean-mode method	Low	Num & Cat
Linear Regression	Low	Num
Standard Deviation	Low	Num
Nearest Neighbor Estimator	High	Num & Cat
Decision Tree imputation	Middle	Char
Auto Associative Neural Network	High	Num & Cat

ISBN: 978-81-931039-5-1

Table II: Comparative Evaluation Embedded Methods to Deal with Missing Values

Method	Computation Cost	Attributes
Case wise Deletion	Low	Num & Cat
Lazy Decision Tree	High	Num & Cat
Dynamic Path Generation	High	Num & Cat
C4.5	Middle	Num & Cat
Surrogate split	Middle	Num & Cat

In the above tables describes large dataset with missing values, complicated methods are not suitable because of their high computational cost .Use simple methods that can reach performance as good as complicated ones.

Depending on the method used for both these steps, the techniques can be grouped into four main categories as given below.

[1]. Deletion of missing values (complete cases and available data analysis), and classifier design using only the complete instances,

[2]. Imputation (estimation and replacement) of missing input values, and after that, another machine learns the classification task using the edited complete set, i.e., complete instances and incomplete patterns with imputed values,

[3]. Use of Maximum Likelihood (ML) approaches, where the input data distribution is modeled by the Expectation-Maximization (EM) algorithm, and the classification is performed by means of the Bayes rule,

[4]. Use of machine learning procedures able to handle missing data without an explicit imputation.

4.3 Other Methods

This section discusses three techniques that are less frequently used. The reason behind their infrequent usage is its poor classification performance when presented with a datasets with missing data. They are,

i). Hot deck imputation

ii). Mean substitution

iii). Regression substitution

Machine Learning Classification Methods

- Ensemble methods
- Fuzzy approaches
- Decision trees
- Support Vector
- Methods (SVM)
- Expectation-Maximization(EM) algorithm
- Mixer Models with EM algorithm

ISBN: 978-81-931039-5-1

V. CONCLUSION

The major purpose of this paper discussed the various approaches used in classification with incomplete data values.

Missing or incomplete data is a usual drawback in many real-world applications of pattern classification. Data may contain unknown features due to different reasons, e.g., sensor failures producing a distorted or immeasurable value, data occlusion by noise, non-response in surveys.

Handling missing data has become a fundamental requirement for pattern classification because an inappropriate missing data treatment may cause large errors or false classification results. It could be seen that both statistical approaches and machine learning approaches have been successful to a certain extent in the problem domain under discussion.

While considering the missing data imputation approaches based on machine learning, artificial neural network algorithms, K-Nearest Neighbor algorithm and Self-Organizing Maps (SOM) are more frequently used. Several variants of SOM like tree-structured SOM are also in existence. EM algorithm is frequently used while considering maximum likelihood based approaches. Decision trees and fuzzy approaches have also been studied. In spite of these studies, it is understood that hundred per cent success is still seen only as a distant possibility because of the numerous factors influencing the relative success of the competing techniques.

Currently, no one method can be used for handling all types of missing data problem and the only right answer, as opined by [3] for missing data procedures, "Rreturn to the old precept that still holds true: The only real cure for missing data is to not have any". However, with the growing database size and complexity in the data attributes, missing value handling procedures is a mandatory process.

Different approaches suit different datasets, which should be selected according to the property of the dataset at hand as well as the requirement on algorithm complexity and efficiency. Moreover, a previous analysis of the classification problem to be solved is very important in order to select the most suitable missing data treatment. The various techniques identified in this study, in future, can be compared with respect to their performance in classification accuracy while provided with incomplete datasets.

Researchers and practitioners often face missing values when applying learned models. This study provides a valuable step toward understanding how best to deal with them, and why.

REFERENCES

[1]. Adèr, H.J. and Mellenbergh, G.J. (Eds.) (2008) Chapter 13: Missing data, Advising on Research Methods: A consultant's companion, Huizen, The Netherlands: Johannes van Kessel Publishing, Pp. 305-332.

[2]. Altmayer, L. (2010) Hot-Deck Imputation: A simple data stepapproach, http://analytics.ncsu.edu/sesug/1999/075.pdf

[3]. Anderson, A.B., Basilevsky, A. and Hum, D.P.J. (1983) Missing data: A review of the literature, P.H. Rossi, Wright, J.D. and A.B. Anderson (Eds.), Handbook of survey research, San Diego: Academic Press, Pp.415-494.

[4]. Chen, J. and Shao, J., (2001) Jackknife variance estimation for nearest-neighbor imputation. J. Amer. Statist. Assoc., Vol.96, Pp

ISBN: 978-81-931039-5-1

[5]. Donders, A., van der Heijden, G., Stijnen, T. and Moons, K. (2006) Review: a gentle introduction to imputation of missing values, Journal of Clinical Epidemiology, Vol. 59, Pp. 1087-1091.

[6]. Fayyad, U.M., Shapiro, G.P. and Smyth, P. (1996) Data Mining and Knowledge Discovery in Databases: An overview, Communications of ACM, Vol. 39, No. 11, P. 27-34.

[7]. Fujikawa, Y. and Ho, T. (2002) Cluster-based algorithms for dealing with missing values, M.S. Chen, Yu, P.S. and Liu, B. (Eds.), PAKDD 2002, LNAI 2336, Springer-Verlag, Pp. 549-554.

[8]. Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2nd edition.

[9]. Lall, U. and Sharma, A., (1996) A nearest-neighbor bootstrap for resampling hydrologic time series, Water Resource. Res., Vol.32, Pp.679–693.

[10]. Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data, 2nd Edition, John Wiley and Sons, New York.

[11]. Martin, T. (2003) A day in the life of a Data Miner, Bulletin of the International Statistical Institute, 54th Session, Vol. LX, Invited Papers, August 2003, Berlin, Germany. Pp. 298-301

[12]. Messner, S.F. (1992) Exploring the Consequences of Erratic Data Reporting for Cross-National Research on Homicide. Journal of Quantitative Criminology, Vol.8, No.2, Pp. 155-173.

[13]. Sancho-Gomez, J., Garcia-Laencina, P.J. and Figueiras-Vidal, A.R. (2009) Combining missing data imputation and pattern classification in a multi-layer perceptron, Inteligent Automation and Soft Computing, Vol. 15, No. 4, Pp. 539-553.

[14]. Scheuren, F. (2005) multiple imputations: How it began and continues, The American Statistician, Vol. 59, Pp. 315-319.

[15]. Zhang, C.Q., et al., (2007) An Imputation Method for Missing Values. PAKDD, LNAI 4426, Pp. 1080–1087.

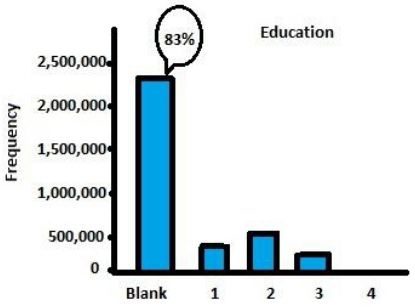
[16]. Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New Jersey

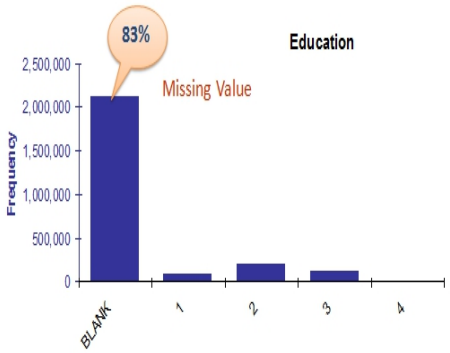
[17]. Zhang, S.C., et al., (2004) Information Enhancement for Data Mining, IEEE Intelligent Systems, Vol. 19, No.2,Pp. 12-13.

[18]. Zhang, S.C., et al., (2004) Information Enhancement for Data Mining, IEEE Intelligent Systems, Vol. 19, No.2,Pp. 12-13

[19]. Zhang, S.C., et al., (2005) Missing is useful: Missing values in cost-sensitive decision trees, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No.12, Pp. 1689-1693.

[20]. Duda RO, Hart PE, Stork DG (2000) Pattern classification. Wiley, New York





International Conference on Recent Advances in Engineering Science and Management
PHD Chamber of Commerce and Industry, New Delhi
30 August 2015, www.conferenceworld.in, www.icraesm.org
ISBN: 978-81-931039-5-1

A SURVEY ON NETWORK SECURITY USING CRYPTOGRAPHIC TECHNIQUE

Dr. L. Sankari[1], R. Keerthana[2]

[1]Associate Professor, [2]Research Scholar, Department of Computer Science,
Sri Ramakrishna College of Arts & Science for Women, Coimbatore

ABSTRACT

Network Security is the most vital component in information security because it is responsible for securing all information passed through network computers. Security in networks refersto all functions in hardware and software, characteristic, feature, operational procedure, accountability, measures, control access, administrative policy and management policy required to provide an acceptable level of protection for Software and Hardware, and information in a network. Only one particular element underlies many of the security mechanisms in use which is called as Cryptography. It is a technology, important for network security this paper focus on study of various research paper in which cryptography technique is the main role.

Keywords: Network Security, Management Policy, Accountability, Access Control, Cryptography.

I. INTRODUCTION

Network is the connection between computers. There is different type of networks connections available such as LAN,WAN, MAN etc. Internet is the largest computer network in the world. World Wide Web is a part of internet. Two kinds of connections present wired connection and wireless connection. In wired connection it provides a plenty of security. There are three ways in wired connection they are Ethernet, phoneLine and Network interface card. Twisted copper pair or coaxial cables are used to connect between systems. Unshielded twisted pair is also used in

Ethernet connections it is costly it is mainly used to connect more no of systems for bulk purpose.

Network interface card a card is used to connect between computers. Wireless LAN uses radio waves to connect devices such as laptops to internet. There are many ways for wireless network connections such as Satellite, Bluetooth and Infrared etc. Satellite is a wireless technology used all over the world. Broadcasting ability in phone and modems are very high. Bluetooth is connected through radio waves. The transmission range lies between 15 -50 feet. It saves the battery and uses it when required. Infrared transmits through Light Emitting diodes. It will not transmit information when obstacles are present.

II. LITERATURE SURVEY

Some of the concepts used in cryptography aredescribed here [1]:

2.1 Cryptography

ISBN: 978-81-931039-5-1

- Plain Text: Any communication in the language thatwe speak- that is the human language, takes the formof plain text. It is understood by the sender, the recipient and also by anyone who gets an access to that message.
- Cipher Text: Cipher means a code or a secretmessage. When a plain text is modified using anysuitable scheme the resulting message is called ascipher text.
- Encryption: The Sender converts plain text in to cipher text using secret key this process is known as encryption.
- Decryption: The reverse process of converting cipher text messages back to plain text is called asdecryption.
- Key: An important aspect of performing encryptionand decryption is the key. It is the key used forencryption and decryption that makes the process ofcryptography secure.

2.2 Purpose of Cryptography

Cryptography Serves Following Purposes

- Confidentiality: The main aim of confidentialityspecifies that only the sender and the intendedrecipient should be able to access the contents of amessage.
- Authentication: Authentication mechanisms help toestablish proof of identities. This process says thatthe origin of the message is correctly identified.
- Integrity: The integrity ensures that thecontents of the message remain the same when itreaches the intended recipient as sent by the sender.
- Availability: The principle of availability states thatresources should be available to authorized parties allthe times.

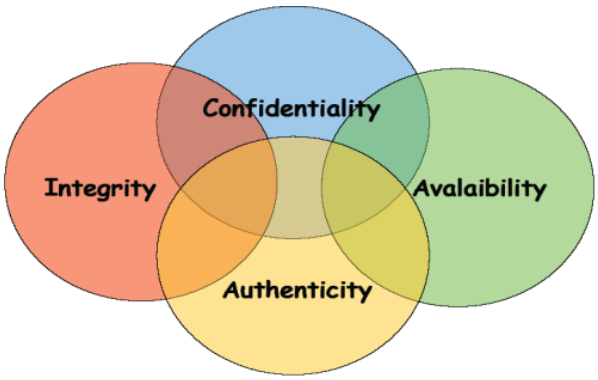


Fig: 2.2.1

- Access Control: It controls and specifies who can access the message.
- Non- repudiation: Non-repudiation will not allow the sender of a message to refute the claim of not sending the message.

2.3 Types of Cryptography

ISBN: 978-81-931039-5-1

Two Types of Cryptography are Studied

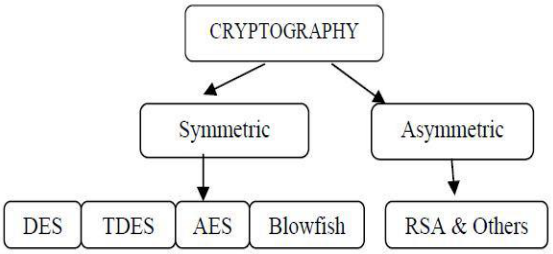


Fig: 2.3.1

- Symmetric Key Cryptography: If samekeyisused for encryption and decryption, then that mechanism is known as symmetric key cryptography.

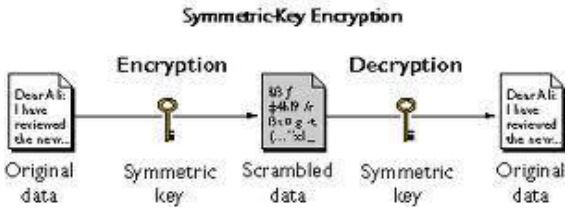


Fig: 2.3.2

- DES: DES stands for data encryption standards. It is a block cipher. In DES we use 64 bit plain text and 56 bit keys. There are three phases present in 1stPhase Initial permutation takes place in 2nd Phase 16 Rounds are performed and in 3rd phase Swapping takes place. In initial permutation it will send the 64 bit plain text to rounds and in each rounds it splits the 64 bit plain text and 56 bit keys into two equal halfs it performs permutation, XOR, substitution etc. Use of 56 bit key is the strength of DES. So strength of DES lies only in its keys.
- TDES: In triple DES there are two types they are Triple DES with 2 keys and Triple DES with 3 keys. In 3 keys it uses three different types of keys for encryption. In two keys 1stis used for Encryption 2nd key for decryption and again uses 1st key to encrypt. This is the working principle of triple DES. Using three Different types of keys it provides high security.
- AES: AES stands for Advanced Encryption standards. It was published by NIST. It is a symmetric block cipher. In AES block cipher. In AES block size is 128 bit and its 128 bit and its key size is 128,192 & 256. In this 128 has 10 rounds, 192 has 12 rounds and 256 has 14 rounds. In general structure the plain text is 128 bit blocks and key in 128 it has 10 rounds. In each round it performs four different stages. Such as substitute bytes, shift rows, Mix column and Add round Key. The plain text of 128 bit blocks are divided into 16 bytes and 1 byte consist of 8 bits and the 16 bit is represented in matrix format. In this 128 bit keys are divided into 16 bytes each consist of 8 bits. Using expansion function it converts to 44 words and each word has 4 bytes.
- Blowfish: Blowfish is a symmetric block cipher. It is developed by Brue Schneider. Blowfish is designed with the following objectives such as fast, compact, simple and secure. It is fast so it encrypts data on 32 bit microprocessor at a rate of 18 clock cycle per byte. Blowfish is compact because it executes in less than 5 Kb memories. Blowfish is simple to use and easy to implement. As it has 448 bit long key to use so it provides high security.

ISBN: 978-81-931039-5-1

- Asymmetric Key Cryptography: Mechanism of using two different keys that is one key for encryption and another key for decryption is known as asymmetric key crypto system.

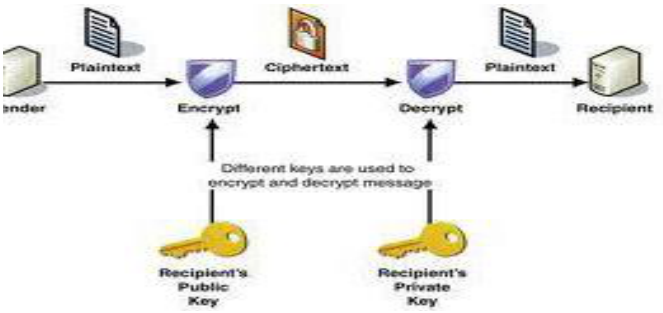


Fig: 2.3.3

- RSA: RSA stands for Rivest-Shamir-Adleman. It was developed by Ron Rivest, Adi Shamir and Len Adleman in 1977. It is used in public key encryption. It is encrypted using binary values so it is highly secure. Four possible approaches to attacking RSA algorithm are Brute force, Mathematical attacks, Timing attacks and Chosen cipher text attacks. It provides high authentication and confidentiality.

III. SECURITY ATTACKS

Two Types of Attacks

3.1 Active Attack

- Information is taken and it also affects the system. Attacks made are detected easily.
- This type of attack takes place through release of message content and traffic analysis.

3.2 Passive Attack

- In passive attack third person can steal the information and the system is not affect. It is difficult to detect. It is better to prevent.

IV. LITERATURE STUDY

- Madhumita panda [1] compared two algorithms RSA and ECC and said ECC is more advantageous than RSA, since it occupied less memory. CPU consumption was low and the key size was also very low.
- Himanshu Gupta &VinodkumarSharma [2] used multiple encryption and conventional encryption technique online transaction takes place over wireless network in secure manner.
- Andrew Simmonds, peter sandilands&louis Van Ekert [3] used a technique to identify the enemy during different attacks and also described that there is no perfect security always. This depends on how a system will react to a successful attack.
- PriyankaBhalia et al [4] uses quantum key Distribution to provide high security in WLANs.
- Ayesha Manzoor et al [5] uses Quadratic scheme to provide secure transmission.

ISBN: 978-81-931039-5-1

Survey Table 1

S.No	Title	Author	Solution
1	Security in Wireless Sensor Networks using Cryptographic Techniques	Madhumita Panda	Comparing ECC, and RSA .ECC is more advantageous than RSA,due to low memory usagelowCPUconsumption and shorter key size
2	Role of Multiple Encryption in Secure Electronic Transaction	Himanshu Gupta	Multiple encryption in Secure Electronic Transaction describes the enhanced security as well as integrity of data confidentialitydue to multiple encryption operations
3	Ontology for network security attacks	Andrew Simmond	A framework for network security based on the concepts that are proved.
4	Frame work for wireless Network security using quantum cryptography	Priyanka Bhatia	Classical cryptographic algorithms is difficult for key management and distribution but use of QKD provides high security.
5	Software quality Assurance in network security using cryptographic techniques.	Ayesha Manzoor	RSA leads to slower speed so quadratic scheme is introduced for secure transmission.

V. CONCLUSIONS

Network Security is the most vital component in information security because it is responsible for securing all information passed through network computers. Network security consists of the provisions made in an underlying computer network infrastructure, policies adopted by the admin in network to protect the network and the network-accessible resources from unauthorized users, and consistent &continuous, monitoring & measurement of its effectiveness (or lack) combined together. We have learnt various cryptographic techniques to increase the security of network.

REFERENCES

[1] American Journal of Engineering Research (AJER) e-ISSN : 2320-0847 p-ISSN : 2320-0936 Volume-03, Issue-01, pp-50-56 www.ajer.org Research Paper Open Access Security in Wireless Sensor Networks using Cryptographic Techniques Madhumita Panda Sambalpur University Institute of Information Technology(SUIIT)Burla, Sambalpur, Odisha, India

[2] International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011 ROLE OF MULTIPLE ENCRYPTION IN SECURE ELECTRONIC TRANSACTION Himanshu Gupta, Vinod Kumar Sharma, GurukulaKangriVishwavidyalaya, Haridwar, India

ISBN: 978-81-931039-5-1

[3] An Ontology for Network Security Attacks Andrew Simmonds1, Peter Sandilands1, Louis van Ekert11Faculty of IT, University of Technology Sydney, PO Box 123, Broadway, NSW 2007,Australia

[4] FRAMEWORK FOR WIRELESS NETWORK SECURITY USING QUANTUM CRYPTOGRAPHY Priyanka Bhatia1 and Ronak Sumbaly2 1, 2 Department of Computer Science, BITS Pilani Dubai, United Arab Emirates

[5] Software Quality Assurance in Network Security Using Cryptographic Techniques Ayesha Manzoor, Sidra Shabbir & Mehreen Sirshar Fatima Jinnah Women University, The Mall Rawalpindi, Pakistan Feb 2015