Print ISSN : 2395-1990
Online ISSN : 2394-4099
www.ijsrset.com

Scientific Journal Impact Factor Value = 3.632

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 3, Issue 7, September-2017

Email: editor@ijsrset.com, ijsrset@gmail.com

International Journal of Scientific Research in Science, Engineering and Technology

[Print ISSN: 2395-1990 | Online ISSN: 2394-4099]

Volume 3, Issue 7, September-2017

International Peer Reviewed, Open Access Journal
Bimonthly Publication

Published By

Technoscience Academy
(The International Open Access Publisher)

Email: info@technoscienceacademy.com
Website: www.technoscienceacademy.com

Advisory/Editorial Board

International Advisory/Editorial Board

CONTENTS

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section:Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering,
Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

Brain Tumour Segmentation and Classification using Convolutional Neural Network in MRI images

Jijith M P[1], Sadhik M S[2], Prof. Linda Sara Mathew[3]

[1,2,3] Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India

ABSTRACT

In brain tumors, gliomas are the most common and aggressive, leading to a very short life expectancy in their highest grade. Thus, treatment planning is a key stage to improve the quality of life of oncological patients. Magnetic resonance imaging (MRI) is a widely used imaging technique to assess these tumors, but the large amount of data produced by MRI prevents manual segmentation in a reasonable time, limiting the use of precise quantitative measurements in the clinical practice. So, automatic and reliable segmentation methods are required; however, the large spatial and structural variability among brain tumors make automatic segmentation a challenging problem. Here we propose an automatic segmentation method based on Convolutional Neural Networks (CNN), exploring small 33 kernels. The use of small kernels allows designing a deeper architecture, besides having a positive effect against over fitting, given the fewer number of weights in the network. We also investigated the use of intensity normalization as a pre-processing step, which though not common in CNN-based segmentation methods, proved together with data augmentation to be very effective for brain tumor segmentation in MRI images. We also try to find out the area of the tumor effected potion in the input image. There are mainly three stages includes. The first stage is pre-processing, second stage is classification via deep neural network and the final stage is the post-processing .

Keywords: Convolutional Neural Networks (CNN), Magnetic Resonance Imaging (MRI)

I. INTRODUCTION

Data mining is the process of extracting hidden knowledge, useful trends and patterns from large databases which is used by organizations for decision making purpose. Data mining deals with the difficulty of extracting patterns from the information by paying suspicious attention to computing, communication and human- computer interface issues. There are various data mining techniques available like clustering, classification, prediction, outlier analysis. Convolutional neural networks are widely used in pattern- and image-recognition problems as they have a number of advantages compared to other techniques. This white paper covers the basics of CNNs including a description of the various layers used. Using traffic sign recognition as an example, we discuss the challenges of the general problem and introduce algorithms and implementation software developed by Cadence that can trade off computational burden and energy for a modest degradation in sign recognition rates. We outline the challenges of using CNNs in embedded systems and introduce the key characteristics of the Cadence Tensilica Vision digital signal processor for Imaging and Computer Vision and software that make it so suitable for CNN applications across many imaging and related recognition tasks.

Clusters are very useful to extract interesting patterns from large data. But as the application grows it generates large amount of data. Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. The growing need for clustering algorithms is required to handle huge size of databases that is common nowadays. Clustering has become an increasingly important task in modern applications such as marketing, bio informatics, spatial analysis, molecular biology as well. In all these applications generates a large amount of data. On the other hand data are originally collected at different sites. This leads us to the requirement of clustering large data sets. Typical clustering algorithms cluster a data set stored in a single site. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction

Emerging data mining applications place many requirements on clustering techniques, motivating need for developing algorithms to handle large data sets. Such as 1) Effective treatment of high dimensionality and type of attributes algorithm can handle, where an object typically has dozens of attributes and the domain for each attribute can be large. Many dimensions or combinations of dimensions can have noise or values that are uniformly distributed. Therefore, distance functions that use all the dimensions of the data may be ineffective. 2) Interpretability of results: It is particularly important to have simple representations because most visualization techniques do not work well in high dimensional spaces. 3) Scalability and usability to large datasets: The clustering technique should be fast and scale with the number of dimensions and the size of input. It should be insensitive to the order in which the data records are presented. 4) Handling outliers and ability to find clusters of irregular shape. 5) Data order dependency: Finally, it should not presume some canonical form for data distribution. These are the main requirements considered while clustering large data sets and is driven by the need of applying algorithms for clustering datasets separated with noise efficiently.

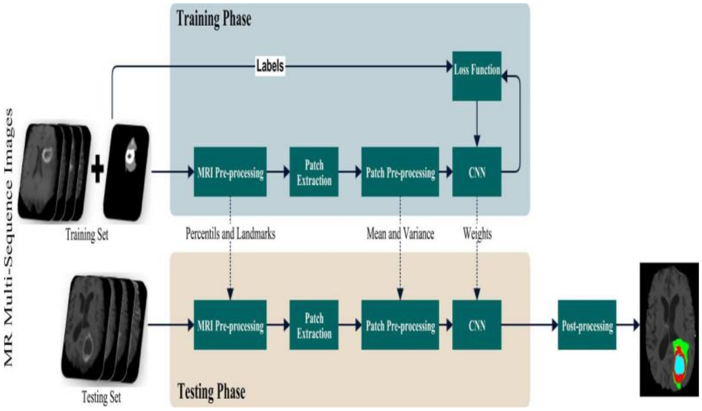


Fig 1: Stages in MRI image Processing

II. LITERATURE SURVEY

Literature survey is the process of referring some base papers that are related to our topic and then choose the best methods to implement our system.

In 2014, Ed-EdilyMohd Azhari[1] proposed gives the details about cells each type of cell has special functions. Most cells in the body grow and then divide in an orderly way to form new cells as they are needed to keep the body healthy and work properly. When cells lose the ability to control their growth, they divide too often and without any order. The extra cells form a mass of tissue called a tumor. Brain tumors are created by abnormal and uncontrolled cell division in brain itself. Generally, if the growth becomes more than 50%, then the patient may not be able to recover. Hence detection of brain tumor at its early stage with its accurate diagnosis is very important. Identification of tumor involves tests like CT and MRI. MRI plays vital role in identifying location, size and type of brain tumor.

An image may be defined as a two-dimensional function f(x, y), where x & y are spatial coordinates, & the amplitude of at any pair of coordinates (x, y) is called the intensity or gray level of the image at that point. Digital image is composed of a finite number of elements, each of which has a particular location & value. The elements are called pixels. For the extraction of useful features from the complex brain structure Magnetic resonance imaging (MRI) is reliable. MRI is very important in order to improve the diagnosis and treatment of brain tumor, by detecting tumor at its early stage. Segmentation of medical images is first important step in their analysis, the segmentation gives organ detection and variation of growth of tissues as a output in medical images. Some segmentation approaches are Global image threshold using Otsu's method, Region Growing, Edge Based Segmentation**,** K-means Clustering, Fuzzy C-means Clustering.

Clustering the process of collection of objects which are similar between them and are dissimilar objects belonging to other clusters. . Region growing is a technique of segmentation in which pixels with similar intensities are grouped in order to find the regions directly. This group of pixels belonging to the region of focus is known as seeds. K-mean is example of exclusive clustering algorithm. In overlapping clustering, one data (pixel) is belonging two or more clusters**.**

Watershed Segmentation: It is one of the best methods to group pixels of an image on the basis of their intensities. Pixels falling under similar intensities are grouped together. It is a good segmentation technique for dividing an image to separate a tumor from the image Watershed is a mathematical morphological operating tool. Watershed is normally used for checking output rather than using as an input segmentation technique because it usually suffers from over segmentation and under segmentation. Morphological image processing (or morphology) describes a range of image processing techniques that deal with the shape (or morphology) of features in an image and morphological operations are typically applied to remove imperfections introduced during segmentation, and so typically operate on bi-level images i.e. binary images.

In 2014, Kailash Sinha[2] describes The structure and function of the brain and researchers using MRI imaging techniques Brain tumor extraction and its analysis are challenging tasks in medical image processing because brain image and its structure is complicated that can be analyzed only by expert radiologists. This paper presents a comparative study of three segmentation methods implemented for tumor detection. The methods include k-means clustering with watershed segmentation algorithm, optimized k-means clustering with genetic algorithm and optimized c- means clustering with genetic algorithm. Traditional k-means algorithm is sensitive to the initial cluster centers. Genetic c-means and k-means clustering techniques are used to detect tumor in MRI of brain images .At the end of process the tumor is extracted from the MR image and its exact position and the shape are determined .Over segmentation and sensitivity to false edges are other difficulties in ordinary k-means method. Determination of exact location and area of brain tumor using k-means method becomes very difficult and hence use of genetic algorithm is suggested. Fittest is searched by the algorithm and hence used in optimization tasks, The implementation of genetic algorithm begins with an initial population of chromosomes which are randomly selected. Chromosome is a long thread of DNA.

The c-means clustering method has been implemented and its performance can be improved by using optimization with the use of genetic algorithm. The combined method results an improvement in segmentation efficiency and higher area of affected region extraction and detection. MRI images were segmented using k-means clustering and Watershed algorithm. The method is implemented using process of two stages. The first stage of the process uses k-means clustering and primary segmentation results are produced for the brain MRI images. Second stage of the process is applied as watershed segmentation algorithm to improve the results of the primary segmentation; and the results obtained are final results.

Datta et al (2011) introduced colour-based segmentation using k-means clustering for brain tumor detection. The developed algorithm shows better result than Canny based edge detection. Nandha et al (2010) designed intelligent system to diagnose brain tumor through MRI using image processing clustering algorithms such as Fuzzy c-means along with intelligent optimization tools, such as Genetic Algorithm (GA), and Particle Swarm Optimization (PSO).

Jobin et al (2012) proposed a method which integrated the k-means clustering algorithm with the marker controlled watershed segmentation algorithm.

Yang et al (2010) presented a new image segmentation algorithm W-SPK (combining watershed and K-means clustering method based on simulated annealing particle swarm optimization) to overcome the shortcomings of watershed and realize fast and accurate image segmentation [9]. Sasikala et al (2006) presented an automatic segmentation of malignant tumor in magnetic resonance images (MRI's) of brain using optimal texture features.

III. PROPOSED SYSTEM

MRI images are altered by the bias field distortion. This makes the intensity of the same tissues to vary across the image. To correct it, we applied the N4ITK method. However, this is not enough to ensure that the intensity distribution of a tissue type is in a similar intensity scale across different subjects for the same MRI sequence, which is an explicit or implicit assumption in most segmentation methods [37]. In fact, it can vary even if the image of the same patient is acquired in the same scanner in different time points, or in the presence of a pathology. So, to make the contrast and intensity ranges more similar across patients and acquisitions, we apply the intensity normalization method proposed by Nyúlon each sequence. In this intensity normalization method, a set of intensity landmarks are learned for each sequence from the training set. And are chosen for each MRI sequence as described represents the intensity at the percentile. After training, the intensity normalization is accomplished by linearly transforming the original intensities between two landmarks into the corresponding learned landmarks. In this way, the histogram of each sequence is more similar across subjects. After normalizing the MRI images, we compute the mean intensity value and standard deviation across all training patches extracted for each sequence. Then, we normalize the patches on each sequence to have zero mean and unit variance

CNN were used to achieve some breakthrough results and win well-known contests. The application of Convolutional layers consists in convolving a signal or an image with kernels to obtain feature maps. So, a unit in a feature map is connected to the previous layer through the weights of the kernels. The weights of the kernels are adapted during the training phase by back propagation, in order to enhance certain characteristics of the input. Since the kernels are shared among all units of the same feature maps, Convolutional layers have fewer weights to train than dense FC layers, making CNN easir to train and less prone to over fitting. Moreover, since the same kernel is convolved over all the image, the same feature is detected independently of the locating—translation invariance. By using kernels, information of the neighborhood is taken into account, which is a useful source of context information. Usually, a non-linear activation function is applied on the output of each neural unit. If we stack several Convolutional layers, the extracted features become more abstract with the increasing depth. The first layers enhance features such as edges, which are aggregated in the following layers as motifs, parts, or objects.

The following concepts are important in the context of CNN:

1) Initialization*:* It is important to achieve convergence. The activations and the gradients are maintained in controlled levels; otherwise back-propagated gradients could vanish or explode.

2) Activation Function: It is responsible for non-linearly transforming the data. Rectifier linear units (ReLU), defined as were found to achieve better results than the more classical sigmoid, or hyperbolic tangent functions, and speed up training. However, imposing a constant 0 can impair the gradient flowing and consequent adjustment of the weights we cope with these limitations using a variant called leaky rectifier linear unit that introduces a small slope on the negative part of the function. This function is defined as where is the leakiness parameter. In the last FC layer, we use softmax.

3) Pooling: It combines spatially nearby features in the feature maps. This combination of possibly redundant features makes the representation more compact and invariant to small image changes, such as in significant details; it also decreases the computational load of the next stages. To join features it is more common to use max-pooling or average pooling.

4) Regularization: It is used to reduce over fitting. We use Dropout in the FC layers. In each training step, it removes nodes from the network with probability In this way, it forces all nodes of the FC layers to learn better representations of the data, preventing nodes from co-adapting to each other. At test time, all nodes are used. Dropout can be seen as an ensemble of different networks and a form of bagging, since each network is trained with a portion of the training data.

5) Data Augmentation: It can be used to increase the size of training sets and reduce overfitting. Since the class of the patch is obtained by the central voxel, we restricted the data augmentation to rotating operations. Some authors also consider image translations, but for segmentation this could result in attributing a wrong class to the patch. So, we increased our data set during training by generating new patches through the rotation of the original patch. In our proposal, we used angles multiple of 90 although another alternative will be evaluated.

6) Loss Function: It is the function to be minimized during training. We used the Categorical Cross-entropy, where represents the probabilistic predictions (after the softmax) and is the target. In the next subsections, we discuss the architecture and training of our CNN.

7) Architecture: We aim at a reliable segmentation method; however, brain tumors present large variability in intra-tumoral structures, which makes the segmentation a challenging problem. To reduce such complexity, we designed a CNN and tuned the intensity normalization transformation for each tumor grade LGG and HGG.

This is supported by the need of setting Dropout with in LGG, while it is in HGG, since the database used for evaluation contained more HGG then LGG cases. Additionally, the appearance and patterns are different in HGG and LGG. Since we are doing segmentation, we need a precise sense of location. Pooling can be positive to achieve invariance and to eliminate irrelevant details; however, it can also have a negative effect by eliminating important details. We apply overlapping pooling with 33 receptive fields and 2 2 stride to keep more information of location. In the Convolutional layers the feature maps are padded before convolution, so that the resulting feature maps could maintain the same dimensions. in all layers with weights, with the exception of the last that uses softmax. Dropout was used only in the FC layers.

8) Training: To train the CNN the loss function must be minimized, but it is highly non-linear. We use Stochastic Gradient Descent as an optimization algorithm, which takes steps proportionally to the negative of the gradient in the direction of local minima. Nevertheless, in regions of low curvature it can be slow. So, we also use Nesterov's Accelerated Momentum to accelerate the algorithm in those regions.

Here we use post preprocessing technique, in this method Some small clusters may be erroneously classified as tumor. To deal with that, we impose volumetric constrains by removing clusters in the segmentation obtained by the CNN that are smaller than a predefined threshold.

IV. CONCLUSION

The proposed brain tumor detection and localization framework detects and localizes brain tumor in MR imaging. In summary, we propose a novel CNN-based method for segmentation of brain tumors in MRI images. We start by a pre-processing stage consisting of bias field correction, intensity and patch normalization. After that, during training, the number of training patches is artificially augmented by rotating the training patches, and using samples of HGG to augment the number of rare LGG classes. The CNN is built over convolutional layers with small 3x3 kernels to allow deeper architectures. In designing our method, we address the heterogeneity caused by multi-site multi-scanner acquisitions of MRI images using intensity normalization. We show that this is important in achieving a good segmentation. Brain tumors are highly variable in their spatial localization and structural composition, so we have investigated the use of data augmentation to cope with such variability.

V. REFERENCES

[1] Ed-EdilyMohd. Azhari1, Muhd. MudzakkirMohd. Hatta1, Zaw ZawHtike1 and Shoon Lei Win2, \Brain Tumor Detection And Localization In Magnetic Resonance Imaging", International Journal of Information Technology Convergence and Services (IJITCS) Vol.4, No.1,February 2014.

[2] Kailash Sinha1, G.R. Sinha., \ E_cient Segmentation Methods for Tumor Detection in MRI Images", IEEE Student's Conference on Electrical, Electronics and Computer Science, 2014.

[3] Pratibha Sharma ,Manoj Diwakar ,sangam Choudhary, \ Application of Edge Detection for Brain Tumor Detection", International Journa lof Computer Applications (0975 { 8887) Volume 58, November2012.

[4] Riries Rulaningtyas1 and Khusnul Ain2, \Edge Detection for Brain Tumor Pattern Recognition".

[5] S. Datta, M. Chakraborty, \Brain Tumor Detection from Pre-ProcessedMR Images using Segmentation Techniques",Special Issue on 2nd National Conference-Computing, Communication and Sensor Network(CCSN) Published by Foundation of Computer Science, NewYork, USA. vol.2, pp.1-5, 2011.

[6] Gopal, N.N., Karnan, M., "Diagnose brain tumor through MRI using image processing clustering algorithms such as Fuzzy C Meansalong with intelligent optimization Techniques", IEEE International Conference on Computational Intelligence and Computing Research(ICCIC),Vol.2, No.3, pp.1-4, 2010.

[7] Christ, M. J., Parvathi, R. M. S., \Magnetic Resonance Brain Image Segmentation", International Journal of VLSI Design and Communication Systems, Vol.3,No.4, pp.121-133,2012.

[8] Wenli Yang, Zhiyuan Zeng, Sizhe Zhang, \Application of Combining Watershed and Fast Clustering Method in Image Segmentation",Computer Modeling and Simulation. ICCMS Second International Conference on, Vol.3,No.,pp.170-174,22-24,Jan.2010.

[9] P.Tamije;V. Palanisamy; T. Purusothaman: “Performance Analysis of Clustering Algorithms in Brain Tumor Detection of MR Images” European Journal of Scientific Research, ISSN 1450-216X Vol.62 No.3 (2011), pp. 321-330.

[10] Ratan, Rajeev, Sanjay Sharma, and S. K. Sharma. "Brain tumor detection based on multi-parameter MRI image analysis." International Journal on Graphics, Vision and Image Processing vol 9.no.3, pp.9-17,2009.

[11] S.K.Bandyopadhyay and D.Saha, “Brain region extraction volume calculation,” UNIASCIT, vol. 1, no. 1, pp. 44-48, 2011.

[12] Gopal, N.N.; Karnan, M., "Diagnose brain tumor through MRI using image processing clustering algorithms such as Fuzzy C Means along with intelligent optimization techniques," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),vol.2, no.3, pp.1-4, 2010.

[13] Amanpreet Kaur; Gagan Jindal “Tumor Detection Using Genetic Algorithm” International Journal on Computer Science and Technology, vol. 4, no.1,pp. 423-427 2013.

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section:Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering, Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

EEG Signal for Diagnosing Diseases using Machine Learning

Aswathy K J[1], Swathi Anil[2], Prof. Elizabeth Issac[3]

[1,2,3] Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India

ABSTRACT

Alzheimer's disease is a chronic neurodegenerative disease that usually starts slowly and worsens over time. Alzheimer’s is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills, and eventually the ability to carry out the simplest tasks. In most people, symptom first appears in their mid-60s. Studies have reported that electroencephalogram (EEG) signals of Alzheimer’s disease patients usually have less synchronization when compare to healthy people. Changes in EEG signals start at early stage but, clinically, these changes are not easily detected. Early detection of Alzheimer’s can have treatments with more positive outcomes. The aim of this paper is to classify Alzheimer’s disease patients using EEG signal processing in order to support medical doctors in the right diagnosis formulation. The proposed system consists of mainly five steps: Signal Acquisition, Pre-processing, Feature Extraction, Feature Selection, and Classification. The Signal Acquisition makes use of EEG dataset, Band-pass-filtering is used in the pre-processing stage to get artifact free signal. The necessary features are then extracted from EEG signals using Wavelet Transform and they are subjected to Principal Component Analysis (PCA) for feature selection. The classification is done using Block Based Neural Network (BBNN). Based on the changes in EEG the structure and internal configuration of BBNN are modified.

Keywords : Alzheimer's disease (AD),EEG, Principal Component Analysis, BlockBased Neural Network.

I. INTRODUCTION

Alzheimer’s disease (AD) is a neuro-degenerative disease, the most common form of dementia, third most expensive disease and sixth leading cause of death in the United States. It affects more than 10% of Americans over age 65, nearly 50% of people older than 85, and it is estimated that the prevalence of the disease will triple within the next 50 years. While no known cure exists for Alzheimer’s disease, a number of medications are believed to delay the symptoms (and perhaps causes) of the disease.The most popular 10 warning signs of Alzheimer’s disease as: Memory loss, Difficulty performing,Problems with language, Poor or decreased judgment, Problems with abstract thinking, Misplacing things, Changes in mood, or behavior, Changes in personality, Loss of initiative.

The progression of the disease can be categorized in four different stages. The first stage is known as Mild Cognitive Impairment (MCI), and corresponds to a variety of symptoms — most commonly memory loss — which do not significantly alter daily life. Between 6 and 25% of people affected with MCI progress to AD every year. The next stages of Alzheimer’s disease (Mild and Moderate AD) are characterized by increasing cognitive deficits, and decreasing independence, culminating in the patient’s complete dependence on caregivers and a complete deterioration of personality (Severe AD) .

Diagnosis of MCI and AD is important for several reasons such as: A positive diagnostic gives the patient and his family time to inform them- selves about the disease, to make life and financial decisions related to the disease, and to plan for the future needs and care of the patients. A negative diagnostic may ease anxiety over memory loss associated with aging. It also allows for early treatments of reversible conditions with similar symptoms (such as thyroidal problems, depression, and nutrition or medication problems).Current symptoms-delaying medications have a given time frame during which they are effective. Early diagnosis of AD helps ensure prescription of these medications when they are most useful. Early diagnosis of AD also allows prompt treatment of psychiatric symptoms such as depression or psychosis, and as such reduces the personal and societal costs of the disease. As research progresses, preventive therapies may be developed. Early diagnosis raises the chance of treating the disease at a nascent stage, before the patient suffers permanent brain damage. Finally, as institutionalization accounts for a large part of health care costs incurred because of AD, by preserving patients' independence longer and preparing families for the needs of AD patients, timely diagnosis further decreases the societal costof the disease. Medical diagnosis of Alzheimer's disease is hard, and symptoms are often dismissed as normal consequences of aging. Diagnosis is usually performed through a combination of extensive testing and eliminations of other possible causes.

Psychological tests such as Mini Mental State Examinations (MMSE), blood tests, spinal fluid, neurological examination, and increasingly, imaging techniques are used to help diagnose the disease databases because of their high computational cost. However, their use is often limited to numeric data.

A traditional method of identifying Alzheimer is visual analysis of the EEG recordings by the trained professionals. It is very costly as well as tedious task to review a 24-h continuous EEG recording, particularly if there are large number of EEG channels to review. Moreover, the detection of Alzheimer by visual scanning of a patient's EEG data is a tedious and time consuming process. In addition, to detect Alzheimer's an expert is required to analyse the entire length of the EEG recordings [5]. Automatic detection is preferred, as complete visual analysis of EEG signal is very difficult. A reliable automatic classification and detection system will ensure that the objectives are met and will facilitate treatment and significantly improve the diagnosis of Alzheimer.

Automating the detection of Alzheimer is very important for assisting neurologists to analyse the EEG recordings. It could also offer solutions for closed-loop therapeutic devices such as implantable electrical stimulation systems. The long-term treatment with anti-Alzheimer drugs, may cause cognitive or other neurological side effects, this could be reduced to a targeted short-acting intervention. Therefore, the development of such automated systems is highly demandable, due to the huge amounts and the increased usage of long-term EEG recordings for proper evaluation and treatment of neurological diseases, including Alzheimer. The failure due to the expert misreading the data and lack of making proper decision would also be narrowed down. There are five stages in the automated diagnosis of Alzheimer such as, signal acquisition, pre-processing, feature extraction, and classification.

The objective of the proposed system is to develop a new method for automatic detection and classification of EEG patterns into three categories MD or Alzheimer's using a Block Based Neural Network (BBNN) and wavelet feature extraction method. The BBNN parameters are optimized using Particle Swarm Optimization (PSO) algorithm.

II. LITERATURE SURVEY

Many automatic Alzheimer detection techniques using machine learning approach are been in UE since 1980s. Among these techniques the areas listed below are prominent areas of research undergoing presently in order to provide more sophisticated and better results.

The EEG was developed as a manse to investigate the mental processes. The brains electrical activity was first recorded and reported by Caton in 1875 in exposed brains of rabbits and monkeys. In 1929, the first measurement of brain electrical activity in humans was reported by Hans Berge. Since then, the EEG signal has been utilized clinically to evaluate neuron's behaviour and functional states of the brain such as (a) different stages of wakefulness (b) sleep or (c) metabolic disturbances.

EEG is an electrophysiological monitoring method which is used to record electrical activity of the brain. EEG measures voltage fluctuations which is caused due to ionic current within the neurons of the brain. EEG refers to the recording of the brain's spontaneous electrical activity over a period of time,[1] as recorded from multiple electrodes placed on the scalp. EEG is most often used to diagnose epilepsy, which causes abnormalities in EEG readings. It can also be used to diagnose sleep disorders, coma, encephalopathy, and brain death. It is measured with electrodes which are placed on standard positions on the head. In clinical and research applications, the name and location of these electrodes are specified by the international standard 10/20 system. This system relies on the relationship between the location of an electrode and the underlying area of cerebral cortex. The '10' and '20' represents the actual distances between adjacent electrodes which is either 10% or 20% of the total front-back or right-left distance of the skull. The range of amplitude for an EEG signal is 5 to 200 µV and frequency ranges from 1 to 30 Hz.

In 2009, Martis*et.al* proposed a novel feature extraction scheme on electrocardiogram (ECG) signal using discrete wavelet transform (DWT)In this research, the discrete wavelet transform (DWT) based on dyadic (powers of 2) scales and positions, was used to make the algorithm computationally very efficient without compromising accuracy. The EEG was subjected to a level 4 decomposition using fourth-order Daubechies wavelet transform.

A data selection algorithm depending on phase congruency to determine Alzheimer from the background EEG was proposed by Logesparan& Rodriguez-Villegas in 2011. The phase congruency denoising was performed by dynamic estimation and compensation for muscle activity in EEG. The approach involved the modification of traditional phase congruency to include the dynamic estimate of muscle activity in the input scalp EEG signal. The authors report that the performance of the data selection algorithm was enhanced to 80% sensitivity for more than 50% data reduction.

Yusof*et.al* proposed a new mutation operation for rapid feature selection by GA. The fittest chromosomes were preserved by normal elitism in GA, which were then evaluated by utilizing the fitness function. The highest fit allele was conserved and the evaluation of fitness of the allele performed based on the frequency of the occurrences. The chromosome that underwent this mutation approach was found to have the highest fitness as it was created based on the fittest alleles.

Akhtar*et.al* [19] have proposed a framework based on Independent Component Analysis (ICA) and Wavelet Denoising to enhance the pre-processing of EEG signals. The Spatially Constrained ICA (SCICA) was used for extracting artifact-only Independent Components (ICs) from the EEG data. The cerebral activity from the acquired artifacts ICs was removed by using WD. The subtractions of the artifacts from the EEG signals were then performed to get clean EEG data. The main benefit of the approach was reported to be the speedy computation as there is no need for identifying all ICs. The approach also achieves effective removal of focal artifacts which can be well separated by SCICA.

In 2007, Lucia *et.al* proposed a novel feature extraction scheme for automated classification of Alzheimer in the human electroence
phalogram-based on principal component analysis.
In 2009, Martis*et.al* proposed a novel feature extraction scheme on electrocardiogram (ECG) signal using discrete wavelet transform (DWT) coefficients followed by PCA. In the method, the principal components of the sub bands of discrete wavelet transformed signal in the compact supported basis space represent the data better than in the time domain. The method provided better results.

In 2012, Acharya*et.al* proposed the use of principal component analysis for automatic classification of Alzheimer EEG activities in wavelet framework. PCA was used for feature selection and it yielded 97% classification accuracy.

The objective of classification is to describe a boundary between the classes and to label them based on their measured features. Guo*et.al* [20] have presented a novel method of automatic Alzheimer detection. This new approach used entropy features derived from Multi Wavelet Transform (MWT), which was then combined with an Artificial Neural network to classify the EEG signals about the presence of Alzheimer. The authors reported that Multi-wavelets achieved better results than scalar wavelets. However, they also place on record that Multi-wavelets produce more number of sub-signals which increase the computational cost. The raw EEG data are decomposed into sub-signals through Multi Wavelet Transform. The formation of feature vector is performed by extracting Approximate Entropy (ApEn) for each sub-signal.

Moon *et.al* in 2001, proposed a novel block based neural network (BBNN) model and the optimization of its structure and weights based on genetic algorithm. The optimized BBNN could solve the engineering problems such as pattern classification and mobile robot control.

In 2007, Jiang *et.al* proposed a novel BBNN model with block wise least squares learning algorithm (BLS). The optimal internal weights of BBNN are found using this BLS algorithm. The method improved the convergence speed with orders of magnitude.

Later in early 2016, BBNN model along with particle swarm optimization algorithm used for optimizing the internal structure of BBNN was proposed by Shadmand*et.al* for classification of Personalized ECG signals. The performance evaluation results show a high classification accuracy of 97%.

III. PROPOSED SYSTEM

A novel technique for the automated analysis of EEG signal is proposed and investigated for classifying the signal as normal or Alzheimer. This method is based on classification of EEG signals using BBNN in which the features are extracted using wavelet transform method. The parameters of neural network are optimized using the Particle Swarm Optimization (PSO) algorithm. This method is proposed in order to eliminate the difficulties and limitations involved in using visual inspection for analyzing the EEG signal.

The proposed method is automatic. Hence it is not subjective and thereby eliminates the need for the visual inspection based method which is subjective. Moreover, the performance of the proposed method is better as compared to the existing visual inspection based method of EEG signal classification. Figure 1 shows the basic architecture of the proposed method.

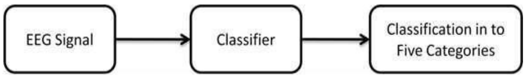


Figure 1. Basic Architecture

An EEG signal is first analyzed and fed to the classifier. The input signal received by the classifier, uses it for classifying depending on the input signals received during the training phase. Proposed system uses a novel classification method which uses the BBNN and the parameters are optimized using PSO.

The proposed system consist of five stages and they are – Signal Acquisition, Pre-processing, Feature Extraction, Feature Selection and Classification. The proposed architecture is as shown in Figure 2. It shows the relationship between each phase with its predecessor phases.

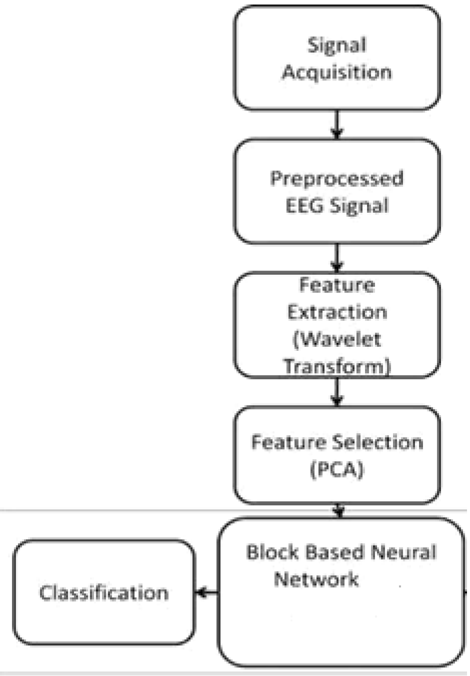


Figure 2. Proposed system architecture

The work addresses the problem of classifying EEG signal as either normal or Alzheimer’s. In the Signal Acquisition phase, only two subsets (set A and set E) from the dataset which was available online are made use of. In Signal Acquisition phase, the EEG segments are filtered by a low pass filter of cut off frequency 40 Hz and a stop band frequency of 50Hz to remove the artifacts from them Both these phases together constitutes the input module.

After obtaining artifact free signals in the preprocessing phase, necessary features are then extracted from EEG signals using Wavelet Transform. The EEG signals thus obtained undergo wavelet decomposition with five scales using db4 wavelet. Among all the wavelets forms db4 wavelet is selected because of its smoothing feature that makes it appropriate for detecting changes in EEG signals. With the use of feature Selection the dimensions of feature vectors are reduced. In the proposed work, the significant sub-bands are subjected to PCA for feature selection. Two principal components are considered from each of these sub-bands and are taken as effective features. These two phases thus together constitute the feature processing module.

The next phase is the Classification phase. For this a BBNN is made use. The features thus selected are then sent to a BBNN for classification. In this work, BBNN is used as a multi class problem. Half the data from each set are selected for training the BBNN while the rest are used for testing. The parameters of BBNN are optimized using a PSO algorithm. The output module is obtained at the classification phase.

The test performance of the proposed method is evaluated. And it is defined **by** four performance parameters such as: sensitivity, specificity, classification accuracy and positive predictive value. Sensitivity defines the ratio of the number of correctly detected positive patterns to the total number of actual positive patterns. A positive pattern indicates that Alzheimer is detected. Specificity is specified as the ratio of the number of correctly detected negative patterns to the total number of actual negative patterns. A negative pattern indicates a detected non Alzheimer's. The total classification accuracy is defined as the ratio of number of correctly classified patterns to the total number of patterns. Positive predictive is given as the ratio of the number of correctly detected positive patterns to the sum of true positives and false positives.

IV. CONCLUSION

Accurate automated Alzheimer detection remains an important challenge and a critical first step in removing the uncertainty associated with when Alzheimer will occur and furthering the understanding of Alzheimer and its causes. In the proposed method, the EEG signals have been classified in five classes. A classification system for this purpose makes use of a Block Based Neural Network (BBNN) with Particle Swarm Optimization (PSO) for optimization of parameters. BBNN addresses a multi class problem. BBNN has been trained with a training dataset and tested using a test data set. Each of these data sets contains a number of EEG records of data set. The features which are extracted from EEG signals have been used as BBNN inputs. Particle Swarm Optimization method is used as BBNN training algorithm and also to optimize the BBNN structure and the weights. The BBNN trained with PSO algorithm presents a high quality system for EEG signals classification. In the proposed method, signal preprocessing signal acquisition phase, preprocessing phase, feature processing which consisting of the feature extraction and feature selection phases and classification module consisting classification of EEG signals using BBNN and the parameters are optimized using PSO algorithm.

V. REFERENCES

[1] H. Berger, "On the Electroencephalogram of Man", Electroencephalography and Clinical Neurophysiology Suppl., vol 28, pp.37-73, 1969.

[2] H. Adeli, S. Ghosh-Dastidar, and N. Dadmehr, "A wavelet-chaos methodology for analysis of EEGs and EEG sub-bands to detect Alzheimer and epilepsy", IEEE Trans. Biomed. Eng., vol. 54, no. 2, pp. 205-211, Feb. 2007.

[3] H. A. Jasper, "The ten-twenty system of the International Federation", Electroencepholography and Clinical Neurophysiology, vol. 10, pp. 371-375, 1958.

[4] M. Steriade, D. A. McCormick, and T. J. Sejnowski, "Thalamocortical oscillations in the sleeping and aroused brain", Science, vol. 262, pp.679-685, 1993.

[5] J. B. Ochoa, "EEG Signal Classification for Brain Computer Interface Applications ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE", [Online] Available: http:nndsp-book.narod.ru/WVT/BZ.pdf

[6] S. Sanei, and J. A. Chambers, "Brain Rhythms", in EEG Signal Processing. New York: Wiley, 2007, pp. 10-12

[7] Lee, B.; Tarng, Y. S. "Application of the discrete wavelet transform to the monitoring of tool failure in end milling using the spindle motor current". International Journal of Advanced Manufacturing Technology, 1999, pp. 238–243

[8] E. Baar, M.Schrmann, C. Baar-Eroglu, and S. Karakas, “Alpha oscillations in brain functioning: an integrative theory", International Journal of Psychophysiology, 26(1-3), pp.5-29, 1997.

[9] Yinxia Liu, Weidong Zhou, Qi Yuan, and Shuangshuang Chen, “Automatic Alzheimer Detection Using Wavelet Transform and SVM in Long-Term Intracranial EEG”, IEEE Transactions On Neural Systems And Rehabilitation Engineering, Vol. 20, No. 6, November 2012.

[10] Sang Woo Moon and SeongGon Kong, “Block Based Neural Networks”, IEEE Transactions On Neural Networks, Vol.12, No.2, March 2001, pp 307 – 317.

[11] Wei Jiang and SeongGon Kong, “A Least Squares Learning for Block Based Neural Networks”, Advances in Neural Netwoks, Vol.14 (SI), 2007, pp 242 –247.

[12] A. Varsavsky, I., Mareels and M. Cook, Alzheimer and the EEG, Boca Raton: CRC Press, 2011.

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section:Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering,
Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

Recent Review of Reversible Data Hiding in Steganography

Rasmi. A*[1], Dr. Mohanapriya M[2]

*[1]Research Scholar, Department of Computer Science and Engineering, Karpagam University ,Karpagam Academy of Higher Education Tamilnadu, India

[2]Associate Professor, Department of Computer Science and Engineering & IT, Coimbatore Institute of Technology, Coimbatore, Tamilnadu, India

ABSTRACT

In the electronic era because of the wide spread usage of network technology, security measures play, a vital role for data transmission between sender and receiver. Steganography is the science of masking data bits in a secure manner using cover file,for various purposes. This paper explores and analyses some of the existing steganographic reversible data hiding methods from its earliest instances through potential future application.

Keywords: Steganography, Reversible data hiding, Spatial domain, Embedding capacity ,data hiding, stego,cover image

I. INTRODUCTION

Steganography is an art and science of concealed communication in a statistically unnoticeable way. It was first used and experienced by Greek people for exchanging secret data between sender and receiver for various purposes. In this type of data hiding the sender hides the secret information to be sent into the digital file where only the intended user can recover it .The word steganography is originated from the greek words “steganos ” and “graphia” means covered writing. Different application areas of steganography are :confidential communication and secret information storing, protection of data alteration ,transport highly private records between International Governments, and terrorists can use this to keep their communication secret and coordinate attack .[1,2,3]. It can be used in military for secret communication purpose. Various data hiding methods are watermarking, steganography and cryptography. The features of the embedding techniques can be defined by features like capacity, robustness and imperceptibility. The vital requirement for steganography is its undetectability and robustness. The main elements of a steganographic system are cover image, embedding algorithm, message, extraction algorithm and stego image .A general form of steganography is shown in figure .1.

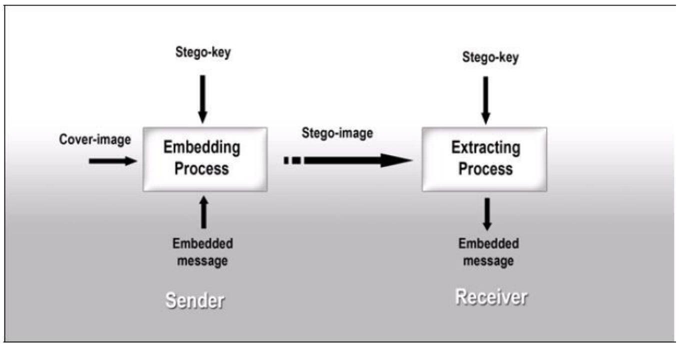


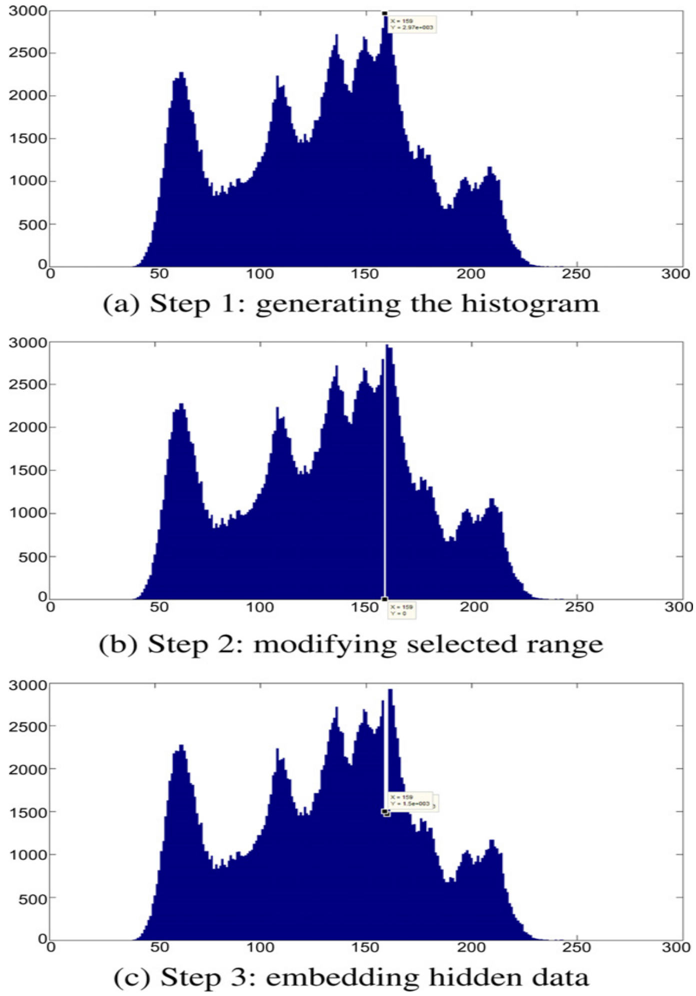
Figure (1) a general form of steganography

Cover file may be image, text, video or audio, which is used as a carrier in the data transfer. The secret data that the sender wants to be sent, is the information, it may be in different format as the carrier file. The secret information is also termed as payload [4,5].The insertion of payload into cover file gives, the resultant stego image. Steganographic techniques can be classified into two types and it named as spatial domain and transform domain, if manipulations are done at the pixel intensity based values, then it comes under spatial domain, while if it is based on the coefficient of pixel value it falls under frequency domain techniques. In spatial domain secret data is embedded directly whereas in transform embedding is in indirect way. Based on the nature of cover file used steganography can be termed as image steganography, text steganography, audio steganography, video steganography and network steganography.[6,7]

II. REVIEW OF LITERATURE

Information hiding can be graded into reversible or lossless data hiding and irreversible data hiding schemes. Most of the commonly used data embedding methods are not reversible based techniques .Reversible data hiding helps to retrieve the exact input data at the extraction phase without altering the content. Reversible data hiding is also known as lossless data hiding because it restores the cover content perfectly .It provides a perfect balance between the image quality and the payload. The major challenges in data embedding are associated with how efficiently inserting can be implemented in a cover file ,without affecting the visual properties .So before hiding data certain parameters has to be checked such as the size of secret data should never exceed the size of cover file ,if it is exceeding it can be easily detectable by the intruder. Embedding capacity means the number of bits can be embedded in a cover file without altering the structure of it .The quality of the stego image can be evaluated by using the peak signal to noise ratio (PSNR) which denotes the maximum power of a signal and the power of the corrupting noise. It is expressed in the logarithmic unit dB. The regular singular scheme first segments an image into non overlapping parts, then divide the parts into 3 sections named as regular(R),singular(S),and unusable(U),and by using some operations we can convert R to S and S to R ,and can insert the secret data into it but U remains as it is, without any change.[8,9,10]

Different types of existing reversible data hiding methods are regular singular scheme, the integer wavelet transform based method, histogram modification scheme and difference expansion scheme. In difference expansion technique the image is divided into pair of pixels, after that inserts one bit of data into each pair. Integer wavelet transform uses the least significant bit technique for high frequency integer wavelet coefficient by selecting a proper threshold value. Histogram based method gives the complete tonal representation of the image, which employs the redundancy information of the cover image to hide the secret message. It mainly depends upon peak value and zero point, then after selecting the peak point ,we are simply incrementing or decrementing 1 in all pixel values which are lesser or greater than peak point value. It prevents the overflow and underflow by applying the modulo addition. In this one it takes the minimum or zero point values of histogram then varying it slightly to insert the information bits .Figure (2) shows histogram modification of Lena image. In this the number of bits can be inserted into an image depends on the peak value of the histogram shifting. The most commonly used image formats are internet are Graphics interchange format (GIF) and Joint photographic experts group (JPEG), because of its better hiding power. Security of data embedding could be improved by using certain factors like proper selection of cover image, reducing the payload distortion and improving the message embedding capacity.



(a) Step 1: generating the histogram

(b) Step 2: modifying selected range

(c) Step 3: embedding hidden data

Figure (2) shows histogram modification technique

Reversible data hiding employs techniques like compression decompression, encryption decryption ,and information embedding and extraction. In prediction based steganography embedding can be done by the technique of predictive coding approach .In this pixel intensity values are predicted using predictor , and prediction error values (EV) are changed to insert secret message.[11,12,13]

III. CONCLUSION

This paper reviews image steganography and reversible data hiding in an effective way and discussing different parameters to increase the performance of the data embedding techniques. It applies different compression and encoding mechanisms to improve the visual quality and efficiency of the reversible data hiding methods.

IV. REFERENCES

[1] Z. Zhao, H. Luo, Z.-M. Lu, J.-S. Pan, Reversible data hiding based on multilevel histogram modification and sequential recovery, Int. J. Electron. Commun. 65 (2011) 814–826.B.

[2] C.-C. Lin, W.-L. Tai, C.-C. Chang, Multilevel reversible data hiding based on histogram modification of difference images, Pattern Recognit. 41 (2008) 3582–3591.

[3] B. Cong, N. Sang, M. Yoon, H.-K. Lee, Multi bit plane image steganography, in: International Workshop on Digitalforensics and Watermarking, vol. 4283 of Lecture Notes in Computer Science, pp. 61–70.

[4] E. Kawaguchi, R.O. Eason, Principle and applications of BPCS steganography, in: Proc. of Multi-media Systems and Applications, in: SPIE, vol. 3528, 1998, pp. 464–473.

[5] V.M. Potdar, E. Chang, Gray level modification steganography for secret communication, in: Proc. of 2nd IEEE International Conference on Industrial Informatics, pp. 223–228.

[6] D Tseng, Y.C. , Chen Y.Y. Pan H.K.:'A secure data hiding scheme for binary images', IEEE Trans. Commun., 2002, 50,pp. 1227-1231 .

[7] Pawan R Sharma, Jitendra Mishra“ A Comprehensive Survey on Data Hiding Technique” IRJET e-ISSN: 2395 -0056 Volume: 02 Issue: 04 July-2015.

[8] Gurpreet Kaur, Kamaljeet Kaur “Digital Watermarking and Other Data Hiding Techniques” IJITEE ISSN: 2278-3075, Volume-2, Issue-5, April 2013 ,181.

[9] Provos, N. &Honeyman, P., “Hide and Seek: An introduction to steganography”, IEEE Security and Privacy Journal, 2003.

[10] Y.K. Lee, L.H. Chen, “High capacity image steganographic model”, IEEE Proceedings on Vision, Image and Signal processing, Vol. 147, No.3,pp. 288-294, 2000.

[11] X. Liao, Q. Wen and J. Zhang, ”A steganographic method for digital images with four-pixel differencing and modified LSB substitution”, Journal of Visual Communication and Image Representation, vol 22, no 1, pp. 18, 2011 .

[12] A. Rashid and M. K. R. Rashid, “Stego-Scheme for Secret Communication in Grayscale and Color Images”, British Journal of Mathematics and Computer Sciences, vol. 10, no, 1 (2015), pp. 1-9.

[13] Sandeep Kaur ,Arunjot Kaur &Kulwinder Singh” A Survey of Image Steganography” IJRECE, Volume 2-Issue 3 June 2014, e-ISSN 2321-3159 p-ISSN 2321-3159.

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering, Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

An Efficient Watermarking Technique Using Genetic Algorithm for Relational Data

Sreesha K S[1], Jincy Easow[2], Prof. Jisha P Abraham[3]

[1,2,3] Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India

ABSTRACT

Watermarking is advocated to enforce ownership rights over shared relational data and for providing a means for tackling data tampering. When ownership rights are enforced using watermarking, the underlying data undergoes certain modifications; as a result the data quality gets compromised. Reversible watermarking is employed to ensure data quality along with data recovery. Reversible watermarking tries to overcome the problem of data quality degradation by allowing recovery of original data along with the embedded watermark information. An optimal watermark value is created through the Genetic Algorithm (GA) and inserted into the selected feature of the relational database. It mainly comprises a data preprocessing phase, watermark encoding phase, attacker channel, watermark decoding phase and data recovery phase. In data preprocessing phase, secret parameters are defined and strategies are used to analyze and rank features to watermark. An optimum watermark string is created in this phase by employing GA an optimization scheme that ensures reversibility without data quality loss.

Keywords: Genetic Algorithm(GA)

I. INTRODUCTION

A watermark is a visible overlay of copyright information usually in the form of text or an image logo added to photos or other digital documents. A watermark protects digital intellectual property, such as photos and artwork, from unauthorized use. It identifies the rightful owner of the work, which discourages other people from using it as their own. While it’s hard to prevent people from doing so, having a well-placed watermark containing a copyright symbol, name and URL of the owner can go a long way to discourage this popular,

- Data Preprocessing
- Watermark Encoding
- Watermark Decoding
- Data Recovery

Watermarking technique provides the protection to the database by embedding information. Invisible watermarking, the watermark information such as text or a logo which identifies the owner of the media, that is visible in the picture or video. Here the objective is to attach ownership or other descriptive information to the signal in a way that is difficult to remove. It is also possible to use hidden embedded information as a means of covert communication between individual. The watermarking technique which are irreversible may causes modification of data undergoes through it at the certain extent. This method tries to overcome the problem of data quality degradation by allowing recovery of original data along with the embedded watermark information. Watermarking scheme modify original database as a modulation of the watermark information, and causes impossible to prevent permanent distortion to the original database, and affects to meet the integrity requirement of many application.

II. LITERATURE SURVEY

Several watermarking techniques are proposed which includes watermarking for images, databases, audio and video. The watermarking is primarily developed for the images, the research in audio is started later. There are less watermarking techniques are proposed for audio compared to the images/video. Embedding the data in audio is difficult compared to the images because the Human auditory system is more sensitive than the Human visual System. In last ten years there is a lot of advancement in watermarking few of them are discussed here. This chapter reviews the literature of information hiding in sequences. Scientific publications included into the literature survey have been chosen in order to build a sufficient background that would help out in identifying and solving the research problems. During the last decade watermarking schemes have been applied widely. These schemes are sophisticated very much in terms of robustness and imperceptibility. Robustness and imperceptibility are important requirements of watermarking, while they conflict each other.

Non-blind watermarking schemes are theoretically interesting, but not so useful in practical use, since it requires double storage capacity and communication bandwidth for watermark detection. Of course, non-blind schemes may be useful for copyright verification mechanism in a copyright dispute. On the other hand, blind watermarking schemes can detect and extract watermarks without use of the unwatermarked. Therefore it requires only a half storage capacity and half bandwidth compared with the non-blind watermarking scheme. The relational data defers from multimedia data in many respects: (i) Few Redundant Data: Multimedia objects consists of large number of bits providing large cover to hide watermark, whereas the database object is a collection of independent objects, called tuples. The watermark has to be embedded into these tuples, (ii) Out-of-Order Relational Data: The relative spatial/temporal positions of different parts or components in multimedia objects do not change, whereas there is no ordering among the tuples in database relations as the collection of tuples is considered as set, (iii) Frequent Updating: Any portion of multimedia objects is not dropped or replaced normally, whereas tuples may be inserted, deleted, or updated during normal database operations, (iv) There are many psychophysical phenomena based on human visual system and human auditory system which can be exploited for mark embedding. However, one cannot exploit such phenomena in case of relational databases.

III. PROPOSED SYSTEM

Advancement in information technology is playing an increasing role in the use of information systems comprising relational databases. A robust and semi-blind reversible watermarking technique for numerical relational data as well as the string data has been proposed that addresses the above objectives. This project proposes one such reversible watermarking technique that keeps the data useful for knowledge discovery. GA - an optimization algorithm is employed in this proposed project to achieve an optimal solution that is feasible for the problem at hand and does not violate the defined constraints. The proposed system consist of four stages: data preprocessing, watermark encoding, watermark decoding and data recovery. The watermark preprocessing phase computes different parameters for calculation of an optimal watermark. These parameters are used for watermark encoding and decoding. The main focus of watermark encoding phase is to embed watermark information in such a way that it does not affect the data quality.

During watermark embedding, data gets modified according to the available bandwidth (or capacity) of the watermark information. The bandwidth of the watermark should be sufficiently large to ensure robustness but not so large that it destroys the data quality. The data owner decides the amount of data modification such that the quality is not compromised for a particular database application before-hand and therefore defines usability constraints to introduce tolerable distortion into the data. After watermarking, the data is released to the intended recipients over a communication channel that is assumed to be insecure and termed as the “attacker channel” in this research domain. The data may undergo several malicious attacks in the attacker channel. The efficiency and effectiveness is described through robustness analysis determined by its response to subset insertion, alteration and deletion attacks. The Watermark decoding phase recovers watermark information effectively for detection of the embedded watermark. Data recovery phase mainly comprises the important task of successful recovery of the original data.

A. Preprocessing Phase

In the preprocessing phase, two important tasks are accomplished:
1) Selection of a suitable feature for watermark embedding
2) Calculation of an optimal watermark with the help of an optimization technique.

Feature Analysis and Selection

For developing a decisive information model of various features of the dataset, all the features are ranked according to their importance in information extraction, subject to their mutual dependence on other features. For this purpose, mutual information (MI), is exploited, that is an important statistical measure for computation of mutual dependence of two random variables. Mutual information of every feature with all other features is calculated by

$$\text{MI (A, B)} = \Sigma a \Sigma PAB(a,b) log \frac{PAB(a,b)}{PA(a)PB(b)}$$

Where MI(A,B) measures the degree of correlation of features by measuring the marginal probability distributions as PA(a), PB(b) and the joint probability distribution PAB(a, b). Then MI of one feature with all other features is computed using the relation. The value of MI of each feature is then used to rank the features. The attacker can try and predict the feature with the lowest MI in an attempt to guess which feature has been watermarked. To deceive the attacker for this particular scenario, a secret threshold can be used for selecting the feature for watermark embedding. In this context, the data owner can define a secret threshold based on MI of all the features in the database. The feature(s) having MI lower than that threshold can be selected for watermarking. The attacker will not attack the features having large MI as in that case the usability of the data will be compromised. Therefore, he will be forced to attack the feature(s) with lower MI without concrete knowledge (due to the use of secret threshold) of which features have been watermarked.

Watermark Creation Using Genetic Algorithm

For the creation of optimal watermark information, that needs to be embedded in the original data, use an evolutionary technique GA. GA is a population-based computational model, basically inspired from genetic evolution. GA evolves a potential solution to an optimization problem by searching the possible solution space. In the search of optimal solution, the GA follows an iterative mechanism to evolve a population of chromosomes. The GA preserves essential information through the application of basic genetic operations to these chromosomes that include: selection, crossover, mutation and replacement. The GA evaluates the quality of each candidate chromosome by employing a fitness function. The evolutionary mechanism of the GA continues through a number of generations, until some termination criteria is met. During watermark creation phase, we employed the following major steps of the GA for getting optimal watermark information:

1) Initial random population of binary strings called chromosomes is generated. Gene values of each chromosome represents l-bit watermark string.

2) Fitness of each chromosome is evaluated by employing a constrained optimized fitness function

3) Tournament selection mechanism is applied to get the most appropriate individuals as parent chromosomes.

4) Genetic operations of crossover and mutation are performed on parent chromosomes to create off-springs. A single point crossover operator is applied to evolve high quality individuals, inheriting parental characteristics, by exchanging information between two or more chromosomes. A uniform mutation operator is applied to bring diversity in population through small random changes in gene values of binary chromosomes. The values of crossover fraction and mutation rate are set empirically

5) Elitism strategy is applied to hire two individuals with best fitness value; as elites to the next generation without genetic changes.

6) Remaining population of the next generation is created by replacing less fit individuals of the previous generation with the most fit newly created off-springs.

7) Steps 2 to 6 are repeated until MIO and MIW reach approximately equal values for a certain number of generations.

8) Both, optimal watermark information string and best fitness value (b) is returned after the fulfillment of the termination criteria.

B. Watermark Encoding Phase

Watermark information calculation is formulated as a CO problem to meet the data quality constraint of the data owner. A GA is used to create optimal watermark information that includes:

1) Optimal chromosomal string (watermark string of length l)

2) b value, b is a parameter that is computed using GA and represents a tolerable amount of change to embed in the feature values.

Once the optimum value of b for each candidate feature A is found, it is saved for use during watermark encoding and decoding. A watermark (bit string) of length l and an optimum value b is used to manipulate the data provided it satisfies the usability constraints. The value b is added into every tuple of the selected feature A when a given bit is 0; otherwise, its value is subtracted from the value of the feature. It is ensured that the mutual information of a feature remains unchanged, when the watermark is inserted into the database. The watermark is inserted into every tuple for the selected feature of the dataset. The data owner can select any number of features for watermark embedding based upon a secret threshold and MI of the feature(s). After finding the optimum value of b, a parameter nr is calculated, that represents the percentage change in the watermark encoding. This parameter is calculated for a tuple r as:

$$nr = Dr * \ell$$

The parameters used in the above equation are nr, Dr, and ℓ where nr is the detected amount of percentage change in encoding, Dr is the recovered data and ℓ is the length of the watermark. Since the length of the watermark is l; nr is calculated and b is inserted l times in the database. The length ℓ of the watermark should be carefully chosen. If it is too small, it will make the watermark fragile against attacks, and if it is too large, it might compromise the data quality because the data gets altered for every bit of the watermark. In this project, the data gets altered for each watermark bit in every tuple. After a number of empirical studies, a length of 16 bits was selected. The watermark encoding algorithm starts the embedding process with the most significant bit MSB of the watermark. For this purpose the algorithm works with one tuple at a time. If the MSB of the watermark is 1, the new value of Dr, denoted by Dwr is calculated using

$$Dwr = Dr - \beta$$

The parameters used here is Dwr, Dr, and β, where Dwr is the original data to be watermarked, Dr is the recovered data and β is an optimized value, here used is 0.16. In order to embed the second MSB of the watermark, the algorithm is again employed using the same procedure, but the updated value Dr of the feature (that has now become Dwr) is used for calculating new values of nr and Dwr. If the algorithm encounters a watermark bit that is 0 then the new value of Dwr, is calculated using

$$Dwr = Dr + \beta$$

C. Watermark Decoding Phase

In the watermark decoding process, the first step is to locate the features which have been marked. The process of optimization through GA is not required during this phase. We use a watermark decoder z, which calculates the amount of change in the value of a feature that does not affect its data quality. The watermark decoder decodes the watermark by working with one bit at a time. In the decoding phase, ndr is calculated and represents the percent change detected in the watermarked data. The value of ndr, nr and nΔr is calculated using the values of tuple r and therefore might be different for every r. The parameter nΔr is computed by calculating the difference between the original data change amount nr and the watermark detected change amount ndr using

$$ndr = Dw *' \ell$$

$$n\Delta r = ndr - nr$$

The decoding phase mainly consists of two steps:

Step 1: For every candidate feature A of all the tuples in D'W, the watermark bits are detected starting from the least significant bit and moving towards the most significant bit. The bits are detected in the reverse order compared with the bits encoding order because it is easy to detect the effect of the last encoded bit of the watermark. This process is carried out using the change matrix nr.

Step 2: The bits are then decoded according to the percentage change values of watermarked data. If nΔr<= 0, the detected watermark bit will be 1. If nΔr> 0 and nΔr < 1, the detected watermark bit will be 0.
The final watermark information is retrieved through a majority voting scheme using
wD< = mode(dt w(1,2,…,l))

D. Data Recovery Phase

After detecting the watermark string, some post processing steps are carried out for error correction and data recovery. The optimized value of b computed through the GA is used for regeneration of original data. The value of a numeric feature is recovered using Dr.

$$Dr=D'wr+\beta$$
$$Dr = D'wr - \beta$$

IV. CONCLUSION

Watermarking is used to enforce ownership rights over shared relational data and for providing a means for tackling data tampering. Reversible watermarking techniques are used to cater to such scenarios because they are able to recover original data from watermarked data and ensure data quality to some extent. A novel robust and reversible technique for watermarking numerical data of relational database is presented. In this project, we considered the Cleveland Heart Disease dataset, in which data includes the ID, age, type, bp, sex, cholesterol, month, right, left, etc. Firstly calculate the mutual information (MI) values between the candidate attribute and the remaining attributes. Identify the attributes that need to be watermarked, those MI values is less than the threshold value selected by the user. Then apply the genetic algorithm to generate the watermarked information that are embedded into the original selected attribute values for watermarking. Thus the encoding phase is completed. For the decoding phase, the reverse operations are performed to separate the original data and the watermarked information. These techniques are not robust against malicious attacks, particularly this techniques that target some selected tuples for watermarking. One of the future concerns of this project is to watermark the shared databases in distributed environments where different members can share their data in different proportions and also extended for non-numeric data stores.

V. REFERENCES

[1]. R. Agrawal and J. Kiernan, “Watermarking relational databases,” in Proc. 28th Int. Conf. Very Large Data Bases, 2002.

[2]. Y. Zhang, B. Yang, and X. M. Niu, “Reversible watermarking for relational database authentication”, Vol.17, 2006.

[3]. Saman Iftikhar, M. Kamran, and Zahid Anwar, “A Robust and Reversible Watermarking Technique for Relational Data”, Vol. 27, No. 4, April 2015.

[4]. X. Li, B. Yang, and T. Zeng, “Efficient reversible watermarking based on adaptive prediction-error expansion and pixel selection,” IEEE Trans. Image Process., vol. 20, Dec. 2011.

[5]. Sonnleitner, “A robust watermarking approach for large databases,” in Proc. IEEE First AESS Eur. Conf. Satellite Telecommun., 2012.

[6]. G. Gupta and J. Pieprzyk, “Database relation watermarking resilient against secondary watermarking attacks,” in Information Systems and Security. New York, NY, USA: Springer, 2009.

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering,
Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

Performance Evaluation of MongoDB and CouchDB Databases

Shyama M Nair[1], Rinu Roy[2], Dr Surekha Mariam Varghese[3]

[1]Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India
[2]Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India
[3]Professor and HOD, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India

ABSTRACT

The rise of new type of databases have seen over the last few years, known as NoSQL databases. NoSQL databases are challenging the dominance of relational databases. Scaling a relational databases on powerful servers are expensive and difficult to handle and Nosql databases are designed to expand horizontally. Also Nosql database is schema less and data can be inserted without any predefined schema. They represent a broad category of databases which allow large quantities of unstructured and semi structured data to be stored and managed. Additionally they are designed to handle high levels of reads and writes while scaling horizontally. There are various open sourced and licenced document oriented NoSQL databases , but all have different mechanism to store data in document format. However it is extremely diligent to decide which is to be used and when. So there is need for performance evaluation of various document oriented databases. This work comprises about a detailed comparative study between MongoDB and CouchDB, two leading document oriented databases.

Keywords: NoSQL, MongoDB, CouchDB , RDBMS.

I. INTRODUCTION

NoSQL, an abbreviation of ‘not only sql’ describes a wide variety of database technologies came into exist in order to overcome the shortcomings of RDBMS and the demands of modern software development. A NoSQL database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations used in relational databases. They provide various advantages over traditional relational databases.Schema agnostic,scalability, performance and high availability are a few features of NoSQL databases.Nosql databases can be broadly classified into four types.They are key-value store, document store, column store and graph based. Among them a document oriented database is designed for storing, retrieving and managing document-oriented data and and there is an additional level of key-value indexing that allows much more efficient queries. The central concept of document-oriented database is the notion of a document.

Choosing right database for right application plays a very critical role as it is the factor that constitutes the platform for analyzing the performance of the application that is under consideration.Some applications might need consistency and availability,some might need availability and partition tolerance and so on. This leads to a complicated situation to choose one database from the many options that needs a good domain knowledge.There are various metrics that are to be considered for performing comparative performance analysis.The metrics include both the qualitative metrics as well as quantitative metrics. Some of the commonly used qualitative metrics are Persistence, Replication, High Availability, Transactions, Rack-locality awareness, Implementation Language, Influences / sponsors , License type and the quantitative measures include size and performance measurements.

Two leading Nosql document oriented databases - MongoDB and CouchDB are used for performance analysis. The metrics that we have taken into consideration are quantitative measures. For comparing the insertion rate (processing time), read / write operations of the MongoDB and CouchDB, a framework written in Javascript, NodeJS, with performance measuring tool Apache JMeter, is used. Some performance evaluation tests have been carried out.Though the database sizes used for the analysis are comparatively smaller, a clear difference in various factors of comparison has been observed.The environment used for conducting these tests was same for both MongoDB and CouchDB.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 introduces proposed work which is followed by section 4 with benchmarking result and analysis. Finally, we conclude and explain our future work.

II. RELATED WORKS

Tudorica, Bogdan George, and Cristian Bucur[2]compares various NoSQL systems. The NoSQL database focused to offer high performance and high availability. Although the SQL and the NoSQL databases are having some shared features, some of their behaviors are not similar in given instances. This paper is trying to comment on the various NoSQL (Not only Structured Query Language) systems and to make a comparison (using multiple criteria) between them. The NoSQL databases were created as a mean to offer high performance and high availability at the price of loosing the ACID (Atomic, Consistent, Isolated, Durable) trait of the traditional databases in exchange with keeping a weaker BASE (Basic Availability, Soft state, Eventual consistency) feature.

Enqing Tang and Yushun Fan[3] compares performance between five NoSQL databases(Redis, MongoDB, CouchBase, Cassandra, HBase) by using a measurement tool - YSCB (Yahoo! Cloud Serving Benchmark)and explain the experimental results by analyzing each database’s data model and mechanism.

Hecht, R., &Jablonski, S. [4]presents a survey on security issues in big data and NoSQL databases and it also evaluate underlying techniques of Nosql databases. Data encryption is lacked in most of the NoSQL databases. To have a more secure database it is essential to encrypt sensitive fields in the database. Due to the high volume, variety and velocity of big data, traditional security models have difficulties in dealing with large scale of data.”Use the right tool for the job" is the propagated ideology of the NoSQL community, because every NoSQL database is specialized on certain use cases.

Leavitt.N suggested that Big data is considered to be large volume of structured and unstructured data. Hence such a large scale of data cannot be effectively managed or exploited using conventional data management tools[5]. To handle this problem, specifically designed alternative database; such as -NoSQL and Search-based systems can be used. The author provided some advanced features of NoSQL databases and showcased how NoSQL databases can live upto their expectations when these new conditions are encountered.

Moniruzzaman, A. B. M., and Syed Akhter Hossain[1]presents classification, characteristics and evaluation of NoSQL databases in Big Data Analytics. This paper provides an understanding of the pros and cons of various NoSQL database approaches; also provides a overview of the non-relational NoSQL databases.

Strauch, Christof, Ultra-Large Scale Sites, and Walter Kriha [6]provides an overview of the motives and rationales,common concepts, techniques and patterns as well as several classes of NoSQL databases (key-/value-stores, documentdatabases, column-oriented databases). There are lot of parameters taken into consideration that includes both the qualitative[7] and quantitative features. In this work we present the qualitative features alone and we propose a system for a streaming application that uses both these databases in order to compare the performances when they encounter various types of queries. This makes sense as we compare two document oriented NoSQL databases in the same environment.With the advent of Big Data, many schema-less, structure-less data were growing prodigiously. So, the effective storage and processing of such data were not possible with the existing RDBMS. The looming of NoSQL databases proved to be one of the best solutions for handling these kind of schema-less data. This work comprises about the various characteristics of NOSQL databases.

III. PROPOSED WORK

In this section, we work on comparing MongoDB and CouchDB - the two leading document oriented databases taken under experimentation.The reason for considering these two databases are to understand for which application, which database suits well.Web applications is taken as a reference application which involves NodeJS.

NodeJS is an open source, cross-platform runtime environment for developing server-side web-based applications. The application is developed using NodeJS and Express. Express is a framework for building web applications on top of Node js. It simplifies the server creation process that is already available in Node. Node allows us to use JavaScript as our server-side language. MongoDB and CouchDB are databases. This is the place where you store information for your web websites(or applications). CRUD is an acronym for Create, Read, Update and Delete. It is a set of operations we get servers to execute (POST, GET, PUT and DELETE respectively).

This is what each operation does:

- Create (POST) - Make something
- Read (GET)_- Get something
- Update (PUT) - Change something
- Delete (DELETE)- Remove something

The resultant JSON file from the application are loaded into MongoDB and CouchDB.

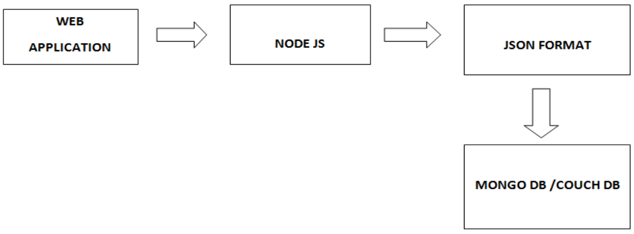


Figure 1: Proposed Framework

The System Architecture in Figure 1 shows that data is accessed through the NodeJS platform. Once the information is captured, they appear in the form of JSON documents.These JSON documents are stored in MongoDB andCouchDB to carry out the performance comparison.

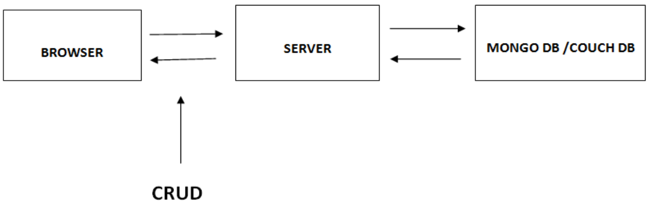


Figure 2: CRUD operations

In Figure 2 CRUD, Express and MongoDB/CouchDB are combine together into a single diagram.

IV.RESULTS AND DISCUSSIONS

We have selected few quatitative features for listing out the comparisons between the two databases. A comparison between the time for insertion, deletion, updation and retrieval are done. The following graphs show a comparative analysis between the two databases.

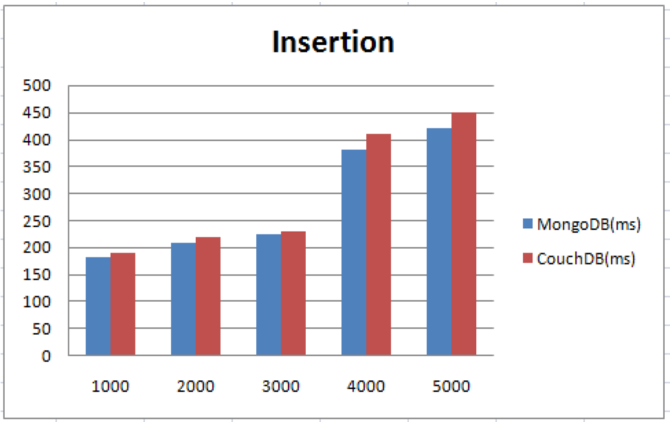


Figure 3: Insertion

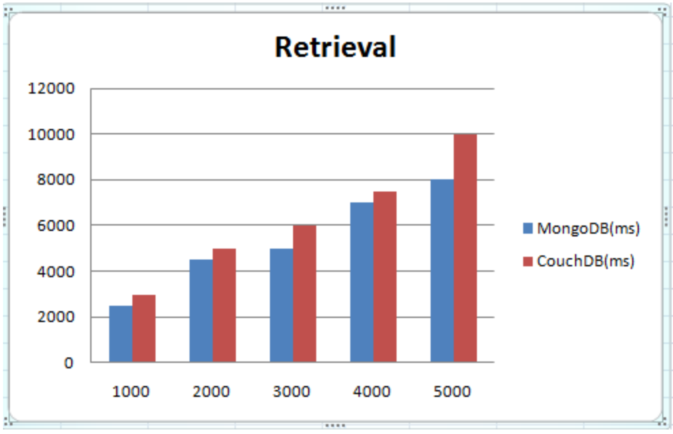


Figure 4: Retrieval

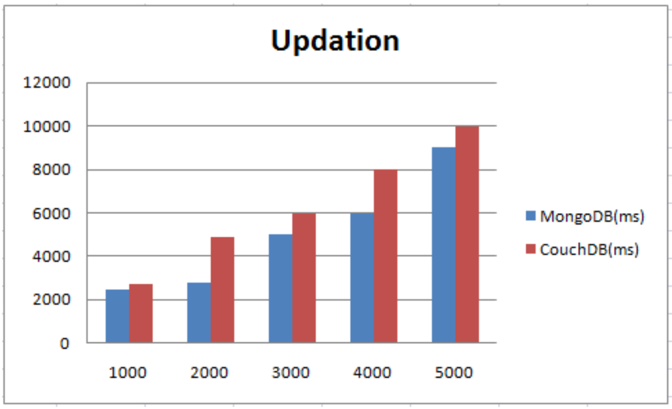


Figure 5: Updation

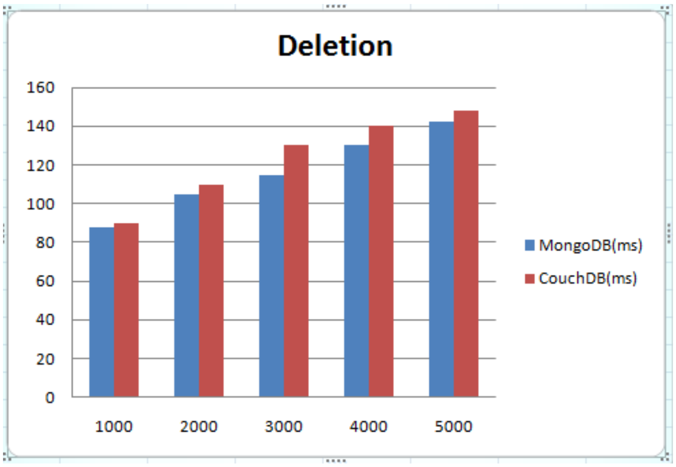


Figure 6: Deletion

The above graphs from Figure 3to 6 shows a comparison between the two databases under study. As discussed,we have taken a web application and performed this comparison.Hence from the graphs, we can conclude that MongoDB performs well for these kind of applications.

V. CONCLUSION

The quatitative features of both the document-store databases are analyzed in this work. It was very tough to provide a comparative analysis of NoSQL databases. Hence comparisons within NoSQL databases are performed. Further, quantitative attributes like size of the data stored in both the databases and how the databases perform when various types of queries encountered are analyzed. The results suggests that MongoDB clearly have an advantage over CouchDB. This is very much evident from the graphs shown in the above section. Since CouchDB doesn't have any options for bulk-importing JSON documents, it proves a failure model for these kind of web applications.

VI.REFERENCES

[1] Moniruzzaman, A. B. M., and Syed Akhter Hossain. "Nosql database: New era of databases for big data analytics- classification, characteristics and comparison." arXiv preprintarXiv:1307.0191 (2013).

[2] Tudorica, Bogdan George, and Cristian Bucur. "A comparison between several NoSQL databases with comments and notes." Roedunet International Conference (RoEduNet), 2011 10th. IEEE, 2011.

[3] Enqing Tang and Yushun Fan, “Performance Comparison between Five NoSQL Databases,” 7th International Conference on Cloud Computing and Big Data 2016.

[4] Hecht, R., & Jablonski, S. (2011, December). NoSQL evaluation: A usecase oriented survey.In Cloud and Service Computing (CSC), 2011 International Conference on (pp. 336-341).IEEE.

[5] Leavitt, N. (2010). Will NoSQL databases live up to their promise? Computer,43(2),12-14.

[6] Strauch, Christof, Ultra-Large Scale Sites, and Walter Kriha. "NoSQL databases." Lecture Notes, Stuttgart Media University (2011).

[7] Sundhara Kumar K B, Senthil Kumar V, Srividya, Mohanavalli.S,"Comparison of NoSQL Databases", Proceedings of National Conference on Communication and Informatics - 2016, Sri Venkateswara College of Engineering, Sriperumbudur.

© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering,
Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

Gradual Class Evolution Detection Using Class Based Ensembles

J Linita Lyle*, Soumya Balan P, Prof. Leya Elizabeth Sunny

Department of Computer Science and Engineering, M A College of Engineering Kothamangalam, Kerala, India

ABSTRACT

The recent advances in hardware and software have enabled the capture of different measurements of data in a wide range of fields. These measurements are generated continuously and in a very high fluctuating data rates. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non- stopping streams of information. The research in data stream mining has gained a high attraction due to the importance of its applications and the increasing generation of streaming information. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non-stopping streams of information. Data Stream classification poses major challenges than classifying static data because of several unique properties of data streams such as infinite length, concept drift, concept evolution and feature evolution. While extensive work has been done in the area of concept drift, concept evolution, a phenomena that induces concept drift has gained little recognition. Class evolution basically focuses on 3 aspects: the phenomenon of class emergence, disappearance and reoccurrence and is an important research topic for data stream mining. Most of the previous works implicitly regard class evolution as a transient change, which is not true for many real-world problems as in many real world applications class evolution is a gradual process. A class-based ensemble approach, namely Class-Based ensemble for Class Evolution (CBCE), is adopted to handle class evolution. By maintaining a base learner for each class and dynamically updating the base learners with new data, CBCE can rapidly adjust to class evolution. A novel under-sampling method for the base learners is used to handle the dynamic class- imbalance problem caused by the gradual evolution of classes. Based on the above concepts of gradual class evolution, a dataset containing records of tweets made in twitter is evaluated at different time stamps by converting the unstructured, dynamic dataset into a more compact form, to evaluate and analyse concept evolution.

Keywords : Data Stream Mining, Concept Drift, Class Evolution

I. INTRODUCTION

The intelligent data analysis has passed through a number of stages. Each stage addresses novel research issues that have arisen. Statistical exploratory data analysis represents the first stage. The goal was to explore the available data in order to test a specific hypothesis. With the advances in computing power, machine learning field has arisen. The objective was to find computationally efficient solutions to data analysis problems. Along with the progress in machine learning research, new data analysis problems have been addressed. Due to the increase in database sizes, new algorithms have been proposed to deal with the scalability issue. More over machine learning and statistical analysis techniques have been adopted and modified in order to address the problem of very large databases. Data mining is that interdisciplinary field of study that can extract models and patterns from large amounts of information stored in data repositories.

With the rapid development of incremental learning and online learning, mining tasks in the context of data stream have been widely studied. Generally, data stream mining refers to the mining tasks that are conducted on a (possibly infinite) sequence of rapidly arriving data records.

The analysis of huge volumes of data is recently the focus of intense research, because such methods could give a competitive advantage for a given company. For contemporary enterprises, the possibility of making appropriate business decisions on the basis of knowledge hidden in stored data is one of the critical success factors. Similar interests in exploring new types of data are present in many other areas of human activity.

In many of these applications, one should also take into consideration that data usually comes continuously in the form of data streams. Representative examples include network analysis, financial data prediction, traffic control, sensor measurement processing, ubiquitous computing, GPS and mobile device tracking, user's click log mining, sentiment analysis, and many others. Data streams pose new challenges for machine learning and data mining as the traditional methods have been designed for static datasets and are not capable of efficiently analysing fast growing amounts of data and taking into consideration characteristics such as:

- Limited computational resources as memory and time, as well as tight needs to make predictions in reasonable time.
- The phenomenon called concept drift, i.e., changes in distribution of data which occur in the stream over time. This could dramatically deteriorate performance of the used model.
- Data may come so quickly in some applications that labelling all items may be delayed or sometimes even impossible.

Out of several tasks studied in data streams, supervised classification has received significant attention. It is often applied to solve many real life problems such as discovering client preference changes, spam filtering, fraud detection, and medical diagnosis to enumerate only a few. The aforementioned speed, size and evolving nature of data streams pose the need for developing new algorithmic solutions. In particular, classifiers dedicated to data streams have to present adaptation abilities, because the distribution of the data in motion can change. To tackle these challenges, several new algorithms, specialized sliding windows, sampling methods, drift detectors and adaptive ensembles have been introduced in the last decade.

As the environment where the data are collected may change dynamically, the data distribution may also change accordingly. This phenomenon, referred to as concept drift, is one of the most important challenges in data stream mining. A data stream mining technique should be capable of constructing and dynamically updating a model in order to learn dynamic changes of data distributions, i.e., to track the concept drift.

For classification problems, concept drift is formally defined as the change of joint distribu tion of data, i.e., p(x,y), where x is the feature vector and y is the class label. Over the past few decades, concept drift has been widely studied. The majority of the previous works focus on the concept drift caused by the change in class-conditional probability distribution, i.e., p(x|y).

In comparison, class evolution, which is another factor that induces concept drift, has attracted relatively less attention. Briefly speaking, class evolution is concerned with certain types of change in the prior probability distribution of classes, i.e., p(y), and usually corresponds to the emergence of a novel class and the disappearance of an outdated class. In some literature, class evolution is also called class-incremental learning or concept evolution.

We put forward a method by which data streams can be analysed more efficiently to detect class evolution.

The data sets is divided into independent units, pre-processed and structured in a way that makes mining tasks easier to be performed. By using this method, small and large datasets may be analysed to gain valuable information in fields like network analysis, medical diagnosis etc.

A. Class Based Ensemble For Class Evolution

Equation (6) suggests that the optimal classification strategy is to assign an example according to the likelihood that it belongs to a class. Therefore, a natural approach to this problem is to maintain a model for each class so that the likelihood can be explicitly estimated. For this reason, the CBCE approach is proposed. Each class-based model (CB model) is maintained for a certain class ci and an example x is classified according to

$$\text{argmax CBMClassify}(x, CBM_i) \qquad (1)$$

where the function CBMClassify returns the likelihood P(x| ci) or scores that can be used to estimate P(x| ci). Depending on the current class evolution state, the CBCE algorithm manages the CB models in mining tasks. Specifically, it may create a new CBmodel for a novel class, inactivate an outdated CB model for a disappeared class and re-activate the CB model when the class reoccurs again. Since the class conditional probability is also likely to change in a real-world data stream, the previously built model for a class could become invalid later. Hence, CBCE also involves a scheme to detect and handle the invalid CB model.

1) CLASS BASED MODEL

A class-based model is one that is specifically constructed for a certain class to get the likelihood (or related score) of a test example. A variety of models are possible candidates for a CB model, e.g., one-class classifier and clustering model. In this work, the CB model is implemented as a binary classifier that is able to output its classification posterior probability. In each CB model, with the one-versus-all strategy, the represented class is the positive class (+1) and the others are the negative one (-1) as a whole. According to Bayesian theory, the posterior probability P(+1|xt) for the positive class at time t is

$$P(+1|x_t) = \frac{P_t(+1)}{P_t(x_t)}.\, P_t(x_t| + 1) \qquad (2)$$

where $Pt(xt)$ is the same for all classes. If the training data are balanced in CB models, P(+1) is a constant 1/2. In this condition, the posterior probability for positive class is proportional to the likelihood of the positive class, i.e., the specific class the CB model is maintained for. In other words, the probability can be used as the score to represent the likelihood for making decisions.

The positive and negative classes are likely to be imbalanced in a CB model. Although class-imbalanced problem has been intensively investigated, most previous studies focus on static class-imbalanced problems. In our case, the prior distribution may change over time, leading to a dynamic class-imbalanced problem. To address this issue, an under-sampling strategy is embedded in each CB model. The sampling probabilities for the positive and negative classes are different. As each CB model acts as an “expert” for its corresponding class, all of the examples received from this positive class are selected. The data size of the negative classes is usually larger than the positive one. Furthermore, the size of each class dynamically changes due to the gradual class evolution. These negative examples are sampled by under-sampling with a dynamic probability, which aims to select the negative data with the same size as the positive ones. Denoting wti as the prior probability of class ci at time t, the probability of sampling the negative examples for ci is calculated as

$$P_i = \min(w_t^i/(1 - w_t^i),\ 1) \qquad (3)$$

In on-line learning, the underlying prior probability wti is hard to be observed. To quickly and accurately estimate wti, it is tracked by the time decay method as:

$$w_t^i = \beta w_{t-1}^i + (1 - \beta)1[y_t = c_t] \qquad (4)$$

Where β (0 < β < 1) denotes decay factor, and 1[yt=ct] = 1 if yt, the true class label of xt is ci and 1[yt=ct] = 0 if yt, the true class label of xt is not ci.

To conveniently apply CBCE in practice, a constant decay factor is used for the prior probabilities of all classes. Since the estimated prior probability will be updated exponentially, it will quickly achieve its underlying value. The appropriate value for β is 0.9, which has been determined after comprehensive experiments.

In the CBCE framework, a CB model is required to provide its output in the form of score and can be updated on the-fly. Quite a few classical base leaners satisfy the first requirement, and logistic regression might be the model that has been mostly investigated with regard to the second issue. Hence, the online Kernelized Logistic Regression is employed in this work as the base learner. It should be noted that CBCE does not necessarily require to establish only one CB model for each class, and in some cases an ensemble model might be more suitable than a single model for a class. For example, if the minority class may comprise small disjuncts of data, a possibly better option for the CB model is to employ cluster over-sampling techniques and build a model for each disjunct of data.

II. CLASS EVOLUTION ADAPTATION

Class evolution has three basic elements, i.e., the emergence of novel classes, the disappearance of outdated classes, and the reoccurrence of disappeared classes.

When a novel class ci emerges at time stamp t, CBCE first estimates its prior probabilitywti, and then initializes a new CB model CBMi for it. The prior probability is initially estimated after receiving the first two examples of this class. Denoting ExampleSize as the example size of the negative classes between these two examples, the prior probability is estimated as follows:

$$w_t^i = 1/(ExampleSize + 1) \qquad (5)$$

Based on the two examples of novel class and the negative examples between them, the CB model is initialized. Next, the CB model participates in classifying the subsequent data stream.

For class disappearance, the approach has to determine the disappearance when a class is shrinking; following this, its CB model should be managed to ensure not to affect the recognition of other classes. Since the evolution state is tracked in CB models, a sufficiently small prior probability threshold, e.g., β1000(β is the decay factor), can be used for disappearance confirmation. That is, if the class has been absent for 1000 consecutive time stamps, it is thus considered to have disappeared. The decision boundary of the CB model, as implemented by binary classifier, merely separates one class from another. In this case, if the class-conditional probability distribution changes or a novel class emerges on the boundary, the original CB model for the disappeared class would be inaccurate and also influence the novel class. Therefore, the CB model of the disappeared class is inactivated in classification. Besides, when a class is considered to have disappeared, its estimated prior probability is set to be 0, which also means its CB model is suspended for updating.

Class reoccurrence means that an example with the label of a disappeared class is received again. Effective handling of class reoccurrence could make use of past training efforts. Once class reoccurrence happens, the model re-estimates the prior probability in the same way as class emergence, and activates the CB model in classification.

B. DESIGN

The most creative and challenging phase of the system lifecycle is the system design. The term design describes a final system and the process by which it is developed. In system design, there is a movement from the logical to the physical aspects of the life cycle.

1) INPUT DESIGN

Input design is one of the most expensive phases of the operation of computerized system and is often the major problem of a system. The decisions made during the input design are: to provide cost effective method of input, to achieve the highest possible level of accuracy, to ensure that input is understood by the user.

2) Data Set Input

UDI TwitterCrawl Dataset, including 50 million tweets posted mainly from 2008 to 2011, is involved. Each record in this data set has its own time stamp and the order of examples in the data stream is completely genuine, without any modification. Since the hashtag roughly describes the tweet’s topic, it was used as the class for each tweet record. If more than one hashtags exist in a tweet, one of them is selected randomly as its label.

Three tweet stream fragments from the whole tweet set are captured by selecting different topics as the classes of interest, i.e., tweet stream a, b, and c.

C. SYSTEM MODULES

The system developed has four modules.

- Initial Setup
- Update CB Model
- Class Evolution Adaptation
- Visualization

1) Initial Setup

Data pre-processing is an important step in the data mining process. Data-gathering methods are often loosely controlled, resulting in out-of-range values impossible data combinations missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

In this module the data stream is pre-processed and converted into a form suitable for performing the mining tasks. The stream is divided into smaller, more compact units.

2) Update CB Model

A class-based model is one that is specifically constructed for a certain class to get the likelihood (or related score) of a test example. A variety of models are possible candidates for a CB model, e.g., one-class classifier and clustering model. In this work, the CB model is implemented as a binary classifier that is able to output its classification posterior probability.

When a new example is received, every CB model will update the estimation of prior probability of its class. For the class that the currently received example belongs to, its CB model uses it for updating directly. For the other CB models, the example is first sampled with the dynamic sampling probability, and then used to update the models as a negative training example.

The learning procedure is summarized in Algorithm 1. When a new example is received, every CB model will update the estimation of prior probability of its class (lines 2 and 5). For the class that the currently received example belongs to, its CB model uses it for updating directly (line 3). For the other CB models, the example is first sampled with the dynamic sampling probability, and then used to update the models as a negative training example (lines 6 and 7).

3) Class Evolution Adaptation

Class evolution has three basic elements, i.e., the emergence of novel classes, the disappearance of outdated classes, and the reoccurrence of disappeared classes. When a novel class ci emerges at time stamp t, CBCE first estimates its prior probability wi, and then initializes a new CB model CBMi for it.

For class disappearance, the approach has to determine the disappearance when a class is shrinking; following this, its CB model should be managed to ensure not to affect the recognition of other classes.

Class reoccurrence means that an example with the label of a disappeared class is received again. Effective handling of class reoccurrence could make use of past training efforts.

For the inactivated CB model of a disappeared class, it can be used again for classification when an example with an old label arrives, which makes CBCE efficient.

This mechanism to deal with the three key components of class evolution is wrapped around each CB model, which equips CBCE to track gradually evolved classes effectively. The procedure of class evolution adaptation is summarized in Algorithm 2. Depending on the change of prior probability, class evolution behaviour can be determined. The active CB models are updated by the sampled data, and the inactive ones are stored additionally in case of class reoccurrence.

4) Visualization

A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may are encoded using dots, lines, or bars, to visually communicate a quantitative message.

The distribution of data and the class behaviour of the data set are expressed using graphs and plots to effectively visualize, analyse and reason about data.

3) OUTPUT DESIGN

Output design generally refers to the results and information that are generated by the system. Outputs of a system can take various forms. The most common are reports, screens displays, printed form, graphical drawing etc. The output also vary in terms of their contents, frequency, timing and format. The users of the output, its purpose and sequence of details to be printed are all considered.

In this project the output is represented through the visualization module in the form of graphs.

III. RESULTS AND DISCUSSION

After getting the tweet streams, the text of each tweet is transferred into the TF-IDF vectors. 242, 247, 242 numerical features are generated, respectively, for the tweet streams a, b, c classes. It can be seen that class evolution may occur frequently in tweet stream. Besides, according to the visualization of tweet streams in the figures below, it can be observed that the class-conditional probability distribution also changes over time in tweet stream.

IV.CONCLUSION

The class-based framework adopted by the system has a number of advantages in comparison to the existing methods. First, since a CB model is specifically maintained for a certain class, it is flexible to be created or removed to adapt to class evolution. This also decouples the whole model, and makes each CB model simple and concentrate on a single class. Second, by using the CB model, only a few of base learners need to be maintained, equal to the number of classes. Third, for massive-volume data streams, the master-slave structure (CB model – ensemble strategy) and the chunk based processing of the learning system is also very convenient for parallelization and distributed implementation.

V. REFERENCES

[1] Y. Sun, K. Tang, L. L. Minku, S. Wang and X. Yao, "Online Ensemble Learning of Data Streams with Gradually Evolved Classes," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 6, pp. 1532-1545, June 1 2016.

[2] J. Gama, I. Zliobait_ e, A. Bifet, M. Pechenizkiy, and A.Bouchachia, "A survey on concept drift adaptation," ACM Comput. Surv.,vol. 46, no. 4, pp. 44:1–44:37, Mar. 2014.

[3] M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B. Thuraisingham, "Addressing concept-evolution in concept drifting data streams," in Proc. IEEE 10th Int. Conf. Data Mining, Dec. 2010, pp. 929–934

[4] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," SIGMOD Rec., vol. 34, no. 2, pp. 18–26, 2005.

[5] S. Wang, L. Minku, and X. Yao, "A learning framework for online class imbalance learning," in Proc. IEEE Symp. Comput. Intell.Ensemble Learn., Apr. 2013, pp. 36–45

[6] Heitor Murilo Gomes, Jean Paul Barddal, Fabricio Enembreck and Albert Bifet, "A Survey on Ensemble Learning for Data Stream Classification" in ACM Computing Surveys, Vol. 50, no. 2, April 2017

[7] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavald. New ensemble methods for evolving data streams. In SIGKDD, pages 139–148, 2009.

www.ijsrset.com

International Journal of Scientific Research in Science, Engineering and Technology
(International Journal Bimonthly Publication)
Print ISSN : 2395-1990, Online ISSN : 2394-4099
www.ijsrset.com

Published by :

TechnoScience Academy

The International Open Access Publisher
Email: info@technoscienceacademy.com
Website: www.technoscienceacademy.com

Email: editor@ijsrset.com, ijsrset@gmail.com