ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/80143, September 2015* ISSN (Online) : 0974-5645

Diagnosis of Disease from Clinical Big Data using Neural Network

S. Sapna[1*] and M. Pravin Kumar[2]

[1]Department of Computer Science, Bharathidasan College of Arts & Science, Erode – 638 116, Tamil Nadu, India; gsrissj@gmail.com
[2]Department of ECE, K.S.R. College of Engineering, Tiruchengode – 637 215, Tamil Nadu, India; pravinkumarmm@gmail.com

Abstract

Big Data is info whose assortment, intricacy and scale need novel procedures, analytics, methods and design to cope with it and mine value and hidden information from it. Big Data requires different approaches with an intention to resolve novel snags or existing hitches in an enhanced way. The greatest challenges is to deal with large dataset with high amount of dimensionality, together in terms of the number of features the data has, as well the number of rows of data that user is dealing with. Neural Networks is a machine learning tool that is capable of performing these tasks. This paper presents an integrative approach to predict the diabetic disease from clinical big data. The clinical database is generally redundant, incomplete, vague and unpredictable. The main objective of integrating is to experiment with different strategies of training neural networks in order to increase the prediction accuracy.

Keywords: Clinical Big Data, Big Data, Diabetes, Neural Networks and Prediction

1. Introduction

The main aim of clinical information science is to require in universe medicinal information from all stages of individual to advance our considerate of treatment and medicinal exercise. The volume of clinical data is likely to increase drastically in the forthcoming years. Health information science utensils comprise among others medical procedures, computers, prescribed medicinal terms and info and communiqué methods[5,6]. Healthcare compensation replicas are varying; eloquent usage and pay for concert are evolving as serious novel issues in recent healthcare environs. Even though yield isn't and may not be a prime promoter, it's crucially significant for healthcare administrations to obtain the accessible methods, structure and utensils to control huge information efficiently[4,15]. The arena of health information science is on the cusp of its maximum stirring retro to date, getting into an innovative epoch wherever technology is beginning to knob huge info, conveying about limitless prospective for info development[2,7]. The large volume of data in health care exceeded our human ability for diagnosing disease without powerful tools. Big Data analytics and neural network technique helps to better understand the objectives of diagnosing and treating patients in need of healthcare.

Huge information in healthcare allude to electronic health records, so huge and composite that they are not easy to manage with out-dated software system and/or hardware; nor will they be simply coped with out-dated or common information managing tools and procedures. Huge records in healthcare is devastating not only due to its size however conjointly due to the variety of information forms and the haste at that it should be coped as it comprises experimental data and medicinal decision support systems[8]. Neural network method is employed to resolve an extensive variety of hitches in diverse application concerning huge capacity of information. It

* *Author for correspondence*

also permits us to introduce the learning and adaptation capabilities[3].

Clinical data contains massive quantity of information and are usually uncertain to arrive at a definite decision. So, in this paper neural network technique is proposed to diagnose the diabetic disease from the clinical diabetic data. This integrative approach provides flexible information capability for medical practitioners in handling ambiguous situations.

2. Big Data

A huge collection of information is more identical to lesser information, however larger in intricacy and adjustable generation methods. But having Big Data means having to set newer technologies and different approach in handling bigger dataset which aims to resolve novel glitches and even resolve ancient hitches in an enhanced mode[17].

International Business Machines[1] estimates that every single day user generates 2.5 quintillion bytes of information, such a lot that 90% of the information within the universe nowadays has been produced in the latter 2 years only. The large information arise from devices that are utilized to collect weather info, posts to societal mass media locations, digital videos and images, mobile phone global positioning system signals and purchase deal archives.

Data is a raw unorganized facts, is in and of itself worthless. Information means potential valuable concepts based on data. Knowledge is what we understand based upon information. Wisdom means effective use of knowledge in decision making. The important enablers for the presence and development of huge information are upsurge in stowage abilities, rise in handling control and obtainability of information. There are enormous capacities of information in the universe. From the start of chronicled time till 2003, User's generated five billion gigabytes of information. In the year 2011, the similar size was generated every 2 days. In the year 2013, the similar quantity of information is produced every ten minutes. Prior to 2012 the US was the largest single contributor to global data. 32% United States, 13% China, 4% India 19% Western Europe and 32% of rest of world. In 2020 the emerging markets are showing the largest increases in data growth 23% United States, 21% China, 6% India, 15% Western Europe and 35% rest of world. In 2012 the amount of information stored worldwide exceeded 2.8 zeta bytes (1021).

By the year 2016 the growing magnitude of all the universal information center's exempted to surpass to 16,000 acreages which is equal in space to a 2 way route stretching from Japanese capital to metropolis over five thousand miles. An estimated 33% of information could be useful if appropriately tagged and analyzed. The amount actually analyzed is only 0.5%. By the year 2020 the overall sum of information stowed is predictable to be 50x greater than these days. In the digital world all that has been generated and stowed are almost partial of which is unguarded.

2.1 V's of Big Data

The volume, velocity, variety and veracity are the five V's of big data. Volume means data at rest. Firms are crammed with ever-increasing information of all sorts, simply increasing terabytes (1012) - even petabytes (1015) - of information. Turn twelve terabytes of Chirps generated every day into better creation sentimentality investigation. Translate 350 billion yearly meter evaluations to well envisage power ingesting. Velocity means knowledge in motion. Occasionally two minutes is just too late. For time-sensitive procedures like catching scam, huge information should be utilized because it creeks into the enterprise so as to increase its worth. Examine five million trade actions generated every day to recognize possible deception. Evaluate 500 million day-to-day call detail archives in period of time to envisage client churn quicker. The newest is ten nano seconds deferment is too ample. Variety means data in many forms i.e., organized and formless information such as audio, text, video, log files, snap streams and further. Novel visions are found once evaluating these information kinds together. Examine 100's of live audiovisual forages from investigation cameras to focus on goals of interest. Abuse the 80% information progress in pictures, audiovisual and records to increase client gratification. Veracity means information in disbelief that is ambiguity because of information discrepancy and dormancy. Value means a perfect understanding of prices and edges. The task is to seek out some way to transmute information into data that has worth, either internally or for creating a trade on view of it. The pictorial depiction of V's of Big Data is shown in Figure 1.

Enormous Information is generated by societal mass media and nets (every one of us are creating information), Technical tools (gathering all kinds of information), Portable expedients (pursuing every entity constantly),

Sensor tools and systems (computing wide range of information). The advancement and novelty is no more ruined by the capacity to gather information. In any case, by the capacity to oversee, examine, condense, envision and find learning from the gathered information in an opportune way and in a versatile manner would surely be the difficulties of data innovation group[1,9].

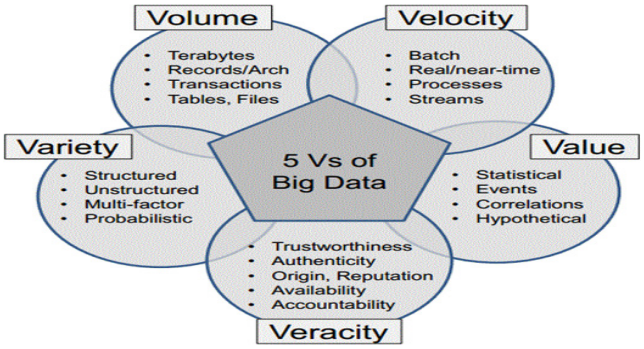

Figure 1. V's of Big Data.

3. Neural Networks

Intelligence implies versatility to circumstance. A person is scholarly in light of the fact that people have the capacity to react adaptively to changes in their interior and outer setting and they utilize their sensory system to finish these practices. Computers aren't intelligent because they are machines which do exactly what they are told to do and nothing more. Neural network mimics the human brain to make the system intelligent. A biological system works to solve a problem by the process of learning and so is Artificial Neural Network (ANN)[10].

3.1 Biological Neural Net

A nerve cell is an exceptional organic cell that procedures data as shown in Figure 2. An apt model or reproduction of the sensory system must have the capacity to create comparable reactions and practices in counterfeit frameworks. The sensory system is made by relatively basic units named neurons, so reproducing their conduct and usefulness ought to be the result. The dendrite receives signals from other neurons (accepts inputs), Soma (cell body) sums all the incoming signals (process the inputs), Axon passes information away from the cell body to other neurons (turn the processed input's into output's). The Synapses are an electrochemical contact between neurons, (The information transmission happens at the synapses).

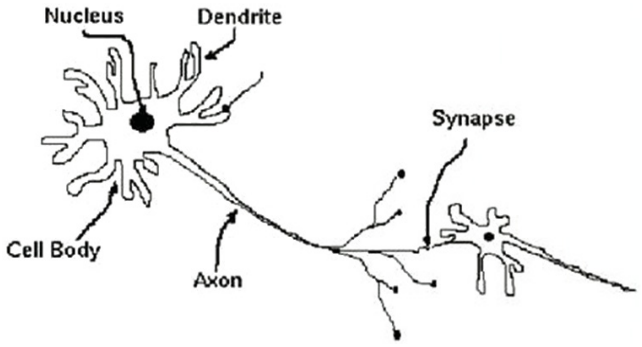

Figure 2. Structure of a biological neuron.

The target of adjusting the retorts on the premise of the data got from the earth is to accomplish a superior state. At the end of the day, the goal of learning in natural creatures is to improve the measure of accessible assets, satisfaction or by and large to accomplish a closer to ideal state[13].

3.2 Artificial Neural Net

An Artificial Neural Net (ANN) is a data handling model that is enlivened by natural sensory systems[3,11,14,16]. It is made out of countless interconnected handling components named neurons. An ANN is intended for a specific use, for example, pattern recognition. ANN utilizes a numerical or computational model for data handling by taking into account a connectionist way to deal with calculation. In most extreme cases an ANN is a versatile system that varies its structure in light of outer or interior information those courses through the system. ANNs assimilate the two primary segments of organic neural nets to be specific Neurons and Synapses - bury neuron associations quality are utilized to store the data. Neural system is a versatile framework made of four principle areas as shown in Figure 3. A node is a unit that activates upon receiving incoming signals (inputs), interconnections between nodes, an activation function (rule) which transforms inside a node, input into output, an optional learning function for managing weights of input-output pairs.

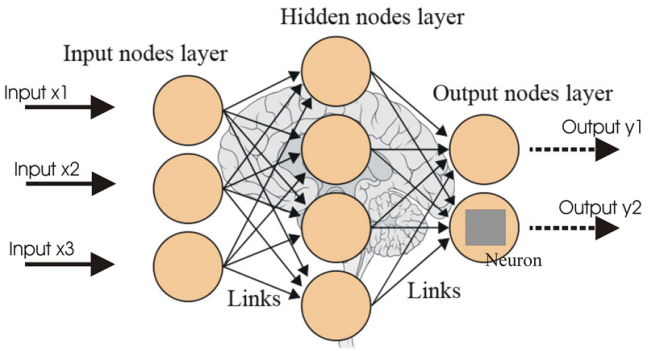

Figure 3. Architecture of Artificial Neural Network.

An ANN is made out of various simulated neurons that are associated together as indicated by specific system construction modeling. The point of the ANN is to change the inputs into applicable yields.

3.3 Association of Biological Net with Artificial Net

The Figure 4 shows the neuron model used to construct the Neural Network[14]. Signals are received by the processing elements. This element sums the weighted inputs. Weight at the receiving end has the capability to modify the incoming signal. Neuron fires (transmits output) when sufficient input is obtained. Output produced from one neuron may be transmitted to other neurons. Weights can be modified by the experience. Both Artificial and biological neurons have fault tolerance.

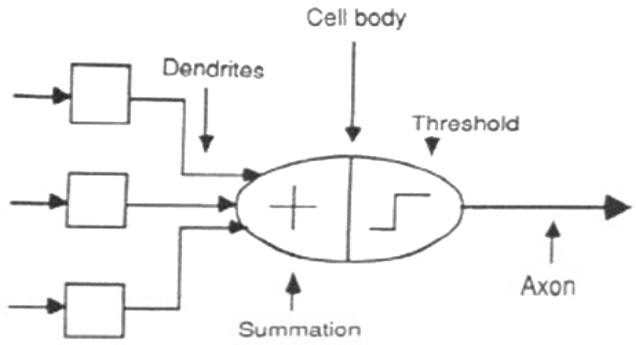

Figure 4. The neuron model.

3.4 Architecture of Neural Net

The architecture of simple neural net is shown in Figure 5. The neuron is the elementary info handling entity of a Neural Net (NN). It comprises of:

- A group of connecting links (synapses), every connection described by a weight - w1, w2,.., w_m.
- A direct combiner which figures the weighted total of the inputs:

$$u = \sum_{j=1}^{m} wjxj \quad (1)$$

- An activation function for restraining the abundance of the yield of the neuron.

$$y=\phi(u+b) \quad (2)$$

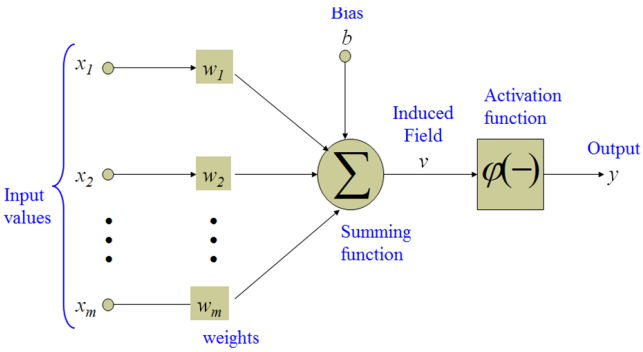

Figure 5. Architecture of simple neural net.

- The bias named b has the influence of applying a change to the weighted aggregate u, v = u + b

$$v = \sum_{j=0}^{m} w_j x_j \qquad \text{where } w_0 = b \quad (3)$$

- The bias b is an exterior factor of the neuron. It can be modeled by including an additional input. The induced arena of the neuron is denoted as v.

The decision of activation function φ decides the neuron exemplary. Amid learning stage, a repetitive system forages its inputs over the net, counting nourishing information nether from yields to inputs; procedure is reiterated till the estimations of the yields don't alter. This state is so-called balance or steadiness. Recurring nets can be trained with the help of back-propagation system. In this technique, at every stride, the initiation of the yield is contrasted with the anticipated initiation and errors are promulgated backward through the net[3,16]. When this preparation procedure is finished, the system gets to be equipped for performing an arrangement of activities.

3.5 Integrated Approach

Gathering and maintaining huge volume of information is one thing, however mining useful information from these collections is even more challenging. These tasks can be accomplished using Neural Networks. There is also a growing requirement to simultaneously gather and process huge amount of data. This situation is creating a need in this field to switch from conventional computers

that process information in sequence, to parallel computers furnished with several processing elements aligned to function in parallel to process information. The neural system technique is a hugely parallel appropriated processor that has a characteristic penchant for putting away experiential information and making it accessible for utilization. A prepared neural system can be considered as a specialist in the class of data it has been given to break down.

4. Disease – Diabetes Mellitus

Diabetes Mellitus is the coercive effect of insulin on the glucose metabolism. Insulin controls starch and obese digestion system in the body. Insulin is produced from the islets of Langerhans. In Latin the word Insula means- "island". Its concentration has wide spread effect all over the body. When the insulin level is not controlled, diabetes will effect. Diabetes type 1 is lack of it whereas diabetes type 2 is the resistance towards it. Not only insulin regulates the glucose in the blood but it is also responsible for lipid metabolism. Insulin production from the beta-cells is mainly controlled by the levels of plasma glucose.

Augmented taking up of glucose by pancreatic beta-cells prompts an attendant increment in metabolism. Physicians will become familiar with other aspects of managing the patient with diabetes, including the importance of postprandial glucose control, diabetes self-management training etc. The sustenance that is eaten is changed over to sugar or glucose which is utilized for vitality. The pancreas conceals insulin which conveys glucose into the cells of our bodies, which thusly delivers vitality for the ideal working of the body. At the point when a person has diabetes, the body either doesn't create sufficient insulin or else can't utilize its own insulin. This reasons sugar to develop in the blood, prompting complexities like coronary illness, stroke, poor dissemination prompting loss of appendages, visual deficiency, neuropathy, kidney letdown, harm in nerves and demise[12].

4.1 General Symptoms of Diabetes

Augmented need for liquids, augmented urination – loss in weight, augmented hunger – tiredness, biliousness and/or queasiness - unclear vision, slow-curative contagions - ineffectiveness in men.

4.2 Diagnostic Tests

4.2.1 Urine Test

A urine test might be utilized to search for glucose and ketones from the breakdown of obese. Nevertheless, a urine analysis alone does not identify diabetes. The following analyses are also carried out to identify diabetes.

4.2.2 Fasting Plasma Glucose Level (FPG)

The normal range of fasting blood glucose is <100 mg/dl. It is done after 8-12 hours of fasting. People having fasting glucose levels from 100 to 125 mg/dl are well-thought-out to have reduced fasting glucose. People with FPG >126 are deliberate to have diabetes mellitus.

4.2.3 Post Pran Dial Plasma Glucose Level (PPG)

A blood sugar test taken after two hours of a meal is known as The Post Prandial Glucose Test or PPG. The normal range for PPG is <140 mg/dl. People with fasting glucose levels from 140-200 mg/dl are well-thought-out to have reduced glucose lenience. Persons with PPG >200 mg/dl are consider to have diabetes mellitus.

The clinical data are usually uncertain and vague to arrive at a conclusion. So, in this paper Neural Network is proposed to diagnose the diabetic disease from the clinical diabetic big data. This integrative approach provides flexible information capability for medical practitioners in handling ambiguous situations.

5. Adaptation of Neural Network

In this paper the diabetic disease dataset is considered for execution, this data set is gathered from various Diabetic Care Centers at Erode. About 570 diabetic patients' data were considered for this prediction and some of which is shown in Table 1. The inputs considered are Age, Fasting Plasma Glucose (FPG), Post Prandial Plasma Glucose (PPG) and the output is D-Diabetic Status. Neural Network is used for training the network. Regression R analysis is performed to evaluate the relationship between yields and targets. An R value of 1 represents a close by correlation, 0 an arbitrary correlation. This paper proposes Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton Back propagation and Resilient Back propagation to diagnose the diabetes disease. BFGS quasi-Newton method is a fast optimization technique and the

Resilient Back propagation is a quick adaptive supervised learning in feed forward artificial neural networks. The Resilient Back propagation is the first order minimizing algorithm. The main ability of this algorithm is to automatically adjust the step length in order to speed up the convergence process. This algorithm is best in terms of accuracy and convergence speed.

Table 1. Practically observed clinical database of Diabetes Mellitus

Inputs				Output
Sl. No.	Age	FPG (mg/dl)	PPG (mg/dl)	D
1	52	95	290	0.7
2	51	109	452	0.91
.	.	.	.	.
570	55	291	357	0.93

The performance of the BFGS quasi-Newton Back propagation when trained with Neural Network using Matlab R2012a is shown in Figure 6 and its regression analysis is shown in Figure 7.

The performance of the Resilient back propagation when trained with Neural Network using Matlab R2012a, is shown in Figure 8 and its regression analysis is shown in Figure 9.

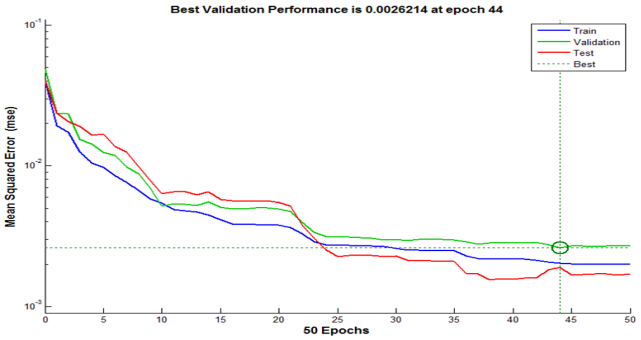

Figure 6. Performance of BFGS quasi-newton back propagation.

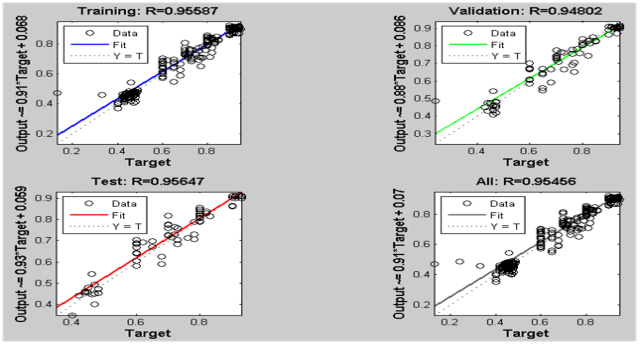

Figure 7. Regression analysis for BFGS quasi-newton back propagation.

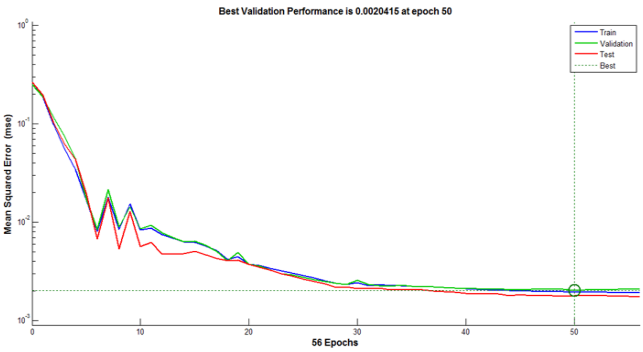

Figure 8. Performance of Resilient Back propagation.

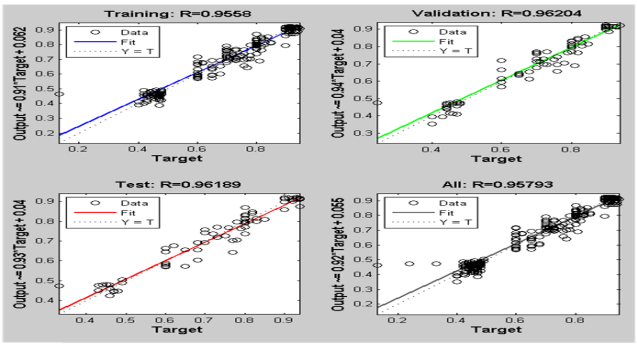

Figure 9. Regression analysis of Resilient Back propagation.

From Figure 6 it is perceived that the finest validation performance 0.0026214 is obtained at the 44th epoch using BFGS quasi-Newton Back propagation and its Regression R-value is 0.95456 which is shown in Figure 7. From Figure 8 it is perceived that the finest validation concert 0.0020415 is obtained at the 50th epoch using Resilient Back propagation and its Regression R-value is found to be 0.95793 which is shown in Figure 9. The results indicate that the Resilient Back propagation algorithm achieves best validation performance and also the Regression value (R) is closest to 1 compared to BFGS quasi-Newton Back propagation which indicates the accurate diagnosing of diabetes disease.

6. Conclusion

In this work an assimilated methodology is bestowed for the prediction of diabetes disease. Two methods of Neural Network, namely BFGS quasi-Newton Back propagation and Resilient Back propagation are proposed for the diagnosis of diabetes. From the result of the proposed methods it is observed that Resilient Back propagation performs better compared to BFGS quasi-Newton Back propagation. Hence it is reasonable to expect a

rapid increase in the understanding of Artificial Neural Network to analyze Big Data efficiently.

7. References

1. Bringing big data to the enterprise #ibmbigdata website. 2012. Available from: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
2. Zou H, Yu Y, Tang W, Chend HWM. FlexAnalytics: A flexible data analytics framework for Big Data Applications with I/O performance improvement. Big Data Research. 2014 Aug; 1:4–13.
3. Jang JSR, Mizutani E, Sun CT. Neuro-fuzzy and soft computing. Englewood Cliff, New Jersey: Prentice-Hall; 1997.
4. LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N. Big Data, analytics and the path from insights to value. MIT Sloan Manag Rev. 2011; 52(2):20–32.
5. Mettler T, Raptis DA. What constitutes the field of health information systems? Fostering a systematic framework and research agenda. Health Informatics Journal. 2012 Jun; 18(2):147–56.
6. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. Journal of Big Data: A Springer Open Journal. 2014 Dec; 1(2):1–35.
7. Augustine PD. Leveraging Big Data analytics and Hadoop in developing India's healthcare services. International Journal of Computer Applications. 2014 Mar; 89(16):44–50.
8. Priyanka K, Kulennavar N. A survey on Big Data analytics in health care. International Journal of Computer Science and Information Technologies. 2014; 5(4):5865–8.
9. Pooja, Jaglan S, Gupta R. Big Data: advancement in data analytics. International Journal of Computer Technology and Applications. 2014 Jul-Aug; 5(4):1466–9.
10. Ramani P, Murarka PD. Stock market prediction using Artificial Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Apr; 3(4):873–7.
11. Russell S, Peter N. Artificial Intelligence: A Modern Approach. Prentice Hall; 2009.
12. Sapna S, Pravin Kumar M. Prediction of uncertainty in clinical database using clustering technique. International Journal of Innovative Research in Technology. 2014; 1(10):98–102.
13. Maind SB, Wankar P. Research paper on basic of Artificial Neural Network. International Journal on Recent and Innovation Trends in Computing and Communication. 2014 Jan; 2(1):96–100.
14. Sivanandam SN, Sumathi S, Deepa SN. Introduction to Neural Networks using MATLAB 6.0. New Delhi: Tata McGraw Hill; 2006.
15. Raghupathi W, Raghupathi V. Big Data analytics in healthcare: Promise and potential. Raghupathi and Raghupathi Health Information Science and Systems. 2014; 2(3):1–10.
16. Birdi Y, Aurora T, Arora P. Study of Artificial Neural Networks and neural implants. International Journal on Recent and Innovation Trends in Computing and Communication. 2013 Apr; 1(4): 258–62.
17. Chhetri B. An approach towards the future of information science. Archers and Elevators International Journal of Multidisciplinary Research. 2015; 3(2):1–6.