National Conference
on
Machine Learning and Artificial Intelligence
August 27-28, 2018

Conference Proceedings

Organized by
Department of Computing
Coimbatore Institute of Technology

In association with
Data Centre and Analytics Lab, IIM Bangalore
&
Analytics Society of India

Social Media Opinion Mining for Iskcon Temple using NLP

Vinodhini Ranganathan*
St John's Research Institute, Bangalore

Vigneshwaran SR
Indian Institute of Science, Bangalore
* Email: vinodhiniranganathan14@gmail.com

Abstract

People visit temples to pray for their own physical, mental and material well being. Very few go to temples for indulging themselves in some charitable social work. In today's world, temples can be categorized into two types: traditional temples that emphasize & preach "bhakti" and contemporary temples that preach & practice "social responsibility with bhakti". One notable TEMPLE (Tranquility Education Medication Purification Love Engagement) that falls in the second category is "ISKCON" (The International Society for Krishna Consciousness). Temples like ISKCON are not only meant to be a place for religious worship, but also as a charitable society that provides free food, free education etc. Such societies see rapid growth in a short span, thereby making it difficult for the management to monitor people's perceptions and identify factors that denigrate their brand equity. With the growing availability and popularity of opinion-rich resources such as online review portals and social media, new opportunities and challenges arise as people can now actively use social mediums to not only share their opinions but also read others' opinions. Therefore, monitoring social platforms is very essential for a temple's feedback management system. By extracting and analyzing the comments from social media platforms, such as Twitter and Facebook, the management can easily identify visitor's pain points and take immediate action to rectify them. In this paper, we develop a systematic approach to extract users' opinions about ISKCON Bangalore from social media and then perform sentiment and emotion analysis to understand visitor's concerns.

A Fuzzy Model to Predict the Basic Cause of Sleep Disorder in Elderly

Richa Pandey
Asst. Professor, Department of Mechanical Engineering
B.I.T. Mesra Ranchi
Email:richapandey@bitmesra.ac.in

Abstract

The troubling sleep disorder in elderly is a major cause of distress and harmful after effects. To predict sleep disorder causes and symptoms and to make their life smooth a study has been conducted and the fuzzy tool is utilized. A primary survey of the elderly who are coming regularly to the hospitals generally for blood pressure, diabetes, depression, stroke and anxiety etc. is done. They are been examined through normal gestures, their way of responding, and the hospitality they wish to get. The care givers who are accompanying them are also questioned and surveyed. The major cause of many diseases has been found to be sleep disorder, sleep apnea, walking at night etc. The different parameters of distress for different elderly are varied and never same. Several causes have been analyzed and some are still too complicated to be handled or known even. A fuzzy Quality Function Deployment (QFD) approach has been analyzed and proposed in this paper to filter needs and focus on the demand of elders to give them Healthy Happy Years (HHY) ahead.

Key Words: sleep disorders; elderly; healthy happy years (HHY)

Predicting All-Cause Hospital Readmissions from Medical Claims Data of Hospitalised Patients

Avinash Kadimisetty, Arun Rajagopalany and Vijendra SK
Evive Software Analytics Pvt. Ltd., Bengaluru, Karnataka, India
Email:avinash.k@goevive.com, arun.rajagopalan@goevive.com, vijendra@goevive.com

Abstract

Reducing preventable hospital readmissions is a national priority for payers, providers, and policymakers seeking to improve health care and lower costs. The rate of readmission is being used as a benchmark to determine the quality of healthcare provided by the hospitals. In this project, we have used machine learning techniques like Logistic Regression, Random Forest and Support Vector Machines to analyze the health claims data and identify demographic and medical factors that play a crucial role in predicting all-cause readmissions. As the health claims data is high dimensional, we have used Principal Component Analysis as a dimension reduction technique and used the results for building regression models. We compared and evaluated these models based on the Area Under Curve (AUC) metric. Random Forest model gave the highest performance followed by Logistic Regression and Support Vector Machine models. These models can be used to identify the crucial factors causing readmissions and help identify patients to focus on to reduce the chances of readmission, ultimately bringing down the cost and increasing the quality of healthcare provided to the patients.

Key Words: Hospital Readmission, Comorbidity, Risk, Classification, Random Forest, Support Vector Machine

Student Specific Smart Question Recommender

Shashank P
Dept. of Computer Science and Engineering
National Institute of Technology Karnataka,
Surathkal, Mangalore, India
Email:shashankp5424@gmail.com

Praveen Kumar Gupta
Dept. of Computer Science and Engineering
National Institute of Technology Karnataka,
Surathkal,Mangalore, India
Email:pvgupta24@gmail.com

K Chandrasekaran
Dept. of Computer Science and Engineering
National Institute of Technology Karnataka,
Surathkal,Mangalore, India
Email:kch@nitk.ac.in

Abstract

A common difficulty that a student faces when it comes to studying is personalised attention and gradual improvement in the learning process. To test one's knowledge and thoroughness in a subject, students take tests and mock exams. These tests and mock exams help the student to analyse and realise his / her strengths and weaknesses. But these tests and mock exams do very less in directly trying to improve the student's knowledge in the subject. Sometimes these mock tests can be either very easy or too difficult.

To solve this problem of biased (too difficult or very easy) set of questions in tests, mock exams or any practice paper, a naive approach is to look at the performance of students after the test, mock exam or practice exam is done. Then based on the performance, set a new question paper accordingly. But this naive way is not efficient because the result of the student's performance will be known only after finishing the test, mock exam or practice question set. This process of selecting questions for the test, mock exam or practice question set from a question bank can be automated.

To solve this problem and improve efficiency of selecting the right questions, the selection of questions can be made dynamic. In dynamic question selection, the questions are selected or recommended while the student is attempting the test, mock exam or practice question set.

In this paper, we have proposed improvements and modifications to three existing reinforcement learning algorithms to build this dynamic personalised question recommendation which will help students to enjoy the learning process while still being challenged.

Key words: multi-armed bandit; epsilon greedy; softmax; upper confidence bound (UCB); Thompson

Disease Predictive Modelling using Machine Learning in Healthcare: A Systematic Mapping Study

Samyak Jain, Mishal Shah
Department of Computer Science and Engineering,
National Institute of Technology Karnataka Surathkal,
Mangalore, Karnataka.
Email:16co254.samyak@nitk.edu.in
Email:shahmishal1998@gmail.com

Abstract

The paper illustrates the approach, methodology and results from a Systematic Mapping Study (SMS) for Disease Predictive Modelling using Machine Learning for Healthcare Industry. The research comprised of studying and analysing 50 Studies in the Healthcare Industry. The various Predictive Modelling techniques that are currently prevalent in industry and academia/ research are analyzed and presented. Also, methods to improve accuracy of Prediction have been discussed, also the factors that impact the prediction modelling and its accuracy are highlighted. Our research indicates that there are varying opinions and that there isn´t a single classifier (model) that comes across as a best-in-class process for disease prediction. Various Studies have offered varying techniques and performance results, but all have concluded that Doctor opinion and analysis cannot be replaced by an algorithm or a set of algorithms and that learning systems need continuous trained or reinforced for better accuracy

Key words: Prediction Modelling, Machine Learning, Prediction Accuracy, Systematic Mapping Study (SMS)

Improving Efficacy and Efficiency of Restaurant Recommendation Mechanism

Saurabh Arora, Jaya Garg, Narayana D, Mallikarjuna Rao

Abstract

The aim of this study is to research and develop a robust and dynamic recommendation system for restaurant selection mechanism. It is observed that current recommendation system used within various popular mobile and web applications are static in nature and are not representative of users' preference.

It is observed that current recommendation system used within various popular mobile and web applications are static in nature and are not representative of users' preference. The project is focused towards understanding the current methodologies in the available options and analyzing the gaps and challenges in the current scenario and mitigating the identified gaps to cater to customers' dynamic palettes.

We created a hybrid Recommendation Model, using R and Python, featuring three Recommendation types (Collaborative Filtering Recommender System, Content-Based Recommender System and Knowledge-Based Recommender System) to ensure the robustness of the recommendation system. User experience or preferences and Restaurant experience over time may vary, Static Recommendation Models fail to capture the degree of change and its impact, thus affecting the efficiency of the model. We captured the trend of the similarity among the Users and similarity among Restaurants over time and incorporated the same in the model, thus ensuring the dynamic nature of Recommendation model.

Key words: Recommendation System, Dynamic, Robust, Similarity, Reviews, Restaurant

Secure Framework for Information Centric Network over IoT Architecture

P Vimala Rani, Nagarathna Ravi, S Mercy Shalinie, K Rajkumar
Computer Science and Engineering, Thiagarajar College of Engineering,
Madurai,India
Email: vimalainfotechhgmail.com, rathnaravi2013@gmail.com, shalinie@tce.edu,
rajkumarkmdu@gmail.com

Abstract

The Internet of Things (IoT) allows billions of physical devices, that are connected to the internet to collect and share data. Along with rapid growth of IoT devices, the security issues also keep growing rapidly. The attackers are able to easily hack in and modify data or even bring the whole network down by various flooding attack leading to Distributed Denial of Service (DDoS). Information centric network (ICN) is a promising internet architecture which is capable of providing secure IoT applications. ICN is reliable approach which natively supports in-network caching, name-based routing, mobility and efficient data delivery model for IoT. In this paper, cognitive approach is deployed in ICN-IoT framework to provide security for IoT devices. In this approach, learning based technique is applied in ICN router to detect the anomalous activities and it is capable of adapting to the dynamic IoT environment. The experimental results shows the effectiveness in terms of increase in detection accuracy and satisfaction ratio of 96.20% and 95% respectively**.**

Key words: IoT, ICN, flooding attack, caching, routing

Intelligent Legal Advisor

Ankit Barsainya
Senior Associate, Cognizant Technology Solutions Pvt Ltd
Pune, India
Ankit.Barsainya@cognizant.com

Manish Naik
Senior Associate, Cognizant Technology Solutions Pvt Ltd
Pune, India
Manish.Naik@cognizant.com

Himmat Patil
Senior Associate, Cognizant Technology Solutions Pvt Ltd
Pune, India
Himmat.Patil@cognizant.com

Abstract

Citizens are usually unaware of the laws, procedures and amendments (e.g. in case of domestic LPG cylinder explosion, consumers are entitled up to 40lakh) hence, they don't have clarity on the strength of their case. Also, there is limited knowledge on expected expenditure, procedures and the associated hassles (Court cases often take ~10 years) before filing a case. This could potentially disrupt an individual's life.

In this paper, we propose Artificial Intelligence based solution that provides realistic approximation of all the expenses and procedures that needs to be followed and the timeline for judgement (more than 30 million cases are still pending) along with the chances of favorable ruling. Such a system will be beneficial in multiple ways. First, Citizens will be well aware of the processes and procedures involved before moving to court, reducing time wasted owing to procedural illiteracy. Second, drastic reduction in the manual research times. Third, cost savings for legal counsel in the initial stages. Fourth, and the most important benefit, end user can prepare themselves mentally and monetarily keeping in mind the nature of the case and expected timeline for resolution. An indirect benefit will be reduction of the burden of Courts as citizens might refrain from filling unnecessary cases and also complete documentation before moving the court rather than moving the court and then looking for relevant documents.

Using Machine learning, we need to build a model that can refer to all the relevant laws, latest amendments, regulations along with all previous court cases that are relevant to current one. This model will study all the above documents, case history and come up with key pointers that had the highest impact on the final judgement. This will be tallied against the key points in current case to predict the case outcome hence enabling the citizens to help with DSS (decision support system) that will reduce logger head of court proceedings. The proposed solution will be to build a dashboard with a 360-degree view of legal proceedings. It will also incorporate chat-bot to help the end user with relevant queries.

Data will be ingested by retrieving historical cases from lower/high/supreme courts and central repository for law maintained by Govt. of India. This information will be consolidated, parsed and stored in a local repository. Machine learning algorithms like Clustering, NLP & Neural Networks will be implemented on this information to predict the outcome

Key words: NLP, Machine Learning, Artificial Intelligence, Decision Support System, Indian Law, Text Analytics

Analyzing Wind Speed to generate Electricity by Wind Turbines

M Barath*, K Suhasini*, V Aswathi*, Vipin Kumar*, S Suvetha*, Dr. D Sudha Devi#

*I year M.Sc(Data Science), #Associate Professor
Department of Computing
Coimbatore Institute of Technology
Coimbatore, Tamilnadu, India

barath1845@gmail.com, suhasinikrishnaswamy699@gmail.com,
aswathiviswanathan21@gmail.com, vpnk999@gmail.com, suvetha1811@gmail.com,
sudhadevi@cit.edu.in

Abstract

Wind turbine plays a crucial role in the generation of electricity. Wind turbines have to be constructed at the right place where the entire wind energy can be utilized in an ideal way. Based on the local wind speed, the amount of electricity a wind turbine could generate can be estimated. The objective of this study is to analyze wind power energy which could help to detect whether wind turbines can be set up in a particular location or not. Analyzing wind speed data at different locations helps to predict the wind power generation by wind turbines. The data available over various locations are considered for this analysis. The annual mean is calculated and is stored as mean wind speed and are analyzed to assess the wind power using Rayleigh Distribution method. The Rayleigh distribution is used to represent a probabilistic based model to estimate the wind power in a given region. Using the moving average method, the prediction for the future wind speed is to be done which would further help in the placement of the wind turbines.

Key words: Wind speed, Wind turbine, Rayleigh distribution, Wind power

Inclusive Education and Artificial Intelligence: A theoretical framework

Dr. Rajesh Panda
Professor, Symbiosis Institute of Business Management, Bengaluru,
Symbiosis International (Deemed University)
Email: rajeshpanda.80@gmail.com, director@sibm.edu.in

Abstract

In the era of Artificial Intelligence (AI), Universities may not be confined to their physical boundaries. Quality education can reach to masses with the help of technology where Universities can develop AI enabled educational tools to impart relevant education to masses. This will make education more inclusive and less expensive. AI can simulate university like atmosphere virtually where students can synchronously attend classes, interact with professors and discuss with other students in real time.

AI based platforms can be exploited to connect educators/universities, Students and employers to make education more relevant. Skill based education can also be imparted through simulations. Assessment of students and personalized coaching can be imparted with the help of machine learning algorithms.

AI will help good universities to reach masses beyond the constraints of their physical classrooms, eliminating the ever pervasive substandard education by multiple local educational institutes. This will make education more affordable, relevant and inclusive geographically and economically.

The paper describes current issues in higher education because of the limited reach of the good quality Universities and proposes AI based Models for adoption by quality Universities/Educational Institutes. This paper deals with the theoretical frameworks of creditable education and narrates the roe of AI in scaling up and making education more inclusive.

Application of Artificial Intelligence in Cognitive Science

Sri Dhanuja R
II year M.Sc Decision and Computing Sciences
(5 year Integrated M.Sc), Coimbatore Institute of Technology

Abstract

In this paper we propose to develop some cognitive science by using Artificial intelligence which could be useful for several organisations. Organisations main aim is to create surplus. In profitable organisations surplus determines profit. In non-profitable organisations surplus determines customer satisfaction, trustworthiness etc. Using Artificial intelligence, we can apply Cognitive science in business to increase profit. We can advertise products in best way and arrange stores to boost sales. One of our main topics will be decision support systems. Decision making process is very important in every organisations . Thus we have to consider the knowledge acquisition stage which is known as the bottleneck in the construction of these systems. We will present a model for learning strategies established by a decision maker for a feature task of categorical judgement of objects described by several attributes. It has been improved for several divisions of organisations to make decision like which sector contributes more, which sector has major problem etc. This Knowledge will be helpful for the organisations to make a decision and plan about advertisement of products and productivity to increase surplus (profit).

Key words: Artificial intelligence, cognitive science, organisation, surplus, decision making knowledge acquisitions

Preserving the Integrity of Big Data in Cloud: A Survey

E Angelin Kanimozhi

Department of Computer Science and Engineering Thiagarajar College of Engineering (TCE)
Madurai, India
angelinkanimozhi@gmail.com

M Suguna

Department of Computer Science and Engineering Thiagarajar College of Engineering (TCE)
Madurai, India
mscse@tce.edu

Dr. S Mercy Shalini

Department of Computer Science and Engineering Thiagarajar College of Engineering (TCE)
Madurai, India
hodcse@tce.edu

Abstract

Big data can be defined as extremely large datasets, which can be structured or unstructured, and difficult to process. The biggest challenge nowadays, is to store and manage big data. There are many technologies available for storing the big data. Storage infrastructures like DAS and NAS were initially used for storing big data. Data analytics solutions like Hadoop, Hive, NoSQL etc. played a major role in big data storage. Cloud computing technology used remote servers to store big data. On storing big data in a cloud, it is important to preserve the data integrity. Data integrity is the property of the data to remain unmodified in the storage location. The cloud users may be concerned about the integrity of their data. Hence the cloud authority provides some methods like Provable Data Possession (PDP) protocols, to guarantee their users about the safety of their data in the cloud. This paper compares several frameworks that have been proposed so far, to ensure the integrity of the big data in a cloud. The methodologies discussed in those frameworks and their limitations are analyzed in this paper. In the end, a new idea is proposed to address the same problem.

Key words: Cloud computing, data integrity, Provable Data Possession (PDP), blockless verification, batch auditing, Remote Data Possession Checking (RDPC), data dynamics

Self-Driving Cars using Markov Models

Shivesh Sinha, Harsh Pandey & Pankaj Rai

Abstract

Self-Driving Car or Autonomous Vehicle are categorized as vehicle that uses a sensor, cameras, artificial intelligence (AI) to travel between destinations without a human Interference. Companies like Audi, Uber, and Google etc. have their labs set up to test and create such Autonomous cars. But recently accidents involving Uber car getting crashed in Arizona shows some of the most catastrophic failures that have taken place. The accident that took place showed the basic failure of these autonomous cars as it was not able to recognize a pedestrian. This has led to a lot of questions of how safe are these autonomous vehicles and if they will not be dangerous once they are on road.

There are two important components of Autonomous Cars – hardware and software. Though we have seen breakthroughs on the hardware side as they are able to deliver mammoth computational power, software still serves as a huge bottleneck and we are still dealing with lot of challenges. Most scenarios currently being developed are for a specific scenario & condition. Our approach involves building a predictive model that adapts to different traffic scenarios. This paper is an attempt to nail the problem using a 'Hybrid approach' which brings more precision and accuracy in Software paradigm. Stochastic models are powerful tools for solving problems which are dynamic in nature, that is, values of random variables changes with time. We model the problem at hand with a collection of random variables indexed with time. It is a two forked problem – one, to predict the driver behavior of the surrounding environment and two, to guide the autonomous car to make a decision. This paper only focuses on prediction of driver behavior of the surrounding entities.We apply the 'Hidden Markov Models (HMMs)'to learn behaviors of surrounding entities. HMMs model will combine prior and posterior knowledge and improve the overall model.

The paper does not talk about pre conceived notions and development of Self Driving car. Neither does it talk about the success & failures of Autonomous Cars. This paper talks about a specific aspect of how AI & Big Data Space will bring more precise software & decision making for Autonomous Vehicles. It combines techniques & information from Radar & Cameras and will use Big Data & Artificial Intelligence to carve out an Intelligent space for Autonomous Vehicle It tries to resolve aspect like Drive Control, Decision Making, Analytics and Data Segmentation. The implementation will bring a change in overall paradigm of Self Driving vehicles.

Key words: Autonomous Cars, Hidden Markov Models, Hybrid Approach

Analysing Lifetime of Components using Gamma Distribution

**K Varshini, S Sudendar, Arjun Kumaran, S Abitha Shree, P Barathi Priya &
Dr. M Marimuthu**

Department of computing
Coimbatore institute of technology
Coimbatore, Tamil Nadu, INDIA

Varshini99.vv@gmail.com, sudendarsude@gmail.com, arjunkumaran126@gmail.com,
subhi.abitha@gmail.com, pbharathip17@gmail.com, marimuthu@cit.edu.in

Abstract

In today's world all the daily needs is connected with electronic gadgets. Electronic gadgets are becoming one of the essential needs. From fans to video gaming everything involves gadgets. Electronic gadgets have made the work of human easier. So, it is important to know the reliability and lifetime of the components that we use. This paper deals with determining the lifetimes of electronic gadgets produced by two different manufacturers using gamma distribution. Two parameters alpha and beta are calculated. A hypothesis is formulated to check if the two means are equal. The result establishes which one among the two manufacturer produces more efficient products.

Key words: Gamma distribution, alpha, beta, hypothesis

Analysis and Prediction of Cancer Rates Before and After Sterlite

C K Ajay Balaji [1], Pujithaa Vijayanand [1], S Karthikeyan [1]
S R Prakatheeswari [1], R Akash [1], Dr. Velvadivu P [2]
1-Student, Department of Computing,
2-Faculty, Department of Computing,
Coimbatore Institute of Technology, Coimbatore

Abstract

This study is to analyse the hazards of Sterlite Copper Plant in Tuticorin region. The main focus of our study is the analysis of cancer rates in Tuticorin. The datasets are obtained with a scope for analysis by considering the rates of cancer causing agents, thus predicting the increase or decrease in the rates of the cancer-causing agents if the Sterlite Copper Plant was not closed in the coming years. The factor that we are taking into consideration is the rates of such agents found in water bodies during the years of 2007-2018. This is implemented by using the trend for the year 2017-18. The predictions are done using regression analysis by taking the factors as the cancer-causing agents and the year. The inferences would be whether the cancer rates have been influenced or has remained the same irrespective of the functioning of the Sterlite Copper Plant.

Key words: Cancer rates, Tuticorin, regression analysis

Fast and Accurate Training of an AI Radiologist on Intel Xeon-based Dell EMC Supercomputers

Lucas A Wilson, Vineet Gundecha, Srinivas Varadharajan, Alex Filby & Quy Ta
HPC & AI Engineering, Dell EMC
Email: (luke_wilson, vineet_gundecha, s_varadharajan, alex_filby, quy_ta)@dell.com

Abstract

The health care industry is expected to be an early adopter of AI and deep learning to improve patient outcomes, reduce costs, and speed up diagnosis. We have developed models for using AI to diagnose pneumonia, emphysema, and other thoracic pathologies from chest x-rays. Using the Stanford University CheXNet model as inspiration, we explore ways of developing accurate models for this problem with fast parallel training on Zenith, the Intel Xeon-based supercomputer at Dell EMC's HPC and AI Innovation Lab. We explore various network topologies to gain insight into what types of neural networks scale well in parallel and improve training time from days to hours. We then explore transferring this learned knowledge to other radiology subdomains, such as mammography, and whether this leads to better models than developing subdomain models independently.

Key words: Deep Learning, Medical Imaging, Radiology, Distributed Training, Benchmarking, Best Practices

Machine Learning for Industrial Extension: Industrial Chatbots

Varshni Devi B, Mounika V
M.Sc Software Systems (Third year)
Coimbatore Institute of Technology
varshnibabu@gmail.com,
mounikavenkatesan1998@gmail.com

Abstract

Machine learning and artificial intelligence has transformed almost every industry. With the huge amount of data and applying machine learning models, we are able to get valuable insights which help in improved production rate and sales. Machine learning and robotics plays an important role and has created a huge impact in our workplace. Automation is being applied in industry to reduce human effort. Breakthroughs in machine learning help in creating natural language processing model. Chatbots is a computer program that communicates with the human via text or auditory. We try to solve industrial issues using chatbots which plays the role of an advisor. It helps us to take decisions regarding stock maintenance, machinery expansion, demand detection, cost cutting, production rate etc. Chatbots analyse these problems and helps us to make right decisions. Thus human biases are removed from the process. In our paper we analyse the data to optimize our decision in the industrial field.The analysis of data is done using machine learning and is used for each prediction. Humans are being replaced by chatbots as they work efficiently, avoid errors, faster response time, increase in accuracy, cost effective, higher volume production and most importantly it saves human time

Key words: Machine learning, data analysis, chatbots, artificial intelligence, robotics, industry

Imperfect Information Game of Contract Bridge using Double Dummy Bridge Problem

Dr. Dharmalingam Muthusamy
Assistant Professor,
Department of Computer Science,
Bharathiar University Arts and Science College,
Modakkurichi, Erode - 638104,
Tamil Nadu, India
Email: emdharma@gmail.com

Abstract

The game contract bridge is one of the most popular card games in the world wide. Bridge game is many interesting aspects such as bidding and playing, winning the number of trick including estimation of human hand strength. The decision made on several stage of the game is purely based on the decision- making that was made on the gradual preceding stage. The imperfect information bridge game is the real spirit of the card game in proceeding further deals. The elman neural network architecture with supervised learning implemented through resilient back-propagation algorithm and back-propagation algorithm to trained data was tested with bamberger point count method and work point count method. The research results reveal that bamberger point count method which was implemented in resilient back-propagation algorithm in elman neural network architecture yields better results than work point count method.

Key words: Elman neural network architecture, Resilient back-propagation algorithm, Double dummy bridge problem, Bamberger point count method, Work point count method.

FormAssist : Deep Learning Methods for Converting Handwritten forms into Digital Assets

Srivathshan KS
Senior Data Analyst PROBYTO Coimbatore, India
Email:srivathshan@probyto.com

Saurav Kumar
Data Analyst PROBYTO Guwahati, India
Email:saurav.k@probyto.com

Shreekanth R
Data Analyst PROBYTO Coimbatore, India
Email:shreekanth.r@probyto.com

Midhilesh E
Data Analyst PROBYTO Coimbatore, India
Email:midhilesh.e@probyto.com

Parvej Reja Saleh
Senior Data Analyst PROBYTO Guwahati, India
Email:parvej@probyto.com

Abstract

Customer agreement are required to follow statuary and legal requirements, which include agreements to be manually signed. In India, paper forms are still prevalent in Banking Industry. The paper forms require customers to fill a template form in capital letters and manually sign by agreeing to the terms. This creates challenge in analytical systems as the data is captured outside the system and requires time to become part of data pipeline. The future of banks is poised to be digital, however we still need historical data for train models for current data applications. This limitation is a known bottleneck in designing data applications for real time decision making. Developing Optical Character Recognition (OCR) with capabilities commensurable to that of human is still not achievable, in spite of decades of excruciating research. Due to idiosyncrasy of individual form, analysts from industry and scholastic circles have coordinated their considerations towards OCR. The work in this paper shows an efficient model to capture offline handwritten forms and convert them into digital records. The model techniques are based on deep learning methodologies and show higher accuracy for our testing set of real application forms of selected Banks. We have experimented with different feature extraction techniques to extract hand written characters in the forms. Our experimentation has evolved over time to find a generalized solution and better results. The final model uses relative position of the characters for extracting characters from the forms and Convolutional Neural Networks (CNNs) to predict the characters. The paper also discusses the serverless architecture to host the *FormAssist* as a REST API with model calibration feature to accommodate multiple types of forms.

Key words: Handwritten Forms, Optical Character Recognition (OCR), Deep Learning, Convolutional Neural Networks (CNNs), Serverless Architecture, REST API

Rainfall Distribution using Joint Probability

Deivarani R
deivarani.cta.cit@gmail.com
Raghul T
Nandhish C
Tharaneeshraj C S
Navaneeth R
Baraneedharan V K

Abstract

This project investigates the combined influence of rainfall and extreme temperatures. The estimation of rainfall may depend on the temperatures and the main focus of our study is to analyse the relationship between temperature and rainfall. The inference would be whether temperature and rainfall are related or not.

AirSense: Smart air Quality Monitoring and Reporting tool using IoT Devices and Cloud Service

Parvej Reja Saleh
Senior Data Analyst PROBYTO Guwahati, India
Email:parvej@probyto.com
Akashjyoti Banik
Data Analyst PROBYTO Guwahati, India
Email:akashjyoti.b@probyto.com
Debalina Banerjee
Data Analyst PROBYTO Guwahati, India
Email:debalina.b@probyto.com
Roshan Kumar Gupta
Data Analyst PROBYTO Guwahati, India
Email:roshan.g@probyto.com
Srivathshan KS
Senior Data Analyst PROBYTO Coimbatore, India
Email:srivathshan@probyto.com

Abstract

In urban areas, exposure to indoor air pollution is expanding because of numerous reasons, including the construction of more tightly sealed buildings, reduced ventilation, the use of synthetic materials for building and furnishing and the use of chemical products, pesticides, and household care products. Indoor air pollution can start inside the building or be attracted from outside. Other than nitrogen dioxide, carbon monoxide, and lead, there are various different toxins that influence the air quality in an encased space. The most susceptible groups to indoor pollution are women and children because women lung sizes are significantly smaller to male counterparts and lung volume to body volume proportion of children is significantly higher than adults. Indoor air pollution monitoring requires equal attention as outdoor pollution. With advent in sensor technology and studies showing harmful effect of indoor air pollution it is important for us to start monitoring the air quality inside our schools, offices, hospitals, home and other places. The nature of air is influenced by multi-dimensional elements including area, time, and unverifiable factors. As of late, numerous specialists started to utilize the big data investigation approach because of headways in big data applications. Sensors build on powerful Arduino board and wi-fi networking units are tested to monitor air quality of three parameters; suspended particles, organic vapors and humidity. These key parameters are monitored over period of time, the time series data is stored in cloud service, and machine learning is applied to find ways to predict and manage air quality. The paper presents the IoT device architecture, cloud application architecture and sample results for an indoor test environment. Mobile and web-based visualizations were created for the data collected from the sensors. An alarm system is also developed to notify the user when the air quality deteriorates to unhealthy level.

Key Words : ndoor Pollution, IoT, Big Data, Machine Learning, Arduino, Visualization, Mobile Application, Gas Sensors

Rainfall Data Analysis using Beta Distribution

**Jeeva K G, Ashvin R, Thejasvin B M, Adnan A, Samuel Naveen V,
Mrs. M Srividhya**

Abstract

Analysation of Rainfall data to check whether it fits to Beta distribution or not. The Rainfall data set taken into account can provide various knowledge depending on users wish. We were able to calculate average monthly rainfall for each year and also the proportion of months in each year where rainfall is above monthly average.The mean and variance of the distribution is calculated along with the estimates of two parameters. It is found out that the Rainfall data obtained fits to the Beta distribution perfectly.

Key Words: Beta distribution, proportion, Mean, Variance and Parameters

Predicting Employee Attrition using Neural Network

Yaswanthan A*, Sridhurkesh S,
MSc (Decision and Computing Science) (5 years Integrated Course)
Coimbatore Institute of Technology, Coimbatore

Abstract

HR attrition is a major problem for any industries (manufacturing or services), retaining key talent and identifying the people who need to be improved was tough task. Losing top key talent and filling them with equivalent talent was paramount challenge and would cause organization a huge loss in terms of time and cost.

To provide solution to this problem, we propose a neural network model to predict the person who most likely to move out of the organization from the set of employee data. The employee data consist of lot of independent variables and decision variable is attrition (Yes/No). With machine learning techniques such as RANDOM FOREST, Neural Network model would be built, this model would be a handy support tool to the HR manager to identify the person who likely to move out of the organization. Based on the impact and insight from this predictive model, insights can be drawn and implement actions to improve the mindset of the person in question. This paper would discuss about the data, data classification, neural network model such as RANDOM Forest, accuracy of the model built and insights drawn based on study.

Key Words: Neural Network, Random forest, predictive model, HR Insights

Stock Market Prediction using Sentimental Analysis in a LSTM Model

Maithreyan K , Mounika V
Msc Software Systems
Coimbatore Institute of Technology
Email:suryamaithreya@gmail.com
Email:mounikavenkatesan1998@gmail.com

Abstract

Machine learning and deep learning are growing at a rapid speed and it is transforming almost every industry. Stock market is also one of them ,Predicting Stock price is considered to be a great challenge ,but with the power of machine learning applied on various data points, the stock price can be predicted. The price of stock depend on various factors that what makes the prediction challenging. We consider the people sentiment about the company from twitter and other social media. We also take input about the various news about the company. Previous price of the stock is also taken as feature .Stock market price prediction is a time series problem so we are using a recurrent neural network a LSTM model. Price of a stock depends a lot upon its price of stock on previous timestamp, so using a LSTM model increase the accuracy. The model combines both people sentiment and stock market trend and provides valuable insight about the stock price. Thus helping us to make decision to buy and sell the right stocks at the right time.

Key Words: Stock price prediction, Machine learning, Sentimental Analysis, LSTM

Is Trading Volume an Appropriate Indicator to forecast the Stock Index?

Dr. T Viswanathan
SIBM, Bengaluru

Prof. Kannadas S
SDMIMD, Mysuru

Abstract

Forecasting the share price is one of the most common decision to be taken by traders and investors. Investors use tools such as fundamental and technical analysis available to forecast the share price. However, these tools are arbitrary to some extent and cannot be tested for its forecasting accuracy. The accuracy of trading may be improved by using a model that provides unbiased estimates of future price. We apply predictive analytics tools to forecast Sensex by using trading volume. We apply Bai and Perron (1998) multiple break point test to determine the structural stability of stock volume and price for the period of observation. Both the models of pure and partial structural breaks are considered and run linear regression model to forecast the market index and checked for model fitness. The nature of joint association of all variables on the price is interpreted by using F-Statistic. We then apply ARIMA model to forecast the index. Residual diagnostics is applied to examine the model fitness. We found ARIMA model provides best forecasting results.

Key Words: Trading volume, Share price, Forecasting, ARIMA, Causal relationship

Distributed Data Analysis with Docker Swarm using R

Aishvarya Sivaram [1], Surya Balamurugan [2], A Kannammal [3], P Aruna [4]
M.Sc Software Systems [1] [2], Professor [3], Assistant Professor [4]
Department of Computing
Coimbatore Institute of Technology, Coimbatore 641014.

Abstract

The Objective of the research work is to propose Docker in R Platform for Distributed Data Analysis with Docker Swarm. Analysing large volume of dataset is a big challenge for Data Analysts which can be overcome by utilizing Docker swarm. Docker swarm is the clustering technique which when combined with UNIX tools is used to dissect information in appropriated volume and schedule jobs on nodes across the cluster. Docker is a container based virtualization which runs map reduce applications on different versions of Linux machine modifications. It also gives reproducible strategy to catch the code setting and make it accessible for later use. This research work is demonstrated for performing Twitter sentimental analysis swarm across swarm nodes. Statistical computation and analysis of data can be done by developing R project within a Docker container which in turn is known as Rocker.

Key Words: Data Analysis, Virtual Machines, Docker swarm, Rocker, R

Customer Perception in Hotel Selection and Feedback Mechanism

Aayushi Kalraa

Research Associate, Indian Institute of Management, Bangalore

Abstract

The hotel industry is witnessing a boom due to increase in number of tourists visiting our country. Thus, in order to woo more customers, there is a need to understand how customers make choices for hotel selection. The amount of differentiation among the hotels is so narrow that it becomes very difficult for tourists to select the most appropriate hotel. It is of great significance for tourists to select residence location in different countries. A complex decision making process is involved in selection of hotel. The current study is aimed to study the different factors that influence the customers to select a hotel. Also, feedback mechanism is developed to identify the factors for the success of a leading corporate hotel in our country.

Key Words: Hotel selection, Factor analysis, ANOVA

Comparison of Dynamic Bayesian Model and Unified Face image Approach using Surveillance Videos

Mrs. A D Chitra [1], Ms. N Priyadharshini Jayadurga [2]
Email: chitracitcta@gmail.com, Assistant Professor, Department of Computing,
Coimbatore Institute of Technology,
Coimbatore-641014, India
Email: priyadharshinijayadurga@gmail.com, MSc Software Systems student,
Department of Computing,
Coimbatore Institute of Technology, Coimbatore-641014, India

Abstract

The economical growth of our society is forcing our lives to be frightened one with increased robbery and theft. The objective of our project is to identify a face from the surveillance video which would help in identifying the person's face. The camera apps in smartphones and digital cameras using image processing to enhance the image quality, video stabilization and noise removal. A Face can be even recognized when they are partially obstructed from view. Image processing allows the use of much more complex algorithms, and hence, can offer both more sophisticated performance at simple tasks, and the implementation of methods which would be impossible by analog means. Face recognition is still a challenging problem in the field of computer vision. It has received a great deal of attention over the past years because of its several applications in various domains. Although there is strong research effort In this area, face recognition systems are far from ideal to perform adequately in all situations from the real world.The software is implemented to recognise human faces using Dynamic Bayesian Network and Unified Face Image approach and obtaining the optimal result. The use of Dynamic Bayesian Model approach models interdependent entities that changes over time. The Unified Face Image is an approach which is generated from the data of surveillance cameras, the different lighting condition in the original frames will introduce non-uniform lighting in UFI's. It may happen that in some cases latter may work better than former approach and vice versa.

Key Words: Surveillance video, face recognition, Dynamic Bayesian Model, Unified Face Image, Interdependent entities, video stabilization, enhance image quality

An Effective Approach of Extracting Local Documents from the Distributed Representation of Text using Doc2vec and LSA

Vikas Chib
Applied Machine Learning
Infineon Technologies
Bangalore, India
Email:Vikas.chib@infineon.com

Ahsan Jafri
Applied Machine Learning
Infineon Technologies
Bangalore, India
Email:Ahsan.jafri@infineon.com

Abstract

All neural embedding models learn distributed representation of text and match the results in the latent semantic space on a given query, but searching documents from the distributed representation will lose the relevance of local representation of a given query. We propose a novel information retrieval system, which uses doc2vec model to give top N similar documents with a relevant ranking using Latent Semantic Indexing to give the top K (documents score is greater than a soft threshold) documents which are the local representation of given query. We can use these K documents to find the most similar ones.

We can show that this 'dual' combination performs better than other traditional information retrieval algorithm or recently developed neural network models

Key Words : Information Retrieval, LSA (Latent Semantic Analysis), Doc2Vec, Embedding

Leveraging Predictive Analytics to Reduce Non-value adding activities in Brick and Mortar stores via use of Smart Cart

Manav Jain
Student
Prin. L. N. Welingkar Institute of Management Development and Research, Mumbai
Mumbai, India
Email:manavjain76@gmail.com
Tanisha Singh
Student
Prin. L. N. Welingkar Institute of Management Development and Research, Mumbai
Mumbai, India
Email:tanishasingh2109@gmail.com
Rohit Kunnath
Student
Prin. L. N. Welingkar Institute of Management Development and Research, Mumbai
Mumbai, India
Email:rhtsmn93@gmail.com
Pranjal Metange
Student
Prin. L. N. Welingkar Institute of Management Development and Research, Mumbai
Mumbai, India
Email:pranjalmetange@gmail.com

Abstract

The conventional shopping culture of purchasing at mom and pop stores, departmental outlets, kirana and brand stores has been diminishing due to the growth of e commerce. Around 2005, several online shopping outfits emerged to cater to the needs of young and tech savvy consumers who were parched on time and energy to shop physically. However, the brick and mortar stores have lost to competition from ecommerce mainly due to factors such as cost, time and need based display of products. It would be of interest to understand if integration of technology into brick and mortar stores to address these 3 factors through the use of smart carts can induce a shift in the consumers who prefer the ecommerce model for shopping.

The study will be conducted in two phases. Phase 1 of the study engages to identify various nonvalue adding activities confronted in the store by the staff through the method of observation and in-depth interviews on 60 store managers and staff members. Stores with 2000 sq ft of area have been considered as sample of the present study. Phase 2 of the present study would involve devising a model based on predictive analytics that can be used to develop a smart cart which will eliminate non-value adding activities identified in phase 1. The study intends to develop a smart cart model that would assist in improving conversion rate from mere selection to purchase, effective advertising, instore experience of the customers and effective shelfing of the products.

Key Words : Smart cart, Predictive analytics, Non value adding activities, Conversion funnel, Retail, In-store Navigation

Machine Learning in Embedded System

Vijay Natarajan
Program Manager
Embedded Product Design
TATA ELXSI Ltd Bengaluru, India
Email:vijayn@tataelxsi.co.in

Abstract

This paper discusses the general opportunities available using the technologies options to port any Machine Learning in Embedded System.

Key Words : Machine Learning, Embedded System Algorithms, Advanced Technologies in semiconductor

Natural Language Processing/ML techniques for Life Sciences Companies

Ch. Yugandhar
Department of Business Analytics
Vignana Jyothi Institute of Management
Hyderabad, TS, India
Email:yugandhar.c@vjim.edu.in

Abstract

This paper presents the success story, findings and learnings related to the implementation of a system that classifies conversations among medical professionals into various categories. A thorough scrutiny of source data, processes, and technological methods was conducted before the system was developed with the intent to implement in production with a well-established maintenance and support infrastructure.

Key Words : Natural Language Processing, Logistic Regression, Random Forest, Production implementation, Machine Learning

Solving Complex Production Scheduling Problem with Machine Learning - A Case Study

Arup Ray
Senior Director (Analytics & Digital ERP)
Cognizant Technology Solutions

Abstract

In this paper author investigates emerging trend of machine learning replacing proven industrial optimizer & heuristics to bring much needed flexibility and adaptability to dynamic business environment. While this optimizers & heuristics had been very successful in replacing manual intuition and experience based approach, the key challenges of this 'black-box' approach is lack of flexibility and adaptability to change in requirements and dynamic business environment. In this case study, a European component manufacturer was using manual scheduling of the moulds leading to significant down time & productivity loss due to suboptimal usage of Press & Moulds. The typical industrial optimizer is expensive & not able to cope up with dynamic demand supply situation of the polymer component manufacturing industry. This paper explores the solution approach using Machine Learning for the mould scheduling leading to optimal usage of moulds & pressing machines during manufacturing of molded components achieving significant reduction in machine down time, increase in productivity & reduction in cost of production.

Key Words : Production Scheduling, Scheduling Algorithms, Optimization, Heuristics, Machine Learning

Automated Catalog Management and Image Quality Assessment using Convolution Neural Networks and Transfer Learning

Mani Garlapati and Souradip Chakraborty

Abstract

Catalogue management is a very important aspect in the field of ecommerce as it helps the visitors in efficiently selecting the necessary interest items. In every retail website, all the items in the catalogue are in a particular order of different categories. In this work, we have developed an entire pipeline where the first task to automatically classify the various orientations (front view, side view, top view etc.) of the images sent by the vendor using Transfer learning. In the second part of our pipeline, we have eased the process of catalogue management with the image quality assessment of the vendor images using Image processing and Transfer Learning. Finally the automatic ordering of items is done as per requirements.

Key Words : Convolution Neural Networks, Transfer Learn- ing, Image quality assessment, Structural similarity index

Community detection based on DNS querying patterns

Aruna Chakkirala
Solutions Architect
Infoblox
Bangalore, India
achakkir@gmail.com, achakkirala@infoblox.com

Abstract

DNS servers capture all the domain querying activity initiated by various source IP addresses. A single user querying a domain has no meaning beyond the purpose it serves. A collective of such calls from the user may reveal something about the user while an analysis of multiple users and their querying patterns can provide useful insights. This data is represented as a weighted undirected graph between the source IP addresses. The nodes of the graphs have edges if and only if they have queried for a common domain. The edge weight is another important factor which quantifies the number of mutually queried domains and hence the affinity between the two nodes. The available data from the DNS logs was transformed into a weighted undirected graph and fed into a community detection algorithm. The community detection was conducted using the Louvain modularity premise, the output looks very promising as it provides a grouping of nodes based on the input graph. This leads to an understanding of similarly behaving source IP addresses.

Key Words : DNS, querying pattern, community detection, modularity

Supplier Selection using Artificial Neural Network & Fuzzy Set Theory

Amit Kumar
Senior Associate, Business Consulting Group
ITC Infotech India Ltd.
Email:amit1509@gmail.com

Abstract

Suppliers are an important part of the supply chain. Selection of appropriate suppliers often becomes the key to success of the supply chain. Thus the evaluation and selection of suppliers becomes a very important activity in any company. Supplier selection is a well-known multi criteria decision making (MCDM) problem. In this paper we propose the use of Artificial Neural Network (ANN) to solve this MCDM problem. Fuzzy set theory has been used to generate the training data set for the ANN. First the criteria's that will be used to assess suppliers is decided upon. Next the importance associated with each criterion is decided and then the performance of each supplier on each criterion is decided. Due to the subjectivity involved in human assessment, it becomes difficult to represent the assessment in pure numbers. At the same time it is quite easy to represent the assessment in terms of linguistic variables. Hence we have chosen Fuzzy set theory to generate the training data set for the ANN, as it allows the decision makers to express their preferences in linguistic terms. These linguistic terms are then converted into fuzzy numbers by using fuzzy membership functions. Fuzzy mathematical operators are then used to determine a fuzzy score for each supplier. These fuzzy scores are then translated into crisp scores. We use the linguistic performance (Poor, Good, Very Good, and Excellent) of the supplier on the various criteria and the crisp score of the supplier for training the ANN. We see that the trained neural network is able to predict the crisp score for a new supplier with a high accuracy. ANN has been used for a regression problem.

Key Words : neural network; supplier selection; fuzzy logic

Psychographic Segmentation Based on Online Reviews

Prof. Lalit K Ojha
IMT Nagpur

Abstract

Online customer reviews are great insights into customer thought process. Prior studied have used online reviews to understand sentiment etc. We looked at TripAdvisor reviews for two properties of same hotel chain to confirm whether the heritage and business travelers accord varying priorities to services. Training and Validation was performed using three different statistical analysis techniques were used to ascertain suitability of one technique over another. The results show that SVM has better prediction capability for classification of results. We also validate the non-overlapping aspect of heritage and business hospitality segments.

REDCLAN - RElative Density based Clustering and ANomaly detection

Diptarka Saha
Walmart Labs

Debanjana Banerjee
Walmart Labs

Bodhisattwa Prasad Majumder
Walmart Labs

Abstract

Cluster analysis and Anomaly Detection are the primary methods for database mining. However, most of the data in today's world, generated from multifarious sources, don't adhere to the assumption of single distribution as their source — hence the problem of finding clusters in the data becomes arduous as clusters are of widely differing sizes, densities and shapes, along with the presence of noise and outliers.Thus We propose a relative KNN kernel density based clustering algorithm. The un-clustered (noise) points are further classified as anomaly or non-anomaly using a weighted rank based anomaly detection method. This method works particularly well when the clusters are of varying variability and shape, in these cases our algorithm can not only find the "dense" clusters that other clustering algorithms find, it also finds low-density clusters that these approaches fail to identify. This more accurate clustering in turn helps reduce the noise points and makes the anomaly detection more accurate.

Key Words: Clustering, Relative KNN − kernel density, Varying density clusters, Anomaly Detection, DBSCAN

Mobile Wallet Fraud –How Machine Learning Solution and Big Data can help Fraud Prevention and Losses

Kavita Dwivedi
Head Data Science, Infinite Sum Modeling

Abstract

One of the growing industries in last 2 years, thanks to Digital India Campaign and of course demonetization, has been the mobile wallet companies and their business. In the last half a decade or so , multiple mobile wallet companies have been growing some at revolutionary pace and some steadily , but what is alarming that the rate at which Fraudulent transactions have grown is much higher than at which the transaction rate has increased.

Digital payments have shown a consistent growth trend in the month of April as well against the number of transactions clocked in March, Reserve Bank of India data show. While cards and IMPS transactions have maintained the similar trend, mobile wallets which had fallen in March 2018 have shown a smart pick up in April 2018, having grown 11% in terms of number of transactions. According to RBI data, in April wallets showed 279 million transactions against 268 million in the previous month. This, however, is still lesser than 310 million transactions wallets clocked in February, the month before full KYC guidelines were enforced by the RBI. Further last year April, wallets had seen more than 320 million transactions, almost 15% higher.

This paper aims at building a robust end to end solution for Fraud Detection and Prevention (FDP) using Machine Learning algorithms. Fraud Detection is challenging as the moment you bring in a robust system to prevent fraud in any FI, the fraudsters work faster to identify the weaknesses or circumvent around its security and continue to commit fraud. Machine Learning algorithms by design are meant to learn from past data and make the system better as it is being used. Continuous usage of this ML algorithms coupled with a Fraud Bureau where multiple industries /companies share fraud list can go a long way in preventing and minimizing fraud losses. We will present and discuss the outcomes of different Machine Learning Algorithms. The paper will also talk about how block chains can revolutionize this industry by giving the power of individual data security in customer's hand.

The paper will also show how use of Big data in cloud environment can help detect and alert the highly suspicious transactions real time thus stopping this transactions to occur. This will require using R, Python, Hadoop, Hive and Spark to meet the requirement of mining petabytes of data in real time. It will also discuss how linking mobile transactions to Awadhi and Biometric security can be beneficial.

The paper will not only aim to look at different types of Fraud , but also suggest how payment services companies can work together with Banks , Other Fintech providers , Telecoms , Healthcare and Insurance to build a strong Fraud Bureau.

This will not only help the company's reputation and help them to grow their business, it will also instill a lot of confidence in the consumer to use the mobile payment services without the fear of being duped.

Key Words: Fraud Management, Mobile Wallet, Machine Learning, Big Data, Payments Fraud, Blockchain

Sentimental Analysis on Twitter Data

B.Ramyaashri
S.Srihareni
S.Subathra
Department of Computing
Coimbatore Institute of Technology

Abstract

The aim of this project is to build technology to detect and summarize an overall sentiment and to fit trend line and time series forecasting for sentiments. Using twitter data, we build models for classifying "tweets" into positive, negative and neutral sentiment. The data for this project is collected on the basis of daily tweets that are based on the political ratings between actors Rajini and Kamal. By comparing all the tweets between Rajini and Kamal, we classify tweets into positive, negative and neutral sentiment and fit a trend line for these sentiments. By this we also infer who has the largest audience.

Key Words : Tweeter, Positive sentiment, Negative sentiment, Neutral sentiment, Trend line, Time series forecasting, Unigram model, Feature based model, Tree kernel based model

Detecting Fraud in Healthcare Insurance Claims

Nidhi Sagar
Data Scientist, DXC Technology

Abstract

Only a fraction of all healthcare claims are fraudulent but using data insights to flag them can help healthcare insurance providers and the U.S. government save millions and tighten their processes.

The United States now spends about $2.6 trillion annually on healthcare (17.5% of GDP). Healthcare fraud is a national problem, prevalent in federal, state and private insurance programs, and billions of dollars have been spent on improper payments over the last decade. The National Health Care Anti-Fraud Association (NHCAA) conservatively estimates that 3 percent of all healthcare spending, or $60 billion, is lost to healthcare fraud. The Federal Bureau of Investigation (FBI) has estimated fraudulent billings to healthcare programs, both public and private, at between 3 percent and 10 percent of total healthcare expenditures.

These statistics illustrate the magnitude of the fraud and abuse prevalent in the system today. By saving money on fraudulent claims, insurance premiums can be reduced, and better quality of healthcare can be provided to those who are in legitimate need.

Data analytics can be a powerful tool that insurers and the government can use to help detect fraud in healthcare. Analytics can be used to identify and review patterns of suspicious behaviour, evaluate data quality, and create and exploit data models to prevent future overpayments or any kind of fraud, waste and abuse.

This paper examines Medicaid fraud from the provider's perspective and proposes a mechanism to detect fraudulent claims before pay-out for additional scrutiny. Identifying potential fraudsters before a claim is made will aid in pre-emptive action rather than post-pay-out analysis and help save millions of dollars for the U.S. government. It could also help in lowering healthcare premiums that continue to rise sharply year over year.

Key Words : Fraud, Healthcare, Analytics, Medicaid

Machine Learning and Artificial Intelligence: Trends and Future Directions

M Hariharan*, AV Karthick**
*Alagappa Chettiar Government College of Engineering and Technology, Karaikudi
**Alagappa University, Karaikudi
(hariharan119494@gmail.com, avk.mba23@gmail.com)

Abstract

Machine learning is one of the recent emerging research areas in computer science and statistics. It is a decision making approach that applied in various domains like social media, health care, finance etc. Machine learning is an interdisciplinary field that integrates Artificial Intelligence, Big data, Cloud Computing and Information Theory. The main cache objective of machine learning is based on the previous experience. Machines are self learned and it predicts the future without error. In Big Data and Cloud Computing the volume of data sizes are very larger, which are processed with low consumption of power with the help of Machine learning. This paper mainly focuses on the current trends and future directions of Machine learning and Artificial intelligence is discussed. Both technological trends will give better results for future generation.

Key Words : Machine learning, decision making, self learned, low consumption.

Customer Intent Based Recommendation System

Mani Garlapati
Souradip Chakraborthy
Sunil Kumar Potnuru

Abstract

A system to provide customer intent recommendations by leveraging voice of the customer like user reviews/ feedback from various platforms (e.g. social media, retail portal etc.). The system uses deep learning and hierarchical clustering to categorize the reviews/feedback from the customer into intent categories by understanding the context and customer sentiment and recommend items from intent categories. Currently retailers are not leveraging customers voice from different channels to make personalized recommendations based on their intent or need. Our algorithm helps to solve this problem by leveraging machine learning and deep learning techniques to identify customer interest categories and interest items in each category. Accordingly, these intent items can be added to the cart based on customers interest.

Key Words : LSTM, Collaborative Filtering, Cosine Similarity, Customer Interest, Context Matching

Loyalty Customer Detection and Personalized offers Using age and Gender Prediction Implemented using Dlib, Keras, Raspberry Pi and Apache Kafka

Sreekiran A R

Abstract

This paper presents an approach to find the loyalty customers in brick and mortar stores without compromising their identity, by assigning them system generated ID's and provide them with personalized offers, suitable for their age and gender using camera based data. Here, a raspberry Pi camera is used as input sensor which can be placed at the bill counters, where it sends high quality video to a high performing server via Apache Kafka [1], where the algorithm checks for the detected faces and see if they match the training data, if match is not found, append the training data with a new ID, the customer's age and gender also will be predicted and the data along with timestamp is sent to a Kafka server and from there to an SQL database which can be used for visualization. The same algorithm can be applied for personalized digital signage, can be used to track people in the store by linking multiple cameras. Thus, this solution has an impact on customer retention, smart advertising and people tracking, also it helps the supply chain for smart inventory management by getting insights using age-gender bucket data.

Key Words : Computer vision, age and gender prediction, operationalizeIoT, Apache Kafka, facial detection and tracking, transfer learning, Raspberry Pi, deep learning, retail solutions

Predictive Analysis on Diabetes using Supervised Machine Learning Algorithms

Roopasri K and Bharanidharan A

Abstract

Diabetes mellitus (DM) results due to insulin deficiencies, which in turn lead to chronic hyperglycemia with disturbances of carbohydrate, fat and protein metabolism. The existing testing system requires a multiple lab assessments and it is a tedious process. Machine learning and computational intelligence techniques play a vital role in transforming healthcare which provides objective decision support tools to assist medical professionals in diagnosis and prognosis the patient conditions. Smart machine learning algorithms are used now-a-days in different industries to replace the costly, repetitive and time consuming tasks. They also capture unforeseen patterns within the complex data set at must faster rate which would have not been seen by human eye and brain. In the present work, pima Indians diabetes data set is taken and employed in the three different machine learning algorithms namely Support Vector Machine (SVM) algorithm, K-Nearest Neighbour (KNN) and Naïve bayes. KNN algorithm is found to have the best accuracy and most suited algorithm. The data set are further analysed using graph maker - plotly and is observed that around 20-30 years of age groups found to have high glucose level and preventive measures to be taken in the young age itself.

Key Words : Diabetes, KNN, Machine learning, Naïve bayes, Supervised learning, SVM

Neural Network Based user Input Recommender for Retrieval-Based Conversational Interfaces

Luckyson Khaidem
Amadeus Software Labs
Bangalore, India
luckyson.khaidem@amadeus.com

Nitin Gupta
Amadeus Software Labs
Bangalore, India
nitin.gupta@amadeus.com

Hari Bhaskar Sankaranarayanan
Amadeus Software Labs
Bangalore, India
hari.sankaranarayanan@amadeus.com

Abstract

This paper discusses the quality issues faced by retrieval-based chat-bots and proposes a solution to improve conversational quality and overall satisfaction for end users. The problem with most chat-bots comes from the inability of the underlying intent classifier to correctly classify every user input text. Since all retrieval-based chat-bots are trained on a limited number of intents, it fails on open-ended conversations. As a result, the best way for such bots to succeed is to maintain the conversational focus around the areas in which the bot understands. The proposed solution revolves around the provision of guidance at every step of the conversation. The guidance is in the form of a recommendation model that attempts to guess the intention of a user as he types first few words by suggesting a list of relevant utterances. The solution makes use of similarity measurement between two pieces of text driven by multiple methods such as Siamese RNN and cosine similarity using word2vec embedding. The recommendation model prioritizes the relevancy of its suggestion by using the conversational context which is modeled as state transitions and past conversations. A conversational agent with the recommendation engine resulted in substantial increase in the average probability estimate of all intents correctly classified i.e. a more accurate conversational agent.

Probability Based Estimator for Estimation of Zero-Inflation Parameter

K M Sakthivel
Department of Statistics
Bharathiar University
Coimbatore - 641046
Tamilnadu, India
Email : sakthithebest@gmail.com

C S Rajitha
Department of Statistics
Bharathiar University
Coimbatore - 641046
Tamilnadu, India
Email : rajitha.sugun@gmail.com

Abstract

In field of research, modeling of count data plays significant and inevitable role. In modeling count data, the assumption of symmetric and normality is not fulfilled due to excess numbers of presence of certain values in the data. Most of the time, the frequency of zero is very high hence the data is over dispersed. Since the traditional standard count models such as Poisson and negative binomial distribution will not provide better fit, hence the researchers started using zero inflated models with an additional parameter called zero inflation parameter to deal with such type of count data. Zero inflation parameter represents the proportion of excess number of zeros in the data which highly influence the accuracy and efficiency of the zero-probability models. In this paper, we have introduced an estimator for zero inflation parameter of zero-inflated negative binomial (ZINB) distribution based on the non zero probabilities and compared the performance of the proposed estimator via a simulation study.

AMS Subject Classification (2010): 62F10; 62P05

Key Words : Zero-Inflated Models, Zero-Inflation Parameter, Maximum Likelihood Estimation, Zero- Inflated Negative Binomial Distribution, Simulation, Mean Square Error.

Pattern Recognition for Stock Index Prediction Using Apriori Algorithm - The Case of NSE NIFTY

S Thirupparkadal Nambi & M V Subha

Abstract

Predicting stock market performance and identifying the influence of various factors on stock prices is an interesting as well as a rewarding activity. The ability to predict its direction will enhance the decision making capability of individual investors, institutional investors and regulating bodies. Hence financial managers always look for more sophisticated tools for prediction. This study is the application of machine learning technique to understand the trend (Bull/Bear) of the popular Indian stock market index NSE NIFTY. It is an attempt to understand the relationship among sectoral indices -which are exclusive performance indicators of various sectors of the industry- and the broad market index NSE NIFTY and to mine association rules using Apriori algorithm. The generated rules would be of immense help for the investors and financial managers to identify the patterns in the sectoral indices that results in the next day's up and down movement of NSE NIFTY.

Key Words : Apriori algorithm, Association rule mining, Predictive analysis, Stock Index Prediction, Stock market data mining

Toxic Pollution Alert using Fog computing and Hybrid Machine Learning Model(RAQ and ARIMA)

D Sudaroli Vijayakumar
Department of Information Technology, Alliance University, Bangalore, India
sudaroli.d@alliance.edu.in

Abstract

Pollution remains as a hot topic discussed globally irrespective of various measures and techniques adopted to improve the air quality. The level of pollution in many cities still exceed the limits of world health organization thus contributing various deadly diseases like cancer, stroke, heart disease and bronchitis. We find lot of compromises in the way the toxic alerts are generated as the alerts are generally generated based on the data that we obtain from the large scientific air monitoring system that are permanently installed at a specific location. The amount of data that these monitoring systems provides is relatively small and thus fatal toxic smog like

occurrence prevails. Toxic smog and similar occurrences can be effectively addressed if we possess a large dataset from multiple sources. Advancements in communication technology creates an opportunity for us to integrate several sources of data through Internet of Things(IoT). This big data can be analyzed effectively with the aid of machine learning algorithms to provide a better recommendation on the toxic level in air. Along with the big data analytics, introduction of fog computing can improve the latency in data. The aim of this paper is to integrate the machine learning approaches along with the fog computing concepts to provide a toxic pollution alert

Key Words : fog computing, cloud computing, IoT, Machine Learning, ANN, RAQ, SVM

Deep AutoEncoder Recommendation Engine using Novel Dense Representation

Anant Gupta
Anant.Gupta@morganstanley.com

Abstract

Traditionally SVD (Singular Value Decomposition) has been used for collaborative filtering. However SVD has certain limitations. It can identify only linear relationships and if the data is very sparse, SVD is not able to approximate the matrix very well. The current methodology for deep collaborative filtering attempts to replicate SVD by using neural networks. This methodology is able to capture nonlinear relationships between a user and an item. But even with this approach, the sparsity of data still exists and when we apply neural networks on sparse data, the chance of overfitting is very high.

To prevent this, we alter the user item matrix form to a simple tabular representation, where each row denotes the interaction between a single user and a single item. Those items which have no interaction with any user need not to be fed into the neural network. This provides a novel way to feed only useful information into the network. This data representation helps us effectively train the neural network with better accuracy.

Additionally this gives us a way to club similar users (and also similar items) into groups. This is because we are not using the autoencoder directly to generate the recommendation but Instead using it to have separate latent representation of users and items using user item interactions. This latent representation can also be used to find similar users and similar item.

Key Words : Deep Learning, Artificial Intelligence, Market Making, Portfolio Construction, Investment Banking, Recurrent Neural Networks, Auto Encoders, Modern Portfolio Theory

Performance Evaluation and Analysis of Tensor Processing Unit against CPU and GPU

Dhileepan Thangamanimaran*
M Sharat Chandar
S Chandia
1 III Year M.Sc. (Software Systems), Department of Computing,
Coimbatore Institute of Technology, Coimbatore, India.
2 III Year M.Sc. (Software Systems), Department of Computing,
Coimbatore Institute of Technology, Coimbatore, India.
3 Assistant Professor, Department of Computing, Coimbatore Institute of Technology,
Coimbatore, India.
*Corresponding author Email: dhileepan123@gmail.com.

Abstract

Tensor Flow is an Open Source Software Library used for Machine Learning Applications. Machine Learning demands profound computation power due to its heavy workloads. Architecture of CPU and GPU restricts it from being an efficient platform for processing of Tensor Flow applications. Tensor Processing Unit abbreviation TPU are tailored specifically for Tensor Flow and Machine Learning Applications. TPUs are said to be 30 times faster than GPU and much more energy efficient. TPUs power Google Datacentres since 2015. This Paper provides a brief comparative study of CPU, GPU and TPU on deep learning models using Tensor Flow Library. We aim to access the performance of the above processing units and determine its scope in real life scenarios and the economic vs efficiency comparison.

Key Words : TPU, TensorFlow, Machine Learning

Real-Time Disease Prediction based on Weather Data:
A Case on Life Sciences Analytics

Garima Makkar
Senior Busines Analyst
Tata Consultancy Servies
Bangalore
Email: garima.makkar@tcs.com

Abstract

Weather is often considered as an unmanageable factor, which has direct consequences on human mortality rates, mental injury, physical health and other health outcomes. Unstable and warm climate plays a major role in driving the global emergence, regeneration and redistribution of communicable diseases like dengue. There is a worldwide pandemic about dengue, which is known to be the most dominant arthropod-borne infection in humans. According to 1995 World Health Organization (WHO) Report, around 50 million of dengue infection cases occur globally every year. One of the important factors which contribute to the spread of dengue is climate change. Events like rainfall, humidity, temperature etc. have well-defined role in the transference cycle. Any changes in these events can lead to increase in incidence of this disease.

The impact of global epidemic on mankind facilitates the need for developing early warning systems (EWS) on infectious disorders with respect to the climate change. Past studies in this context takes only the historical weather statistics into account. However because of increasing incidence and climate variability, these traditional systems are likely to get outstripped. In this paper, a new methodology is proposed to predict the number of dengue cases that are likely to occur on the basis of real time weather data. Our analysis is universally applicable, and enables comprehensive scenarios of daily dengue cases to be explored using real-time weather API, enabling dengue control measures to be effectively targeted, timed and implemented.

Key Words : Weather, Real-time, Disease, Dengue, Early Warning System, Climate, Forecast

AI-powered Counterfeit Products Detection Solution for Consumer Brands

Anshul Garg
Kumar Shubham
Sanket Patil
Karthik Bettadapura

Semantics Dept., Dataweave Pvt. Ltd., Singapore
Email:anshul.garg@dataweave.com
Email:shubham@dataweave.com
Email:sanket@dataweave.com
Email:karthik@dataweave.com

Abstract

With the rise in e-commerce adoption by consumers worldwide, the sale of counterfeit products on open marketplaces is an increasingly challenging problem faced by shoppers, online retailers, and consumer brands. Recent surveys have reported that over a third of shoppers have had counterfeits delivered to them on placing an order online. These defective products can have a severe impact on the brand value of retailers and consumer brands, damaging consumer trust and potentially having an adverse impact on sales volumes. If one considers the large amount of products being sold over retailer websites, the problem of identifying counterfeits only magnifies.

In this paper, we propose an efficient and robust method of detecting counterfeit products at scale using computer vision and deep learning applied on online product catalog data. The approach entails identifying small variations between two images. Essentially, the original brand manufacturer's image on its own website is compared with images listed on online marketplaces by third party sellers to detect variations, which indicate whether a product listing is legitimate or not. We use pre-trained Convolutional Neural Network (CNN) based models to take advantage of transfer learning and further fine tune them on internal data to focus on fine grained image features. In addition, we use several image processing and matching techniques based on image signatures, key points, and descriptors to achieve better accuracy.

DataWeave's Counterfeit Products Detection solution is powered by datasets that we have built in house over the years, consisting of millions of products collected from thousands of retail websites across geographies. The dataset has hierarchical information pertaining to retail taxonomy.

Detecting counterfeits on e-commerce websites where the volume of data can be in the range of terabytes is a major challenge. Our technology platform efficiently stores data pertaining to millions of processed images in Internet Archive's ARC files and maintains the indexes in lucene search engines. The entire pipeline is automated and connected via big data technologies like Kafka and orchestrated across servers using Celery.

Key Words : Counterfeit detection, CNNs, Transfer learning, Kafka, Machine Learning, Image signatures, Solr

Fraud Detection using Dual Distance Method

Dinker G Mattam
Analytics Data Labs
DXC Technology
Bangalore, India
dinker.g.mattam@hpe.com

Abstract

Fraud is a significant problem faced by companies across the world. In securities trading, fraud can lead to major losses and regulatory sanctions. Detecting fraud from millions of transactions through manual effort is next to impossible. Machine learning provides a scalable alternative through automation or partial automation of the fraud detection process.

This paper discusses a novel unsupervised anomaly detection technique to identify fraud transactions. We present a distance-based method, which we believe has advantages over other related conventional techniques, and is faster than density based methods. It is a semi-automated process, which leverages on domain knowledge to choose the input attributes. We also add an ensemble transformation to capture various types of outliers. We implement the system on a real world dataset and identify suspicious trades. We think the technique is suitable for fraud detection in a wide range of applications not limited to financial services domain.

Key Words : fraud; anomaly; outlier; detection; machine learning; unsupervised; distance; clustering; automation

The Amazon's A9 Algorithm

Mr. R RAJKUMAR [1], Ms. N PRIYADHARSINI JAYADURGA [2],
Ms. R KALAIVANI [3], Dr. P. ARUNA[4]
[1] rajkumarrajagopal7@gmail.com, MSc Software Systems student,
Department of Computing,
Coimbatore Institute of Technology, Coimbatore-641014, India
[2] priyadharshinijayadurga@gmail.com, MSc Software Systems student,
Department of Computing,
Coimbatore Institute of Technology, Coimbatore-641014, India
[3] kalaivani251197@gmail.com, MSc Software Systems student,
Department of Computing,
Coimbatore Institute of Technology, Coimbatore-641014, India
[4] aruna2023@gmail.com, Assistant Professor, Department of Computing,
Coimbatore Institute of Technology,
Coimbatore-641014, India

Abstract

In the Amazon's website the products are displayed based on the ranking. The highest ranked product is displayed first in the Amazon's website. This ranking method is known as A9 Algorithm which is the search engine and search advertising technology. It is a search algorithm in which the products are ranked initially. A9 Algorithm delivers result through a two-step process:- First, they pull the relevant results from their massive "catalog" of product listings. Then, they sort those results into an order that is "most relevant" to the user. The products are ranked based on by considering 25 factors. The study on the existing system reveals that the initially given word in the search bar is first matched with either the products name or its brand. If the given search key is found matched then the relevant products are displayed. Else the product that match the keyword are displayed and those keyword are added as a new keyword to the products displayed while searching. When a product is searched rank is added to the corresponding product that is being searched which is done dynamically. The method used in ranking a product are Customer reviews, pricing, title, description and feedback. Initially the products are displayed based on categories. When the customer logs in for the second time the most frequently searched products are displayed in the screen. There is no age limiting factor which restricts the display of adult products. In the proposed system, we would enhance the existing system by introducing age limiting factor duringsign up. This enhancement promotes safe search amongst younger generation.

Key Words : Search Engine, Ranking algorithm, A9 algorithm, Limiting factor

Fault prediction in aircraft engines using Regularized Greedy Forests

Akshata
Kishore Moharir

Abstract

Aircraft engines maintenance is very challenging and costly task. The main objective is to ensure a proper operation of the aircraft engines, in all conditions, with a zero probability of engine failure, while taking into account aging of the aircraft. Aircraft maintenance can be improved if an efficient procedure for the prediction of failures is implemented. Several variables such as the core speed, the pressure , the fan speed of the aircraft , etc. are measured and taken in to consideration while predicting the faults , together with environmental variables such as the outside temperature, altitude, aircraft speed, etc. In this paper we demonstrate the application of regularized greedy forest algorithm which directly learns decision forests via fully-corrective regularized greedy search using the underlying forest structure on an aircraft engine data involving failures of aircraft engines. The prediction accuracy obtained with the Regularized Greedy Forests presents a significant improvement over the state-of-the-art ensemble classifiers. Moreover, the performance of three ensemble classifiers with different characteristics – Random Forest, Regularized Greedy Forest, and XGBoost- was compared in terms of their prediction accuracy. We achieved higher accuracy and smaller models using regularized greedy forests on jet engine data set compared to state of the art ensemble methods.

Key Words : Aircraft engine maintenance, Fault prediction, Random Forest, Regularized Greedy Forest, XGBoost, Normalization, and Recursive Feature Elimination

Intelligent IT – Driving IT through A2I

Vinod Sundararaju Antony
Kavitha Rajendran
Ankur Joshi

Abstract

With the data explosion, both structured and unstructured, each organization is striving towards gathering meaningful insights to either enhance the productivity of their employees (insights from organizational data) or to strengthen their competitive position in the market by providing customer centric products and services. IT services is a potent use case for such a scenario as there is a wealth of data developed in the IT lifecycle across development, quality assurance (QA) and production (live). The IT resources (people, monitoring tools) consume data at each stage in an isolated manner and generate insights which are focused on enhancing that particular phase (development, QA or production incidents) of the IT lifecycle. The paper aims at finding an integrated and collaborative AI-ML (artificial intelligence-machine learning) solution which can break these silos and provide a comprehensive and intelligent solution driven by the principle of *A2I – Automate, Augment and Invent*, aimed at improving both employee productivity and customer centricity of an IT organization.

Key Words : Artificial Intelligence (AI), Machine Learning (ML), Automation

Understanding Importance of Artificial Intelligence from the view of Retail Industry

Prof. Dhruv Brahmbhatt
Faculty of Management GLS University Ahmedabad, India
dhruv.brahmbhatt@glsuniversity.ac.in

Prof. Jaineel Shah
Faculty of Management GLS University Ahmedabad, India
jaineel.shah@glsuniversity.ac.in

Abstract

The retail and consumer goods industry faces substantial challenges, endangered by a demand for personalisation, an unoptimized business process constrained by human error and an inability to fully anticipate customer demands. These problems leave consumers unfulfilled in the search for an experience that reflects their continually developing desires. Then you have technology, like Artificial Intelligence which presents both a challenge and the key solution to facing these challenges. This paper identifies the key importance of the Artificial Intelligence and understand the significance of the Technology to enhance the growth of Retail Industry. Various expert gives their varied view on the application of Artificial Intelligence in business applications and concluded the positive effect of it. Artificial Intelligence can be considered for Automated Checkout, Sales Forecasting, Optimising Energy Usage, Churn Rate Minimisation and IoT Instore analysis for the enhancement in to the retail industry. Overall Artificial Intelligence is the future for the Retail Sector.

Key Words : Artificial Intelligence, Retail Industry, Consumer Behaviour, Business Intelligence

Thank You