IV. INTERNATIONAL CONFERENCE ON DATA SCIENCE AND APPLICATIONS 2021 (ICONDATA'21)

PROCEEDINGS BOOK

Volume 2

Abstract Book (DRAFT)

PREFACE

In the current information age, data is the basis of all intelligent systems. From the agricultural society to the industrial society, from there to the information society, the transformation process is moving towards Industry 4.0.The ability to store and process large amounts of data on computers has led to an increase in the capabilities of the products and services produced. The statistical and artificial learning studies based on meaning deriving from data have paved the way for intelligent systems in all sciences.

4th International Conference on Data Science and Applications (ICONDATA’21) has been organized on June 4 - 6, 2021 as online.

The main objective of ICONDATA'21 is to present the latest data based researches from all disciplines. This conference provides opportunities for the different areas delegates to exchange new ideas and application experiences face to face, to cooperate between different disciplines from both natural and social sciences and to find global partners for future collaboration.

All paper submissions have been blind and peer reviewed and evaluated based on originality, technical and/or research content/depth, correctness, relevance to conference, contributions, and readability.

20 selected papers presented in the conference that match with the topics of the journals will be published in Data Science and Applications and Veri Bilimi Dergisi.

Looking forward to see you in ICONDATA 2021,
Dr. Murat GÖK
Editor

COMMITTEES

Honorable Committee

Prof. Dr. Suat Cebeci	Rector of Yalova University
Prof. Dr. H. Tamer Dodurka	Rector of Istanbul Rumeli University
Prof. Dr. Şükrü Beydemir	Rector of Bilecik Şeyh Edebali University
Prof. Dr. Hüseyin Çiçek	Rector of Muğla Sıtkı Koçman University

Conference Chair

Prof. Dr. Murat Gök	Yalova University

Organizing Board

Assoc. Prof. Dr. Emre Dandıl	Bilecik Şeyh Edebali University
Assoc. Prof. Dr. Hüseyin Gürüler	Muğla Sıtkı Koçman University
Asist. Prof. Dr. Faruk Bulut	Istanbul Rumeli University
Hasibe Candan	Yalova University
Melike Bektaş	Istanbul Rumeli University
Aykut Durgut	Balıkesir University
Abdullah Yavuz	Istanbul Rumeli University

Science Board

Prof. Dr. Abderrahmane Bouda	National Maritime Superior Institute, Algeria
Prof. Dr. Abdullah Uz Tansel	Baruch College, New York, USA
Prof. Dr. Ahmet Sabri Öğütlü	Harran University, Turkey
Prof. Dr. Ayhan İstanbullu	Balıkesir University, Turkey
Prof. Dr. Ayşe Çetin	Yalova University, Turkey
Prof. Dr. Dursun Aydın	Muğla Sıtkı Koçman University, Turkey
Prof. Dr. Latif Taşkaya	Muğla Sıtkı Koçman University, Turkey
Prof. Dr. İlhan Tarımer	Muğla Sıtkı Koçman University, Turkey
Prof. Dr. Ecir Uğur Küçüksille	Süleyman Demirel University, Turkey
Prof. Dr. Erhan Akyazı	Marmara University, Turkey
Prof. Dr. Eray Can	Yalova University, Turkey
Prof. Dr. Erman Coşkun	Bakırçay University, Turkey
Prof. Dr. Gufran Ahmad Ansari	Crescent Institute of Science & Technology, India
Prof. Dr. Hacer Gümüş	Kocaeli University, Turkey
Prof. Dr. Kaka Shahedi	Sari Agricultural Sciences and Natural Resources University, Iran

Prof. Dr. Kamel Harchouche	University of Sciences and Technology of Houari Boumédiene, Algeria
Prof. Dr. Mustafa Öztaş	Yalova University, Turkey
Prof. Dr. Müfit Çetin	Yalova University, Turkey
Prof. Dr. Nour El Islam Bachari	University of Sciences and Technology of Houari Boumédiene, Algeria
Prof. Dr. Ramazan Bayındır	Gazi University, Turkey
Prof. Dr. Sinan Şen	Yalova University, Turkey
Prof. Dr .Osman Çakmak	Istanbul Rumeli University, Turkey
Prof. Dr. Mine Aksoy Kavalcı	Yalova University, Turkey
Prof. Dr. Şakir Taşdemir	Selçuk University, Turkey
Prof. Dr. Tamer Kahveci	University of Florida, Florida, USA
Prof. Dr. Vilda Purutçuoğlu	Middle East Teknik University, Turkey
Prof. Dr. Naci Genç	Yalova University, Turkey
Prof. Dr. Kadriye Tuzlakoğlu	Yalova University, Turkey
Prof. Dr. Zuhal Oktay Coşkun	Izmir Democracy University, Turkey
Assoc. Prof. Dr. Demet Aydınoğlu	Yalova University, Turkey
Assoc. Prof. Dr. Can Coşkun	Izmir Democracy University, Turkey
Assoc. Prof. Dr. Murat Yabanlı	Muğla Sıtkı Koçman University, Turkey
Assoc. Prof. Dr. Abdülkadir Tepecik	Yalova University, Turkey
Assoc. Prof. Dr. Ferhat Sayım	Yalova University, Turkey
Assoc. Prof. Dr. Mehmet Selçuk Mert	Yalova University, Turkey
Assoc. Prof. Dr. Akın Özçift	Celal Bayar University, Turkey
Assoc. Prof. Dr. Deniz Kılınç	Celal Bayar University, Turkey
Assoc. Prof. Dr. Demet Aydınoğlu	Yalova University, Turkey
Assoc. Prof. Dr. Eyüp Akçetin	Muğla Sıtkı Koçman University, Turkey
Assoc. Prof. Dr. Hasan Erdinç Koçer	Selçuk University, Turkey
Assoc. Prof. Dr.M. Gökhan Genel	Yalova University, Turkey
Assoc. Prof. Dr. Nevin Güler Dinçer	Muğla Sıtkı Koçman University, Turkey
Assoc. Prof. Dr. Orhan Kesemen	Karadeniz Teknik University, Turkey
Assoc. Prof. Dr. Ali Hakan Işık	Mehmet Akif Ersoy University, Turkey
Assoc.Prof. Dr. Hamir Erdemi	Yalova University, Turkey
Assoc. Prof. Dr. Ufuk Bal	Muğla Sıtkı Koçman University, Turkey
Assoc. Prof. Dr. Gamze Yüksel	Muğla Sıtkı Koçman University, Turkey
Assoc. Prof. Dr. Ümit Ünver	Yalova University, Turkey
Assoc. Prof. Dr. Adnan Taşgın	Atatürk University, Turkey
Assoc. Prof. Dr. Züleyha Özer	Balikesir University, Turkey

Senior Assist. Prof. Dr. J. Amudhavel	VIT Bhopal University, India
Assist. Prof. Dr. Serhan Mantoğlu	Yalova University, Turkey
Assist. Prof. Dr. Abit Balin	Istanbul University, Turkey
Assist. Prof. Dr. Levent Civcik	Konya Technical University, Turkey
Assist. Prof. Dr. Adem Tuncer	Yalova University, Turkey
Assist. Prof. Dr. Uğur Bekir Dilek	Yalova University, Turkey
Assist. Prof. Dr. Nida Gökçe Narin	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. Alper Kürşat Uysal	Eskişehir Teknik University
Assist. Prof. Dr. Aytaç Pekmezci	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. Ayşe Özlem Mestçioğlu	Istanbul Okan University, Turkey
Assist. Prof. Dr. Kubilay Ovacıklı	Istanbul Rumeli University, Turkey
Assist. Prof. Dr. İrfan Kösesoy	Kocaeli University, Turkey
Assist. Prof. Dr.Enis Karaaslan	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. Mithat Çelebi	Yalova University, Turkey
Assist. Prof. Dr. Beste Hamiye Beyaztas	Bartın University, Turkey
Assist. Prof. Dr. Burcu Okkalıoğlu	Yalova University, Turkey
Assist. Prof. Dr. Duygu Yıldırım Peksen	Yalova University, Turkey
Assist. Prof. Dr. Erdoğan Camcıoğlu	Istanbul Rumeli University, Turkey
Assist. Prof. Dr. Eyüp Çalık	Yalova University, Turkey
Assist. Prof. Dr. Fahriye Zemheri Navruz	Bartın University, Turkey
Assist. Prof. Dr. Gözde Mert	Nişantaşı University, Turkey
Assist. Prof. Dr. Gül Yücel	Yalova University, Turkey
Assist. Prof. Dr. Güncel Sarıman	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. Güneş Harman	Yalova University, Turkey
Assist. Prof. Dr. Gürcan Çetin	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. Halit Karalar	Muğla Sıtkı Koçman University, Turkey
Assist. Prof. Dr. İsmail Kırbaş	Mehmet Akif Ersoy University, Turkey
Assist. Prof. Dr. M.Bahar Başkır	Bartın University, Turkey
Assist. Prof. Dr. Melih Ağraz	Middle East Teknik University, Turkey
Assist. Prof. Dr. Muhammed Kürşad Uçar	Sakarya University, Turkey
Assist. Prof. Dr. Murat Okkalıoğlu	Yalova University, Turkey
Assist. Prof. Dr. Naveed Islam	Islamia College University, Pakistan
Assist. Prof. Dr. Nazlı Güney	Istanbul Rumeli University, Turkey
Assist. Prof. Dr. Osman H. Koçal	Yalova University, Turkey
Assist. Prof. Dr.Ali İmran Vaizoğlu	Muğla Sıtkı Koçman University, Turkey

Assist. Prof. Dr. Sait Ali Uymaz	Selçuk University, Turkey
Assist. Prof. Dr. Salih Ergüt	Istanbul Rumeli University, Turkey
Assist. Prof. Dr. Serdar Neslihanoğlu	Eskişehir Osmangazi University, Turkey
Assist. Prof. Dr. Serdar Birogul	Düzce University, Turkey
Assist. Prof. Dr. Süleyman Uzun	Sakarya Uygulamal Bilimler University, Turkey

CONTENTS

Algorithm Overview and Design for Mixed Effects Models

Burcu KOCA[1]*, Fulya GOKALP YAVUZ[1]

[1]*Middle East Technical University, Faculty of Arts and Sciences, Department of Statistics, Ankara, TURKEY*

Abstract

Linear Mixed Model (LMM) is an extended regression method that is used for longitudinal data which has repeated measures within the individual. It is natural to expect high correlation between these repeats over a period of time for the same individual. Since classical approaches may fail to cover these correlations, LMM handles this significant concern by introducing random effect terms in the model. Besides its flexible structure in terms of modeling, LMM has several application areas such as clinical trials, genetics, neurosciences, economy, etc. However, the statistical inference procedure of the model may not always generate closed form solutions of the parameter estimations. Therefore, a large number of estimation techniques and computational strategies are adapted in LMM such as Expectation Maximization algorithm. Also, even the main inferential tool is likelihood method for the LMM, the implementation of the method may change depending on the data structure (balanced/unbalanced), covariance structure or the distributional assumptions. It is possible to see these methods in many different sources, but it is not always easy to see which one will be used in what kind of situations and in what direction the results will change. In this study, we systematically categorize these algorithms and compare them in terms of efficiency and time for longitudinal data sets.

Keywords: *Linear Mixed Model, Longitudinal Data, EM, Algorithm Design, Efficiency*

This work has been supported by Middle East Technical University Scientific Research Projects Coordination Unit under grant number AGEP-109-2019-10094

Predicting Monthly Streamflow Using a Hybrid Wavelet Neural Network: Case Study of the Çoruh River Basin

Mehmet Şamil Güneş [1]*, Coşkun Parim[1], Doğan Yıldız [1], Ali Hakan Büyüklü[1]

[1] Department of Statistics, Yildiz Technical University, Istanbul, TURKEY

Abstract

In this study, a hybrid model combining discrete wavelet transforms (WTs) and artificial neural networks (ANNs) is used to estimate the monthly streamflow. The WT-ANN hybrid model was developed using the Daubechies main wavelet to predict the streamflow for three gauging stations on the Çoruh river basin one month in advance, with different combinations of air temperature, precipitation, and streamflow variables, and their wavelet transformations. Four different hybrid WT-ANN models were generated and compared with four different conventional ANN models. The dataset was chronologically divided into training, validation, and testing data. The results indicated that the WT-ANN hybrid models performed better than the traditional ANN models for all three stations. Furthermore, the chronologically divided dataset was used to examine the effects of changes in hydrological data over time on model performance. In conclusion, model performances in the training period deteriorated during the validation and testing periods due to structural changes in the hydrological data.

Keywords: *Streamflow, Artificial neural network (ANN), Wavelet transform (WT), Air temperature, Precipitation*

* İletişim e-posta: msgunes@yildiz.edu.tr

The Psychological Impact of the Covid-19 Outbreak Among Infected Patients

Mehmet Tahir HUYUT[1]*, Süleyman SOYGÜDER[2]

[1]Erzincan Binali Yıldırım University, Department of Basic Medical Sciences, Biostatistics and Medical Informatics AD, Erzincan TURKEY

[2] Van Yüzüncü Yıl University, Faculty of Agriculture, Department of Biometrics and Genetics, Van, TURKEY

Abstract

The novel coronavirus (2019-nCoV) or severe acute respiratory syndrome (SARS-CoV-2) emerged as an epidemic and rapidly spread into a global pandemic. This epidemic and the measures taken to combat it negatively affected the mental health of societies and individuals. Although there may be high levels of anxiety, depressive and distress symptoms in the general population, some groups may be more vulnerable than others to the psychosocial effects of the pandemic. People suffering from the disease, those at high risk of infection; people with preexisting medical, psychiatric or substance use problems, and healthcare providers in particular, are at high risk of adverse psychosocial outcomes. This study focused on the psychological impact of the epidemic among those who suffer from the disease. For this purpose, a comprehensive literature review was conducted to determine the post-treatment anxiety-depression levels of infected individuals who were healed. The variables that were thought to affect the level of Anxiety-Depression were determined and the individuals were asked by phone call. With the confirmatory-factor analysis, basic components were determined, independent factors were found and the attributes of the variables were extracted. The relationship structure between the determined attributes was examined using the Multiple-Correspondence-Analysis. Anxiety-level was found to be highly correlated with post-discharge sleep disturbance, while mostly obese individuals and asthmatic patients had sleep disorders. The number of hospitalizations-days was the variable that increased the anxiety-level the most. While headache and asthma-presence were found to be highly correlated, these variables did not affect the level of anxiety. The variables most positively associated with depression level were cardiovascular disease, number of days of hospitalization, travel history and receiving O2 support. It was observed that diabetes, loss of smell and smoking increased the level of depression. In addition, it was observed that individuals with high fever after discharge received more oxygen support.

Keywords: *Covid-19, Anxiety, Depression, Multiple Correspondence Analysis, Factor Analysis*

Time Series Prediction using Dendritic Neuron Model Trained by a Robust Learning Algorithm

Ayşe YILMAZ[1]*, Ufuk YOLCU[2]

[1]Ondokuz Mayıs University, Faculty of Arts and Sciences, Department of Statistics, Samsun, TURKEY

[2]Giresun University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Giresun, TURKEY

Abstract

Time series prediction is a crucial problem encountered in a variety of fields. Although there are a great number of methods in the literature, they can be reviewed under two basic title as probabilistic and non-probabilistic methods. Especially, as computational-based time series prediction models, different kinds of artificial neural networks (ANNs) have been widely and successfully used in the literature. While some of them use additive aggregation function, some of them use multiplicative aggregation function in the structure of their neuron models. Dendritic Neural Networks, proposed in recent years, have also both additional and multiplicative neuron models, together. It is inevitable that the prediction performance of such an artificial neural network will be negatively affected by the outliers that the time series of interest may contain due to the neuron model in its structure. In this study, a robust learning algorithm is proposed for dendritic artificial neural network. The proposed robust learning algorithm uses Huber's loss function as a fitness function. The iterative process of the robust learning algorithm is carried out by particle swarm optimization, an artificial intelligence optimization algorithm. The performance of the dendritic artificial neural network trained with the proposed robust learning algorithm has been demonstrated by the analysis of different real-life time series and the analysis of the contaminated time series obtained by injecting different numbers and scales of outliers. The obtained results show that the dendritic artificial neural network trained by the proposed robust learning algorithm produces satisfactory predictive results in the analysis of time series with and without outliers.

Keywords: *Dendritic Neuron Model, Time Series Prediction, Particle Swarm Optimization, Robust Learning Algorithm, Huber's Loss Function*

Machine Learning Methods in the Prediction of Inpatient Length of Stay

Beste Kaysi[1], Ozgur Gumus[2]

[1]*Ege University Computer Engineering Department, Institute of Science and Technology*

[2]*Ege University, Department of Computer Engineering*

Abstract

Length of stay is defined as the time between admission to the hospital and discharge. Hospitals have a limited number of facilities depending on the number of beds and devices they have. Effectively managing the hospital stay of patients is very important in terms of sufficient and efficient use of limited resources. The short length of stay times in the hospital increases the rate of continuous use of the beds, providing treatment opportunities for more patients and increasing the earnings of the hospitals. However, the health of patients who are discharged too early without full recovery for profit may deteriorate over time, causing them to stay in the hospital for longer periods in the future. The long length of stay causes the patients in need of treatment not to be provided with health care and cause a decrease in the quality of the health service provided to the patients. For these reasons, the length of stay in the hospital is also regarded as an important indicator of the quality of healthcare services provided to patients, the efficient use of resources and the success of hospital management. In this study, the predictive performance of the decision tree, support vector machine, random forest, naïve bayes and gradient boosting machine learning algorithms were evaluated in the estimation of the length of stay of hospitalized patients using the data obtained from the Medical Information Mart for Intensive Care III (MIMIC-III) database. 23Thus; It is aimed to ensure the minimum waiting time, efficient use of beds and devices, and beneficial and sufficient treatment by regulating the patient flow by successfully predicting the length of the stay in hospitals.

Keywords: *Length of Stay, Machine Learning, MIMIC-III, Prediction*

Completion of missing temperature data using Adaptive Network-Based Fuzzy Inference System (ANFIS)

Okan Mert KATİPOĞLU[1*]

[1]Erzincan Binali Yildirim University, Faculty of Engineering, Department of Civil Engineering, Erzincan, Turkey

Abstract

Data loss can occur due to instrument failures, environmental factors, and changes in the measurement method during measurements of meteorological data such as temperature, wind speed, precipitation, and humidity. The completeness of temperature data, which is one of the main inputs of climatic and meteorological studies, is of great importance for the reliability of the study. In this study, it was aimed to complete the missing temperature data in the Horasan meteorology observation station using the Adaptive Network-Based Fuzzy Inference System (ANFIS). For this reason, Sarıkamış, Tortum, and Ağrı meteorology observation stations that are closest to the station and have the highest correlation coefficient were used as inputs to the ANFIS model. Monthly average temperature data between 1968 and 2017 (50 years) were used in the ANFIS model. In the established model, 80% of the data (1900 between 1968-2007) were used for training and 20% (476 between 2008-2017) for testing. In the ANFIS model, variables were tried by dividing them into sub-sets between 3 and 8. The most suitable ANFIS model was determined according to the error values and determination coefficients of the training and test results. As a result of the study, the model with 4 sub-sets, a hybrid learning algorithm, and 300 epochs was selected as the most suitable model.

Keywords: *Temperature, Missing Data, ANFIS, Horasan*

Real-Time Facial Expression Recognition Using Convolutional Neural Network

Hüseyin Gürüler [1], Mehmet Osman Devrim [2]

[1]*Muğla Sıtkı Koçman University, Information Systems Engineering*

[2]*Düzce University, IT Department*

Abstract

Facial expressions play an important role in how people communicate with each other. For this reason, facial expression recognition methods based on machine learning algorithms have been studied in recent years and are developing day by day. Different types of information such as static images, video images or sound are used in the facial expression recognition problem. In this study, a system working on real-time video images was developed to perform facial expression recognition. In the developed system, seven universal emotions are defined: real-time facial expression recognition, anger, disgust, happiness, sadness, astonishment, fear, and neutral state, with snapshots taken from the webcam using a convolutional neural network.

Keywords: *Facial expression recognition, Convolutional neural network, Deep learning, Real time emotional recognition*

Modeling The Risk of Introducing Non-Native Species Through Ship Hull Biofouling by Percent Cover Calculation

Adel Kacimi [1], Abderrahmane Bouda [2], Bilel Bensari[3], Nour El Islam Bachari[3]

[1]*National School of Marine Sciences and Coastal Planning (ENSSMAL)*

[2]*National Maritime School (ENSM)*

[3]*Houari Boumedienne University of Science and Technology (USTHB)*

Abstract

Biofouling of ship hulls is considered as one of the most important vectors for the transfer of aquatic invasive species. These species cause widespread impacts to native environments and ecological communities, in addition to financial costs for various industries. Targeted monitoring and effective adaptive management require knowledge on the likelihood of new introductions by non-indigenous species (NIS). In this study, we develop a model of the introduction and invasion of NIS for the port of Arzew, based on the length of stay of vessels in the ports of call, the geographical position of these ports, the ship's speed, the effectiveness of the antifouling system and the antifouling strategy used in port of origin. We calculated the biofouling accumulation of all ships calling at the port of Arzew for 2016 using spatial modeling to highlight the most relevant information. We identified areas that represent a high risk of species invasion according to their respective ecoregions of origin; the type of vessel that most likely promotes the establishment of non-native species by comparing the environmental similarity of the origin ecoregions with the environmental characteristics of the Arzew port obtained from satellite imagery. We show that over one year, 738 vessels called at the port of Arzew, inflicting a very high risk of invasion from six coastal ecoregions, in particular. These results can be used for invasive species management purposes, such as: the application of specific regulations to vessels of a certain tonnage that most favor the transfer of non-native species, as well as their ecoregions of origin that have a great environmental similarity with the Western Mediterranean, in order to minimize the transfer of these species.

Keywords: *Modeling, Hull biofouling, Non-indigenous species, Maritime traffic, Ecoregion, Exotic species*

Removal of Mercury II from Aqueous Solutions by Adsorption on a Natural Adsorbent

Cigdem Oter[1]

[1]*Van Yüzüncü Yıl University, Faculty of Science, Department of Chemistry, Van, TURKEY*

Abstract

Heavy metals are major pollutants in marine, soil, industrial and even treated wastewater. Most of the point sources of heavy metal contaminants are industrial wastewater from mining, metal processing, tanneries, pharmaceuticals, pesticides, organic chemicals, rubber and plastics, timber, and wood products. Heavy metals are transported by flowing waters and contaminated water sources downstream of the industrial site. Therefore, toxic heavy metals must be removed from wastewater before disposal. As most of the heavy metals discharged into wastewater are toxic and carcinogenic, they pose a serious threat to human health. Mercury is an extremely toxic heavy metal. Mercury spillage is extremely dangerous because it destroys brain tissue, lungs, and could degrade protein leading to toxic effects; it mainly affects the kidney and nervous systems and may cause some ailments and illnesses. In addition, mercury is a mutagen, teratogen and carcinogen that causes embryonical, cytochemical and histopathological events. Therefore, removal of mercury from aqueous solutions, especially drinking water, is very important in hydrometallurgical and wastewater treatment. Various methods have been proposed to remove Hg (II) ion from wastewater. Adsorption method is used as a low-cost, effective, and efficient technique for removing toxic heavy metals from wastewater. Researchers have turned to inexpensive adsorbents such as herbal waste. They used materials such as tea waste, sawdust, oiled coffee beans, tree ferns, chitosan, olive oil waste, orange juice waste, rice husks, algae, and dried herbs as adsorbents. In this study, it is aimed to remove Hg (II) ions from wastewater by using ground rice grains as adsorbents. The effects of contact time, pH, temperature, and initial concentration of mercury on adsorption were investigated using the batch method. Langmuir, Freundlich, Temkin and Dubinin-Radushkevich adsorption isotherm models were examined to analyze equilibrium data. It has been determined that the Langmuir isotherm, which provides the best correlation in Hg (II) adsorption, is the isotherm model that best describes the adsorption equilibrium process. In adsorption studies examining pseudo first order kinetic model, pseudo second order kinetic model and intra-particle diffusion model; it was determined that the adsorption process was compatible with the pseudo second order kinetic model. As a result of the analysis of thermodynamic parameters, it was concluded that the adsorption process is a self-progressing and endothermic process. The data obtained show that rice grains can be used as a cheap, useful, and effective adsorbent for the adsorption of Hg (II) from wastewater.

Keywords: *Adsorption, Mercury II, Isotherm, Kinetics, Thermodynamics*

On Text Clustering Algorithms

Duygu Selin Turan[1], Burak Ordin[1]

[1]*Ege University, Faculty of Science, Department of Mathematics, İzmir, TURKEY*

Abstract

Text mining can generally be defined as the process of obtaining previously undiscovered patterns or information from unstructured text documents. With the developing technology, the growth of databases and the fact that most of the data accumulated in databases are text data increase the importance of text mining. In general, the text mining process consists of 6 steps: creating the text collection, text preprocessing, feature selection, text transformation, data mining, evaluation and interpretation. In this study, the data set named 270 Köse Yazisi, which is one of the data sets of the Kemik natural language processing group, was used. The selected data set was preprocessed using Zemberek natural language processing library. Later, unstructured text data was transformed in structured form with the help of term-document matrix (TFIDF). In this study, in which clustering technique, one of the data mining techniques, is processed, two clustering algorithms are used: Classical k-means algorithm and Incremental k-means algorithm. The mean square error, one of the internal quality measurement methods, was used to evaluate the two algorithms run on the structured data set. As a result of the evaluations made, when the error values are examined, the efficiency of the incremental k-means algorithm is seen.

Keywords: *Machine learning, Data mining, Clustering problem*

Artificial Neural Network Modeling for The Prediction of Kinetic Parameters of Single-Stage Degrading Polymers

Gamzenur Özsin[1], Melis Alpaslan Takan[2], Ayşe Eren Pütün[3]

[1]*Bilecik Şeyh Edebali University, Department of Chemical Engineering, Bilecik, TURKEY*

[2]*Bilecik Şeyh Edebali University, Department of Industrial Engineering, Bilecik, TURKEY*

[3]*Anadolu University, Department of Chemical Engineering, Eskişehir TURKEY*

Abstract

Pyrolysis is one of the most promising thermochemical methods that could be used to convert plastics into energy products and monomers to recover chemical feedstocks and energy. The development of precise mathematical approaches in order to estimate thermal degradation of polymeric is necessary for proper process design and monitoring of pyrolysis processes. Therefore, this study aims to develop an effective artificial neural network (ANN) model to estimate the pyrolysis of two commercial polymers as polystyrene (PS) and polyethylene terephthalate (PET). For this purpose, thermal degradation behaviors of PS and PET were investigated by thermogravimetric analysis (TGA) at different heating rates. For both PS and PET, thermograms at 5, 10, 20, and 40 °C/min showed only a single pyrolysis zone, indicating pyrolysis reactions of these polymers occur in a single stage. Since the structure of ANN is suitable to analyze these polymers, the obtained results presented the whole process in detail. Moreover, the values of the kinetic parameters of PS and PET pyrolysis have been calculated at different conversions degrees different iso-conversional methods. The computational results between the experimental data and ANN predicted values present that the studied neural structure is a well-designed approach for modeling complex nonlinear systems such as thermal degradation of thermoplastics and for the determination of the kinetics of the process.

Keywords: *Pyrolysis, Artificial neural network, Kinetics, Polystyrene, Polyethylene terephthalate*

Intestinal Tissue Antioxidant Enzyme Changes of Van Fish Exposed to Pesticide Toxicity

Aslı Çilingir Yeltekin*

**Van Yüzüncü Yıl University, Faculty of Science, Department of Chemistry, Van, TURKEY*

Abstract

The use of pesticides is increasing day by day to increase the yield of agricultural products. The widening of this situation affects all living things negatively by disrupting the natural balance. Fungicides containing tebuconazole are mostly used for wheat and its derivatives. Therefore, in this study, it was aimed to investigate the effects of the fungicide that is the main ingredient of tebuconazole, which is widely used worldwide, on Van fish. Antioxidants are very important defense systems for the immune system in metabolism. In this respect, the exchange of antioxidant enzymes is of great importance. In the study, Van fish were divided into concentration and control groups. Each group was administered tebuconazole at concentrations (2.5 mg / L) at 24, 48, 72 and 96 hours. In the study, antioxidant enzymes were analyzed by spectrophotometric methods. As a result of the study, it was determined that the levels of SOD, GSH-Px and CAT, which are important parameters of the antioxidant defense system of Van Fish, decreased over time with its fungicidal effect.

Keywords: *Antioxidant, Van Fish, Toxicity*

Repellent Activity of Some Plant Extracts Against Wheat Weevil, *Sitophilus Granarius* L. (Coleoptera: Curculionidae)

Pervin Erdoğan[1]

[1]*Sivas University of Science and Technology, Faculty of Agricultural Sciences and Technology, Sivas, TURKEY*

Abstract

Sitophilus granarius L. (Coleoptera: Curculionidae) is one of the most important pests causing loss of economically important crops in stored grains. In general, chemical pesticides are used to control S. granaries. Researchers have focused on alternative methods of pest control due to the side effects of chemical pesticides. In this context, most studies have been done with plant extracts. Plant extracts have been used as insecticides since ancient times. In this study, Tagetes patula L. (Asteraceae), Tanacetum vulgare L. (Asteracea), Aleo vera L. (Liliaceae), Hyoscyamus niger (Solanaceae), Lantana camara L. (Verbenaceae), Allium sativum L. (Amaryllidaceae), Capsicum annuum L. (Solanaceae) and Tanacetum parthenium L. (Asteracea) plant extracts to determine the repellent effect on S. granarius. For this purpose, filter papers that are cut in the same size and divided in half are placed in 9 cm petri dishes. The papers are fixed in the middle with tape. Half of the paper was left as control, and the other half was given the concentrations of the extracts (2.5, 5, 7.5, 10%) prepared by means of a micropipette (3 µL of each concentration). After the application, the plates were kept to dry. Then, 20 pieces of 1-3 years old adult were left in the middle of each petri dish Petri lids were closed and left in a dark environment. The counts were recorded 24 hours later by counting individuals in the applied and control section under light. Pure water was used for control. Experiment in four replicates laboratory (24 ± 1 ° C and 60% ± 65 humidity) conditions. A commercial preparation called Gamma-t-ol obtained from the extract of Melaleuca alternifolia (Maiden & Betche) (Myrtaceae) was used as a positive control. The experiment was conducted under laboratory conditions. As a result of the counts, the strongest repellent effect was determined in the concentrations of the Gammat-ol preparation, this value followed by high concentrations of L. camara and other plants, respectively. The lowest repellent effect was determined in T. patula extract.

Keywords: *Plant extracts, Wheat weevil, repellent effect*

Machine Learning Models and Statistical Dexketoprofen Pharmaceutical Dosage of Approaches Application In The Form

Pervin Erdoğan[1]

[1]Sivas University of Science and Technology, Faculty of Agricultural Sciences and Technology, Sivas, Turkey

Abstract

Minimizing the raw material cost required for pharmaceutical production and reducing the work-time burden is important for the sectors working in the relevant field. Continuing the trials with manpower takes time and imposes costs on the company. Recently, with increasing technological innovations, more efficient ways are being sought in drug discovery and machine learning models have come to the fore. In this study, statistical studies were conducted with the limited data set of the product called Dexketoprofen, which is an orally disintegrating tablet, and normality test, t-test, Mann – Whitney U, ANOVA, Kruskal – Wallis Test were applied. Considering the statistical test results, machine learning models, k-NN, SVR, CART, BAGGING, RF, GBM and XGBOOST were used, and a new formulation was created that was not tested after the optimum values were estimated on the formulation.

Keywords: *Machine learning, Optimization, ANOVA, Student t test, Mann Whitney U, Kruskal Wallis Test, Dexketoprofen*

Examining the Causal Links Among Economic Complexity, Globalization and Financial Development in Turkey by Using the Fourier Granger Causality Test

Alper Karasoy[1]

[1]*Afyon Kocatepe University, Faculty of Economics and Administrative Sciences, Department of Economics, Afyonkarahisar, TURKEY*

Abstract

Economic complexity is a measure that shows the productive capability of an economy by considering the activities that it is able to develop. In this context, investigating the causal linkages among globalization, economic complexity, and financial development in Turkey may provide new insights into whether the globalization and financial development processes that it has experienced affected its economy's complexity. As the empirical literature on this subject is very limited, this study aims to fill this gap for the Turkish case. In this regard, this research utilizes the Fourier co-integration and the Fourier Granger causality tests to observe the interlinkages among the abovementioned indicators for the 1970-2017 period in Turkey. The main reason for employing these tests is to account for the unknown (structural) break numbers, dates, and forms. The co-integration test results show that a long-run association exists between globalization, economic complexity, and financial development in Turkey for the sample period. Further, the causality analysis indicates that bi-directional causalities exist between globalization and financial development, and between economic complexity and globalization. Besides, the causality test results also show that there is a uni-directional causality running from financial development to economic complexity. These results show that globalization causes economic complexity in Turkey both directly and indirectly through financial development. Also, financial development in Turkey causes economic complexity. These findings are used to propose some policy suggestions for Turkey.

Keywords: *Economic complexity, globalization, financial development, Turkey, the Fourier Granger causality test*

Use of Persuasive Technology in The Health Field

Arzu Kurşun[1], Ceren Türkdoğan Görgün[2]

[1]Giresun University, Vocational School of Health Services, Department of Medical Services and Techniques, Giresun, Turkey

[2] Giresun University, Keşap Vocational School, Department of Management and Organization, Giresun, Turkey

Abstract

Persuasive technology defines technologies designed to change users' behavior and attitude. Researchers have started to examine the unique abilities that technology can have to support individuals change their behavior. This emerging area of research is addressed as technology-specific abilities that can persuade users to perform certain behaviors and go beyond making behaviors the user wants to perform more easily and instead focus on encouraging users to think and act differently and this topic is becoming increasingly common in the field of health behavior change. There has been an increasing growth in the availability of digital technologies within the health care management. This flow of technology has permitted people to self-observing an excessive amount of health indexes. The aim of this study is, to reveal the usability of persuasion technologies in the field of health by reviewing existing studies in order to evaluate the capacity of persuasion technology tools to influence health behaviors. A database search was managed to determine relevant articles. Then, articles were reviewed using the persuasive systems as a framework for analysis. When the studies are examined; it has been determined that persuasion technology was used in subjects such as pain assessment, diabetes, healthy nutrition, physical activity, prevention of smoking, obesity, cancer patients, organ donation applications and studies on the use of persuasion technology in the field of health were found to be insufficient.

Keywords: *Persuasive Technology, Health Informatics, Health Behavior Change, Digital Technology*

Investigation of Inhibition Effect of Busulfan and Carfilzomib Chemotherapeutic Drugs on Paraoxonase (PON1) Enzyme Activity

Hakan Söyüt[1], Yakup Ulutaş[2], Ekrem Köksal[2]

1 Bursa Uludag University, Faculty of Education, Bursa, TURKEY

2 Erzincan Binali Yıldırım University, Faculty of Arts and Sciences, Erzincan TURKEY

Abstract

Paraoxonase-1 (PON1) is a circulating antioxidant enzyme found in cell membranes and bound to high density lipoproteins (HDL). PON1 is a lactonase (LAC). Lactones make up their primary substrate. It is this catalytic capacity that allows PON1 to reduce lipid peroxides in the cell and circulating lipoproteins [1]. In addition, PON1 has an esterase activity that disrupts organophosphate xenobiotics and nerve agents. PON1 is synthesized primarily by the liver and to a lesser extent in the kidney and colon, and then HDL-dependent blood is released. As an antioxidant molecule, PON1 plays an important role in lipid metabolism and the control of inflammation [2]. PON1 enzyme activity is affected by inflammation changes and oxidized low-density lipoprotein (LDL) levels. PON1 has been shown to protect against oxidative stress by hydrolyzing oxidized phospholipids, maintaining HDL integrity and function, and preventing LDL oxidation. It also exhibits atheroprotective properties by reducing the capacity of macrophages to oxidize LDL [3].

In this study, the inhibition effects of some chemotherapeutic drugs (Busulfan and Carfilzomib), which are commonly used in chemotherapy, on human serum PON1 enzyme activity were investigated. K_i constants were found as 0.042 ± 0.012 mM, 0.043 ± 0.008 mM.

Keywords: *Paraoxonase, Inhibition, Busulfan, Carfilzomib*

Analysis of Football Data with Classification and Decision Tree Methods: Logistic Regression and CART Algorithm

Duygu Topcu[1], Özgül Vupa Çilengiroğlu[2]

[1]Dokuz Eylul University, Institute of Science and Technology, Izmir, TURKEY

[2]Dokuz Eylül University, Faculty of Science, Department of Statistics, İzmir, TURKEY

Abstract

Football is one of the most followed sports in the world and Turkey. This prevalence of football is used in information technologies and with the developing data science, match statistics can be determined easily. The most important issue in football competitions is the match result. There are many different criteria (the number of goals scored, the number of cards the team has received, the weather, play away, etc.) that affect the match result. In this study, the data obtained from Turkey Football Federation Super League 2019-2020 season matches with classification and decision tree method were analyzed with. In the matches played, the red or yellow cards received by the host and rival teams, the number of foreign players in the teams and the number of goals scored were brought into a categorical format and determined as independent variables. Depending on these variables, the home team's winning or losing situation was modeled using Logistic Regression and Decision Tree algorithms. Two separate models have been created within the scope of this study. In the first model, the variables of rival yellow card, host red card, rival red card and number of foreign players in the home team were used. In the second model, the red card received by the rival team and the goals scored variables were used. The results of the created model were evaluated with various statistical criteria.

Keywords: *Match Result, Football, Machine Learning Algorithms, Logistic Regression, CART*

Analyzing Twitter Users Behavior Against COVID-19 Vaccines with Natural Language Processing Techniques

Hilal Tekgöz[1], Halil İbrahim Çelenli[1]

[1]IBSS Consulting, Research and Development Department, İstanbul, TÜRKİYE

Abstract

COVID-19 is causing a global crisis that affects many areas of human life around the world. The most effective solution to prevent this global epidemic is the development of vaccines and antiviral drugs. Health organizations conduct various vaccine studies in order to cope with the global pandemic and prevent the spread of the pandemic and put these vaccine studies on the market. During the pandemic period, social media provides an important communication network with a large user base. Throughout the COVID-19 process, people have shared their ideas about the pandemic and vaccines on social media. In this study; Natural language processing techniques have applied on Twitter, one of the social media platforms, using data including the posts of people about Pfizer / BioNTech, Sinopharm, Sinovac, Moderna, Oxford / AstraZeneca, Covaxin, Sputnik V. vaccines. This paper aims to study the behavior of people in a different region of the world to COVID-19 vaccines. Sentiment analysis techniques have used on tweets. In addition, topic modeling and word embedding techniques have applied to the tweets sent for each vaccine, visualizing a series of words that best describes each vaccine group and words inferring relationships with similar meanings.

Keywords: *Natural Language Processing, COVID-19 Vaccine, Sentiment Analysis, Topic Modeling, Word Embedding*

Sentiment Analysis of Articles About Education in Pandemic Published on The ERIC Database

Abdullatif Kaban[1], Ömer Bilen[1]

[1]*Ataturk University, Faculty of Applied Sciences; Department: Department of Information Systems and Technologies*

Abstract

Sentiment analysis is a calculation method that automatically analyzes the value of large amounts of text. Basic sentiment analysis involves extracting and counting emotionally loaded keywords (hate, love, happiness, sadness, etc.) from texts. The idea of "extracting meaningful expressions from renewable and usable data", which is one of the basic principles of big data and data mining, has increased the popularity of sentiment analysis, which is one of the important research topics in recent days. Sentiment analysis is a general definition that is given to the processes of defining and classifying the opinions/expressions specified with a piece of text in order to evaluate the attitude of an author or an article towards a certain subject as positive, negative, and neutral. This study, it was aimed to analyze the sentiment analysis of the articles on education during the pandemic scanned in the ERIC database. For this purpose, using the python programming language, the ERIC database was scanned with the web scraping method and a data frame was generated from the article information. Sentiment analysis was performed on the summary of the 1253 articles obtained by using artificial intelligence capabilities. The abstracts obtained were examined on the stems by removing the suffixes of each word after the cleaning process used in this analysis method. The 20 most frequently repeated words in the reviewed abstracts are learning (f=1896), education (f=1723), pandemic (f=1428), covid (f=1364), online (f=1335), teacher (f=1108), school (f=1005), study (f=941), teaching (f=854), university (f=670), research (f=611), course (f=602), experience (f=505), challenge (f=485), educational (f=453), face (f=449), social (f=430), article (f=422), data (f=414), and time (f=411). As a result of the sentiment analysis, it was concluded that 87.6% of the article summaries contain positive (f=1098), 11.2% neutral (f=140) and 1.2% negative (f=15) meanings.

Keywords: *Sentiment analysis, Artificial intelligence, Pandemic, Education*

Fitting Random Survival Forest Model For Time To Event Data: An Application On Evaluation of Factors Affecting Longevity Of Arabian Racehorses

Hülya Özen [1]

[1] *Eskişehir Osmangazi University, Faculty of Medicine, Department of Biostatistics, Eskişehir, TURKEY*

Abstract

In this study, it is aimed to introduce the Random Survival Forest (RSF) method and use it on an application dataset. Cox Regression model is often preferred to define risk factors in survival analysis, but this model has certain limitations and assumptions. RSF, on the other hand, is an unbiased and low-variance nonparametric machine learning technique that is successful in determining important variables. Dataset, which is obtained from Turkish Jokey Club website, is consist of Arabian horses that were born in or later 2003. In the study, the end of the racing career of the racehorses was defined as the event of interest. Fifteen different variables that were thought to have an effect on the duration of the racing career were included in the model. A suitable RSF model was fitted by tunning some parameters. The error rate of the RSF model (0.1681) was determined to be lower than the error rate (0.1884) obtained from the Cox Regression that is fitted with the same variables. In determining the risk factors affecting the longevity of racehorses, variable importance measures and the average minimal depth approaches were used. According to the common result obtained from these approaches, the most important risk factors affecting the longevity of racehorses are obtained as earnings, total number of starts, gender, first running distance and age at first start to professional career. This study shows that RSF method, which does not contain parametric assumptions, provide results with a lower error rate than conventional methods used in survival analysis and can be used as an alternative method in the field of health sciences.

Keywords: *Random survival forests, Survival analysis, Risk factor*

Software Defect Prediction Using Transformers

Shefitjon Bregu[1], Yusuf Kartal[2], Kemal Özkan[2]

[1]*Eskisehir Osmangazi University, Department of Mathematics and Computer Science, Eskisehir, TURKEY*

[2]*Eskisehir Osmangazi University, Computer Engineering, Eskisehiri TURKEY*

Abstract

Software quality has been a great interest of researchers in this new technology era. There are a lot of variables that define the quality of software including here correctness, reliability, quality of product, scalability, and the most important one, the absence of bugs in the software. Day by day, projects are growing bigger and bigger, resulting in a huge circulation of developers working on the same project in turns, and thereby, this fact causes an increase in the number of bugs, making the project less scalable and reliable, and consequently, less qualitative. To help improving Software reliability, Software Defect Prediction that is widely defined as the process of predicting error events in the software has found broad usage lately in the technological world. Normally, Software Defect Prediction locates the faulty locations for a developer to prioritize testing efforts only on that part of the code. As of today, researchers are heavily focused on achieving the predictive model through static code metrics, which try to make the code analysis without executing it. This static analysis has shown to be successful but at the same time have a lot of limitation on the information learned, and such a more dynamic or hybrid analysis is introduced to this paper. Studying the behavior of code through syntactic and semantic structures provided us an improved predictive model compared to the state-of-the-art model. In this paper, it is made possible to create a framework called software defect prediction using Transformers, which starts the journey of achieving the predictive model through parsing abstract syntax trees of the code and extracting them as vectors. The next step will be learning semantic and syntactic information from the encoded AST, continuing with feeding the AST to the transformer, which with the help of the attention mechanism generates more significant features for a more accurate predictive system. This study aims to improve the F1-measure of the state-of-the-art method and comparing it with the data of the proposed method, concludes this aim to be feasible.

Keywords: *Software Defect, Transformers, Machine Learning, Deep Learning, Abstract Syntax Tree*

Evaluation of Slake Durability Index Values with Post-Cycle Mass Loss Amounts

Hüseyin Ankara[1], Fatma Çiçek[2]

[1]Eskişehir Osmangazi University, Faculty of Engineering – Architecture, Department of Mining Engineering, Eskisehir, TURKEY

[2]Çukurova University, Faculty of Engineering – Architecture, Department of Mining Engineering, Adana, TURKEY

Abstract

Slake durability index test was proposed by Chandra, Chandra and Franklin. This test has been taken place in rock mechanics as one of the recommended tests by ISRM (The International Society for Rock Mechanics) in 1981 and standardized. This test has been accepted by ASTM (The American Society of Testing and Materials) in 1987 and standardized. The purpose of the test, which usually practices to clastic rocks as wetting and drying cycles is to determine an index value indicating wear and the resistance to dissolution in water. In this study, 12 cycles of Slake Durability Index Test was applied on representative sphere test samples with uniform size / mass and smooth surface prepared from massive, laminated marl and white tuff rock samples. The change between index values calculated after the cycles was examined and interpreted with the amount of mass loss in the cycles.

Keywords: *White Tuff, Laminated Marl, Massive Marl, Slake Durability Index*

Small Pelagic Fish Catches In The Central Algerian Coast (SW Mediterranean) Combined With Satellite Observation And Meteorological Data

ALI Lamia[1], Bacharı Nour El Islam[1]

[1]*Spatial Oceanography Laboratory - Houari Boumediene University of Science and Technology (USTHB), Bab Ezzouar, ALGERIA*

Abstract

The Algerian central coast contributes largely to the national fisheries production, which has been the subject of considerable exploitation in recent years. Nevertheless, the exploitation of fishery resources on the Algerian coast is conditioned by meteorological variations. In our study, we have combined three major databases to understand the functioning of small pelagics on the Algerian coasts. First, we conducted an analysis of meteorological factors, mainly average temperature and wind speed, to define the fishing effort in the study area. Secondly, remote sensing was used to measure monthly abiotic parameters between 2008 and 2018 (Sea Surface Temperature (SST) and chlorophyll (Chl-α)), and thirdly, combine them with monthly small pelagic catches in 2016 and 2017. SST and Chl-α vary according to the season and are negatively correlated (r_2 = 0.45). Sardine (*Sardina pilchardus*), is very strongly related to SST (r_2 = 0.9), and moderately related to Chl-α (r_2 = 0.43), while horse mackerel (*Trachurus trachurus)* is weakly related to SST (r_2 = 0.23), and strongly related to Chl-α (r_2 = 0.57). Bogue (*Boops boops)* is weakly correlated with SST (r_2 = 0.23), and Chl-α (r_2 = 0.21). Round sardinella (*Sardinella aurita*) is negatively correlated with SST (r_2 = 0.42), and weakly correlated with Chl-α (r_2 = 0.12). All these species are seasonal, except Round sardinella, which is positively correlated with Chl-α and negatively correlated with SST, i.e. when Chl-α increases and SST decreases. This phenomenon is present in the winter season or is related to upwelling.

Keywords: *Small pelagic, SST, Chl-a, Remote sensing*

Dynamic Mode Decomposition for Covid 19 Data

Gamze Yüksel[1]*, Nida Gökçe Narin[2]

[1]*Muğla Sıtkı Koçman University, Faculty of Science, Mathematics, Muğla, TURKEY*

[2] *Muğla Sıtkı Koçman University, Faculty of Science, Statistics, Muğla, TURKEY*

Abstract

In this study, the number of people infected with Covid 19 disease was considered for 3143 cities of the USA for a period of approximately 1 year. The data were taken from Github. The Covid 19 pandemic is a dynamic model that changes over time. It is possible to extract the space-time patterns that dominate the dynamic activity and thus predict the future-state of dynamic models with dynamic mode decomposition. The Dynamic Mode Decomposition is based on singular value decomposition. The data is preprocessed and stabilized before dynamic mode decomposition is applied to the data. The city-based patterns of the number of people infected with Covid 19 disease in the USA were extracted by applying dynamic mode decomposition to the stationary data. Thus, the future-state estimating of the number of people to be infected from this pattern was predicted and the performance of the predictions was determined by the confidence interval.

Keywords: *Dynamic Mode Decomposition, Singular Value Decomposition, Machine Learning, Estimation*

A Study on Anomaly Detection with Boosting in Imbalanced Datasets

Engin Yıldıztepe[1], Nihat Akıllı[2]

[1]*Dokuz Eylül University, Faculty of Science, Statistics, İzmir, TURKEY*

[2]*Dokuz Eylül University, Institute of Science and Technology, Data Science, İzmir, TURKEY*

Abstract

The purpose of anomaly detection may be to exclude an incorrect measurement or to identify a situation that is vital to detect. For this reason, anomaly detection is widely used in different areas such as fraud detection in banking transactions, fault detection in critical systems, detecting network intrusion in digital networks, and medical diagnosis. Anomaly detection in labeled data can be considered as a classification problem. In this case, the anomalies form the minority class. However, as the imbalance ratio between classes increases, the performance of the classification methods may decrease. In this case, which is called imbalanced classification, it becomes difficult to determine the anomalies accurately that form the minority class. In related studies, data-level and algorithm-level solutions have been proposed to increase the classification performance in imbalanced classes. One of these solutions, SMOTE, is a technique based on the artificial reproduction of minority class data by resampling in the training set. In this study, a combination of resampling and ensemble learning methods are examined in imbalanced data with R statistical programming language. In the application, the models trained without solving the class imbalance problem are compared with the models trained by eliminating the class imbalance problem. The solution using the XGBoost algorithm with SMOTE outcomes more successful results. In cases where anomaly detection is handled as an imbalanced classification problem, it is observed that solving the class imbalance problem before training with appropriate techniques can increase the anomaly detection performance.

Keywords: *Anomaly detection, Imbalanced classification, XGBoost, SMOTE*

The Effect of Strong and Weak Unidimensional Item Pools on Computerized Adaptive Classification Testing Criteria

Ceylan Gündeğer[1], Sümeyra Soysal[2]

[1]*Aksaray University, Faculty of Education*

[2]*Necmettın Erbakan Unıversıty, Faculty Of Educational Sciences*

Abstract

In this study, the effects of dimensionality, which is one of the item pool characteristics in computerized adaptive classification testings (CACT), on average test length, average classification accuracy, bias, RMSE and mean absolute error was investigated. For that purpose firstly, the ability parameters (thetas) of 1000 individuals were derived from N(0,1) in the range of -3,3. CACT simulation was carried out under manipulating different ability estimation, item selection and classification methods using item pools of 500 items derived to show strong and weak unidimensionality, which are two different representations of unidimensionality. The data were simulated based on a 3-parameter logistic model (3PL), and item factor loadings ranged from 0.30-0.50 in data representing weak unidimensionality, whereas it ranged from 0.60-1.00 in strong one-dimensional structures. The study is conducted on 25 replications. As ability estimation methods Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE); as item seleciton methods Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) and as classificaton criteria Sequential Probability Ratio Test (SPRT) and Confidence Interval (CI) methods were handled. So, 2x2x2x2=16 conditions were investigated in this research. A comparison was made on average test length, average classification accuracy, bias, RMSE, mean absolute error values for the conditions. According to the results, it was provided to determine which classification criteria works in which dimensionality. Results from simulation studies in R will be completed and discussed considering the literature.

Keywords: *Individualized computerized classification tests, Individualized computerized tests, Dimensionality*

The Effect of Pseudo-Guessing Parameters on Computerized Adaptive Classification Testing

Sümeyra Soysal[1,] Ceylan Gündeğer[2]

[1]*Necmettın Erbakan Unıversıty, Faculty Of Educational Sciences*

[2]*Aksaray University, Faculty of Education*

Abstract

This was aimed at examining the effect of pseudo-guessing parameter on average test length, average classification accuracy, bias and RMSE in different conditions of computerized adaptive classification testing (CACT). Item discrimination and difficulty parameters were generated from a normal distribution N(1, 0.2) and a standard normal distribution N(0,1), respectively. Manipulated factor included pseudo-guessing parameters (0.10, 0.20, 0.25, 0.30) generated from beta distribution. The latent trait parameters were generated from a standard normal distribution N(0, 1). Twenty-five replications were conducted under each simulation condition. As ability estimation methods Expected a Posteriori (EAP) and Weighted Likelihood Estimation (WLE); as item seleciton methods Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) and as classificaton criteria Sequential Probability Ratio Test (SPRT) and Confidence Interval (CI) methods were handled. So, 2x2x2x2=16 conditions were investigated in this research. A comparison was made on test length, classification accuracy, bias, RMSE values for the conditions. According to the results, it was provided to determine which classification criteria works in which dimensionality. Results from simulation studies in R will be completed and discussed considering the literature.

Keywords: *Individualized Computerized tests, Individualized computerized classification tests, Luck parameter, Item response theory*

Comparison of The Influence Diagnostic Techniques in The Inverse Gaussian Liu Regression Model

Şahin Canalp[1], Y. Murat Bulut[2]

[1]Eskişehir Osmangazi University, Institute of Science and Technology, Eskişehir, TURKEY

[2]Eskişehir Osmangazi University, Faculty of Arts and Sciences, Department of Statistics, Eskişehir, TURKEY

Abstract

When the dependent variable's distribution is skewed, one of the used regression methods is the inverse Gaussian regression (IGR) model. The maximum likelihood estimation (MLE) method is generally used to estimate the IGR model's unknown parameters. The efficiency of the MLE method decrease when the dataset includes influence observation(s). It is crucial to detect influence observations in the dataset as the negative effects on the estimation values reduce the model's explainability, and the parameter estimations do not be consistent. Some of the influence diagnostic techniques, which recommended for the IGR model in the literature, are Cook distance, modified Cook distance, covariance ratio, DFFITS, DFBETAS, and Welch distance (Amin et al., 2020). These influence diagnostic techniques are proposed based on the MLE method. Another situation that negatively affects the MLE in regression analysis is a linear relationship among explanatory variables, which is called multicollinearity. In the literature, the biased estimators that give more effective results than the MLE have been proposed to solve the multicollinearity problem. One of these biased estimators is the Liu estimator proposed by Liu (1993). Akram et al. (2020) have proposed the inverse Gaussian Liu regression model (IGLRM) using the Liu estimator in the IGR model. In this work, we will define influence diagnostic techniques for the IGLRM. Monte Carlo simulation has been done to show the effectiveness of the proposed methods.

Keywords: *Influence Diagnostics, Inverse Gaussian Regression Model, Liu Estimator, Multicollinearity*

Determination of Elemental Affinities in Soma Lower (kM2) Coal Seam in Soma Basin Using Agglomerative Hierarchical Clustering Algorithm

Mete Eminağaoğlu[1], Rıza Görkem Oskay[2], Ali İhsan Karayiğit[3]

[1]*Dokuz Eylül University, Faculty of Science, Department of Computer Science, İzmir, TURKEY*

[2]*Hacettepe University, Başkent OSB Vocational School of Technical Sciences, Ankara, TURKEY*

[3]*Hacettepe University, Department of Geological Engineering, Ankara, TURKEY*

Abstract

Coal, as most common fossil-fuel, has been subjected to several geochemical and mineralogical studies due to presence of potentially toxic elements. In these studies, statistical methods (e.g., correlation coefficient, cluster and factor analyses) are mainly used for determination of toxic elements affinities in coal. In environmental concerns, the statistical methods are recently topic of discussion due to correlations between some elements with minerals that cannot be chemically affiliated. Nevertheless, microanalyses methods as like scanning electron microscopy-energy dispersive spectrometer (SEM-EDX) or electron microprobe (EPMA) and machine learning algorithms more commonly applied in determination of affinities of some toxic elements in coal. This study aims to correlate geochemical and mineralogical data of lower (kM2) coal seam in the Soma coalfield with Bray-Curtis, Cosine and Tanimoto similarities and different similarity measures like Pearson correlation co-efficiencies. The results of similarity measures evaluated using agglomerative hierarchical clustering algorithm (average linkage) and elements grouped in several clusters. Most of identified elemental groups, expected a few of them, based on Canberra, Chebyshev, Bray-Curtis and Tanimoto measures do not appear to be in agreement with mineralogical and geochemical data. Nevertheless, elements affiliated with aluminosilicate elements (e.g., Al, K, B, and Cs) are grouped in together, and elements (e.g., S, As, Mo and U) related with redox conditions in coal formation environment are located in the same group according to Pearson correlation co-efficiencies and cosine similarity. In addition, this data is in agreement with SEM-EDX and XRD data of studied coal samples. These results imply that cosine similarity could be an alternative for Pearson correlation coefficiency's method in coal studies. As a result, more detailed studies using both similarity measures should be conducted in future, and these similarity measures should always be correlated with SEM-EDX and XRD data.

Keywords: *Coal geochemistry, Element, agglomerative hierarchical clustering, Similarity measures, Soma basin*

Optimization Studies of Alfuzosin Tablet Formulation Using Machine Learning Models

Atakan BAŞKOR[1], Burcu MESUT[2], Buket AKSU[3], Yıldız ÖZSOY[2]

[1] *Bahçeşehir University, Institute of Science and Technology, Big Data Mining Analysis and Management, Istanbul, TURKEY*

[2] *Istanbul University, Department of Pharmaceutical Technology, Pharmacy Technology, Istanbul, TURKEY*

[3] *Altınbaş University, Faculty of Pharmacy, Department of Pharmacy Technology, Istanbul, TURKEY*

Abstract

The aim of this study is to develop and optimize the tablet formulation of extended-release directly compressible alfuzosin (ALF) hydrochloride (HCl) using machine learning models. During the formulation studies, first of all, critical quality parameters (CQAs) were determined and fishbone methodology, one of the risk assessment tools, was used to determine critical material and critical process parameters (CFPs). In-process control tests, analysis and dissolution studies were carried out. The test results were evaluated with machine learning models k-NN, SVR, CART, BAGGING, RF, GBM and XGBOOST algorithms and the program was trained according to these data. The program introduced new tablet formulations that had not been studied before, and the dissolution test results of this formulation were quite similar to the results of the reference product than other formulations. In conclusion, this study showed that using machine learning models in solid dosage formulation development provides many benefits and advantages for the pharmaceutical industry.

Keywords: *Alfuzosin, Machine learning models, Mathematical modeling, Algorithm*

Return Transmissions in Cryptocurrency Markets and The Implications of Covid-19 Outbreak on Dynamic Correlations

Arda Doğruöz[1], Selin Karatepe Yurdal[1], Çağrı Levent[1]

[1]*Yalova University, Faculty of Economics and Administrative Sciences, Economics, Yalova, TURKEY*

Abstract

This paper examines both return spillover dynamics among cryptocurrencies and time-varying conditional correlations between pairs of cryptocurrencies in the framework of a VAR-DCC-GARCH model. The dataset includes daily returns of Bitcoin, Ethereum, Ripple, Chainlink, Dash, Neo, Litecoin, Eos and Binance Coin for the period from 28/03/2019 to 05/03/2021. All price series are in US Dollars and obtained from investing.com. We used the logarithmic differentiation of closing prices to calculate the daily returns for the respective market. The empirical results indicate that the return spillovers among cryptocurrencies are mostly uni-directional. Nonetheless, we identified bi-directional linkages between Ethereum and Binance Coin, Ethereum and Litecoin, Ethereum and Chainlink, Binance Coin and Neo. The identification of these linkages reveals evidence that Ethereum is the main return transmitter by effecting all other cryptocurrency markets except Dash which means Ethereum returns could be used in forecasting other cyrptocurrencies' returns. More importantly, significantly estimated parameters of lagged mean spillovers indicate that all the markets are interrelated via return transmissions. These interactions among markets constitute a complex network structure where each market is connected with another directly or indirectly, and hence provide strong evidence supporting the progress of market integration in cryptocurrency markets. Besides, an investigation of time-varying conditional correlations between pairs of cryptocurrencies provided further evidence on co-movements or inter-dependencies within the markets. Finally, time-varying conditional correlations are found to be higher during COVID-19 period for all pairs of cryptocurrencies than pre-COVID-19 period. This result implies that the connectedness of cryptocurrencies is strengthen during the pandemic, which could be related to the fear originated herding behavior.

Keywords: *Cryptocurrency, return spillovers, time-varying correlation, DCC-GARCH, COVID-19*

Supplier Evaluations Using Intuitionistic Fuzzy Modeling

Ayşenur Akın Vargeloğlu [1], M. Bahar Başkır [2], Hamza Gamgam [1]

1Gazi University, Faculty of Science, Department of Statistics, Ankara, TURKEY

2Bartın University, Faculty of Science, Department of Mathematics, Bartın, TURKEY

Abstract

Fuzzy logic-based approaches are widely used in modeling real-life problems involving uncertainty. In fuzzy set theory, uncertainties arising from human perception are examined using the degree of belonging. Intuitionistic fuzzy sets, as an extension of fuzzy sets, enables the uncertainties to be examined in more detail by using the non-belonging information. As an alternative of classical methods, fuzzy regression analysis is utilized in modeling studies performed in fuzzy environment. Besides, in intuitionistic fuzzy regression analysis, uncertainties between system components are examined with their belonging and non-belonging degrees. Due to the intuitionistic fuzzy set-based regression analysis, system-uncertainties are evaluated in more realistic-approach. In this study, modeling of supplier evaluations in a company is investigated using least square-based intuitionistic fuzzy regression analysis. The performance of the intuitionistic fuzzy regression was compared with classical and fuzzy regression models. The intuitionistic fuzzy, classical, and fuzzy model estimators were evaluated using the root of mean residual square-criterion. In addition, the model validation of intuitionistic fuzzy regression was examined by cross-validation method. Fuzzy and intuitionistic fuzzy regression calculations were obtained with user-friendly code windows written in Matlab. As a result of comparing, the success of model performance formed by intuitionistic fuzzy regression for supplier evaluations was revealed.

Keywords: *Intuitionistic Fuzzy Set, Fuzzy Regression, Supplier Evaluations, Data Science*

Value-At-Risk Prediction: A Comparison of Historical Simulation Extensions

Serdar Neslihanoğlu[1]

[1]*Eskişehir Osmangazi University, Department of Statistics*

Abstract

This research assesses the extensions of the historical simulation (HS) methods, which is the most popular and model-free method for predicting the value-at-risk (VaR) of a portfolio. For this purpose, the performances of the HS, the Monte Carlo simulation (MCS) (model-based method) and the filtered historical simulation (FHS) (which is a combination of HS and MCS methods) methods for predicting the VaR of a portfolio are compared in this research. The portfolio analysis is implemented using daily data from January 2017 to January 2021, including stock markets and firms' indices, gold prices, oil prices, US government securities funds, and treasury bonds. While predicting the VaR of the value-weighted portfolios through the proposed methods, the volatility of the daily and multiday returns forecasts of these portfolios is modelled by the GARCH model with normal and student-t distributions, respectively. The empirical findings favour the FHS approach, which outperforms the others interms of VaR predictability in both time forecasts.

Keywords: *Filtered Historical Simulation, GARCH model, Monte Carlo Simulation, Value-at-Risk, Historical Simulation*

Development of Dinçer Platform for the Integrated and Management of Big Data in the Logistics Sector

Batuhan M. ALİOĞLU[1], Cemil ÇELİK[1], Sinan BUGAN[1], Mustafa TEMİZ[1]

[1]*Dinçer Lojistik A.Ş., R&D Center, Istanbul, TURKEY*

Abstract

Developing technology and new communication systems cause vital changes and transformations for the logistics sector as in every sector. These transformations have created certain imperativeness on companies and enabled companies to turn to technology. Dinçer platform is a system that enables the data required in the storage and transportation processes to be received from the customer in full time, accurately and systematically and to analyze, process, combine and provide feedback as a result of all these operations. If big data is not managed well with technology, it leads to failure to provide a sustainable service and causes customer losses for companies. Also, the inability to respond quickly to the needs of the customers, the lack of strong communication infrastructure and the inability to make improvements at the rate of increasing data sizes formed the main need subjects of the study. The scope of the study, it is aimed to develop the Dinçer Platform in a company that provides 3rd party logistics services with its customers, warehouses and transfer centers. In the first stage of the study, system analysis and preliminary studies were carried out. The scope of the study, conceptual analyzes were made by determining the needs of customers, warehouse (WMS) and transport management system (TMS) and which data to be used from which fields were determined. In the second stage of the study, as a database; SQL, in the design of the screens; Angular js, in the processing of incoming data and sending it to services; NET Core worker services and Quartz, as software languages C # technologies was used in portal developments. Sub screens have been developed on the database and platform by using data mining techniques for processing and analyzing big data in databases. Also, data pre-processing techniques have been applied to ensure that the data is consistent, smooth and fast before data mining techniques. In the third stage of the study, API data integrations were made with the WMS and TMS used by the company and a user-friendly interface was designed. As a result of the study, a platform that provides productivity increase was developed increase by increasing customer satisfaction, reducing labor and time loss, and reducing unexpected costs by the R&D Center.

Keywords: *Big Data, Logistics, Platform, Transport Management System, Warehouse Management System*

Classification of High Dimensional Data After Dimension Reduction with PCA

Yıldırım DEMİR[1]

1Van Yüzüncü Yıl University, Faculty of Economics and Administrative Sciences, Department of Econometrics, Van, TURKEY

Abstract

Models that can predict the class of new cases are created by using classification algorithms in data mining. These models provide easier understanding and analysis of data. However, the presence of a large number of low-impact variables in the data set to be classified poses a problem. Dimension reduction method, which is a process in which variables are reduced, provides easier interpretation of high dimensional data sets with a certain loss. In this study, 7 different types of Dry beans with 17 variables, Decision Tree, Naive Bayes and Support Vector Machines classification algorithms were used. First the raw data set and then the size of the data set was reduced with Principal Component Analysis and the data set was classified and the performances of the classification algorithms according to three different measurement values were examined. Analyzes were made with the WEKA program. Considering the weighted averages in the analysis results, the highest and lowest classification accuracies for raw data were obtained from the Naive Bayes algorithm with Roc area (99%) and accuracy (89.7%), respectively. In addition, in the data set reduced with PCA, the highest and lowest classification accuracies were obtained from the Naive Bayes algorithm with Roc area (98.7%) and accuracy (89.5%). Although the size of the data set was reduced by PCA, the classification algorithms performed very well.

Keywords: *Dimension reduction, PCA, Classification*

Distributed Intrusion Detection System Using NSL-KDD and SK-DIST on Apache Spark

Mohamed Seghir Othman Djedıden[1], Hicham Regueg[2], Zoulikha Mekkakıa Maaza[3]

[1] Laboratoire SIMPA, Faculté des Mathématiques et d'Informatique Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, Oran, ALGERIA

Abstract

Network security is very important in today's data communications environment. Hackers can create many successful attempts to crash networks and web services by the unauthorized intrusion. Intrusion detection systems (IDS) becomes an essential part of building a computer network to capture attack at an early stage because IDS works against all intruder attacks. With the massive data generated in computer networks. The main task of IDS has become more complicated. Most existing IDS are deployed on a single server and have encountered several problems since the volume of data to be analyzed is larger. In our previous work, we create a distributed and powerful IDS based on an optimized version of the Scikit-learn library named SK-Dist. The proposed IDS has been tested and evaluated on the UNSW-NB15 dataset and is effective in terms of accuracy, scalability and fault tolerance. In this research, we aim to prove the effectiveness of our previous approach by applying it to the NSL-KDD dataset. This approach consists of creating a distributed IDS which supports big data analysis and which ensures better detection accuracy while using the minimum number of features. The proposed IDS is a combination of the features selection methods (Chi2 and RFE) included in the Scikit-learn library, the classifiers integrated into the optimized Scikit-learn library named "Sk-Dist" (supports the distribution of the classifier random Forest in a cluster) and the Apache Spark framework to provide the processing cluster which is more suited to big data analysis. The first step of our approach is the data preprocessing, in this step we will load the NSL-KDD data (training and testing data) in Apache Spark, then we will convert the threes categorical features (protocol_type, service, flag) to numerical features using the integrated encoder of the Scikit-learn library named Label-Encoder and since we are targeting a Binary-classification we will convert the label which contains the 39 subcategories of attacks to a binary value (1 for attacks and 0 for normal traffic), after this, all features will be formatted to the same scale using the Scikit Learn methods named Min-Max –Scaler. The second step is the feature selection where the aim is to reduce the number of features to a minimum while ensuring better detection accuracy. in this step, we have opted for two selection methods integrated into the scikit-learn library (Chi2 and RFE). The Scikit-learn library offers classifiers that do not support distribution on a cluster, so in the third step, we integrated the Sk-Dist package with its distributed version of the Random Forest (RF) classifier. Using this package, our IDS will be distributed in a Spark cluster, which makes it more available and fault tolerant. The RF classifier has a set of parameters, the choice of the values of these parameters directly influences the accuracy of the model. The challenge is to find the best combination of parameters that ensures the best performance of the model. This is why for the optimization of the parameters we have opted for a new package named Hypopt, the main advantage of this package compared to the functions integrated in Sickit-learn is its ability to execute the optimization loop on a cluster which makes it faster. To evaluate the performance of our approach, two parameters are used: the accuracy and the weighted score f1. These two metrics will be used for comparison with other related work. For the comparisons, we have ignored some related work because they only used custom subsets of NSL-KDD hence the impossibility to compare our performances with theirs, or they used traditional

Keywords: *Intrusion Detection, Machine Learning, Big Data, Distributed Computing, Apache Spark, Scikit-learn, NSL-KDD*

An Enhanced Round Robin Cpu Scheduling Algorithm with Variable Quantum Time

Alaa Fiad [1], Zoulikha Mekkakia Maaza 1

[1] Laboratoire SIMPA, Faculté des Mathématiques et d'Informatique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, ORAN, ALGERIA

Abstract

An operating system is a program that manages the computer hardware. CPU scheduling is one of the fundamental features that an Operating System (OS) needs to perform multitasking; it is the basis of multiprogramming systems. It refers to a set of policies and mechanisms to control the order of work to be performed by a computer system. It is made by the part of the operating system called the scheduler, using a CPU scheduling algorithm. One of the most used scheduling algorithms is the Round Robin (RR); it is framed mainly for a time-sharing system. The mechanism being that each process is executed in First Come First Serve (FCFS) order in the given time slice, and those greater than that time quantum are sent to the back of the ready queue, where the remaining processes are waiting for their execution. The selection of time slice is important as it affects the performance of the algorithm. The existing improvement RR algorithms do not consider different parameters to define the (TQ) value, which affects the performance of the OS. The proposed algorithm is an improved version of the RR algorithm using a dynamic time quantum (TQ) considering the arrival time and the burst time of each task. The main contribution of this approach is: 1) Take into account parameters ignored in existing algorithms like (the arrival time and burst time of the task). 2) Minimising the average waiting time of the tasks. 3) Minimising the average turnaround time of the tasks. 4) Maximizing the CPU utilization, ensuring optimal use of resources. The performance of our proposed algorithm has been analysed through two cases (same arrival time, different arrival time) and compared with Dynamic Average Burst time Round Robin (DABRR), An Efficient Dynamic Round Robin Algorithm for CPU scheduling (EDRR). In the case of same arrival time: the tasks arriving in the ready queue are arranged in the ascending order of their burst time, using the burst time of the middle task, the ready queue is divided into two sub-queues as follow: tasks having burst time less than the burst time of the middle task are stored in the first queue (light task queue), and the rest in the second queue (heavy task queue). The tasks in the first queue are executed first following these steps: The CPU is allocated to tasks using RR scheduling with a small time unit, called time quantum equal to the average burst time of existing tasks in the queue, the process is repeated until the queue is empty. After all the tasks in the first queue have been executed, the same steps are applied to schedule the tasks of the second queue. In the first example the average waiting time is 94.6 ms in the proposed algorithm and 102.8 ms in EDRR. The average turnaround time is 154.4 ms in the proposed approach and 162.6 ms in EDRR. When different tasks arrive in the ready queue, the TQ is set equal to the burst time of the first task, when different tasks are in the ready queue the same steps the same steps mentioned in the previous algorithm are applied. The average waiting time is 85.2 ms in the proposed approach, 97.8 ms in DABRR algorithm and 94.6 ms in EDRR, the average turnaround time is 144.8 ms in the proposed approach, 157.4 ms in DABRR algorithm and 154.2 ms in EDRR. The results show that the proposed approach gives better performance in terms of minimizing the average waiting time, and turnaround time compared to DABRR and EDRR. These improvements ensure significant time savings and better allocation of resources.

Keywords: *Operating system, Task scheduling, Round Robin, Waiting time, Turnaround time*

Development Of A Land Cover Mapping Method For An Algerian Steppe Region (Djelfa Region) Using Multispectral Remote Sensing

Saida Sadi[1], Fateh Karim Amghar[1], Nour El Islam Bachari[2]

[1]M'Hamed Bougara University of Boumerdès, Boumerdes, ALGERIA

[2]University of Science and Technology Houari Boumediene, Bab Ezzouar, ALGERIA

Abstract

The rangelands of the Algerian high steppe plains cover a large area, about 20 million hectares, or 8.4% of the total area of the country. These steppes play a fundamental role in the agro-pastoral economy. Unfortunately, they are, currently, very threatened. For some decades, these ecosystems have been experiencing profound changes that are manifested by the modification of composition, structure, as well as functioning of these steppe landscapes. These changes are linked to anthropic actions and variability of climatic factors. The rate of degradation is increasing causing an ecological and socio-economic imbalance. The objective of this work is to use remote sensing techniques to map the land use of an Algerian steppe region (Wilaya of Djelfa) where desertification has affected biodiversity, soil and water resources. The study is based on the exploitation of Landsat 8 OLI image data from the month of April and May of the year 2020. The methodology adopted is based on the combination of unsupervised classification, identification of soil elements using their spectral properties, photo-interpretation of very high-resolution images available free online, and then field verification to discriminate the different land use classes. The results of the spectral behavior analysis showed that we can distinguish and describe the relationship between the different elements of the soil surface based on the principle that each element on the soil surface is characterized by a spectral signature of its own and thus map the main land cover classes of the study area.

***Keywords**: Land cover, steppe, remote sensing, spectral signature*

Can you write the Script of Happiness Artificial Intelligence?:Augmented Scriptwriter and the Changing Psychology of Scriptwriting

Seçkin Sevim[1], Bilgen Aydın Sevim[2]

[1]*Marmara University, Faculty of Fine Arts, Film Design and Direction, Istanbul, TURKEY*

[2]*Sakarya University, Faculty of Art, Design and Architecture, Visual Communication Design, Sakarya, TURKEY*

Abstract

The script is the first step in the filmmaking process. Professionals involved in the production and screening chain of the cinema industry, from director to actor, producer to distributor, are in search of a creative and original script. This quest is also a brief summary of the history of the art of cinema for one hundred and twenty-five years. The expectations of the audience and the industry put great pressure on the scriptwriters. Scriptwriting is an intrapersonal communication process that requires creative solutions to storytelling problems. This asocial life spent in front of the computer can bring many physiological and psychological disorders. There is no guarantee that this process will always be successful. A scriptwriter may have to shelve the script written with great effort. Artificial intelligence, which has become a part of daily life, also works to overcome these problems in the field of scriptwriting based on artistic creativity. Using the cooperation of artificial intelligence in scriptwriting has come to the fore with Plots Unlimited and Collaborator softwares. These programs were released in the 1990s when machine learning and algorithms were not powerful enough. In recent years, new softwares have emerged with the help of big data enabling more successful algorithms to be created and machine learning getting stronger. Using a data set of thousands of films, these programs offer various options to overcome the "writer blockage" experienced in the scriptwriting process. Thanks to the collaboration of these softwares designed as a kind of co-author, it is aimed to reduce the creativity, budget and time pressure on the scriptwriter. The aim of this study is to discuss the psychology of scriptwriting that has changed with the introduction of artificial intelligence in the digital age. Deepstory, Script R and Master Writer softwares were selected within the scope of the appropriate sample and analyzed by document analysis method. Before artificial intelligence, a scriptwriter was a lonely professional who tried to solve storytelling problems on his own. The scriptwriter of the digital age, on the other hand, turns into an augmented scriptwriter that optimizes his creativity thanks to his collaboration with artificial intelligence.

Keywords: *Artificial Intelligent, Augmented Scriptwriter, Intrapersonal Communication, Author Blockage*

COVID-19 Literature Search Engine with Natural Language Queries

Hilal Tekgöz[1], Halil İbrahim Çelenli[1]

[1]IBSS Consulting, Research and Development Department, Istanbul, TURKEY

Abstract

COVID-19 is causing a global crisis that affects many areas of human life around the world. Researchers and doctors are conducting various studies to cope with the global epidemic and prevent the spread of the epidemic and publish these studies on academic literature platforms. Therefore, the importance of academic literature platforms has increased during the pandemic process. Correct queries should be made in order to reach the desired studies on the academic literature platforms. Academic studies on COVID-19 are increasing day by day, reading each study is a challenging process for doctors and researchers. Academic studies on COVID-19 have been increasing day by day, reading each study has been a challenging process for doctors and researchers. It is useful to use important keywords in the search step in academic literature search platforms to find the right articles. However, with natural language queries, the right articles can be accessed quickly and easily. In this study; A BERT model has designed out of 150 thousand academic studies collected on COVID-19, SARS-CoV-2, and coronaviruses. While designing the model, academic studies have pre-processed and embedding vectors have produced. Cosine similarity method has used on embedding vectors to help users find the right articles. The using natural language queries on the web application, the articles with the highest score have shown to the users.

Keywords: *Natural Language Processing, COVID-19 Literature Search Engine, Cosine Similarity, BERT*

Covid 19 Forecasting with Artificial Neural Networks and ARIMA Model

Büşra Çetin[1], Nida Gökçe Narin[2]

[1]Muğla Sıtkı Koçman University, Institute of Science and Technology, Statistics, Muğla, TURKEY

[2]Muğla Sıtkı Koçman University, Faculty of Science, Statistics, Muğla, TURKEY

Abstract

Sars-Cov-2 virus, which emerged in Wuhan, China in December 2019, spread rapidly all over the world. Despite the vaccines developed, it has not been fully controlled yet. The restrictions imposed due to the pandemic have been stretched for economic reasons in many countries. Uncertainty in the number of cases in countries also affects the normalization process. For this reason, analyzing the situation that has emerged with the effect of the policies followed up to now and estimating the future course of the pandemic maintains its importance in terms of measures to be taken and restriction decisions. In this study, Germany, France, Italy, Ireland, Poland, Russia, and Turkey for the number of cases daily, the number of patients, number of deaths, and so on. It is aimed to select the best model to predict the possible course of the Covid-19 pandemic with ARIMA models and Artificial Neural Network models. RMSE, MAPE, MAE, MASE, AIC, and BIC criteria were used in model selection.

Keywords: *Covid-19, ARIMA, Artifical Neural Networks, Forecasting*

Detecting Smokers Using Artificial Neural Networks

Levent Civcik[1], Osama Alkayal[2]

[1] Konya Technical University, Technical Sciences Vocational School, Department of Computer, Konya, TURKEY

[2] Konya Technical University, Department of Electrical & Electronics Engineering, Konya, TURKEY

Abstract

In the last few years, deep learning has been used in many application areas and has been giving outstanding results. Deep learning is one of the machine learning classes that use deep structures and hierarchical learning approaches that have largely been developed since 2006. Due to their structure that consists of many layers between the input and output layers, deep learning approaches are quite successful in the pattern classification of nonlinear information. In this paper, using various images, it is aimed to determine whether a person is smoking or not using artificial intelligence approaches. Under this scope, an artificial neural network that uses deep learning approaches was developed to classify the images and determine whether the people in the images are smoking or not. More than 1000 colored images have been used to train and test the neural network.

Keywords: *Deep learning, Image Processing, Artificial Intelligence, Artificial Neural Networks*

Prediction of Covid 19 Spread with Box-Jenkins Models

Nida Gökçe Narin [1], Gamze Yüksel[2]

[1]*Muğla Sıtkı Koçman University, Faculty of Science, Statistics, Mugla, TURKEY*

[2]*Muğla Sıtkı Koçman University, Faculty of Science, Mathematics, Mugla, TURKEY*

Abstract

The Sars-Cov-2 virus, which affects approximately 128 million people worldwide, has still not been brought under control, despite vaccination studies and global restrictions. The fact that many countries have loosened restrictions for economic reasons has brought the second and third waves in the epidemic. Analyzing the spread of the epidemic with time-series approaches and making predictions will help officials in the decision-making process. In this study, cases, and number of deaths, the highest first 20 countries (respectively the United States, Brazil, India, France, Russia, United Kingdom, Italy, Turkey, Spain, Germany, Colombia, Argentina, Poland, Mexico, Iran, Ukraine, South Africa, Peru, Czech Republic, and Indonesia) will be modeled with Box-Jenkins methods using daily Covid 19 data. As a variable, the number of new cases per day, the number of patients recovering, and the number of deaths for each country will be considered. Model selection will be made according to RMSE, MAPE, MAE, MASE, AIC, and BIC criteria.

Keywords: *Covid-19, Box-Jenkins Models, Prediction*

The Contribution of Explained Artificial Intelligence (XAI) to the Covid-19 Virus: A Systematic Literature Review

Zeynep Aytaç[1]

[1]*Aksaray University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Aksaray, TURKEY*

Abstract

According to World Health Organization, as of March 2021, there are approximately 2 million 796 thousand deaths, and approximately 127 million 877 thousand cases caused by Covid-19 virus worldwide. For more than a year, scientific research and vaccine studies have been carried out to fight coronavirus. However, due to the mutation of the virus and rapid spread of its variants, health systems are in a difficult situation and concerns are increasing. In this study, it is aimed to reveal the contributions of Explained Artificial Intelligence (XAI) methods in terms of diagnosing coronavirus and reducing the rate of spread, with a systematical literature review. Scientific articles published after 2019 in Web of Science, Scopus and Science Direct databases based on Explained Artificial Intelligence and Covid-19 keywords, which can reveal the reliability of Artificial Intelligence, were examined and findings are discussed.

Keywords: *Expainable Artificial Intelligence (XAI), Covid-19, Coronavirus, Pandemic*

Stroke Detection from CT Images

Gamze Yüksel[1], Hakan Sökün[2]

[1]*Mugla Sıtkı Koçman University, Faculty of Science, Mathematics, Muğla, TURKEY*

[2]*Muğla Sıtkı Koçman University, Institute of Science and Technology Artificial Intelligence USA, Muğla, TURKEY*

Abstract

In this study, a machine learning method was developed to determine cerebral palsy on Computed Tomography (CT) images. Cerebral palsy has more than one type. In this study, open-source data sets belonging to the type of hemorrhagic cerebral palsy were used. It was determined whether there was a type of hemorrhage causing cerebral palsy from the images obtained by the CT method, and it was classified. First of all, the data set was separated from the noisy samples and a set was composed by combining the images with the same attributes. After that, hemorrhagic and non-hemorrhagic brain CT images were separated by binary classification using machine learning methods such as singular value decomposition and principal component analysis. Here, it is aimed to determine whether there is cerebral palsy in the CT image by extracting the features (eigenvectors) that characterize the CT images with singular value decomposition and principal component analysis methods.

Keywords: *Stroke Detection, Classification, Singular Value Decomposition, Principal Component Analysis*

Application for Automatic Image Colorization Based on Deep Neural Networks

Emre Dandıl[1], Bilal Aral[1]

[1]*Bilecik Şeyh Edebali University, Faculty of Engineering, Computer Engineering, Bilecik, TURKEY*

Abstract

Image and video colorization stands out as a popular application area in image processing and computer vision. Especially in recent years, deep neural networks are widely proposed for image coloring. There are different methods in image coloration such as guided by an expert with manual, semi-automatic and fully automatic. In this study, a fully automated method based on deep neural networks is proposed for colorization of gray-level images. Encoder and Decoder models are used in the structure of the proposed deep neural network. The proposed deep neural network is trained on the open datasets ImageNet and VisualGenome, and an application with a user interface for image colorization is developed. Within the scope of the proposed study, in the experimental studies carried out on the developed application, the gray-level images are successfully colored in the LAB color space.

Keywords: *Image Colorization, Deep Neural Networks, Encoder and Decoder, Software Development, ImageNet, VisualGenome*

Servo Motor Control with PLC and HMI Panel

Levent Civcik[1], Alperen Aksin[2]

[1] Konya Technical University, Technical Sciences Vocational School, Department of Computer, Konya, TURKEY

[2] Konya Technical University, Department of Electrical & Electronics Engineering, Konya, TURKEY

Abstract

Servo motors, step motors, asynchronous motors are among the widely preferred electric motors in the industry. Among these motors, servo and stepper motors are used more in applications that require position control. Servo motors are preferred in the industry to make high performance position control.

In the designed application, it is aimed to drive two servo motors sequentially with the commands received from the HMI panel. This application is made with OMRON servo motor driver and servo motor, ladder diagram is written in SYSMAC STUDIO. The start and stop signals of the engines are made with the buttons positioned on the prepared screen. Movement is made as Relative mode . Thus, the motor is provided to go to the desired point in a linear (linear) way even though the movement is rotary to the specified distance. In addition, real-time speed and position values are shown on the indicators on the HMI panel during the movement.

Keywords: *Servo motor, PLC, HMI Panel*

Categorical Principal Component Analysis and Application with Depression Data Set

Canan Demir[1], Sıddık Keskin[2], Hamit Mirtagioğlu[3], Yıldırım Demir[4]

[1]Van Yüzüncü Yıl University, Vocational School of Health Services, Van, TURKEY

[2]Van Yüzüncü Yıl University, Faculty of Medicine, Department of Biostatistics, Van, TURKEY

[3]Bitlis Eren University, Faculty of Arts and Sciences, Department of Statistics, Bitlis, TURKEY

[4]Van Yüzüncü Yıl University, Faculty of Economics and Administrative Sciences, Department of Statistics, Van, TURKEY

Abstract

Categorical Principal Component Analysis (CATPCA) is a multivariate statistical analysis method used to reveal the correlations between independent variables that affect the dependent variable, as well as dimension reduction and visualization. Considered geometrically, it aims to graphically represent variables and categories in a lower dimensional space rather than real space. Thus, results are interpreted with several components instead of many variables. In the method, after determining component loads for each dimension in multiple nominal and multiple non-nominal variables, eigenvalues and total explained variance are calculated. In Categorical Principal Component Analysis, categorical variables are digitized with various transformations and the loss function allows the application of multivariate analysis methods. Thus, relationships in real space with minimum loss can be shown in a lower dimensional space. In the literature review, it was observed that there was almost no Turkish literature on the subject, and in this study; It was aimed to make an application in the depression data set in order to explain the method, to mention the basic concepts and to contribute to the understanding of the subject.

Keywords: *Depression, Optimal Scaling, Multiple Nominal, Size Reduction*

Filtering Best Models Containing Missing Covariates by Using T-step Occam's Window

Sezgin Çiftçi[1]

[1]*Insurance Department, Başkent University, Ankara, TURKEY*

Abstract

Model averaging methods are used by constructing a model space in order to reduce the uncertainty about best fitting model selection that may be problematic in estimation. Since, handling a large model space is challenging in numerical calculations, Occam's window method is a popular tool for reducing the model space. However, this method is easy to apply when only the data is fully observed. If the data contains missing parts, the reduced model space may vary upon the imputation methods for handling the missingness. The focus of this study is reducing the inconsistency of filtering the best fitting models containing missing covariates by using an adaptive method called T-step Occam's window. It starts with constructing *t* model spaces with different sizes by obtaining *t* different datasets of which the missing parts are re-imputed *t* times and performing Occam's window for every obtained dataset. Then, the model space is filtered by choosing arbitrarily *s* most frequent common models occurring in these model spaces. The method is applied under different *t* and *s* values to a suitable dataset used in a reference model selection study for observing the consistency. Then, the model fits are compared with the ones in the reference study.

Keywords: *Bayesian approach, Missing data, Model filtering, Occam's window*

Relations Among Efficient Elements of a Set and Some Scalarizing Functions

İlknur Atasever Güvenç[1]

[1]Eskişehir Technical University, Faculty of Science, Department of Mathematics, Eskişehir, TURKEY

Abstract

In the literature, there are several approaches for solutions of set-valued optimization problems. One of them is set optimization approach and is based on comparing values of given objective map. Some ordering relations are required to do this comparison. Karaman et al. defined two partial order relations on family of nonempty and bounded sets by using a cone. In addition, they defined two scalarizing functions which help to find solutions of a set optimization problem with respect to these relations. They reduced the given set optimization problem to a scalar problem via these scalarizing functions and presented optimality conditions. In this study, some properties of these scalarizing functions are given. It is shown that instead of boundedness, cone boundedness of a component of scalarizing functions implies that scalarizing functions take value of greater than $-\infty$. In addition, some conditions are presented for scalarizing functions to be equal or greater than a real number. By using monotonicity properties of the scalarizing function some necessary and sufficient conditions are presented for weakly efficient elements of a set. In order to give necessary and sufficient conditions for weakly minimal elements of a set by means of second component of the scalarizing function, compactness of the set is required. Moreover, it is shown that conditions can be given by means of the first component without assumption of compactness. These conditions are given via scalarizing functions which are useful tools for set optimization approach and enable us to check whether a point is weakly efficient point of a set or not.

Keywords: *Weakly minimal element, Weakly maximal element, Scalarizing function*

Investigation of Virus-Host Interactions Between Sars-Cov-2 and Human Proteins by Composition Moment Vector Feature Method

Firdes Gul Korkut[1], Murat GÖK[2]

[1]*Yalova University, Institute of Science and Technology, Department of Computer Engineering, Yalova, TURKEY*

[2]*Yalova UniversitY, Department of Computer Engineering, Yalova, TURKEY*

Abstract

Covid19 disease, which has caused mass deaths in the last two years worldwide, has been declared as an epidemic (pandemic) disease by the World Health Organization (WHO). The Sars-CoV-2 virus, which causes Covid19, has had a devastating effect on humanity due to its spread rate and deadly effect. In this study, we predicted the places where Sars-CoV-2 virus proteins interact with host (human) cells using machine learning methods. In our study, we used a protein interaction dataset consisting of 30,046 negative and 20,365 positive data. To predict protein interaction sites, we first encoded the protein sequences with the Composition Moment Vector feature coding method. Then we classified the numeric features vectors we obtained with machine learning algorithms. According to the experimental results we obtained, the k-Nearest Accuracy algorithm gave the best performance with 66.9% accuracy, 0.765 sensitivity and 0.663 F-score values.

Keywords: *Composition moment vector, Protein coding, Classification*

Prediction of Anti-CoV Protein Sequences Using Machine Learning Methods

Hasibe Candan[1], Murat Gök[2]

[1]Yalova University, Institute of Science and Technology, Department of Computer Engineering, Yalova, TURKEY

[2]Yalova UniversitY, Department of Computer Engineering, Yalova, TURKEY

Abstract

The coronavirus, which caused a worldwide epidemic, infected more than 167 million people in the world and caused the death of 3.5 million people. The treatment for Covid-19 disease caused by the coronavirus has still not been found. Detection of anti-coronavirus (anti-CoV) peptides with functional activity is of great importance in the treatment of the disease. In this study, we used machine learning methods to detect anti-CoV proteins. First, we quantified the protein sequences using the Amino Acid Conjugation (AAC), Conjoint Triad (CT), and Amino Acid Pair (AAP) protein coding methods. Next, we classified it with k-Nearest Neighbor, Random Forest, Naive Bayes and BayesNet algorithms. According to the experimental results we obtained, the Naive Bayes algorithm gave the best performance on protein data encoded with the AAC method with 84.66% class accuracy, 0.810 sensitivity, 0.883 specificity, and 0.695 Mathew Correlation Coefficient.

Keywords: *Protein encoding, Amino acid combination, Naive bayes, Classification*

Development of Electrochemical Aptasensors That Can Determine Phosphate Ions

Elif Esra Altuner[1], Veli Cengiz Özalp[2], Mahmut Deniz Yılmaz[3], Havvanur Tatlı[4]

[1]Sen Research Group, Department of Biochemistry, University of Dumlupinar, 43000 Kutahya, TURKEY

[2]Medical School, Department of Medical Biology, Atilim University, 06830, Ankara, TURKEY

[3]Department of Bioengineering, Faculty of Engineering and Architecture, Konya Food and Agriculture University, 42080 Konya, TURKEY

[4]Selcuk University, Kulu Vocational School, Konya, TURKEY

Abstract

Biosensors are highly selective devices used in agriculture, pharmaceutical chemistry, tissue chemistry, and many other branches. Aptasensors are biosensors in which aptamers are used as affinity molecules. Enzymes act as catalysts in sensors, and artificial enzymes have been sought due to their high cost and lack of stability under certain conditions. In this study, an aptasensor that can determine the amount of phosphate and give a signal depending on the peroxidase activity has been developed. For this purpose, an artificial enzyme has been searched to provide the most appropriate quality to the aptasensor format used in this study. First, chitosan-cobalt (II) (CTS-Co (II)), and then when the cobalt ion was tested with chitosan with separate measurements, the desired result could not be obtained, and the last commercially available palladium carbon (Pd / C) was studied and peroxidase (HRP) activities were examined. The most favorable peroxidase enzymatic results were found in Pd/C. After Pd/C imitates the HRP enzyme, aptasensor study was started. In the studies, 3.3'5.5 'tetramethylbenzidine (TMB) was chosen as the substrate. First, mesoporous silica nanoparticles (MSNPs) were synthesized TMB is confined in MSNPs and attached with phosphate ions by covalent bonds. Thus, TMB release was accelerated in the presence of phosphate ions. In the characterization analysis of aptasensor measurements, the screen image with transmission electron microscopy (TEM) and particle sizes with DLS & ELS were examined. Assistance was received from FTIR, ICP-OES, and TGA for CTS-Co (II) and CTS and commercial Pd/C. Also, studies were carried out to develop sensors with cyclic voltammetry (CV), detection limit (LOD), limit of quantification (LOQ) in electrochemical applications, and an aptasensor based on phosphate sensitivity was developed.

Keywords: *Biosensor, artificial enzyme, Pd/C, CTS-Co (II), TMB (ox), Aptamer, MSNPs*

Properties of Fluorescence and Electrochemical Applications in the Determination of Schiff Bases

Elif Esra Altuner[1], Havvanur Tatlı[2]

[1]Sen Research Group, Department of Biochemistry, University of Dumlupinar, 43000 Kutahya, TURKEY

[2]Selcuk University, Kulu Vocational School, Konya, TURKEY

Abstract

Schiff bases are the condensation products of aldehydes or ketones with a primary amine. Schiff bases are preferred as a good ligand in the determination of these metals, since Schiff bases form solid complexes with many metals and especially with transition metals. It is necessary to determine and know all the chemical properties of Schiff bases and metal complexes used in many fields. When the fluorescence properties of the molecules are examined, flatness in molecules, inhibition of rotation, conjugation and increase in the number of rings generally increase the fluorescence efficiency. The most intense fluorescence rays give compounds containing aromatic rings that allow the transition of low energy π → π * in their structure. In addition, aliphatic and alicyclic aromatic rings that contain many conjugated double bonds in their structure also show fluorescence properties. Since n → π * low energy electronic transitions in simple nitrogen-containing heterocyclic rings, the transition between states becomes easier and the excited singlet state easily turns into the excited triplet state. The transformation from the triplet state to the basic state occurs by emitting phosphorescence, and where the phosphorescence is present, the fluorescence decreases or disappears. When a ligand giving a chelate complex complexes with a cation, its rigidity increases and thus fluorescence emission increases. The strongest and most beneficial fluorescence occurs in compounds containing aromatic functional groups with low energy π → π * transitions. Substituents in a luminescent compound that can delocalize the (bileş) electrons of the compound usually increase a possible light transmission between the excited singlet state and the ground state. This result also increases the fluorescence. The fluorescence and electrochemical properties of some Schiff bases and metal complexes in various techniques have been studied in this review considering all these properties.

Keywords: *Schiff base, Electrochemistry, Fluorescence*

The Positive Effect of The Pandemic on Diversification Strategy in Bist 30 Investments: An Assessment within The Framework of Portfolio Optimization with Bist 30 Shares

Gamze Vural[1], Serkan Nas[2]

[1]Çukurova University, Academic Data Management System, Adana, TURKEY

Tarsus University, Department of Management Information Systems, Mersin, TURKEY

Abstract

The Covid-19 pandemic has had vital effects in many areas all over the World. The first impact on financial markets was the severe decrease in stock markets all over the World. In Borsa Istanbul, in February 2020 there have been sudden decreases but as of March 2020 the increase in index stock with the increase in the transaction volume has started and stock market investors have again started to gain. In fact, after November 2020 the most significant increases in BIST transaction volume in its history have been recorded and as of March 2021, the transaction volume has nearly tripled the pre-covid-19 period. However, in investment, it is a more convenient approach to evaluate the return and risk together. Risk is defined as the deviation from expected and can be calculated as the standard deviation of the return of a financial asset. Creating a portfolio makes it possible to get higher returns at a certain risk level or to get a certain return at a lower risk level. The risk of the portfolio is based on the concept of diversification, and diversification is based on the logic of bringing assets together where the correlation between their returns is negative or weak. Markowitz laid the foundations of modern portfolio theory with the optimum portfolio theory. In the years 1952-1959, he developed and found out the solution to the portfolio selection problem with the mean-variance method. The theory combines probability and optimization techniques in case of uncertainty. The average variance model of Markowitz mathematically can show which asset portfolio in what proportion should be in the optimum portfolio by aiming to minimize the risk at a certain level of return or by maximizing the return at a certain risk level. The significant increase in the number of investors and transaction volume in 2020 has motivated this study. In the study, it was tried to determine how the characteristics of the optimum portfolio which was selected from BIST 30 stocks in 2020 where the risk was minimized, varied compared to previous periods. For this reason, weekly return data of the stocks included in BIST30 for the period 2016-2020 were used, to determine the optimum portfolio weights (wi) and to minimize the coefficient of variation (portfolio risk/portfolio return) with the assumption that there was no short sale within the framework of the Markowitz modern portfolio model. When the structure of the optimum portfolios in the 2016-2019 period is analyzed, it is seen that the optimum portfolios consist of a small number of stocks in these years. Besides, the analysis shows us that a few of these shares entails high weights. In 2016, the optimum portfolio includes only 8 out of 30 shares. The three shares with the highest weight constitute 83% of the portfolio. In 2017, the optimum portfolio was formed with 11 shares, with the highest weight of three shares constituting 63%, and four shares constituting 74% of the optimum portfolio. In 2018, 6 out of 30 stocks were included in the optimum portfolio, and the weight of the first three stocks was found to be 77% in the optimum portfolio. In 2019, the market situation looks worse in terms of creating diversification. The optimum portfolio consisted of only four stocks and the two stocks with the highest weight had 81% share in the investment. In 2020, 16 out of 30 stocks were included in the optimum portfolio. Also, unlike previous years, it is observed that the weights do not concentrate on a few stocks. It has been determined that the shares with the highest proportion have a 15% share in the portfolio, and the total proportion of the first three stocks is around 38%. This situation can be interpreted as Borsa Istanbul is in a better position in terms of offering diversification opportunities in 2020. In the next stage of the study, the analysis to be made with the stocks in BIST 100 will clarify more this interpretation. It is observed that the falling asset prices in the first stage due to the pandemic shock makes stock investment attractive. Especially, during the pandemic period, the increase in the time remaining out of work and the individuals' being able to have more time for investing in stocks cause an increase in the number of new investors and the number of funds transferred to the stock market. It can be said that the increase in transaction volume and the number of new investors in 2020, especially after the second half, positively affected Borsa Istanbul. It is of great importance for the development of the market that this situation is not temporary and an increase in transaction volume, so an increase in diversification will be sustainable.

Keywords: *Pandemic, Portfolio Optimization, Diversification,BIST*

Estimation of Incomplete Precipitation Data Using the Adaptive Neuro-Fuzzy Inference System (ANFIS) Approach

Okan Mert Katipoğlu[1]

[1]Erzincan Binali Yıldırım University, Faculty of Engineering, Department of Civil Engineering, Erzincan, TURKEY

Abstract

To plan and manage water resources effectively, many meteorological and hydrological data such as precipitation, streamflow, evaporation, temperature, humidity, and infiltration must be at least 30 years and continuous. In this study, it was aimed to complete the missing precipitation data at the Erzincan precipitation observation station by using the adaptive neuro-fuzzy inference system (ANFIS). For this reason, Bayburt no.17089, Tercan no.17718, and Zara precipitation stations no.1716 were used, which are the closest to Erzincan station 17094 and have the highest correlation coefficient. In the ANFIS model, monthly total precipitation data (52 years) between 1966 and 2017 were used. In the model established, 80% of the data (1968) were used for training and 20% (492) for testing. In the ANFIS model, variables were tried by dividing them into sub-sets between 3 and 8. The most suitable ANFIS model was determined according to the error values and determination coefficients of the training and test results. As a result of the study, 3 sub-sets, hybrid learning algorithm, trimf membership function, and model with 600 epochs were selected as the most suitable model.

Keywords: *Precipitation, Completion of Missing Data, ANFIS, Erzincan*

University Students' Privacy Concerns Towards Social Media Platforms: Whatsapp Contract Change

Ceren Çubukçu[1], Cemal Akturk[2]

[1]Maltepe University, Engineering and Natural Sciences Faculty, Computer Engineering, Istanbul, TURKEY

[2] Gaziantep Islam, Science and Technology University, Engineering and Natural Sciences Faculty, Computer Engineering, Gaziantep, TURKEY

Abstract

The recent Covid-19 pandemic has changed our lives drastically. As a result, we have started to communicate online more than ever. As a result, social media and social messengers have started to take more part in our daily lives. Facebook, Twitter, Instagram, Youtube, Whatsapp and Telegram are among the most popular social applications that have been used worldwide. In January 2021, Whatsapp announced a change in its data privacy policy for the end users located in Turkey. This change has created a great global awareness as well as concern on data privacy and security. This study will mainly research whether this change made people switch to other applications such as Telegram or Signal. It will also analyze whether there is a gender difference in using social applications. The study uses the survey method for data collection. A total of 489 students filled out the survey studying in 8 different universities. The data of this research have been examined using frequency analysis. This study distinguishes from others in the literature by especially focusing on the contract change of Whatsapp and users' behaviors towards this contract change.

Keywords: *Social Messengers, Whatsapp, Frequency Analysis, Privacy Policy, Covid-19*

[1] İletişim e-posta: ceren.cubukcu@gmail.com

Information Extraction from Invoice Images Using Character Vectorization Technique

Adem Akdogan[1], Resmiye Nasıboglu[2]

[1]*Dokuz Eylül University, Institute of Science and Technology, İzmir, TURKEY*

[2]*Dokuz Eylul University, Faculty of Science, Computer Science, Izmir, TURKEY*

Abstract

The transfer of physical documents to digital media has started with the development of technology. Especially keeping the documents physically makes it difficult to reach the desired information quickly. In addition, as the number of documents increases, physical storage and preservation become more difficult than digital storage. However, since the number of documents is too high for large companies, it is often difficult for these companies to transition to digital environment. In this work, the invoices, which are generally the most physical documents in companies, were examined. With the help of artificial intelligence, it is aimed to obtain data that has critical importance such as invoice amount, invoice number, invoice date. Tesseract (Optical Character Recognition) engine was used for the analysis of invoice images. The training process is carried out with the attributes both obtained from the Tesseract engine and calculated features. At this point, character vectors have been added to increase success. Overall success has increased with this procedure. Random Forest, Gradient Boosting Machine, Extreme Gradient Boosting, K-Nearest Neighbor, AdaBoost and Decision Tree algorithms were used for training. A total of 9910 invoices were used for 80% training and 20% testing. The F1 score value of the Random Forest model, which is the main model, was obtained as 0.91.

Keywords: *Tesseract, Character Vectorization, Machine Learning*

Analysis and Segmentation of X-ray Images of COVID-19 Patients using the k-means Algorithm

Ahmet SAYGILI[1]

[1]Tekirdağ Namık Kemal University, Çorlu Engineering Faculty, Computer Engineering Department, Tekirdağ, TURKEY

Abstract

Medical imaging techniques have been used frequently in computer-aided systems in recent years. Computer-aided automatic diagnosis systems created using medical images are of a nature that will benefit medical professionals in their decisions. Image processing methods are also used to create these diagnostic systems. The main purpose of image processing is to reveal meaningful information in the image. For this purpose, many studies such as cancer detection, tumor detection, anomaly detection from medical images are carried out. In this study, the process of determining the anomalies on the X-ray images by segmentation is carried out. The X-ray images used are images of COVID-19 patients and healthy individuals. As it is known, COVID-19 has affected the whole world since the end of 2019. All health authorities are working hard to detect this virus that causes the death of millions of people. When diagnosing COVID-19, radiological imaging is often used. Our main goal in this study we have done is to realize an application that can support healthcare professionals in their decisions with the help of X-ray images. For this, segmentation process has been performed on X-ray images with k-means clustering algorithm. Segmentation is the process of clustering similar structures on an image. In this way, it is aimed to make COVID-19 anomalies more prominent on X-ray images. The obtained results of the studies have shown that there are significant differences between the segmentation of the lung image of a patient with COVID-19 and the segmentation of the lung image of a healthy individual. Lung X-ray images of 5 healthy and 5 COVID-19 (+) were used in the study. The future goal of the study is to perform a comparative analysis with different segmentation techniques and different imaging methods. In addition, it is aimed to create an automatic diagnosis system that will enable the detection of COVID-19.

Keywords: *COVID-19, Segmentation, Image Processing, X-ray, k-means, Medical Images*

The Effect of Message Queues on the Communication of IoT Devices

Ahmet Toprak[1], Abdül Halim Zaim[2], Feyzanur Sağlam Toprak[2]

[1]Istanbul Commerce University, Computer Engineering, Istanbul, TURKEY
[2]Istanbul Commerce University, Computer Engineering, Istanbul, TURKEY
[2]Turkey Finans Participation Bank Inc, Information Technologies, Istanbul, TURKEY

Abstract

Nowadays, IoT (Internet of Things) devices have reached quite high numbers. This situation brought with it the presence of high density, unstructured data. The difficulties in obtaining, processing, storing and visualizing this data made it necessary to use the components of the big data system. High density data should be taken from IoT devices instantly, meaningful data should be obtained from these unstructured data, the meaningful data should be stored and presented at the request of the user when needed. In this article, a model is designed to process the data obtained from IoT devices and transmit them instantly to the end user. In the study, unstructured data collected primarily from IoT devices were subjected to data pre-processing steps. Significant words were determined from the data obtained after the data pre-processing steps. For this purpose, the Helmholtz Principle has been applied. After meaningful word detection, it is directed to both Rabbit MQ messaging queue and IBM MQ message queue separately to instantly process data on the subject of each meaningful word. Apache Storm topology was used to instantly receive and process the messages transmitted to the queues. According to the results obtained, messages sent to the IBM MQ message queue are consumed 30% faster than the Rabbit MQ message queue.

Keywords: *Apache Storm, Big Data, Elasticsearch, Hadoop, Helmholtz Principle, Ibm Mq, Internet of Things, Rabbit Mq*

Autonomous Self-Parking Robot using A-Star In VREP

Umut ÇELEBİ[1], Muhammad Umer KHAN[1], Mert Anıl DEMİRHAN[2]

[1]*Department of Mechatronics Engineering, Atılım University, Ankara, TURKEY*

[2]*Department of Electrical and Electronics Engineering, Çankaya University, Etimesgut, Ankara, TURKEY*

Abstract

Due to recent technological developments, the concept of self-parking is now realized. Most well-known car producers have already introduced that feature to their latest vehicles. Probably, the most critical elements that defines the success or failure of this feature is based upon determining the free slot and defining the route to that slot. To find a free slot, vision-based systems are being used by car producers. These systems do not always guarantee the solution and may find some challenges. Due to the limited view of the vision sensor and incapability of detecting disrupted lanes, free slot detection may fail. In order to improve the reliability of the system, the proposed approach tries to support the vision-based system with the defined maps that are already available. The proposed approach discusses the path planning algorithm, its efficiency (energy consumption), and compares its performance for three different ways of parking: parallel, backward, and forward.

Keywords: *Self-Parking, A-Star, Path Smoothing, Energy Consumption*

Design of An Attendance System Based on Face Recognition

Ahmet Ali Ünsal[1], Serkan Ballı[2]

[1]Muğla Sıtkı Koçman University, Faculty of Technology, Department of Information Systems Engineering, Muğla, TURKEY

Abstract

Students' participation in the lesson; It helps them to learn efficiently and to increase their success level. In addition to these, the high attendance rates create a suitable environment for teachers to teach the lesson in a motivated and more enthusiastic way. The most common practice known to be done with the intention of increasing course attendance is polling. Polling operations on classic papers have disadvantages such as loss of time, division of courses and processing of false information into the system. There are also some problems during the retention of polling information and access to polling information. Today, many technologies such as RFID, fingerprint, iris and face recognition-based, wireless communication, polling retrieval have been developed and put into use. The installation cost of the systems of most of the developed methods is high and has some advantages and disadvantages. In this study, it is aimed to adapt the polling process to today's technology with the help of modern technological infrastructures. With this design, "Face Recognition based polling system" has been developed. Thanks to the developed design, polling, data tracking and reporting operations can be performed. Students must first be introduced to the system and a data set must be created. The generated dataset is taught to the machine. When the student comes to the camera, he / she will be recognized by the face data taught in advance and will be registered in the database by specifying the date and time in the form of "entered the lesson" or "left the lesson". Students who are in a certain part of the course will write "var" in the lesson that day. Data stored in the database can be submitted as a report by making the necessary arrangements. Polling system with facial recognition system, not only in polling in schools, but in workplaces, etc. it can be used wherever polling is needed. It can be developed over time and used in many areas such as forensic cases, location detection, recognizing people on the street.

Keywords: *Face Detection, Face Recognition, Data Set, Image Processing, Open-CV LBPH, Haar Cascade*

An Approach for Airfare Prices Analysis with Penalized Regression Methods

Selim Buyrukoğlu[1], Yıldıran Yılmaz[2]

[1]*Çankırı Karatekin University, Faculty of Engineering, Computer Engineering Department, Çankırı, TURKEY*

[2]*Recep Tayyip Erdogan University, Faculty of Engineering and Architecture, Computer Engineering Department, Rize, TURKEY*

Abstract

This paper focuses on analysing airfare prices which are affected by the set of features, such as free luggage, departure-arrival time, etc. Also, at present, the number of passengers preferring to use the airline is increasing with each passing day. Thus, correctly analysing the airfare prices is essential to raise awareness of passengers. Some researchers have applied different kinds of Machine Learning (ML) models in order to analyse the airfare prices. However, to the best of our knowledge, penalized regression methods have not been applied to analyse the airfare prices. Ridge, Lasso and Elastic Net regressions are penalized regression methods. This paper proposes an approach that combines a public data set and penalised regression methods. This dataset consists of 1814 one-way flights of Aegean Airlines from Thessaloniki (Greece) to Stuttgart (Germany). The results and findings reveal that the proposed approach is potentially valuable in the analysis of datasets consisting one-way of flights since it achieves promising results for the airfare price prediction.

Keywords: *Airfare Price, Analysis Model, Penalised Regression Methods*

Site Selection for a New Housing Project to be Built with an Integrated Fuzzy AHP and Fuzzy EDAS Model

Şura Toptancı[1], Ezgi Aktaş Potur[2]

[1]Eskişehir Technical University, Faculty of Engineering, Department of Industrial Engineering, Eskişehir, TURKEY

[2]Gazi University, Faculty of Engineering, Department of Industrial Engineering, Ankara, TURKEY

Abstract

The construction sector is a sector where it is necessary to make strategic decisions to maintain profitability of the companies because of the high investment costs of the projects, the limited resources used, involving various risks on a project basis, the competitive and dynamic pattern of the sector. After the feasibility study, construction companies have to make the most appropriate selection among alternative locations according to project-specific criteria in order to achieve a successful project management process while making their investments. As real life problems are mostly decided under uncertainty, the use of precise expressions complicates the decision-making process. Fuzzy Multi-criteria Decision Making (Fuzzy MCDM) techniques using linguistic expressions are suitable for solving such problems. In this study, it is aimed to determine the most suitable site for a new housing project of a construction company operating in Ankara. In order to determine the most feasible solution, 5 alternative sites in the city are evaluated according to the criteria determined by the literature research and the opinions of the decision makers. In this study, an integrated model that addresses uncertainty within the scope of Type-1 fuzzy set theory has been developed, and the Fuzzy AHP (Fuzzy Analytical Hierarchy Process) method is used to determine the importance weights of the evaluation criteria and the Fuzzy EDAS (Fuzzy Evaluation based on Distance from Average Solution) method is used to decide the most appropriate alternative according to these evaluation criteria. The effectiveness of the proposed integrated model is demonstrated with an application and this study contributed to the company making a correct investment.

Keywords: *Fuzzy AHP, Fuzzy EDAS, Multi Criteria Decision Making, Housing Project, Site Selection*

A non-smooth dual formulation of ε-Insensitive Weighted Least Square Support Vector Regression

Aykut Kocaoğlu[1]

[1]*Department of Electrical and Energy, Dokuz Eylul University, İzmir, TURKEY*

Abstract

Least Square Support Vector Machine (LSSVM) is a well-known and powerful tool for both classification and regression tasks. It employs a regularized l_2 error loss function with equality constraints and forms a Quadratic Programming (QP) problem in dual. However, it is lack of sparseness, as all data points are used to determine the output function. It is also lack of outlier robustness since it employs l_2 error loss function. In this paper, an ε-insensitive Weighted Least Square Support Vector Regression (ε-WLSSVR) with equality constraints is introduced to improve outlier robustness by weighting error terms as well as sparseness using the ε-insensitive l_2 error loss function and its dual problem is formulated as a non-smooth, indeed piecewise quadratic, optimization problem. This non-smooth problem with L optimization parameters, same as the number of the training samples, is solved by the Sequential Minimal Optimization (SMO) algorithm based on the second-order like information. The effectiveness of the proposed ε-WLSSVR model is validated by a number of real-world benchmark datasets.

Keywords: *Weighted least squares support vector regression, Sequential minimal optimization, Non-smooth optimization*

Breast Cancer Diagnosis using Geometrical Descriptors Obtained from Adaptive Convex Hulls of Suspicious Regions

İdil Işıklı Esener[1], Şükriye Kara[2], Semih Ergin[2]

[1]*Bilecik Seyh Edebali University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Bilecik, TURKEY*

[2]*Eskiehir Osmangazi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Eskisehir, TURKEY*

Abstract

Breast cancer is the leading cause of cancer-related deaths among women worldwide as well as being the most frequently diagnosed cancer type. The breast cancer-caused mortality rate can be reduced with early diagnosis, which is known to be most effectively provided by mammography. Computer-Aided Diagnosis (CAD) systems for breast cancer help to increase the sensitivity of diagnosis by giving radiologists the opportunity of re-evaluation, and therefore gain importance in reduced mortality rate. In this paper, a new approach for geometrical feature extraction is proposed for a CAD system for breast cancer diagnosis and is verified on a subset of the publicly available Mammographic Image Analysis Society digital mammogram database. In the detection phase, initially, adaptive median filtering is applied for noise reduction; artifact suppression and background removal is realized via morphological operations, and pectoral muscle removal is executed using a region growing algorithm. Then, Chan-Vese active contour modeling is utilized for the ROI detection. Thereupon, the center of gravity (CoG) of each ROI is determined, and a convex image is created by specifying 92 points, called as edge points, on the boundary curves of the related ROI. In the feature extraction stage of the diagnosis phase, the angles between each pair of edge points and the CoG, the Euclidean distance between edge points and the CoG, and the Euclidean distance between each pair of edge points are computed. These geometrical descriptors are utilized in the classification stage via the Random Forest classifier using the five-fold cross-validation technique. As a result, breast cancer diagnosis is achieved by an accuracy of 70.13%. Analyzing the overall confusion matrix constructed in the classification stage, it is clearly seen that although healthy and benign diagnoses are mixed, malignancy is diagnosed well by the proposed geometrical descriptors.

Keywords: *Digital Mammography, Computer-Aided Diagnosis, Feature Extraction, Geometric Descriptor*

A Sound Based Method for Fault Classification with Support Vector Machines in UAV Motors

Ferhat Yol[1], Ayhan Altinors[1], Orhan Yaman[2]

[1]Department of Electronics and Automation, Vocational School of Technical Sciences, Firat University, Elazig, TURKEY

[2] Department of Digital Forensics Engineering, Technology Faculty, Firat University, Elazig, TURKEY

Abstract

In this study, a machine learning-based method is proposed for Brushless DC (BLDC) motors used in unmanned aerial vehicles (UAVs). The most common 4 different failure conditions in BLDC motors were determined by the literature review and the failure conditions were applied on the motors. Sound recordings were taken from the motors for each fault condition. First of all, the sound was recorded while the robust motor was running at constant speed. Then, a fixed time-length sound recording was taken for 4 fault classes at the same speed and a dataset was created. This dataset consists of five classes, including the case of no failure. Mean Filter, Average Polling, and Normalization processes were applied, respectively, to reduce the data size on these voices. Then, the Chi2 Method was used for feature extraction. In the next step, the Support Vector Machine (SVM) algorithm was used to classify the obtained features. In classification, 96.70% accuracy was calculated with the Cubic SVM algorithm.

Keywords: *Brushless DC Motor, Sound classification, Fault detection, Chi2 Feature Extraction, SVM, UAV*

Machine Learning Application for Fault Detection in Power Distributed Network

José Eduardo Urrea Cabus [1], İsmail H. Altaş [1]

[1]*Department of Electrical and Electronics Engineering, Karadeniz Technical University, Trabzon, TURKEY*

Abstract

This paper presents a comparison between a variety of protections approaches using Machine Learning (ML) algorithms for fault detection over a power distributed generation system. A modified version of the IEEE 34 bus test feeder with two PV systems and a generator installed in nodes 840, 848, and 890 as DGs are considered in this study. System simulations have been done using PowerFactory DigSILENT software, where 3-phase voltages and currents pre and post-fault are collected, and Python software for the data-mining analysis. Simulation results validate that using feature extraction techniques and wavelet packets transform selection algorithms effectively can achieve a high identification accuracy by removing the less relevant features from consideration, preventing the ML algorithms from overfitting or underfitting the dataset.

Keywords: *Data mining, Fault diagnosis, Feature extraction, Machine Learning, Power reliability*

***** Corresponding author: joseeduardourrea@gmail.com

Comparison of Feature Selection and Classification Methods Performances for Microarray Data of Leukemia, Cervical and Prostate Cancer

Özlem Arık[1], Erdem Karabulut[2]

[1]*Kutahya Health Sciences University, Faculty of Medicine, Department of Biostatistics, Kütahya, TURKEY*

[2]*Hacettepe University, Faculty of Medicine, Department of Biostatistics, Ankara, TURKEY*

Abstract

One of the research areas of the bioinformatics, which covers the fields of statistics, mathematics, computer, genetics and biology, is gene analysis. Low n (sample number) and high p (feature number) parameters in microarray data obtained by DNA microarray technology used in gene analysis negatively affect the performance values of machine learning algorithms. Therefore, feature selection and data mining methods are very important in the modeling of microarray data. With the frequently used classification methods in data mining, the data is parsed by using the common properties of the data and the classification of the new data is decided. The main purpose of this study is, to obtain classification models and measure their performance by making feature selection on data sets belonging to three different cancer types obtained from NCBI-GEO database. Feature (gene) selection was performed using rf, lasso, rfe and limma feature selection methods on the microarray gene expression data of leukemia, cervical and prostate cancers. Classification models were obtained through naive bayes, support vector machines, k-nearest neighbor, artificial neural networks and deep learning methods in the data sets for which feature selection was made. The performances of the models were measured by accuracy, sensitivity, specificity and area under the curve criteria. In general, in the study conducted through the R program, the performance values of the classification models, created after feature selection with lasso and limma methods, were found to be higher. Among the classification models obtained, while classification success of deep learning method was better, that of artificial neural networks was lower.

Keywords: *Cancer, Microarray, Feature Selection, Classification, Data Mining*

Workload Forecasting of Warehouse Stations using Classical Time Series Methods and XGBoost

İrem Kalafat[1]*, Mustafa Hekimoğlu[1], Ahmet Deniz Yücekaya[1], Nilay Ay[2], Habib Gültekin[2]

[1]*Kadir Has University, Faculty of Engineering and Natural Sciences, Industrial Engineering, Istanbul, TURKEY*

[2]*Dogus Technology, Machine Learning and Artificial Intelligence, Istanbul, TURKEY*

Abstract

Effective management of warehouse processes is essential in order to maintain high-level service quality and keep the costs at optimum. Each item passes through many workstations during their journey from the entrepot to the shipping area in spare part warehouses. Estimating the workloads in workstations in advance allows personnel assignment to relevant workstation optimization and the increase of the warehouse performance. On the other hand, inaccurate estimations cause personnel shortages at stations, delays on shipment commitment dates, and disruption in warehouse activities. In this paper, time series forecasting models are used to estimate the load in each workstation for a better operation. The proposed methodologies are applied to an automotive spare part warehouse in Turkey. Each workstation is analyzed separately, and the best fitting standard time series model is determined to predict daily loads. Finally, an XGBoost model is presented, and the performance metrics of all approaches are compared. The proposed research includes part acceptance, storing, order picking, and packaging operations and their substations, which have not been considered in previous studies.

Keywords: *Machine Learning, Warehouse Management, Workload Prediction, Time Series, XGBoost*

* Corresponding author: irem.kalafat@stu.khas.edu.tr

Comparison of Boosting Algorithms' Performances on Diabetes Prediction

Hilal Koçak*, Gürcan Çetin[1]

[1]*Muğla Sıtkı Koçman University, Technology Faculty, Information Systems Engineering, Muğla, TURKEY*

Abstract

Diabetes is a serious disease that affects millions of people and plays a leading role in the development of many deadly diseases. Diabetes Mellitus that it's ful name, manifests itself when the amount of sugar in the blood rises above normal due to the unbalanced secretion of the insulin hormone. High blood sugar for a long time may result with permanent damage to the whole body, especially the cardiovascular system, kidneys and eyes. Because of these dangers, early detection of diabetes and appropriate treatment is vital. Rapidly developing machine learning algorithms has laid out many new perspectives in the medical health. With reference to this, many researches for the detection of diseases are based on machine learning techniques in their background. Especially boosting algorithms are pervasively applied to predict life- threatening illnesses such as diabetes. In this study, we made detailed exploration and comparison of boosting algorithms on diabetes dataset with applying knowledge discovery methods. The dataset was taken from National Institute of Diabetes and Digestive and Kidney Diseases through Kaggle. The boosting algorithms analyzed are Gradient Boosting, Adaptive Boosting, XGBoost, LightGBM and CatBoost. 20-fold cross validation was used to examine robustness of the models. Performances of boosting algorithms were interpreted by drawing their receiver operating characteristic curves and comparing average accuracy values. Among all of the boosting algorithms, it has been observed that the CatBoost algorithm slightly gives the highest results.

Keywords: *Data Science, Boosting Algorithms, Comparison, Machine Learning*

A Comparison of Classification Methods for EMG Based Hand Gesture Recognition

Burcu Melis Toprak[1], Selda Güney[1]

[1]Baskent University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Ankara, TURKEY

Abstract

Nowadays, machine learning and pattern recognition methods are quite impressive and open to improvement, which are used in many researches. A lot of studies are focused on the application of these methods to electromyography-based motion recognition. In this study, this recognition based application was used for recognizing 6-7 hand movements. The aim of the study is to provide a brief overview of machine learning methods for electromyography-based hand gesture recognition along with an analysis of a model based on weighted k-Nearest Neighbor. An armband is used to receive surface electromyography (sEMG) signals consisting of 8 channels. For the feature selection, the model is trained using principal component analysis (PCA). In the classification stage, weighted k-Nearest Neighbor (k-NN) is proposed as a classifier. Also the proposed method is compared with decision tree, Naive Bayes and Linear Discriminant Analysis. The precision offered by the proposed model is 97.8% and it has been observed that it is a model open to improvement.

Keywords: *Electromyography (EMG), surface Electromyography(sEMG), weighted k-Nearest Neighbor, Hand Gesture Recognition, Principal Compnent Analysis*

Estimating the Gini Coefficient using Machine Learning Algorithms for OECD Countries

Tuba Koç[1], Pelin Akın[1]

[1]Çankırı Karatekin University, Faculty of Science, Statistics, Çankırı, TURKEY

Abstract

Inequality of income refers to an unequal distribution of income among people, in which one person's share of total income is lower than that of others. The Gini coefficient is a widely used measurement in terms of effectively analyzing income distribution. In this study, random forest, support vector algorithms, and multiple linear regression model, which are among the machine learning algorithms, were applied to estimate the Gini coefficient of Organization for Economic Co-operation and Development (OECD) countries. When the models were compared according to performance criteria, the best model was found as random forest model with the highest $R^2 = 0.7085$ and the smallest RMSE = 0.0264. According to the random forest model results, it is the tax revenue variable that has the greatest impact on the Gini coefficient. The country with the highest Gini coefficient is Mexico and the lowest is the Slovak Republic. Also, it has been observed that the lowest tax income value belongs to Mexico.

Keywords: *Support vector machine, Random forest, Gini coefficient, OECD*

Malicious Urls Detection Using Ensemble Classifier

Kübra Köksal [1], Buket Doğan [2], Zehra Aysun Altikardeş [2,3]

[1] *Marmara University, Institute of Pure and Applied Sciences, Department of Computer Engineering, Istanbul, TURKEY*

[2] *Marmara University, Faculty of Technology, Department of Computer Engineering, Istanbul, TURKEY*

[3] *Marmara University, Vocational School of Technical Sciences, Department of Computer Technologies, Istanbul, TURKEY*

Abstract

Recently, malicious websites have been playing an important role in most cyber-attack and fraud cases. Malicious URLs are sent to unsuspecting users via email, text messages, pop-ups or advertisements. Clicking or crawling such URLs causes unsafe e-mail accounts, launching phishing campaigns, downloading malware, spyware, and ransomware, resulting in serious monetary losses. Therefore, it has become an important problem to effectively detect and prevent these threats. The standard and fastest way to identify malicious URLs is to compare URLs with blacklists. However, blacklists are never exhaustive and lack the ability to detect newly created URLs. Considering the current needs and deficiencies of blacklist-based methods, a machine learning based classification approach was used in this study to combat malicious URLs. In the study, the URL data set of the Canadian Cyber Security Institute (ISCX-URL-2016) was studied, which contains 79 lexical features obtained from benign and malignant URLs. There are five different URL types in the dataset: benign, spam, phishing, malware and defacement. A binary classification process using harmless, malicious labels and a multi-classification process using five different labels information was performed on a total of 7781 benign, harmless and 28,917 malicious URL records. Random Forest algorithm, one of the machine learning methods, used together with 10-fold cross validation to validate the success of the applied method, and an average accuracy value of 99.42% for the binary classification problem and 95.68% for the multiple classification problem was obtained. Thus, a model proposal with a high-performance rate is presented to protect from maliciously designed websites in a dynamic environment, where new ones join the system every day.

Keywords: *Malicious URL, cyber security, machine learning, outlier data, random forest*

Realization of Turkish Sign Language Expressions with the Developed Humanoid Robot

Mehmet Gül[1]

[1]Şırnak University, Faculty of Engineering, Computer Engineering, Şırnak, TURKEY

Abstract

Today, technology solutions, which are increasing day by day, have started to make itself felt in many areas of life. Especially, the widespread use of humanoid robots, one of the latest outputs of technology, in education, health, and many other vital areas are among the most obvious examples to be given. There are successful examples such as the use of humanoid robots as auxiliary equipment in the language learning process in education, for instance as an educational tool in sign language learning. In a series of studies on Turkish sign language, professional humanoid robots have been used so far. Within the scope of the study, some expressions in Turkish sign language were made with the humanoid robot, whose design and software were developed uniquely. A high success rate was achieved in all of the statements made. The aim of the study is to develop a fully developed humanoid robot that includes original design and software elements and to use the developed humanoid robot as an educational tool if desired, or to use it as a useful tool such as enabling hearing and speech impaired individuals to participate more in daily life.

Keywords: *Sign Language, Humanoid Robot, Human-Robot Interaction*

Predicting Title of Given Text Using Deep Learning Methods with LSTM

Mohamed Barre Omer[1], Mustafa Cem Kasapbaşı[1]

[1] *Engineering Faculty Computer Engineering Department, Istanbul Commerce University Istanbul, TURKEY*

Abstract

Nowadays, tremendous text data resources are everywhere in the form of books, news journals, websites, and many more. Exploring and gaining insight into text data is very crucial. The titles give a summary of the article and history, getting a coherent semantically, and syntactically title is quite a challenging task. In this study, a deep learning system namely the LSTM (Long Short Term Memory) neural system is proposed for predicting the title of a given text (PTT) which is a natural language generation system. Deep neural network architecture recently gains popularity, which is easier than previous statistical models for generating text. In this study publicly open news dataset is used from Kaggle called news summary for headline generation. A 500 hundred news summary subset is chosen out of 98403 records for efficiency and less processing power requirements. Firstly stop words are removed as preprocessing then punctuations are corrected and text is transformed to lower case. Later Porter Stemmer is used to obtaining stems of the words in the text. After tokenization, it is divided into 16-word-length sequences. In order to feed numerical values to LSTM word embedding is utilized. The proposed LSTM model generated high-quality titles according to human evaluation based on results we get from Rough Recall Oriented (ROUGE) as for ROUGE 1 Average_ Recall: 0,69886, Average_Presicion :0,69924, Average_F1:0,69905 as for ROUGE 2 Average_Recall:0,69874, Average_Precision :0,69895, Average_F1:0,69884, as for ROUGE L Average_Recall:0,69829, Average_Precision:0,69829 Average_F1:0,69829.

Keywords: *NLP, Text Generation, Deep learning, LSTM, Rouge*

A Data Science Application to Compare Pre and Post Pandemic Online Car Marketplace Data

Mehmet Arın Zeyneloğlu[1], Tolga Kaplan[1], Bekir Çetintav[2],

İsmail Kırbaş[3]

[1]Arabam.com, R&D Unit, Istanbul, TURKEY

[2]Burdur Mehmet Akif Ersoy University, Faculty of Arts and Sciences, Department of Statistics, Burdur, TURKEY

[3]Burdur Mehmet Akif Ersoy University, Faculty of Engineering, Department of Computer Engineering, Burdur, TURKEY

Abstract

Arabam.com is a website where car advertisements are listed. Corporate and individual members post their ads on the website by entering the information. It serves 5 million unique visitors with more than 100 thousand ads every month. It is thought that the pandemic, which seriously shakes the flow of daily life around the world, also affects the vehicle sales/ads processes. In order to better understand these effects and develop new strategies accordingly, a study has been planned in which data belonging to the pre and post pandemic periods are analyzed using various data science methods. Various analyzes have been made on 11 million advertisement data, which includes approximately 14 thousand car models. Some of the prominent findings are as follows. The 100 best-selling models are largely the same in the pre- and post-pandemic era. In the post-pandemic period, the average duration of announcement has increased. While the cluster of car models having short length of stay in ads becomes smaller, the clusters of the models having longer length of stay expand. The most important features for the ad time length (target variable) in the pre-pandemic period, the variables of price, km, model, year and HP yield similar results in the post-pandemic period.However, the importance of the fuel type variable has decreased while the importance of the member type (individual, commercial) variable has increased in the post-pandemic period. The findings of the study, which included a comprehensive data science application, are reported to the relevant units in the company.

Keywords: *Online Car Marketplace, Pandemics, COVİD-19, Data Science*

***** İletişim e-posta: arin.zeyneloglu@arabam.com

Early Prediction of Sepsis from Clinical Data Using Machine Learning Algorithms

Beste KAYSI[1], Özgür GÜMÜŞ[1]

[1]Ege University, Faculty of Engineering, Computer Engineering, İzmir, TURKEY

Abstract

Sepsis is a clinical syndrome caused by an overreaction of the immune system against infection. For this reason, it can cause extremely serious clinical consequences as it affects the physiological and biological structure of the whole body. Today, the incidence of sepsis syndrome is increasing all over the world and continues to be one of the deadliest clinical events encountered despite the developments in the field of health. Early diagnosis and treatment of sepsis is of critical importance in order to prevent serious clinical conditions such as mortality and organ failure. In this study, early-stage prediction of sepsis was made with machine learning algorithms using clinical data obtained from Physionet 2019 Challenge dataset. The prediction success of the Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Extreme Gradient Boosting, and Light Gradient Boosting algorithms used in the study was evaluated with Precision, Recall, and Accuracy metrics. As a result of the studies carried out, the most successful prediction accuracy was obtained with the Extreme Gradient Boosting algorithm. With the high results obtained, it has been seen that the early prediction of sepsis can be successfully performed with machine learning algorithms.

Keywords: *Sepsis, Machine Learning, Clinical Data*

Big Data AI System for Air Quality Prediction

Roba Zayed[1], Maysam Abbod[1]

[1]*Department of Electronic and Electrical Engineering, Brunel University London, UK*

Abstract

Air Quality has become an exploratory area for many researchers from different disciplines in respect to the global warming, climate change, health impact theories and others. There have been adequate evidence urging humanity to act before real crisis and the explosion of consequences of air pollution, are accelerating community and leaders' fears. Several machine learning approaches have been used with different parameters combined for air quality prediction and it is becoming more complex due to the complexity of air components. Traditional machine learning approaches are not sufficient for air quality prediction accuracy as mentioned in the literature. This research aims at collating to measure and predict certain air gases in selected areas and various locations with high traffic through machine learning modelling to develop air quality modelling with appropriate accuracy which could have extraordinary impact in decision making that is aligned with the air quality status for cities.

Keywords: *Big Data, AI, Air Quality, Prediction*