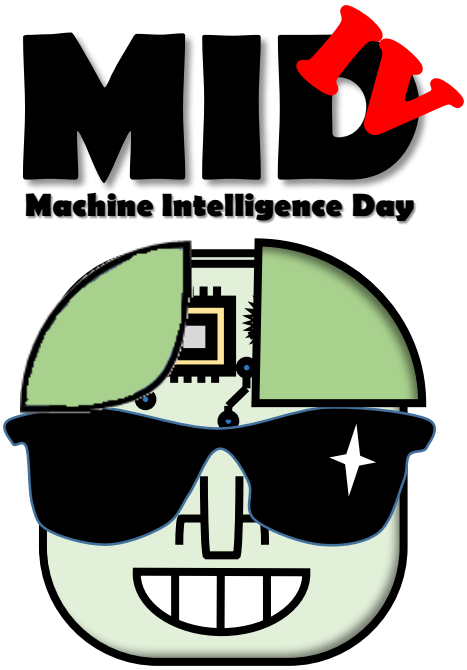
Proceedings of the fourth

Machine Intelligence Day

December 9 2022



Hosted by
Seidenberg School of Computer Science and Information Systems, Pace University
New York, New York

edited by Sung-Hyuk Cha
D. Paul Benjamin

Proceedings of the 4th Machine Intelligence Day

Hosted by:

The Seidenberg School of Computer Science and Information Systems

Published by Pace University Press

Available online
http://csis.pace.edu/~scha/MID2022

Table of Contents

Preface **S.-H. Cha and D. P. Benjamin**

Invited Speaker **Damian M. Lyons**, Fordham University

Wide-Area Visual Navigation for Autonomous Mobile Robots A-1

Video and/or Poster Presentation Competition

R. Kroening and C. Scharff Utilizing Natural Language Processing to Classify Legal Cases A-2

K. A. LoPiccolo and F. Parisi Modeling the Potential Impact of Government Regulation on Cryptocurrency Prices A-3

Y. Liang, N. Singhal, and D. P. Benjamin High-dimensional Spaces Motion Planning for Robotic Arm in Dynamic Environment A-4

L. Pavani, M. Spanburgh, and Y. Shah Product Sentiment Analysis in Ecommerce A-5

K. Sahabir and S. Li Classification of Song Emotions with Neural Networks A-6

R. Jennings Modeling the Climate Change of Countries Through DNN and LSTM A-7

D. Bakhitov and S.-H. Cha Optimal Output Layer Configuration for Multiple Class Artificial Neural Networks A-8

N. Ghagada Anomaly Detection on ATP Tournaments using Autoencoders A-9

M. A. De Castro Theodore, Artificial Synesthetic Alchemist A-10

N. Singhal Automatic and fast Generation of Virtual Street level Views with a High Level of Realism A-11

N. Singhal Image Inpainting removing things or persons and reproducing background textures using Deep Learning A-12

A. Bruev Rules of Correct and Pleasant Data Visualization A-13

R. Jennings Weighted Weak Learners for Random Forests A-14

S. S. Huda Optimizing Scaling Factor to Predict Breast Cancer A-15

E. B. Allen Dynamic Range Entropy A-16

M. Ali On the Scaled Principal Component Analysis A-17

Preface

We are very pleased to have the opportunity to organize the fourth Machine Intelligence Day 2022. Machine Intelligence Day is an annual New York based conference hosted by Seidenberg School of Computer Science and Information Systems at Pace University. It occupies a unique place among conferences, presenting both new research and exceptional student papers, providing opportunities for both faculty and student participation. The purpose of Machine Intelligence Days is to provide a learning and sharing experience on recent developments in Artificial Intelligence, Computer Vision, Data Mining, Machine Learning, and Pattern Recognition. The conference is welcoming to a range of participants, open to both researchers in the field and students. While experts give talks, they are targeted at audiences in general computer science with an eye dedicated towards students. We have strived to publish well-written abstracts that present important original research results and/or open problems relevant to Machine Intelligence.

Professor Damian M. Lyons from Fordham University delivered the invited talk entitled, "Wide-Area Visual Navigation for Autonomous Mobile Robots." We are grateful to him.

16 abstracts were selected in the Proceedings. 15 participated the recorded video presentation competition and 15 participated the poster presentation competition. 5 minute rapid oral presentation for each poster is included in the program prior to the poster competition. We would like to express our gratitude to all the contributors and participants. Finally, we hope that you will benefit from this conference and its proceedings.

S.-H. Cha and D. P. Benjamin

Wide-Area Visual Navigation for Autonomous Mobile Robots

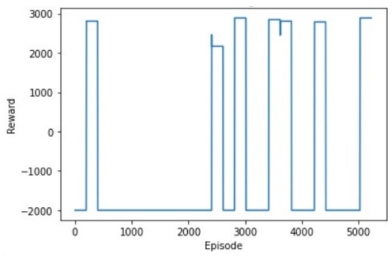
Damian M. Lyons
Department of Computer & Information Science, Fordham University,
Bronx, NY, USA
dlyons@fordham.edu

Navigation, travelling from a current spatial location to a desired final location, is a necessary skill for any mobile robot platform. A common approach to this problem is to construct a map of the entire spatial region to be navigated and use the map to plan navigation paths. This works best for static, well explored environments. For robots that must operate outdoors in all weathers and seasons, these assumptions are not as valid. I will present our work in developing an alternate approach to navigation that can be used alone or together with mapping approaches to handle such challenges.

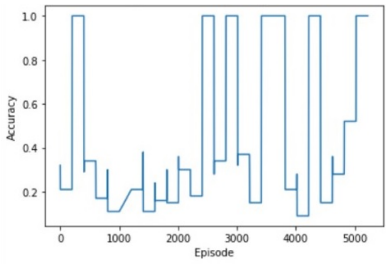
Utilizing Natural Language Processing to Classify Legal Cases

Ronald Kroening and Christelle Scharf
Department of Computer Science, Pace University New York, NY
{rk24279n,cscharff}@pace.edu

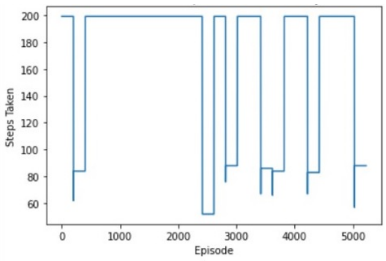
The field of legal technology as it relates to artificial intelligence is a field relatively new as compared to other fields that have been researched in artificial intelligence. One reason is the density of the task of automated legal analysis and its difficulty. My project seeks to utilize natural language processing to classify legal documents, with the goal to have a tool that lawyers can use to automate certain aspects of their work. The dataset contains 2600 examples of Pretrial court data from a public NYPD database, from January 2020 to December 2021. Exploratory data analysis was performed on the data set manually to expand the given case arrest charges and rewrite them in proper context. Afterwards, each case was classified into one of four potential categories: personal crimes, larceny, property crimes, and statutory crimes. The model utilizes K means clustering as well as a Q-learning table. The initial clusters are formed based on principal component analysis and natural language processing. Next, a Q-Learning Agent is trained on the documents being featured, having to recluster the dataset after each episode. The model runs on 26 subsets of the original 2600 examples of 100 each, split up to risk hindering the model, with appropriate measures taken to ensure integrity. Each subset executed 200 episodes each having 400 steps. The results found the model able to reach 100



Q-Learning



Highest Accuracy per Episode



Number of Steps to reach Accuracy

Modeling the Potential Impact of Government Regulation on Cryptocurrency Prices

Kylie A. LoPiccolo[1] and Francis Parisi[2]
[1]Department of Information Systems, Pace University, New York, NY, USA
[2]Department of Computer Science, Pace University, New York, NY, USA
klopiccolo@pace.edu, fparisi@pace.edu

First conceptualized in late 2008, followed by its first transaction in January 2009, cryptocurrency has captured the attention of investors and regulators, alike. In this study, we consider the potential impact regulation might have on cryptocurrency pricing by using intervention analysis [1, 2]. Intervention analysis considers how 'events' affect the data in a time series [3].

Over the past few years several articles have speculated on the idea of cryptocurrency regulation. Moreover, political leaders have raised concerns President Trump in 2019, and recently President Biden and Federal Reserve Chair Powell.

This project requires a fundamental understanding of what cryptocurrency is, how news, political, and market events can influence the price of assets, in particular cryptocurrency prices, and an understanding of time series analysis, data modeling concepts, and coding (R) to conduct the analysis.

While several countries consider regulation, from soft regulation (Japan) to more rigid standards, we study the effect of other news or events on cryptocurrency prices. Moreover, the impact regulation has had on other markets and asset pricing, serves as a proxy. The outcome of this study provides an understanding of the potential impact regulation may have on cryptocurrency pricing, and a method to quantify the impact.

References

[1] P. J. Brockwell and R. A. Davis *Time Series Theory and Methods* 2nd ed, Springer, New York 1991.

[2] G. E. P. Box and G. C. Tiao, Intervention Analysis with Applications to Economic and Environmental Problems, *Journal of the American Statistical Association*, **70** (349): 7079, 1975.

[3] R. S. Tsay, D. Pena and A. E. Pankratz, Outliers in Multivariate Time Series, *Biometrika*, **87** (4): 789-804, 2000

High-dimensional Spaces Motion Planning for Robotic Arm in Dynamic Environment

Yaobin Liang, Nishank Singhal, and D. Paul Benjamin
Department of Computer Science, Pace University New York, NY
{yliang2,ns74172n,dbenjamin}@pace.edu

This study is about solving the motion planning problem in the dynamic uncertain environment with high-dimensional redundant manipulators robotics. The concept of a dynamic environment refers to an environment involving static and moving obstacles, as well as obstacles whose position at any current or previous given time can be known, but not their full trajectory, due to internal or external uncertainties. Motion planning is the process of determining how to manipulate the robot to reach the desired configuration without collisions. Motion planning in real-time is important for the high degrees-of-freedom (DOF) articulated robots, since it permits different parts of the robot to move simultaneously, avoiding collisions with robot itself and other obstacles. Motion planning is challenging for dynamic environments with moving obstacles, particularly when they have dynamic constraints and require computing solutions fast and accurate enough to be useful for real-time implementation. The deep learning approach is adopted in this study to speed up the planning process. Finally, a deep neural network architecture that considers the robotic arm mechanism and its shape will be designed and tested in a simulated dynamic environment to solve the high-dimensional motion planning problems.

Product Review Sentiment Analysis in E Commerce Applications

Lalitha Pavani, Matthew Spanburg, and Yash Shah
Department of Computer Science, Pace University New York, NY
{ls81743n,ms06702p,ys52635n}@pace.edu

As we are aware that the E-commerce market continues to grow and is expected to surpass $5 trillion by 2026, the needs of consumers and businesses are constantly evolving. Companies face difficulty in maintaining high customer satisfaction levels to stand out in the market. E-commerce businesses are using sentiment analysis to overcome this issue and better understand their customers.

Sentiment Analysis is a matter of natural language processing that might influence customer behaviour about particular products by systematically analysing the reviews about service quality, product quality and price regulation.

The goal of our project is to apply sentiment analysis on the Amazon product dataset and analyze the review comments of the customers. Machine learning models were applied and results compared as to which model was better in terms of indicating whether the product category was doing good in terms of customer sentiment.

References

[1] Sentiment Analysis in E-commerce: Benefits & Top 3 Applications
https://research.aimultiple.com/ecommerce-sentiment-analysis/

Classification of Song Emotions with Neural Networks

Kavindra Sahabir and Sukun Li
Department of Mathematics and Computer Science
Adelphi University, Garden City, NY, USA
kavindrasahabir@mail.adelphi.edu, sli@adelphi.edu

As the end of the year approaches, users of popular music apps such as Spotify flood social media with graphics detailing their Year Unwrapped. This is a breakdown of the kinds of songs that each user listens to throughout the year, with superlatives such as most listened to artist and most listened to songs. Recently, a new classification has been included in these breakdowns: the most popular emotions of the songs users listened to throughout the year. Music emotion classification is a major topic that has been studied by industries and research organizations. These types of classifiers recognize music into specific categories based on the emotional information that it conveys to the listener. Multiple factors are taken into account when classifying songs by their emotions. When the classifier is used with a recommendation system, it can allow users of sites such as Spotify to find music similar to the music they usually listen to.

In our project, we aimed to perform classifications of four essential emotions using artificial neural networks. Our chosen dataset is from the RAVDESS, which studies a freely accessible, multimodal dataset of vocal phrases and expressions performed by voice actors. Each recording contains an ID detailing its emotion; we trained our models to classify the emotions of each recording based on their Mel-frequency cepstral coefficients (MFCCs), are numerical decompositions of the frequency spectrum of audio files. Therefore, using neural networks, we implemented our model to recognize the emotion of each recordings based on the MFCC features for the recordings.

References

[1] S. R. Livingstone and F. A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, *PLoS ONE* textbf13 (5): e0196391, 2018.

Modeling Climate Change of Top 12 GDP Countries Using DNN and LSTM

Ray Jennings III
Department of Computer Science, Pace University, New York, NY, USA
rj07609p@pace.edu

Climatologists have studied the change in air and ocean temperature and have come up a set of risk-factors as the primary causes of climate-change. Reputable data sources have put together datasets for climate-change risk factors. Many of these factors are attributed to population- specifically overall population size, urban population size, educational level, life expectancy, poverty rate, population density, land usage types, energy consumption of renewable and non-renewable sources, greenhouse gas emissions. In a previous publication [1] I looked at using the Long Short Term Memory model [2] to create a time series prediction model based on global warming world temperatures. Within this new work, I take into consideration 30 climate-change features and use a deep neural network which includes stacked and bidirectional LSTM layers among others as shown below. For each of the top 13 countries based on GDP [3], a multivariate, time-series based dataset, with a dimensionality of 30, was created for each of the 13 countries. Each countrys dataset was used with the LSTM based model and a future climate change prediction is made.

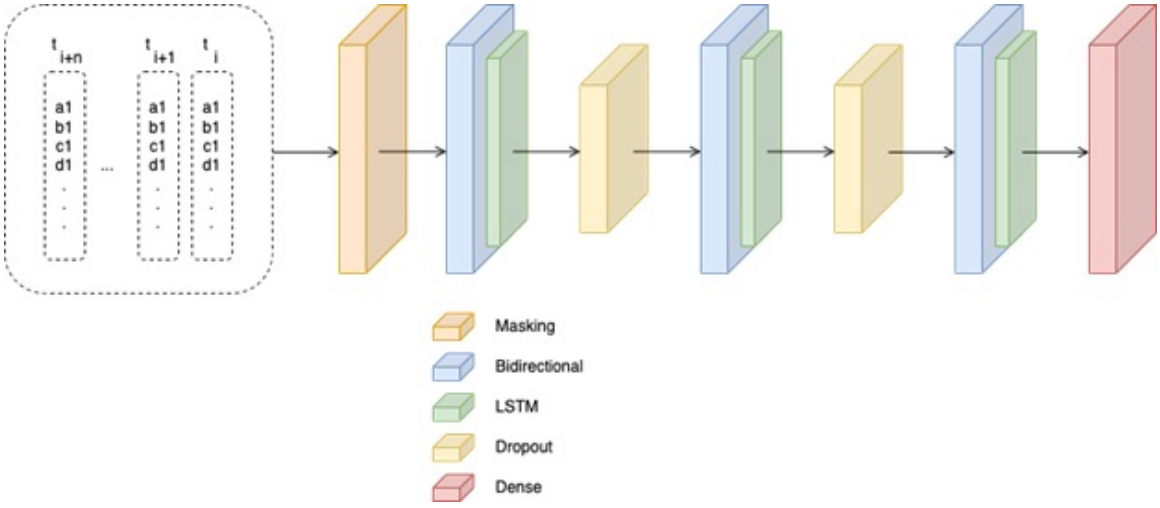


Figure 1: Multivariate Deep Neural Network Model

References

[1] R. Jennings and K. LNU, A Design and Implementation of Music And Image Retrieval Recommendation System based on emotion. in Proc. of the 51st *Southeast Decision Sciences Institute Conference*, February 2022

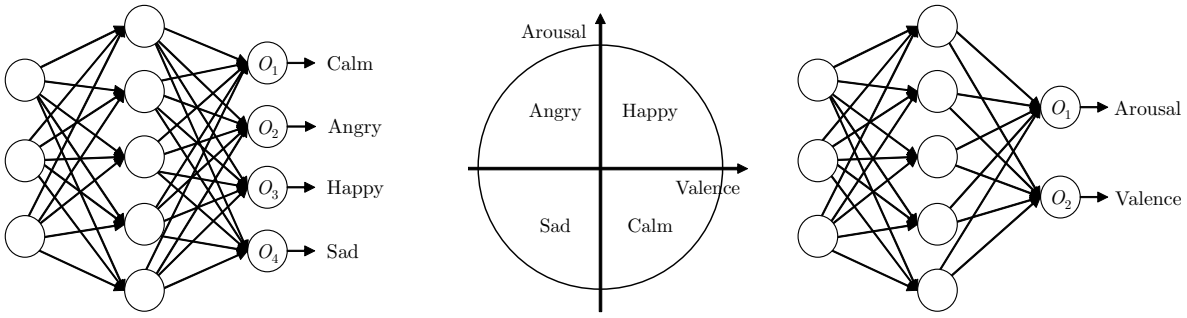
[2] S. Hochreiter ad J. Schmidhuber, *Neural Computation*, 9 (8), pp 17351780, 1997.

[3] World Population Review, https://worldpopulationreview.com/countries/countries-by-gdp

Optimal Output Layer Configuration of Artificial Neural Networks for Multi-class Classification

Dmitrii Bakhitov and Sung-Hyuk Cha
Department of Computer Science, Pace University New York, NY
{db26938n, scha}@pace.edu

Typical artificial neural networks (ANN) for multi-class classification problem have the number of output neurons equal to the number of classes as shown in Figure (a). They use the traditional one-hot encoding method to configure the output layer neurons. This paper suggests that the simpler ANN model with the fewer number of output neurons than the number of classes may perform better or compatible results. For example, ANN with 2 output neurons can solve 4 class emotion classification problem by the Russell's circumplex model as shown in Figure (b) and (c).



(a) ANN with 4 output neurons

(b) Russell's circumplex model

(c) ANN with 2 output neurons

Figure 1: Output neuron configuration for ($m = 4$) class classification

Experiments were conducted using the standard MNIST digit image data set [1] to test whether fewer number of output neurons is better or compatible and to find an optimal output layer configuration. First, Convolutional Neural Networks (CNN) with the traditional 10 one-hot encoding method resulted in the accuracy of 98.9%. Two methods of reducing k, the number of output neurons were examined. The first obvious method is the minimal method which uses the ($k = \lceil \log m \rceil$) number of output neurons. 4 output neurons are minimal to represent 10 classes. However, it did not produce better results. The next proposed method uses the following minimization technique.

$$\text{minimize } k \text{ such that } \binom{k}{\lfloor \frac{k}{2} \rfloor} \geq m$$

It encodes each class to binary number with $\binom{k}{\lfloor \frac{k}{2} \rfloor}$ number of 1's where the encoded binary number has roughly half 1's and 0's. For example of digit dataset, 5 output neurons are needed and it resulted in 99.2% accuracy rate. The experimental results suggest that simpler ANN is better and the number of output neurons can be fewer than the number of classes.

References

[1] Y. LeCun, C. Cortes, and C. J.C. Burges, *The MNIST Dataset Of Handwritten Digits*, https://yann.lecun.com/exdb/mnist/

Anomaly Detection on ATP Tournaments using Autoencoders

Niyati Ghagada
Department of Computer Science, Pace University, New York, NY, USA
ng59819n@pace.edu

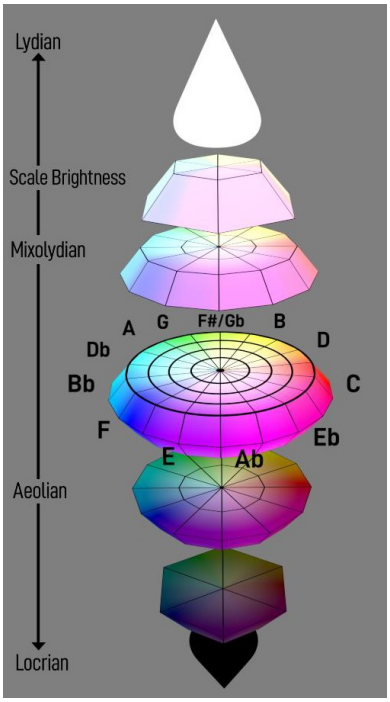
Anomaly detection is an important technique in the field of Machine Learning and part of Data Analytics. ATP Tennis Tournaments have been a huge deal among the enthusiastic tennis players. There are several analyses on predicting winners, betting on a player, or picking the G.O.A.T among the well-known top three legends. However, when it comes to anomaly detection, there are attempts on it, but they are not able to detect as many outliers efficiently due to missing data or other reasons. For my project, I am going to use the ATP Tennis dataset from Kaggle from the year 2012 to 2017 to better analyze the events and prove the concept. As well as use the autoencoder algorithm to detect the anomalies. Autoencoder is one of the most vital algorithms in the field of Deep Learning and an efficient method to perform anomaly detection algorithm. In this work, I will be using the ATP Tennis Tournament dataset to perform unsupervised anomaly detection using autoencoders concept to produces results that show anomalies in those years of tournaments.

Theodore, Artificial Synesthetic Alchemist

Mark Anthony De Castro
Department of Computer Science, Pace University, New York, NY, USA
markanthony.decastro@pace.edu

A lot of AI generated art are from General Adversarial networks that are trained to retrieve image properties from text. This paper aims to contribute retrieval through musical properties and instead of a deep neural network, this paper uses reinforcement learning to imprint musical properties that acts as a query and weight.

Theodore, is an AI agent with induced artificial synesthesia. Synesthesia is the activation and stimulation of one or more sensory and cognitive pathways. The main contributions of this paper are image retrieval through musical properties, a mapping function that bridges music and art, harmonic alignment of colors through musical imprinting and a reinforcement learning environment derived from Pythagorean intonation on which agent actions are equipped with a geometric inter-disciplinary filter.



References

[1] T. Y. Kim, B. H. Song and S. H. Bae, A Design and Implementation of Music And Image Retrieval Recommendation System based on emotion. *Journal of the Institute of Electronics Engineers of Korea CI*, **47** (1): 73-79, 2010

[2] T.Nicholson, Outliers in Multivariate Time Series, *Fundamental Principles of Just Intonation and Microtonal Composition*, 2018

Automatic and fast Generation of Virtual Street level Views with a High Level of Realism

Nishank Singhal
Department of Computer Science, Pace University New York, NY
ns74172n@pace.edu

The main aim of the research is to create likely to be implemented for the automatic generation of virtual street-level views of the city environment using the data set of a 2- D map such as Open Street Map (OSM) with available speci

cations in the data, then extracting the layout of Buildings placement, road structure, people, cars, trees, etc. To make the layout more realistic from a single-view map, investigating deep generative models with adversarially learned layout shapes prior to 3-D shape construction and reconstruction. Now we have the aerial image in the form of an Open Street Map (OSM) and street view shapes from the above model will help in creating cross-view image synthesis and segmentation by investigating conditional Generative Adversarial Network (cGAN). Using the results of all the above and the depth image, an investigated GAN network will generate segmented images, consequently generating a Google-like virtual Streetview. The objective of this thesis work is to give bases for development of an 'APPLICATION' which can generate a dataset of virtual street-level views that is very near to realism and that can further be used to warm-up the training of deep learning models (GANs). This thesis work would provide the future researchers in the

eld and to the professionals making fantasy world, for the gamers to feel the realism of a particular place or a city. Also, the problem of scarce availability of dataset would be addressed.

Image Inpainting removing things or persons and reproducing background textures using Deep Learning

Nishank Singhal
Department of Computer Science, Pace University New York, NY
ns74172n@pace.edu

This research focuses on creating an application with image in-painting, which can allow us to update our clicked photo by selecting the object we need to remove with any distortion of our current image after removing that part from the image, which makes the image incomplete. Now the in-painting Deep learning Algorithm is inspired by Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, Jiaya Jia - MAT: Mask-Aware Transformer for Large Hole Image Inpainting [1]. modified to handle incomplete images on both continuous and discontinuous large missing areas in an adversarial manner. Image inpainting makes it possible to erase elements present in an image and replace them with a plausible background, in particular by reproducing textures when the area to be filled is relatively large and by propagating linear structures such as contours.

References

[1] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang and J. Jia, MAT: Mask-Aware Transformer for Large Hole Image Inpainting, in Proc. of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2022, pp. 10748-10758

Rules of Correct and Pleasant Data Visualization

Artur Bruev
Department of Computer Science, Pace University, New York, NY, USA
ab93404n@pace.edu

Data visualization is a skill, became one of importance for several professions, such as: Marketing specialist, Business intelligence analyst, Analytics manager, Data Scientist. Tables and dry figures are not as attractive to listeners as visually presented information [1]. Unity of connections between math and arts methods, makes it as unique, as looks simple for viewers. Despite all that, to create and operate graphs, plots, and bars, its important to understand several basic techniques.

Coordinate system can be presented in different ways. Cartesian (using x and y-axis) and Curved-Circular are most used, as they create visualizations for different areas. Supposed on needs of client, data can be presented different ways. Distribution charts, like histograms, violin, and box plot, represent how the data distributes along several categories. Correlation charts, like bubble and 2D histograms, showing relationship between several variables. Evolution visualizations showing change of variable trough the space or the time. Geographical spots present data supposed to coordinates on map. Pie charts is one of the popular charts and one of the less useful. They can present only 3-5 values and be difficult to understand correct. The good figures, there will still be differences in quality, and some good figures will be better than others [2]. 3D visualization become a 'bad tone, as it presents variables incorrect (cause of 3D effect) and can be useful in small sectors such as physics and biology. Sorted bars are more comfortable for eyes, as showing more accurate result. Colors of data visualization, becoming a new problem like for people who trying to create new plots, for clients too. Designers choose colors by "opposite rule" (when color staying in 80 degrees to each other) or same palette (colors and tons staying in-close to each other). These rules can be difficult for people with troubles of color detection. Separating of "RGB Circle" by lines to center, helping to understand place of basic colors. Small circle (with White area) is a color can be useful for background. Other sectors helping to identify contrast colors. The data source for correct data visualization will be storytelling with data, #simple_text: Think about solely using the number-making it as prominent as possible-and a few supporting words to clearly make your point. #tables are great for just that-communicating to a mixed audience whose members will each look for their particular row of interest. #graphs interact with our visual system, which is faster at processing information. We want to take a discerning look at the visual elements that we allow into our communications [3].

References

[1] T. Y. Kim, B. H. Song and S. H. Bae, A Design and Implementation of Music And Image Retrieval Recommendation System based on emotion. *Journal of the Institute of Electronics Engineers of Korea CI*, **47** (1): 73-79, 2010

[2] T.Nicholson, Outliers in Multivariate Time Series, *Fundamental Principles of Just Intonation and Microtonal Composition*, 2018

Weighted Weak Learners for Random Forests

Ray Jennings III
Department of Computer Science, Pace University, New York, NY, USA
rj07609p@pace.edu

The predictions from a random forest are based on the aggregate votes from the set of weak learners that comprise the random forest. The aggregation is commonly done by using majority voting where each weak learner has a vote of 1 for the class that it predicted. The class with the largest number of votes is used for the final prediction. The weak learners can sometimes be created with varying degrees of performance. This can, at times, be caused by the data selected during the bootstrapping where some weak learners might not see the entire set of classes. Previous works have shown that using a weighted vote instead of a majority vote can frequently improve the accuracy of a random forest. One technique [1] is to use only a subset of the weak learners where the subset is chosen based on a similarity measure between instances already seen. This technique requires additional execution time for the similarity calculation, plus additional memory to save the information at every leaf node. Another technique used [2] is to weight each of the weak learners by the performance based on the OOB data. I propose a method where a weight for each weak learner is assigned as: $\text{weight}_i = (\text{accuracy}_i)^p$.

Values for p typically range from 1 to 10. The accuracy of each weak learner is computed during the validation phase. Like [2], the final predicated class is the one with the highest score. The best value of p is highly dependent upon the variance of the dataset. The higher the value of p puts more emphasis on the higher accuracy trees and less emphasis on the lower accuracy trees. When the accuracy is known for a non-weighted random forest, then an appropriate value for p can be selected. Multiple datasets were chosen to test this proposal. Each test uses a 5 k-fold cross validation. The table below shows the accuracies for a standard random forest (RF), this version of the weighted random forest with $p = 5$ (WRF) and Scikit-learns Random Forest Classifier (SKL) [4] used as a benchmark. The *Sonar Rock or Mine* dataset [3] was used.

	5 trees			10 trees			15 trees			20 trees			25 trees		
	RF	***WRF***	***SKL***	***RF***	***WRF***	***SKL***	***RF***	***WRF***	***SKL***	***RF***	***WRF***	***SKL***	***RF***	***WRF***	***SKL***
#1	0.8293	0.8781	0.7805	0.8293	0.8049	0.8537	0.6829	0.7805	0.8781	0.8293	0.8537	0.8537	0.8293	0.8049	0.7805
#2	0.7805	0.7805	0.7317	0.8537	0.8781	0.8781	0.7805	0.8293	0.7805	0.7317	0.7561	0.8537	0.8293	0.8049	0.8781
#3	0.6585	0.7805	0.7073	0.7805	0.8293	0.6585	0.8049	0.8537	0.8537	0.8049	0.8537	0.78050	0.8293	0.9268	0.7561
#4	0.8049	0.8049	0.7317	0.7805	0.9024	0.8049	0.7317	0.8537	0.7561	0.7805	0.8293	0.8049	0.7561	0.7805	0.7561
#5	0.7805	0.8293	0.8781	0.7805	0.7805	0.7317	0.8293	0.8049	0.7805	0.8537	0.9756	0.8293	0.8293	0.8781	0.8293
μ	0.7707	**0.8146**	***0.7659***	0.8049	**0.8390**	***0.7854***	0.7659	**0.8244**	***0.8098***	0.8000	**0.8537**	***0.8244***	0.8146	**0.8390**	***0.8000***

References

[1] M. Robnik-Šikonja, Improving Random Forests, *Machine Learning: ECML*, Springer Berlin Heidelberg, 359–370, 2004

[2] M. El Habib Daho, N. Settouti, M. El Amine Lazouni, and M. El Amine Chikh, Weighted vote for trees aggregation in Random Forest, in Proc. of *International Conference on Multimedia Computing and Systems* (ICMCS), 438-443, 2014

[3] Connectionist Bench (Sonar, Mines vs. Rocks). UCI Machine Learning Repository.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Optimizing Scaling Factor to Predict Breast Cancer

Syed Shahyan Huda
Department of Computer Science, Pace University, New York, NY, USA
sh66091n@pace.edu

As we try to decipher meaningful data behind our day-to-day analysis, we are often exposed to data that suffers from large volumes. One of the most common problems of machine learning is Breast Cancer is often considered the most common type of cancer among women. It has a mortality rate of about 2.5 % and especially higher in women of color. Multiple socio-economic reasons as well as genetic conditions can cause the cancer. In near futureThis research aims to improve an existing Breast Cancer prediction model first published in 1992 by Dr. William H. Wolberg at University of Wisconsin. The last prediction model was published in 1995 by W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian called "Computer-derived nuclear features distinguish malignant from benign breast cytology." The accuracy of the prediction model back in 1990 was about 93.5 % followed but the paper in 1992 with accuracy about 95.7 %. The project aims to use Normalization techniques to improve the accuracy of breast cancer prediction. The classification was done using the KNN (nearest neighbor) model as it was used in past research papers. Normalization method used were Standard, Min- Max Scaling, Mean Absolute Deviation, Median Absolute deviation which sometimes is known as MAD around Mean/Median. All these normalization techniques use different statistical parameters. The independent s study had about 569 instances of patient tissue samples processed using XCYT Software which test for 30 different distinct features. This raw data was processed in Python to train, split and test the data model using classification methods followed by various normalization techniques. All the various results were studied and researched concluded the best accuracy was given by Min- Max scaling of about 97.66 % much better than the past published models and models tested in the research. In near future, AI and ML can play a significant role in healthcare especially disease prediction and prevention.

Dynamic Range Entropy: A Novel Method For Measuring Entropy In A System

Emmet Allen
Department of Computer Science, Pace University, New York, NY, USA
ea54290n@pace.edu

Entropy is a concept that has been used time and time again, when it comes to understanding the amount of disorder in a dataset, the construction and evaluation of nodes in a classification tree, or as the basis of many famous reduction algorithms, entropy as a concept has proved itself to be the defector way to measure disorder. Though, the formula used for Entropy is pretty standard throughout all of the above uses, most commonly geared towards the use of Shannon Entropy formula to compute the measure of disorder in a system [1].

We have been able to make plenty of breakthroughs within the Machine Learning community solely on Shannons Entropy formula, however with this report I propose a novel entropy formula to reinvigorate the same curiosities and breakthroughs that were had with Shannons Entropy formula many years ago. The formula that I am proposing, Dynamic Range Entropy formula is given in eqn (1).

$$E(X) = \sum^{b} p(x_i) log_b(\frac{1}{x_i}) \tag{1}$$

The data set that is being used to compute the various uses and differences of Dynamic Range Entropy when compared to the standard Shannons Entropy in both Entropy measurement and Information Gain is the University of California Irvine Machine Learning repository Mushroom Classification data set [2]. Entropy values and Information Gain values for the data set were produced where average Shannon Entropy for single attribute against data set = 1.9, average Dynamic Range Entropy for single attribute against data set = 1.11, average Information Gain using Dynamic Range Entropy = -0.35, and average Information Gain using Shannon Entropy = 0.30. Concluding that the Dynamic Range Entropy formula can be promising when determining the Entropy and Information Gain for a data set to build out decision trees based on these values.

References

[1] C. E. Shannon, *A Mathematical Theory of Communication* The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[2] A. A. Knopf and G. H. Lincoff, *Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms*, New York 1981.

On the Scaled Principal Component Analysis

Mudassir Ali
Department of Computer Science, Pace University, New York, NY, USA
ma83671n@pace.edu

As we try to decipher meaningful data behind our day-to-day analysis, we are often exposed to data that suffers from large volumes. One of the most common problems of machine learning is when continuous variables have wide ranges, the model then has a considerable number of values to consider when making predictions. A supported practice to deal with wide ranges is to scale the data. There are various ways to scale such as min max scalar and logarithmic (log) scalar. Min max scalar is a method to bound all the values in the data between zero and one. Logarithmic scalar applies log based ten to each value minimizing the values. With these two scaling methods the model has an easier process in making predictions.

Another issue with volume is dimensionality, if a dataset has too many columns it can be difficult for a machine learning model to learn the patterns of the data and properly give a prediction because of the background noise in redundant and unnecessary columns. To improve model performance, against large datasets particularly, principle component analysis (PCA) is a reduction method that can positively affect model prediction outcome. PCA absorbs all the information within the dataset and returns a smaller dataset (comprising of continuous values) but retains the valuable information for modeling purposes.

After the results of PCA, the data still consists of ranges that can be uncomfortable for the model to efficiently predict. Thus arrives the hypothesis: is it more effective to scale the data first and then apply PCA (option 1)? Or should we perform PCA and then scale (option 2)?

I have taken two datasets from Kaggle, one predicting churn on banking customers (classification), and one predicting the mean radius of cancer cells within patients (regression.) Following the hypothesis of both the continuous and classification problems, it came to my understanding that the preferable method was option 1, scaling then PCA. The metrics showed that option 1 performed favorably compared to option 2. During the classification, the accuracy and roc auc score was larger and during the regression the r squared and mean squared error was also better for option 1. The type of scaling did not impact the end results, as both min max and log scaling out performed non scaling. However, there are several other factors that can be explored to further analyze this hypothesis, such as other scaling methods outside of min max and log and data sets that require much more cleaning than Kaggle datasets.

Index of Participants

Mudassir Ali A-17
Emmet Byron Allen A-16
Dmitrii Bakhitov A-8
D. Paul Benjamin A-4,CC
Artur Bruev A-13
Sung-Hyuk Cha A-8,CC
Teryn Cha ..... PC
Soon Ae Chun ..... PC
Mark Anthony De Castro A-10
James Geller ..... PC
Yegin Genc ..... IS, PC
Niyati Ghagada A-9
Jonathan Hill ..... Dean
Syed Shahyan Huda A-15
Ray Jennings III A-7,A-14
Ronald Kroening A-2
Sukun Li A-6,PC
Yaobin Liang A-4
Kylie A. LoPiccolo A-3
Damian M. Lyons ..... IS,PC
Melanie Madera ..... LA
Francis Parisi A-3,PC
Lalitha Pavani A-5
Kavindra Sahabir A-6
Christelle Scharff A-2
Yash Shah A-5
Juan Shan ..... PC
Nishank Singhal A-4,A-11,A-12
Matthew Spanburgh A-5
Lixin Tao ..... PC

A - Authors CC - Conference Chairs PC - Program Committee LA - Local Arrangement