CE888 – Assignment 1 – Project 1: Auxiliary feature learning for small dataset regularization (February 2019)

Sylvia Tattersfield Marty – st18580

***Abstract*— Data has a big impact on the outcome of the desired tasks that it creates, one of the biggest issues machine learning and deep learning come across is having insufficient data or insufficient training examples. While different solutions have been suggested and explored this research paper will focus on a proposed project method for using auxiliary feature learning for small dataset regularization. This will be done through the implementation of both an autoenconder and a neural network. The autoencoder will work to learn and create feature selection, this features will then be used to train the neural network which will be a discriminative classifier. The project will be tested in three different data sets, which will be manipulated to include less data and increase the amount of data as part of an experiment to properly evaluate the performance of the method proposed, as the amount of data increases. Another variation that will be included is the impact of the attributes on the datasets on the final classification, and how focusing on certain datasets will allow for better classification results. Lastly the method will be evaluated by comparing the results obtained to those obtained with other classifiers such as decision trees.**

***Index Terms*— Autoenconders, Deep Learning, Multi-task Learning, Neural Network, Supervised Learning.**

I. Introduction

IN the following report a proposed method for using auxiliary feature learning for small dataset regularization is suggested. The anticipated method will consist of two main parts, an autoenconder used for feature selection and a neural network used for classification. The autoenconder will focus on receiving the input data and making a feature extraction for the representation of the input data which will be fed to the neural network as inputs, for it to become a discriminative classifier. One of the main premises of this project is the amount of dataset, the overall data selected and the approach towards its manipulation is explained on further sections, however it must be mentioned that the project will focus on the difference in performance when utilizing different sizes of datasets, therefore the experiments that will be done will follow an data incrimination plan. Another of the variations will be the manipulation of the attributes within the input data and the impact that this can have on the overall performance of the trained classifier. This approach is attempting to tackle one of the greatest issues when regarding data science and processing, insufficient amount of data or training examples. Several approach have been done previously, and have been successful, however the current approach will not be evaluated with current research but with other state-of the-art classification methods. Finally this results will help evaluate the performance of using an autoenconcder for feature selection to then use for training for a discriminative classifier.

II. Background research

Machine learning (ML) usually consists in the training of a single or multiple models to perform a specific task, this performance is then evaluated for the model to be fine-tuned until the performance no longer improves. [1] However there are multiple approaches which complement ML and can increase efficiency and the overall execution depending on the desired task.

A. Multi-Task Learning

One of these approaches is Multi-Task Learning (MTL), which consist of utilizing the information obtained from the training signals of the related or auxiliary tasks and using this information to allow the model to make a better generalization of the original task. [1] This approach is actually fairly common, since every time more than one loss function is being optimized in the model, MTL is being executed. [1]

Multi-task learning is based on trying to exploit the results of certain features in the model by allowing them to be used in other auxiliary tasks. Firstly common features are found in the foremost layers of the network, while individual tasks are solved in later branches of the network. This can be done through what can be known as an encoder-decoder structure. [2] Auxiliary tasks are less important or might even be irrelevant to the overall main application of the model, yet despite being unrelated they are used to find a tougher and more robust representation of the input data which is then used for the main task and as a consequence improve the performance of the network. One of the main characteristics of these related tasks is that they should be easy enough to be learned and that they must require little to no effort to obtain the labels or annotations. The use of these tasks is to force the network to generalize a bigger amount of tasks and by having

This paragraph of the first footnote will contain the date on which you

the auxiliary tasks the network restricts the parameter space during optimization and are therefore used as a regularization measure. [2]

In the encoder-decoder structure the auxiliary tasks explained previously, enact as a specialized decoder to the representation supplied by the encoder. The encoder favours the learning of the features in common, which are in turn exploited by the rest of the tasks, which enhance the performance of both the auxiliary tasks and the overall network. [2]

B. Deep Learning

Multi-task learning can also be used in deep learning methods, one of this methods is the Deep Neural Network, which consists of a neural network with a hierarchy of layers, which is used to extract representations form raw input data. Each hidden layer has an output that can be considered as a feature extraction, each output is used to construct advanced representation of the original data, and this is known as feature learning. Once a respectable representation of the features is obtained the classification can be done satisfactorily. [3] Deep Neural Networks are only one of the methods of deep learning, this approaches are important since according to [3] studies have shown, that for many applications, using a deep learning approach can outperform a standard machine learning method. One of the main issues of the deep learning approaches is overfitting, which will be discussed to a bigger extent later in the section.

C. Autoencoders

Another approach of Deep learning includes autoenconders, which play an important part in the extraction of the representation of input training patterns. [4]The representation that the autoenconder creates is an abstract representation which includes informative features to demonstrate a large set of data. [4] An autoenconder compresses and decompresses data, however they are data specific, which means that the data that they will be able to compress must be similar to the one they were trained on, and the decompressed data will always loose resolution, when compared to the original input data. [5] Autoencoders encompass the previously discussed terms, they follow a deep learning architecture and follow multi-task learning in the distance or loss function. They can also be used for classification and feature selection, especially after sparse regularization of the hidden outputs which allows for a high learning performance. [4]

D. Previous approaches

Recently several methods have bene proposed to deal with the problem of having a small data sets or a limited number of training examples. [3] One of the approaches is Deep Neural Networks, since they have a large number of parameters and therefore a great ability to classify complicated tasks, but it can lead to overfitting, especially when training samples are not enough. [3] One of the solutions for both overfitting and dealing with small data sets is data augmentation, which is basically to increase the size of the data set. The basis of data augmentation is to create new samples from the original samples by applying transformations and then using these new samples to enlarge the data set. [3] Some of the possible transformations include rotation, translation and flipping, to name a few, this allows the network to learn different variations of each of the features, which leads to a better rate of training and prediction effectivity. [3]

Another approach is the use of stacked autoenconders, this approach consists of a neural network built with autoenconders. This method has a specialized training process, it has the feature selection training and a fine-tuning section. Firstly the feature selection training uses unsupervised learning to pre-train the layers of the network, the target output of this phase of each autoenconder is later used as an input. After the pre-training, the parameters are fine-tuned, this time using supervised learning, afterwards a classification layer is added, which uses the output of the last hidden layer as inputs. The network is then trained using backpropagation expecting to minimize the classification error. [3]
To complement the previous approaches it must be mentioned the impact of semi-supervised learning, since it provides a framework to influence or manipulate unlabelled data. This means that unlike purely supervised learning, semi-supervised learning can improve their performance using both labelled and unlabelled examples. [6] According to [6] studies have proven that in certain cases, semi-supervised learning can almost match the performance of the pure supervised learning approaches, even when labels have been discarded from the dataset.

III. Methodology

The proposed project, as stated previously, will focus on the creation of a method that allows for the use of auxiliary tasks to regularize a network, this will be done through the implementation of an autoencoder for feature selection and a neural network for classification. The main premise of this approach is that there is important information contained in the features. The programming language in which the algorithm will be developed will be python, and it will make use of different libraries such as numpy, pandas, sckit-learn, seaborn and other libraries that will be used for the plotting of results and for the overall implementation of the algorithm. The main goal of this project is to utilize a small portion of each of the proposed data sets in order to verify that the features learned through the autoencoder can be used to train a standard discriminative neural network and progressively verify this by analyzing the whole data sets. Finally the results will be compared to those obtained with other classifiers such as decision trees or Support Vector Machines.

Three different data sets were selected from the UC Irvine Machine Learning Repository, the data sets that were selected are Breast Cancer Coimbra Data Set, Heart Disease Data Set and Autistic Spectrum Disorder Screening Data for Adolescent Data Set.

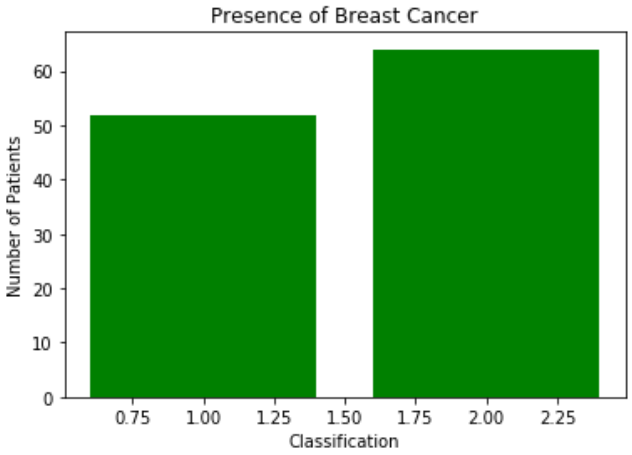

Fig. _. Bar graph of the distribution of patients with breast cancer and the control group for the Breast Cancer Coimbra Data Set.

A. *Breast Cancer Coimbra Data Set*

The breast cancer Coimbra data sets, is a clinical data which was obtained at the Faculty of Medicine of the University of Coimbra as well as University Hospital Centre of Coimbra, in the city of Coimbra, Portugal. The clinical research was observed in 64 patients with breast cancer, and a control group of 52 healthy persons. The data consist of 10 attributes or predictors, shown in table _, all the attributes are quantitative; also a final attribute, a binary variable that indicates either the presence or absence of breast cancer. All the attributes are anthropometric and where all gathered through routine blood tests. The distribution of the data can be seen in figure __.

TABLE I
QUANTITATIVE ATTRIBUTES:

Attribute	Quantity
Age	years
BMI	(kg/m2)
Glucose	(mg/dL)
Insulin	(µU/mL)
HOMA	
Leptin	(ng/mL)
Adiponectin	(µg/mL)
Resistin	(ng/mL)
MCP-1	(pg/dL)
Labels:	
1	Healthy controls
2	Patients

Attributes with corresponding units for each of features in the test for the breast cancer diagnosis on for the Breast Cancer Coimbra Data Set.

B. *Heart Disease Data Set*

The heart disease data set is a clinical data which consists of data from 5 different sources: the Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. However the data selected for the development of this project will be focused specifically on the data from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. This database contains 76 attributes but only 14 of the 76 attributes are used, this can be seen in table 2. The classification of the results is based on whether the patient has a heart disease or not, if there is no presence of heart disease the value is 0, yet there are 4 different categories for the presence of heart disease. The cataloguing of the patients who do have heart diseases varies from 1 to 4, depending on the results of the heart disease. In figure __ we can see the graphical representation of the dataset distribution among the 5 possible classifications depending on the presence of heart disease.

TABLE 2
QUANTITATIVE ATTRIBUTES:

Attribute	Quantity
Age	years
Sex	(1 = male; 0 = female)
Chest Pain Type	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
Resting Blood Pressure (trestbps)	(mm Hg)
Cholesterol	(mg/dl)
Fasting Blood Sugar (fbs > 120mg/dl)	(1 = true; 0 = false)
Resting Electrocardiographic Results (restecg)	-- Value 0: normal -- Value 1: having ST-T wave abnormality -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Maximum Heart Rate Achieved (thalach)	
Exercise Induced Angina (exang)	(1 = yes; 0 = no)
Oldpeak	
Slope	-- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
Number Of Major Vessels Colored By Flourosopy (ca)	(0-3)
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Diagnosis Of Heart Disease (num)	-- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing

14 of the 76 possible attributes for each of the features in the test for the heart disease diagnosis with their corresponding units for the Heart Disease Data Set.

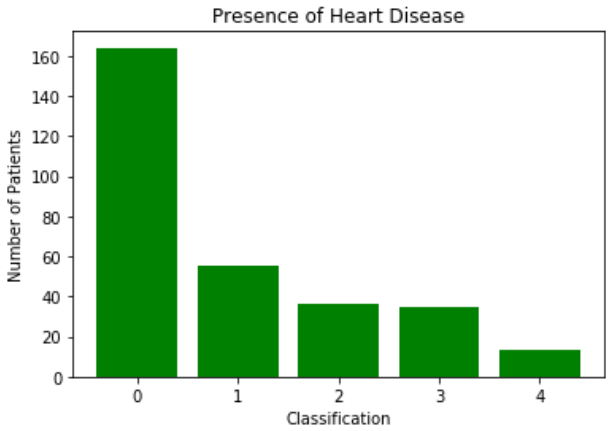

Fig. _. Bar graph of the distribution of patients with heart for the Heart Disease Data Set.

C. Autistic Spectrum Disorder Screening Data for Adolescent Data Set

The autistic spectrum disorder screening data for adolescent data set is a clinical data set which was recorded in Auckland, New Zealand, they proposed a new data set which consists of 20 features that are used or determining influential autistic traits in adolescents, 10 of these features are behavioral, while the other ten are individual characteristics which have been proven to help in the detecting of autism, these attributes can be seen in table 3. The data sets consists of 104 instances and the final result consists of a predicted diagnosis of Autism Spectrum disorder or ASD, in figure __ we can see the graphical representation of the dataset distribution among the results of the predicted diagnosis of the clinical condition.

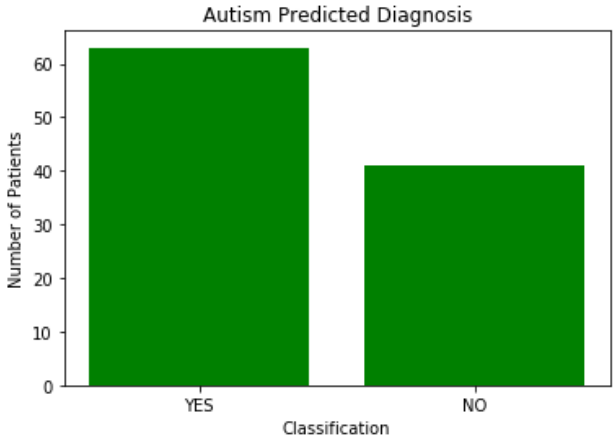

Fig. _. Bar graph of the distribution of the predicted diagnosis of patients with ASD for the Autistic Spectrum Disorder Screening Data for Adolescent Data Set.

IV. Experiments

For the proposed project the approach that will be followed will be the creation of an autoencoder followed by a neural network. The objective of the autoencoder is to learn feature selection from the input data, features which in turn will be fed to the neural network as inputs, this neural network will work as a discriminative classifier. The experiments will focus on two variants, one regarding data and the other regarding the attributes of the datasets. The main goal of this project is to analyse the feature extraction in small datasets, reason for which the experimentation will focus on the first instance using a small portion of the overall data set, a possible example is 30% of the whole data set, and compare the results as a bigger data set is used, a possible combination of the of the amount of data incrementation can be 50% followed by 75% and finally 100% of the datasets. Another possible variation of the experiment can be the manipulation of the attributes, by focusing in only certain attribute and eliminating some the results can change drastically, this approach will be n interesting take on the current proposed project. Each variation or manipulation of the experiments will be done on each of the three selected data sets, explained previously.

TABLE 3
QUANTITATIVE & QUALITATIVE ATTRIBUTES:

Attribute	Quantity
Age	years
Gender	(m = male; f = female)
Ethnicity	List of possible ethnicities (String)
Born with Jaundice	Boolean (yes or no)
Family Members with PDD	Boolean (yes or no)
Who is completing the test?	List of possible family relations (S
Country of Residence	List of possible ethnicities (String)
Used the Screening App Before	Boolean (yes or no)
Screening Score (Results)	Final numeric score obtained basec
Class/ASD predicted diagnosis	Boolean (yes or no)
Question 1 Answer	Binary (0,1)
Question 2 Answer	Binary (0,1)
Question 3 Answer	Binary (0,1)
Question 4 Answer	Binary (0,1)
Question 5 Answer	Binary (0,1)
Question 6 Answer	Binary (0,1)
Question 7 Answer	Binary (0,1)
Question 8 Answer	Binary (0,1)
Question 9 Answer	Binary (0,1)
Question 10 Answer	Binary (0,1)

20 possible attributes for each of the features in the test for the predicted diagnosis of ASD with their corresponding units for the Autistic Spectrum Disorder Screening Data for Adolescent Data Set

V. Discussion

Once the autoenconder and the neural network are done and have been trained and tested appropriately the evaluation of the proposed method as a whole will consists in comparing the results obtained from each of the experiments, with each of the datasets and their respective increments in the amount of data that was fed as inputs, with results obtained with other classifiers. Some of the possible classifiers that might be used form comparison is a decision tree, or a Support Vector Machine among others.

VI. Conclusion

It can be concluded that the proposed project will greatly depend on the data sets, as well as the training of the autoencoder, if the feature selection of the autoencoder is done appropriately, then the classification of the neural net will be achieved with a high accuracy. Nevertheless the neural network still must be trained and tuned accordingly to ensure the best possible outcomes. The main issue will be the data pre-processing and the creation of the autoenconder, since the neural network can be done fairly easily with the libraries provided by python.

VII. References

[1] S. Ruder, "An Overview of Multi-Task Learning," *Insight Centre for Data Analytics*, 2017.

[2] L. Liebel and M. Körner, "Auxiliary Tasks in Multi-task Learning," *Computer Vision Research Group, Chair of Remote Sensing Technology*, 2018.

[3] J. Xue, P. P. Chan and X. Hu, "EXPERIMENTAL STUDY ON STACKED AUTOENCODER ON INSUFFICIENT," in *Proceedings of the 2017 International Conference on Wavelet Analysis and Pattern Recognition*, Ningbo, China, 2017.

[4] J. Liu, C. Li and W. Yang, "Supervised Learning via Unsupervised Sparse Autoencoder," *IEEE Access*, vol. 6, pp. 73802 - 73814, 2018.

[5] F. Chollet, "The Keras Blog," Pelican, 14 May 2016. [Online]. Available: https://blog.keras.io/building-autoencoders-in-keras.html. [Accessed 19 February 2018].

[6] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk and I. J. Goodfellow, "Realistic Evaluation of Deep Semi-Supervised," *Google Brain*, 2018.