Introduction to Data Science Tools & Techniques

# Assignment No. 1

## Ali Nauman – 18L1863

## Question No. 1

## Comparative Analysis – Classification Problem

In this we shall be experimenting with three different classification algorithms on three different data sets. All the below mentioned data sets have been extracted from the *UCI Machine Learning Repository*.

### EEG Eye State Data Set

All the data in this data set has been extracted from one continuous EEG measurement with the *Emotiv EEG Neuroheadset*. The duration of the measurement was 117 seconds. The eye state was detected via a camera during the EEG measurement and added later manually to the file after analyzing the video frames. '1' indicates the eye-closed and '0' the eye-open state. All values are in chronological order with the first measured value at the top of the data. This data set has a total of 15 Attributes and a total of 14980 instances for the algorithms to be implemented.

### Iris Plants Data Set

This data set is probably one of the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. It has a total of 4 numeric, predictive attributes and class. The class distribution is at 33.3% for each of the 3 classes.

### Pima Indians Diabetes Data Set

This data set was collected by using the ADAP learning algorithm to forecast the onset of diabetes mellitus. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. This data set includes 8 plus class attributes, and a total of 768 instances.

### Algorithms Utilized

### Naïve Bayes

It is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

### Multilayer Perceptron

A multilayer perceptron is a class of feed-forward artificial neural network. It consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.

### Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

### Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set** | **Naïve Bayes** | **Multilayer Perceptron** | **Random Forest** |
| **EEG Eye State** | 46.769% | 54.806% | **93.612%** |
| **Iris Plants** | 96% | **97.333%** | 95.333% |
| **Pima Indian Diabetes** | **76.302%** | 75.391% | 75.781% |

### Conclusion

As we can see from the results, the *Iris* data set has worked perfectly on all the three algorithms with the *Multilayer Perceptron*, giving an excellent accuracy on the data set. For the *Pima Indian Diabetes* data set, all the three algorithms gave almost the same accuracy. However, the *Naïve Bayes* overshadowed the other two. In the case of *EEG Eye State*, *Naïve Bayes* and *Multilayer Perceptron* performed extremely poorly. But *Random Forest* gave an excellent result, with *93.612%* accuracy. Overall, *Random Forest* has performed tremendously well on all the three data sets.

## Comparative Analysis – Clustering Problem

In this we shall be experimenting with three different clustering algorithms on three different data sets. All the below mentioned data sets have been extracted from the *UCI Machine Learning Repository*.

### Tamilnadu Electricity Board Hourly Readings Data Set

This data set is collected from the real time readings from residential, commercial, industrial, agriculture sectors to find the accuracy consumption in Tamil Nadu around Thanajvur. The data set has a total of 5 attributes and 45781 instances of data.

### Absenteeism at work Data Set

The database was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. This data set is a collection of employee absenteeism data. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research. The data set (Absenteeism at work - Part I) was used in academic research at the Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management. It has a total of 21 attributes and 740 instances.

### Final settlements in labor negotiations in Canadian industry Data Set

The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc.) in Canada in 87 and first quarter of 88. The data was used to learn the description of an acceptable and unacceptable contract. Data was used to test 2tier approach with learning from positive and negative examples. It has a total of 16 attributes and 57 instances.

### Algorithms Utilized

### Simple KMeans

KMeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

### FarthestFirst

Farthest first is a variant off KMeans that places each cluster center in turn at the point furthest from the existing cluster centers. This point must lie within the data area. This greatly speeds up the clustering in most cases since less reassignment and adjustment is needed.

### MakeDensityBasedClusterer

It refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The data points in the separating regions of low point density are typically considered noise/outliers.

### Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set** | **KMeans** | **FarthestFirst** | **MakeDensityBasedClusterer** |
| **Tamil Nadu Elect.** | 0: 50% 1: 50% | 0: 76% 1: 24% | 0: 50% 1: 50% |
| **Absenteeism** | 0: 54% 1: 46% | 0: 62% 1: 38% | 0: 54% 1: 46% |
| **Labor** | 0: 84% 1: 16% | 0: 82% 1: 18% | 0: 77% 1: 23% |

## Questions No. 2

## Smart Flood Monitoring and Management System – Data Design and Software & Hardware Requirements

## Introduction

Floods are common natural disasters that cause severe devastation to any country. Due to global warming issues and extreme environmental effects, flood has become a serious problem to the extent of bringing about negative impact to the mankind and infrastructure. Timely and active detecting and monitoring of a flood event are critical for a quick response, effective decision-making and disaster reduction. A proper state-of-the-art flood forecasting system should be in place to detect the anomalies beforehand.

Flood risk management generally involves flood monitoring, risk assessment and forecasting of the flood inundation areas. Flood monitoring can be performed by using satellite images which allow detecting river overflows and provide data for flood damage modelling and assessment.

## Model Design

### Weather Information

Flood monitoring and control require measurement and awareness of the water level, velocity, precipitation. Input data for precipitation forecast are meteorological data and weather forecasts as the most important components of a flood forecasting and early warning system.

### Mining Data

Flood forecasting is based on mining historical data and specific domain knowledge to deliver more accurate flood forecasts. Effective flood monitoring makes use of space and ground-observed data received from satellites and terrestrial stations. This data may constitute of images, terrain information, and environmental information i.e. soil type, drainage network, catchment area, rainfall, and hydrology data.

### Physical Modelling – Hydrodynamic Models

Hydrodynamic Models represent the motion of water flow describing the motion of fluid substances in physics. These are the simplified conceptual representations of the hydrological cycle. These can be further divided into two models to gather the required data: 1) *stochastic hydrological black-box models* that define input-output feed based on stochastic data and utilize mathematical and statistical concepts to link a certain input to the model output; and 2) *conceptual or process-based models* that represent the physical processes in the real world.

### Sensor Technology

Sensor technology is widely utilized for detecting water level variations in different flood prone zone. There are several equipment that can be combined together to send notification using water level sensor to the emergency management authority. Water level sensor, a Programmable Logic Controller (PIC), a Global Communication and Mobile System (GSM) modem, a power supply and Subscriber Identity Module (SIM). Use of GSM Technology, can allow to trigger the water level severity based on water sensor in real time via SMS service.

### System Design

For an effective transmission of information in communicating to the victims and protecting and preventing loss of civilian casualties, SMS service serves as a proper alert communication tool.

* Water level sensors – to provide real-time information to the flood control center
* Sensors can be used to detect normal water level and the dangerous threshold level
* Microcontroller setup – to transmit SMS service to the flood control center
* Creation of a setup to process the signals from analog to digital
* Graphical user interface in the flood control center with all the processed data
* Web Based Network to transmit information alongside the SMS service from the sensor information
* The sensor readings can be transmitted at an interval of every 30 minutes to the flood control center of the water level

### Data Information

Data is collected from the sensors transmitted to the main control unit via the microcontroller setup. Following are the most important attributes we can depend, which are used to check for the flood disaster,

* Weather Information
* Tides
* Season
* Months
* Moisture Level
* Value of the signal
* Minimum flood depth
* Maximum flood depth
* Monsoon period

We can use these attributes to classify flood data with normal and up-normal condition. These attributes would easily justify the under-risk conditions and non-risky conditions.

### Machine Learning Algorithms

We can implement multiple machine learning algorithms on the above generated data, such as,

* Random Forest
* Bagging
* Decision Trees

We can construct a confusion matrix to examine the proposed machine learning techniques for the classification of flood data. The True Positive (TP) and True Negative (TN) illustrate the most correct classification of positive instance and negative instance, respectively. False Positives (FP) shows negative symbols while, False Negatives (FN) demonstrates the positive instance. There are a number of measurements which can performed on this confusion matrix, such as,

* Accuracy
* Precision
* Recall
* F1-Score – to determine how good our model performance is.

## Punjab Government – Flood Relief

This is an online dashboard constructed by the Punjab Government, to visualize and see the flood activities. The main purpose of this portal is to provide the mass community with flood warnings and real-time flood forecasts. It constitutes a *Flood Disaster Report* for all the districts. The design of the portal covers majorly all the potential information and it is designed in a way for the user to explore information district wise as well. Following are the important features of the web-portal,

* **Flood Disaster Report** – This is constitutes on a whole and also district wise, where information is provided about, 1) Total number of deaths till date, 2) Total number of persons injured till date, 3) Cattle head perished, 4) Number of relief camps established till date.
* **Inventory Data** – This covers information about the essentials in place a flood is observed such as, Strength of available Bulldozers, Boats, Tents, Life Jackets, Life Ring, De Watering Set, Generators, Plastic Mat, and Mosquito Net etc.
* **Livestock Data** - It constitutes information about active Flood Relief camps, Deputed Veterinary Officers, Count of Perished animals (Large/Small), Vaccination availability, and Date etc.
* **Diseases Data** – This contains information about the total number of patients and information about diseases and the people affected by a particular disease. It also informs about the vaccinated areas in particular with timestamps.
* **Agriculture Data** – This informs about the type of agriculture land and what is grown where. It informs whether the land is affected or no damage has been done so far.
* **Relief Activities** – This data constitutes about all the relief activities conducted so far with timestamps and the information about the particulars provided in the relief setup.
* **Preliminary Losses/Damages** – It contains information about the count of damages and losses in relevance to human life and animal displacement as well.

### Improvements

This setup on an initial basis is very good and seems to be updated with a consistent regularity. More features can be implemented to include weather information and the live sensor readings of the river banks in particular. This can be improved to achieve more advance and better application. Authorities in responsible can upgrade their system by adding GPS module alongside the sensors to track the equipment that is located in various field such as riverbank or low-lying area. The profile of the SMS service can be enhanced and spread on a national level, providing accurate and certain information. Inclusion of timestamps with each of the data reading collected is an important measure.

Time series architecture can be implemented. It provides massive scalability and performance: An effective time series database would enable an application to scale easily to support millions of IoT devices or time series data points in a continuous flow and perform real-time analysis.

This would improve the performance of the machine learning algorithms in place. Thus, enhancing the outcomes and results. In return, giving us better prediction metrics to work on.

## Conclusion

This report shows a basic analysis of the idea, as more data metrics can be made part of this to increase productivity. An increase in informative data would allow us to predict the flood areas better and we can plan better to cater the disaster with a potential decrease in casualties.

## Questions No. 3

### Comparative Analysis between Commercial and Open-Source Analytical Tools

In this we shall be presenting a comparative analysis between the *Open Source Analytical Tools* and *Commercial Analytical Tools*. We shall be discussing the potential and possibilities of both types of tools.

Data is money in today’s world. However, most of the data is unstructured and hence it takes a process and method to extract useful information from the data and transform it into understandable and usable form. There are multiple tools available both as open-source and commercially, which help us in transforming the data in a more readable format.

In *Open Source Analytical Tools*, the following are the most in use,

1. RapidMiner
2. Weka
3. R-Programming
4. Knime
5. Orange
6. Pentaho

In *Commercial Analytical Tools*, we have the following tools at our disposal,

1. Octoparse
2. Content Grabber
3. ParseHub
4. Mozenda
5. Import.io
6. SPSS

All of the mentioned tools are equally important for a variety of different statistical and data mining techniques in practice.

Weka, in particular, was developed for analyzing data from the agricultural domain. The tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. It is mostly used for predictive modeling.

Whereas, SPSS offers advanced statistical analysis. It has a vast library of machine-learning algorithms, text analysis, open-source extensibility, integration with big data and seamless deployment into applications. It surveys in order to find statistical results and correlation between the data.

Most of these tools are easy to use. It depends on the choice of the user and the problem in hand. Most of these tools offer very good expertise in visualizations as well.

### Detailed report on available resources and advantages of “Profillic”

Profillic is a web-portal for *AI research and source code to supercharge the projects*. It allows users to explore state-of-the-art in machine learning and AI research. It is an upcoming platform where users can easily pick top technical talents. For this it proposes to use role-specific quizzes with machine learning or fast and accurate skill-based screening because resumes and coding challenges are poor and inefficient signals of technical ability.

The research content is very good and it is increasing on a daily basis, with the source code as well. A user can also add a source code of the paper implementation. It allows to accelerate scientific research and innovation for humankind by unlocking information bottlenecks using artificial intelligence. This portal is designed in guiding research scientists through the arduous R&D lifecycle. Furthermore, support is and will be added to find research sponsorships and R&D funding. Connect with researchers interested in similar problems in machine learning, deep learning, natural language processing, computer vision, data mining, neural networks, artificial intelligence, data science and explore working together on projects.