BRAC UNIVERSITY

Classification of Arsenic Contamination in Water using Machine Learning

A Thesis submitted in partial fulfillment of the

Requirement for the degree Bachelor of Science

In

Computer Science & Engineering

By

Yeasir Hossain Leon (ID: 13301095)

Adib Mosharrof (ID: 13341001)

Supervisor:

Prof. Mohammad Zahidur Rahman

Co-Supervisor

Moin Mostakim

Date: 14th January 2014

Declaration

This is to certify that the thesis entitled “Classification of Arsenic Contamination in Water using Machine Learning”, which is submitted by Yeasir Hossain Leon (ID: 13301095) and Adib Mosharrof (ID: 13341001) in partial fulfillment of the requirement for the award of degree of Bachelor of Science in Computer Science & Engineering to the Department of Computer Science & Engineering, BRAC University, 66 Mohakhali C/A, Dhaka, 1212, comprises only their original work and due acknowledgement has been made in the text to all other material used. The result of the thesis has not been submitted to any other University or Institute for the award of any degree or diploma.

Approved By:

Supervisor: Prof. Mohammad Zahidur Rahman

Co-Supervisor: Moin Mostakim

Acknowledgement

Firstly, we would like to thank our supervisor Prof. Mohammad Zahidur Rahman and Co-Supervisor Moin Mostakim Sir for guiding us through our bachelor thesis. We have learned a great amount working on this thesis and much of that is due to them.

We would like to thank Asian Arsenic Network for providing us their survey datasheets on arsenic contamination for the purpose of our research.

Finally, we thank our family, friends, and all the teachers for their motivation, inspiration and support.

Abstract

Arsenic is a semi-metal element in the periodic table that is odorless and tasteless. It enters drinking water supplies from natural deposits in the earth or from agriculture and industrial practices. In South Asian countries, especially in Bangladesh, arsenic contamination is a big concern for a mass population because the main sources of drinking water are shallow and deep tube wells. This causes deadly effects to humans as it causes different types of diseases and can also lead to cancer.

An NGO, Asia Arsenic Network, has performed laboratory tests on samples of arsenic contaminated water from some areas of Bangladesh, and the resulting data has been provided to us. There are 11 features in the data, and one output feature, arsenic level, which has 5 classes.

Introducing Machine Learning, a branch of Artificial Intelligence, into the arsenic contamination data will help to produce a better diagnosis of this threat. Algorithms like Neural Networks and Support Vector Machines have been applied on this dataset and the performances of each algorithm has been analyzed to find out which algorithm performs best in the classification of arsenic contamination in the data set provided. Error analysis has been done using precision, recall and F1 score.

Table of Contents

Chapter 1

Introduction

Arsenic is a big issue in South Asian countries, especially in Bangladesh. With the use of technology, a lot of the hard and difficult tasks could be made simple and newer relations and information could be discovered. With such hopes the initiative to perform a research on arsenic contamination and applying machine learning on it was taken. This research tries to classify the levels of arsenic based on many features of the water that were tested in the laboratory. Two algorithms were used and an analysis and comparison between the two were carried out to understand which algorithms are performing well for this type of data. Applying artificial intelligence on some raw unprocessed data was an avenue of interest for us.

1.1 What is Arsenic

Arsenic is a semi-metal element in the periodic table. It is odorless and tasteless. It enters drinking water supplies from natural deposits in the earth or from agricultural and industrial practices [1].

1.2 Arsenic Contamination

Arsenic contamination of groundwater is often due to naturally occurring high concentrations of arsenic in deeper levels of groundwater [2].

The main source of drinking water in most of the rural areas of Bangladesh is comprised of deep and shallow tube wells. Since arsenic is present in groundwater, and since the water is directly collected without any filtering, arsenic contamination poses a great threat to a mass population.

The table below shows some of the diseases that can be produced due to arsenic

Kidney damage and failure	Anemia	Low blood pressure
Shock	Headaches	Weakness
Delirium	Increased risk of diabetes	Adverse liver and respiratory effects, including irritation of mucous membranes
During development, increased incidence of preterm delivery, miscarriage, stillbirths, low birth weight, and infant mortality	During childhood, decreased performance in tests of intelligence and long-term memory	Skin lesions

Regular exposure leads to cancer and other toxic health effects, including cardiovascular disease, neurological problems, and developmental disorders. A 581-participant study conducted by the Harvard School of Public Health confirmed the association between arsenic exposure and DNA methylation, a biological process that causes many debilitating and fatal diseases [3],[4] .

Johns Hopkins Bloomberg School of Public Health has linked arsenic exposure to high blood pressure and the New York University School of Medicine cites it as a contributor to cardiovascular disease [5].

Also, most toxic chemicals adversely affect the male reproductive system and arsenic is one of them; Chinese research has confirmed it reduces semen quality [6].

There are many NGOs working to mitigate the problem of arsenic contamination. Many tube wells are tested for arsenic contamination and the marked green for safe and red for unsafe. These NGOs collect and record data on arsenic contamination and day by day more data are being collected and processed. Some NGOs even do further lab testing on these water to get even more information about the water and which elements are contributing heavily towards the contamination.

1.3 Goals

Classifying arsenic contamination using machine learning algorithms like Neural Networks (NN) and Support Vector Machines (SVM). A comparison in the performance of the algorithms will be done for the classification purpose to see which algorithm fits best in this scenario.

Chapter 2

Literature Review

2.1 What is Machine Learning

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [7].

The first real definition of Machine learning was created by Arthur Samuel in 1959, which defined machine learning as follows

> Field of study that gives computers the ability to learn without being explicitly programmed [24].

Later Tom M. Mitchell provided a widely quoted and more formal definition:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.[8]

The core of machine learning deals with representation and generalization. Representing the data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the ability of a machine learning system to perform accurately on new, unseen data instances after having experienced a learning data instance. The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences) and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases. The performance of generalization is

usually evaluated with respect to the ability to reproduce known knowledge from newer examples.
There are different types of machine learning, but the two main ones are:

- Supervised Learning
- Unsupervised Learning

2.2 Supervised Learning

Supervised learning is the machine learning task of inferring a function from labeled training data [9].

A simple analogy to supervised learning is the relationship between a student and a teacher. Initially the teacher teaches the student about a particular topic. Teaching the student the concepts of the topic giving answers to many questions regarding the topic. Then the teacher sets an exam paper for the student to take, where the student answers newer questions.

To relate this to machine learning, the system learns from the data provided which contains the features and the output as well. After it has done learning, newer data is provided without outputs, and the system generates the output using the knowledge it gained from the data on which it trained.

2.3 Unsupervised Learning

In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution [10].

2.4 Algorithms

Since there are so many algorithms for machine learning, it is not possible to use all of them for analysis. For this research paper, only two algorithms were selected, Artificial Neural Networks (AAN) and Support Vector Machines (SVM).

2.4.1 Neural Networks

NN (Neural Network) are models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected 'neurons' that can compute values from inputs by feeding information through the network [11].

In the history of machine learning, NN has been greatly used for problems like handwritten character identification, data clustering, computer vision, and many more.

NN are comprised of individual processing units called neurons that can compute values from inputs provided to them. Many neurons together form a network that can perform certain tasks.

2.4.1.1 Architecture of a NN

A NN is comprised of network layers. Each layer consists of many neurons. The first layer is called the input layer, the final layer is called the output layer and all the layers in between are called hidden layers. Data or values are passed between layers via synapses. The synapses store parameters called weights that manipulate the data before passing them forward.

A simple network structure of NN is given in figure 1.

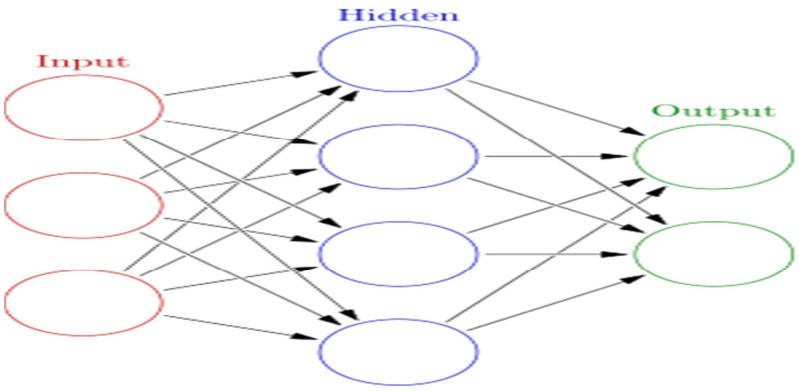


Figure 1: Neural Network Structure

2.4.1.2 How NN works

There are three important parts in a NN, which are

1. The interconnection pattern between different layers of neurons
2. The learning process for updating the weights of the interconnections
3. The activation function that converts a neuron's weighted input to its output activation.

Each neuron has an activation function. A particular computation is done in the activation function. The value produced from the activation is passed on to the next layer using the weights between the layers. These weights are learning parameters that learn from the training data.

2.4.1.3 Feed Forward

Feed-forward NNs allow signals to travel one way only; from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. Feed-forward NNs tend to be straight forward networks that associate inputs with outputs [12].
The inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The sum of the products of the weights and the inputs is calculated in each node, and if the value is above some threshold (typically 0) the neuron fires and takes the activated value (typically 1); otherwise it takes the deactivated value (typically -1). Neurons with this kind of activation function are also called artificial neurons or linear threshold units [13].

2.4.1.4 Backpropagation

In order to train a neural network to perform some task, we must adjust the weights of each unit in such a way that the error between the desired output and the actual output is reduced. This process requires that the neural network compute the error derivative of the weights (**EW**). In other words, it must calculate how the error changes as each weight is increased or decreased slightly. The backpropagation algorithm is the most widely used method for determining the **EW**.

The backpropagation algorithm is easiest to understand if all the units in the network are linear. The algorithm computes each **EW** by first computing the **EA**, the rate at which the error changes as the activity level of a unit is changed. For output units, the **EA** is simply the difference

between the actual and the desired output. To compute the **EA** for a hidden unit in the layer just before the output layer, we first identify all the weights between that hidden unit and the output units to which it is connected. We then multiply those weights by the **EA**s of those output units and add the products. This sum equals the **EA** for the chosen hidden unit. After calculating all the **EA**s in the hidden layer just before the output layer, we can compute in like fashion the **EA**s for other layers, moving from layer to layer in a direction opposite to the way activities propagate through the network. This is what gives back propagation its name. Once the **EA** has been computed for a unit, it is straight forward to compute the **EW** for each incoming connection of the unit. The **EW** is the product of the EA and the activity through the incoming connection [14].

2.4.2 SVM

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [15].

A more formal definition is that a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [16].

SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classier to data that have no obvious fixed-dimensional vector space representation [17].

2.4.2.1 Gaussian Kernel

The Gaussian kernel computed with a support vector is an exponentially decaying function in the input feature space, the maximum value of which is attained at the support vector and which decays uniformly in all directions around the support vector, leading to hyper-spherical contours of the kernel function. The SVM classifier with the Gaussian kernel is simply a weighted linear combination of the kernel function computed between a data point and each of the support vectors. The role of a support vector in the classification of a data point is tempered with alpha, the global prediction usefulness of the support vector, and K(x,y), the local influence of a support vector in prediction at a particular data point.

2.5 Cross validation

Cross validation is the step where the best parameters for the algorithms are selected. The problem of overfitting and underfitting is discovered using cross validation. Normally a machine learning problem has many input features, so it is not possible to visualize the data or the problems that might be occurring. Using cross validation, such problems can be identified via learning curves. The two main problems encountered are underfitting and overfitting.

2.5.1 Underfitting

Underfitting occurs when the algorithm cannot properly fit the training set. The curve produced is probably not complex enough for the classification purpose. A synonym to underfitting is high bias.

To identify the presence of underfitting, learning curves need to be plotted. A learning curve with the training error and cross validation error needs to be plotted. If both the training error and

cross validation are high and there is a small gap between the curves, it can be positively inferred that the algorithm has underfit the training set.

Figure 2 shows a learning curve indication underfitting

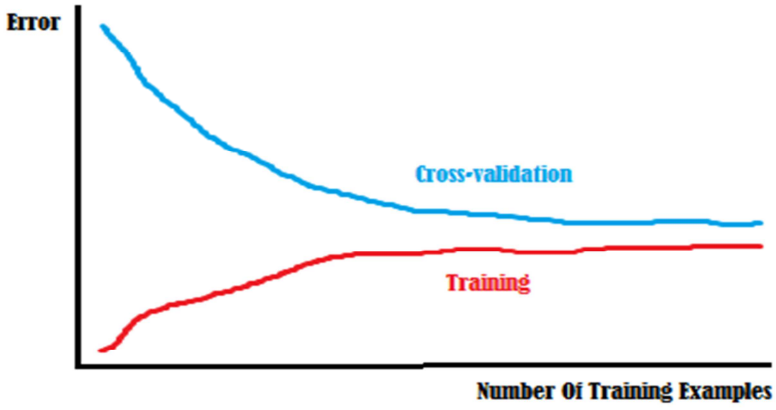


Fig 2: learning curves showing underfitting

2.5.1.1 Solving underfitting

To solve the case of underfitting, a more complex hypothesis is required. In order to solve the underfitting problem, the following steps could be taken

- Adding more features
- Increasing the degree of polynomial features
- Decreasing lambda

In this research, for the data set available, the first option, increasing the number of features is not possible. There are only 11 features in the data, and no new features can be generated from this dataset.

Increasing the degree of polynomial features is another solution, but this works best when the number of rows of data is large. For this research, the data set does not have many rows of data. Increasing the degree of polynomial features could result in a lower accuracy of the algorithms.

A very high value of lambda results in an underfitting problem. The reasons why this happens will be described in the section where regularization is defined in detail. In order to solve this, the optimal lambda is generated by comparing the cross validation error with an array of lambda values. The lambda value that produces the minimum cross validation error is considered to be the optimal lambda.

2.5.2 Overfitting

Overfitting occurs when the algorithm fits the training set a bit too well and does poorly in the test set. The algorithm fit the training set a bit too much, thus it was not able to generalize for unseen examples in the test set. A synonym to overfitting is high variance.

Figure 3 shows a learning curve sample for the case of overfitting.

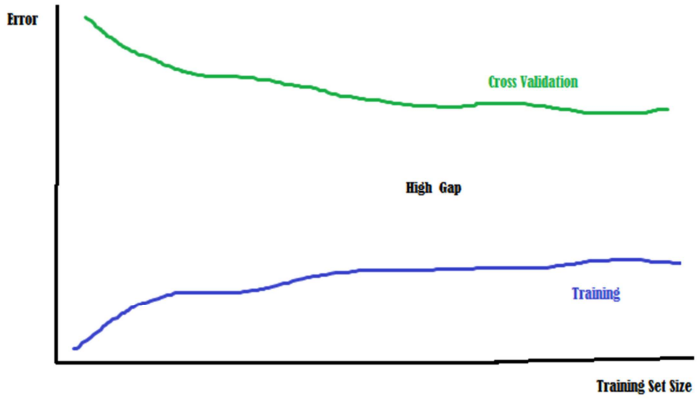


Fig 3: learning curves showing overfitting

From the curve it can be clearly seen that, the training error is very low, but the cross validation error is very high. This is the indication that the algorithm has overfitted the training set.

2.5.2.1 Solving Overfitting

To ameliorate the problem of overfitting, two steps can be taken

- Reducing the number of features
- Regularization

The dataset could be inspected by hand and a handful of features could be dropped. This might produce a better result, but in most cases this is not feasible as there is no measure on how to go about eliminating columns of data. By accident, some important features could be removed, thus hindering the output of the classification further.

To reduce the headache of selecting features by hand, model selection algorithms could be used to select the best features.

2.5.3 Regularization

Another technique that tackles the problem of overfitting is regularization. The basic principle behind regularization is that, a small penalty is given to each feature with a higher order polynomial. In this way the effect each term has on the final output is reduced. This helps in producing a curve that is less bouncy.

Figure 4,5 shows a graph without regularization.

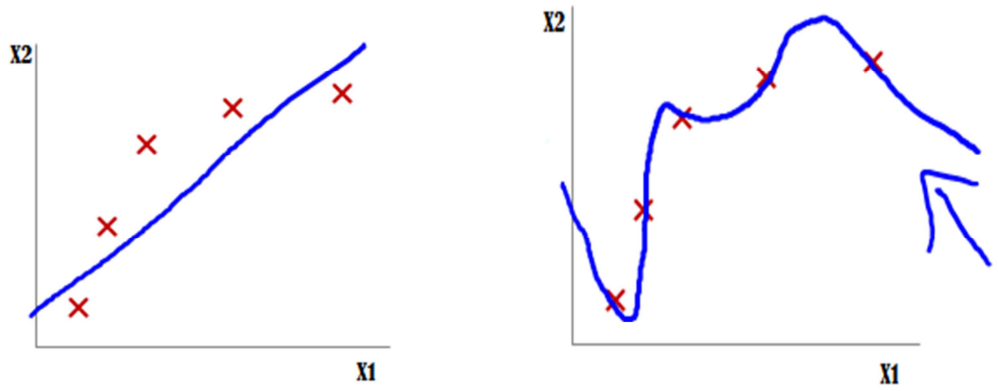


Fig 4, 5: Figure without regularization

After applying regularization correctly, a smoother and better curve to fit the data can be produced.

Figure 6 shows a case where regularization was able to curtail the effects of overfitting.

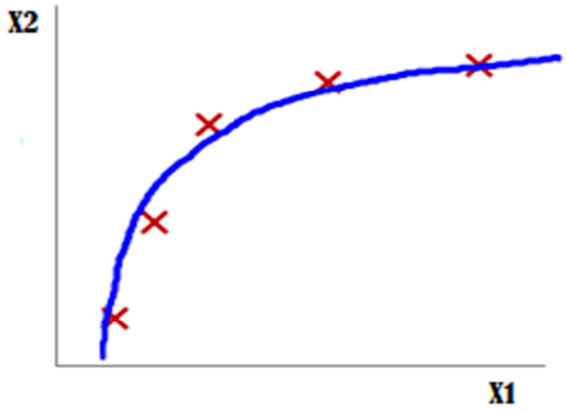


Fig 6: Figure with regularization

Unfortunately, if regularization is done with a very high value, the overfitting problem will transform into an underfitting problem. The complex bouncy curve will be transformed into a

flat line. Regularizing with a small value will not be able to trim the effects of overfitting to a great deal.

2.6 Related Work

Comparison between NN and SVM has been done on different types of machine learning problems. Some common problems are cancer detection, license plate detection, stock market prediction.

In the field of cancer the diagnosis of malignant and benign breast cancer is of great interest. In this paper [18] the above task has been done. 9 input features were fed into three different types of neural network algorithm, which were

- Radial Basis Function (RBF)
- Probabilistic Neural Network (PNN)
- Generalized Neural Network

Each of the algorithms was used to classify the data and then the results of each algorithm were compared.

The characters of a license plate are also analyzed greatly. In this paper [19] the characters of car license plates were classified and recognized using SVM.

In this paper [20] data from the Nigerian Stock Market was collected and the uncertainties in the market were predicted using Naive Bayes and SVM. SVM was implemented in the weka platform.

Data on squall-thunderstorms were collected in this paper [21] and fed into a multi-layer perceptron. Using backpropagation the errors were calculated and weights updated simultaneously to predict squall storm day and no squall storm day.

In the case for water contamination, there have been a few researches on water quality [22]. Factors like pH, temperature, conductivity, turbidity and oxygen were used to classify water quality using SVM. For the quality of drinking water this paper [23] performed a multiclass classification of five classes

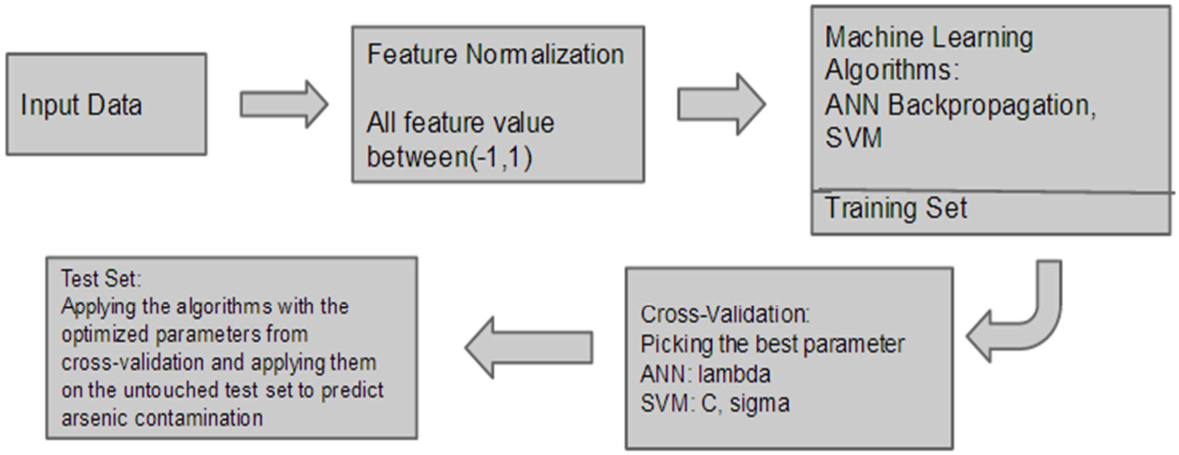
- Excellent
- Good
- Bad
- Very Bad

using Kumar's method. Algorithms like kNN, PART, NN were used to classify only two classes using features of water that were measured in real time like dissolved oxygen, nitrate, pH and temperature.

Chapter 3

Methodology

Figure 7: shows the methodology diagram. All the steps in sequential order are given



3.1 Data Collection

Asia Arsenic Network (AAN), NGO Registration No. 1609, has provided us with data on arsenic contamination. AAN works on arsenic mitigation in many of the highly arsenic contaminated areas. In one of their projects, Survey on Water Sources in Hatila and Tamta Union in Chandpur District, they collected around 7000 samples of water these areas and sent around 5% of these to the lab for further testing. The water sources were groundwater (Shallow and Deep tube wells) and rainwater. The sampling period was from 15th September to 15th November 2012, and the

analysis period from 15th September to 16th November 2012. The following elements were tested in the environment laboratory of AAN:

Property	**Process**	**Device**
pH	Membrane Electrode	Hanna pH meter
TDS	Conductivity electrode	Hanna TDS meter
Chloride	Mohr's Titration	HACH DR/2010 Spectrophotometer
Color	Platinum-Cobalt	HACH DR/2010 Spectrophotometer
Nitrate	Cadmium-reduction	HACH DR/2010 Spectrophotometer
Ammonium	Nesslar	HACH DR/2010 Spectrophotometer
Phosphate	Molibdenum-blue	HACH DR/2010 Spectrophotometer
Iron	Phenanthroline	HACH DR/2010 Spectrophotometer
Manganese	PAN	HACH DR/2010 Spectrophotometer

As a result of this analysis some raw data of arsenic contamination was generated. Around 417 rows of data were tested and figure 8 shows a sample of the column headers of the data.

Option ID	Village	V. ID	Ward	Union	Year of Instllation	Depth (ft)	Date of Sampling	Date of Analysis	Type of Water Device	pH	Color (Pt-Co NCASI)	TDS (mg/l)	Chloride (mg/L)	Ammonium (NH_4) (mg/l)	Nitrate (NO3) (mg/l)	Phosphate (PO_4) (mg/l)	Iron (Fe) (mg/l)	Manganese (Mn) (mg/l)	Arsenic Level
1	Uttar Balasid	3	3	Uttar Tamta	1969	150	15.09.12	17.09.12	0	7.50	11	250	36	0.01	3.20	17.28	3.10	0.18	2
11	Uttar Balasid	3	3	Uttar Tamta	2010	120	15.09.12	17.09.12	0	7.70	9	400	142	0.75	2.60	8.08	2.32	0.52	2
22	Uttar Balasid	3	3	Uttar Tamta	1998	200	16.09.12	18.09.12	0	7.70	20	420	178	3.50	0.01	9.40	3.72	0.02	2
33	Uttar Balasid	3	3	Uttar Tamta	1998	80	17.09.12	18.09.12	0	7.40	19	470	178	6.40	0.60	9.90	3.46	0.34	2
41	Uttar Balasid	3	3	Uttar Tamta	2012	195	17.09.12	18.09.12	0	7.40	5	390	178	0.35	1.10	1.50	1.30	1.52	1
44	Uttar Balasid	3	3	Uttar Tamta	2010	98	17.09.12	18.09.12	0	7.90	26	450	178	8.15	1.20	11.30	3.66	0.12	2
46	Uttar Balasid	3	3	Uttar Tamta	2002	90	4.11.12	6.11.12	0	7.70	32	820	142	0.50	0.00	0.10	12.75	0.62	3

Figure 8: sample of the column headers of the raw data.

For our research, all the columns were not included, as many rows seemed irrelevant so, they were dropped. Figure 9 shows the columns that were included for this research. That includes Depth, Type of water device (Shallow tube, Deep tube wells), pH, Color, TDS, Chloride, Ammonium, Nitrate, Phosphate, Iron, Manganese, and the level of arsenic in the water.

Depth (ft)	Type of Water Device	pH	Color (Pt-Co NCASI)	TDS (mg/l)	Chloride (mg/L)	Ammonium (NH_4) (mg/l)	Nitrate (NO3) (mg/l)	Phosphate (PO_4) (mg/l)	Iron (Fe) (mg/l)	Manganese (Mn) (mg/l)	Arsenic Level
150	0	7.50	11	250	36	0.01	3.20	17.28	3.10	0.18	2
120	0	7.70	9	400	142	0.75	2.60	8.08	2.32	0.52	2
200	0	7.70	20	420	178	3.50	0.01	9.40	3.72	0.02	2
80	0	7.40	19	470	178	6.40	0.60	9.90	3.46	0.34	2
195	0	7.40	5	390	178	0.35	1.10	1.50	1.30	1.52	1

Figure 9: samples of the column headers that were included in the research.

The frequency distribution of the output class, arsenic level, is shown in figure 10.

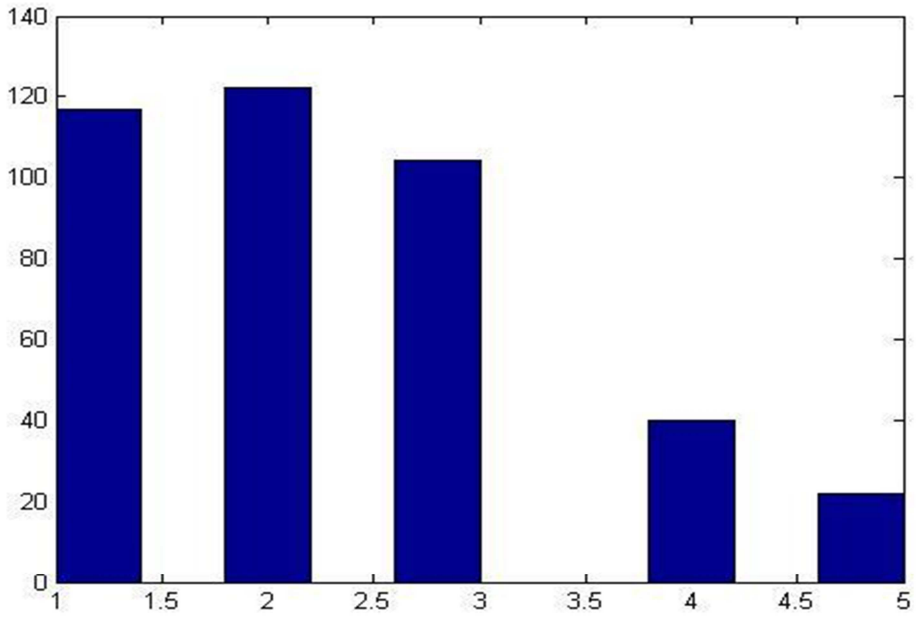


Figure 10: frequency distribution of the output class (1-5)

Figure 10 shows that there is skewness in the classes. The classes from 1-3 have a very high frequency. It dominates over the low frequency of the classes 4, 5. Since there is very little data in the dataset, this will indeed pose a problem for classifying the classes 4 and 5. It will be a common picture that the algorithms will not be able to classify these classes. Since there are a lot of rows for the first three classes, the algorithms will be able to learn about these classes effectively, thus will be able to predict these classes with greater accuracy. A frequency distribution of the classes helps with these types of analysis even before implementing the algorithms.

Another important feature in the data distribution is the skewness of each class. The distribution of each of the input features is shown in the normal distribution in figure 11.

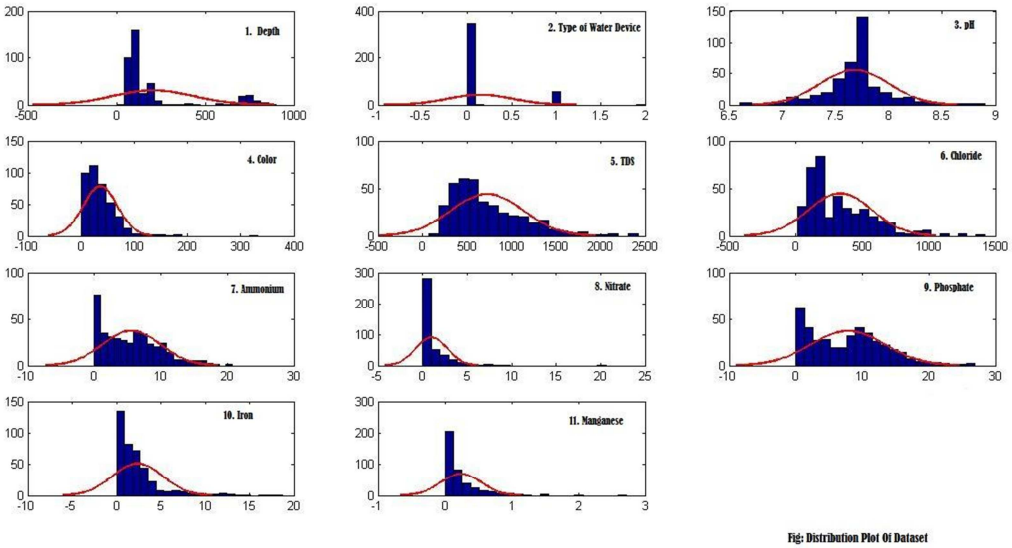


Figure 11: distribution plot for each of the feature in the dataset used in research.

Figure 11 shows that there is skewness and that there are outliers in each of the columns. There is a heavy distribution of a particular type of value in the case of the following features, type of water and manganese. From the figures we can see that there is somewhat an even distribution in the following columns, pH, color, TDS, but there are outliers present in almost all of the features.

3.2 Data pre-processing

3.2.1 Handling missing values and special cases

As mentioned above, there were few missing rows in the data. For the columns depth, and type of water device the missing values were substituted with the average value of the whole column. Features like iron, phosphate. nitrate, ammonium and color had BDL (Below Detection Level) in many rows. This type of value was prominent especially in the column, Nitrate, having about 120 rows containing BDL. All the rows containing BDL were substituted with 0.

3.2.2 Mean Normalization

Different columns have values in different ranges. Some columns like Depth, TDS have values in the range of 100-800, while Manganese and Nitrate have values in the range of 0.1-1. Using values in such different ranges will greatly decrease the efficiency of the machine learning algorithms. In order to mitigate this problem, mean normalization has been applied. The formula used is as below:

$$\mathbf{X = Xn - \mu n / S}$$

Where **Xn** is the column, **μn** is the average value of the column **Xn,** and **S** is the range of values in the column **Xn** like **Xn(max) - Xn(min).**

After performing mean normalization, the values of all the columns fall in the range of $-1 < X < 1$. If mean normalization is not used, some features with values in the range of 1000 will dominate over other features that are in the range of 1-10 or 0-1. Without mean regularization, the contour plot would be somewhat like figure 12

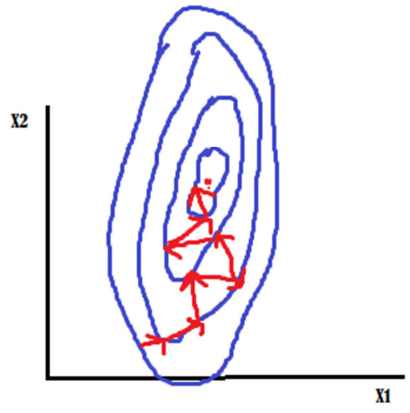


Fig 12: Contour plot without mean normalization, algorithm overshoots to find optimal global minima.

When trying to minimize the cost function, the algorithm will overshoot and the number of iterations needed to find the global minimum will increase, thus increasing the time.

Performing mean normalization will produce a contour plot similar to figure 13, and the number of iteration needed to find the global minimum will be optimal.

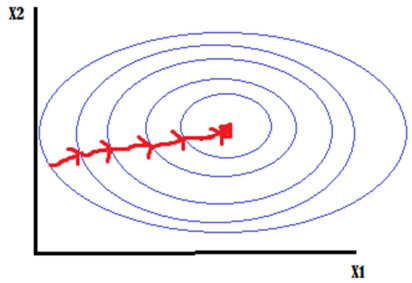


Fig 13: contour plot after mean normalization, algorithm find the optimal global minima, with least iteration

3.3 Applying the Algorithm

The algorithms are applied on the training set. The algorithms try to learn from the data by minimizing the cost function.

3.3.1 Formula of NN

The hypothesis is calculated using feed forward algorithm. The formula for the cost function is

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)}\log((h_\theta(x^{(i)}))_k) - (1-y_k^{(i)})\log(1-(h_\theta(x^{(i)}))_k)\right],$$

In the above formula, the abbreviations of the notations used are as follows

m : number of training examples

k : number of classes in the output

h(x) : hypothesis

J : cost function

3.3.2 Formula of SVM

LIBSVM package was used to implement SVM.

The Gaussian kernel function used in SVM is shown below.

$$K_{gaussian}(x^{(i)}, x^{(j)}) = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{k=1}^{n}(x_k^{(i)} - x_k^{(j)})^2}{2\sigma^2}\right)$$

In the above formula, the abbreviations of the notations used are as follows

Sigma = defines the smoothness of the curve.

X(i) = actual point

X(j) = landmark

3.4 Cross Validation

After training has been done, the parameters that have been learned by the algorithms will be optimized in the cross validation dataset. Learning curves are plotted and the algorithms are further debugged here.

3.5 Types of Analysis

Following cross validation, the algorithms are tested on the untouched 20% test set. Since this 20% of the test set data is selected randomly there is a slight variance of the accuracy of the algorithms each time they are run.

For tackling overfitting, regularization was used. To produce the best parameter for regularization, an array of values were selected and the value which produced minimum cross validation error was selected as the optimal theta.

Running the algorithms returns an accuracy of the classification. Unfortunately, just the accuracy is not enough to describe the performance of the algorithm. The presence of skewed classes is one of the major reasons.

To further evaluate the performance of the algorithms, precision/recall and F1 scores were evaluated.

Chapter 4

Results

4.1 Cross validation results of NN

For NN, the parameter that is to be optimized is lambda which is the regularization parameter. Different values of lambda are used in the training dataset for training the parameter theta; this trained theta values are fed in to the cost function for calculating the training set error, the same parameter is used in the cross-validation data to calculate the cross-validation error. The lambda value with minimum cross-validation error is used in the test set data for training test set parameter. Figure 14 shows the cross-validation curve of lambda.

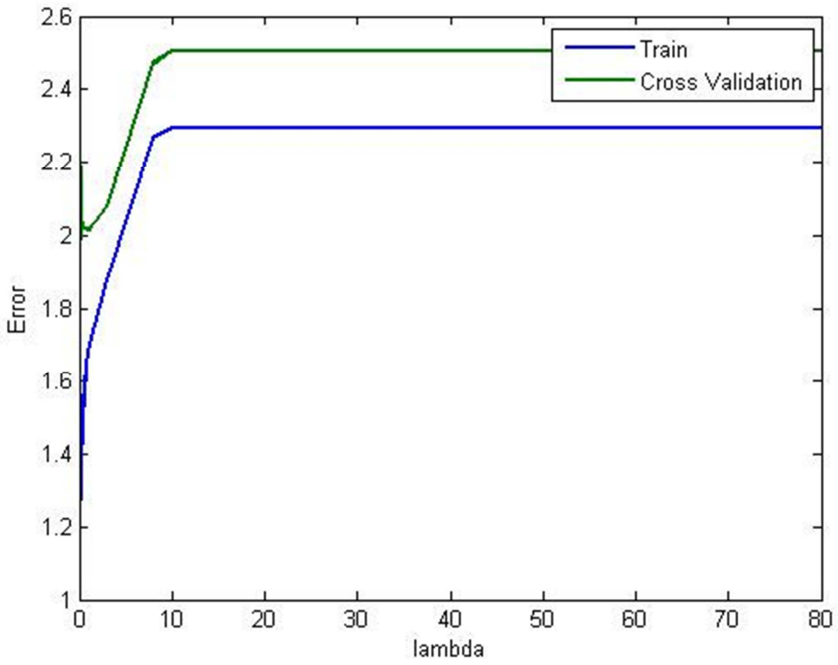


Figure 14: cross-validation curve of lambda in NN

Figure 15 shows the learning curve of NN when the regularization parameter is set to 0.

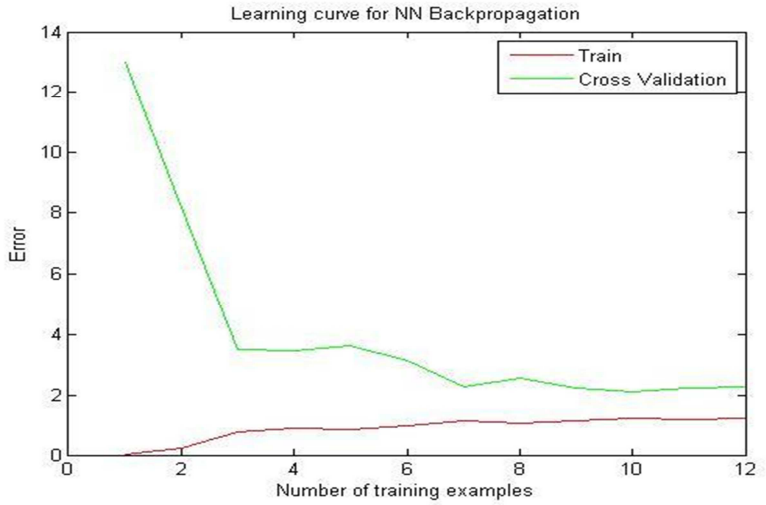


Figure 15: learning curves when regularization parameter (lambda) is set to 0

Figure 15 clearly shows that there is an underfitting problem. The two curves are close to each other, which is the sign that NN has underfit the data.

Cross validation is used to find the lambda which produces minimum cost. Since the data set is randomly selected, the optimal value of lambda differs each time.

Figure 16 shows the learning curve with the optimal lambda that is select from cross-validation

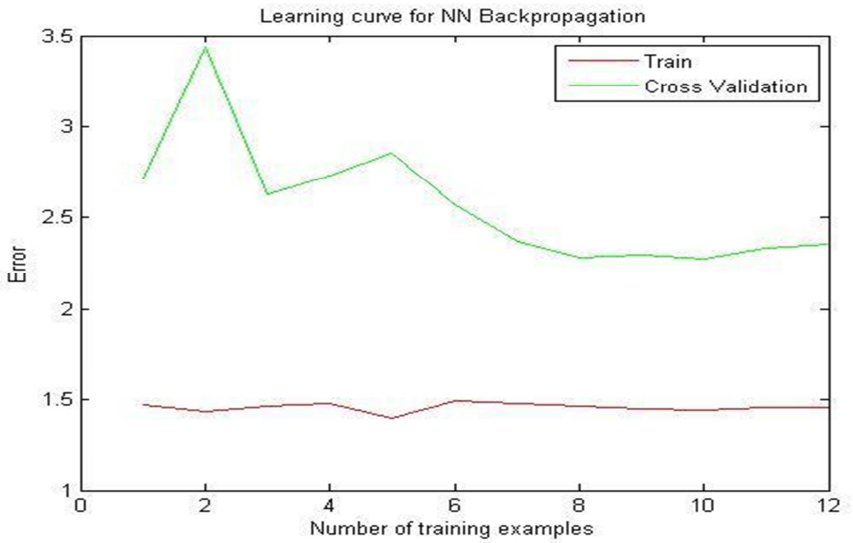


Figure 16: learning curve with optimal lambda, that is found from cross-validation.

Surprisingly, figure 16 shows that now there is an overfitting problem. With the optimal lambda, NN has overfit the data.

4.2 Cross validation results of SVM

For SVM, the parameters that are to be optimized are C and sigma. The formula for C is

C = 1/ lambda

C is the regularization parameter so higher the value of C defines high variance, and lower the value defines high bias.

Sigma defines the smoothness of the kernel function. A high value of sigma will make the features vary smoothly. Similarly a low value of sigma will make the features vary less smoothly.

Using cross validation the optimal value for sigma and C is found using the graph in figure 17.

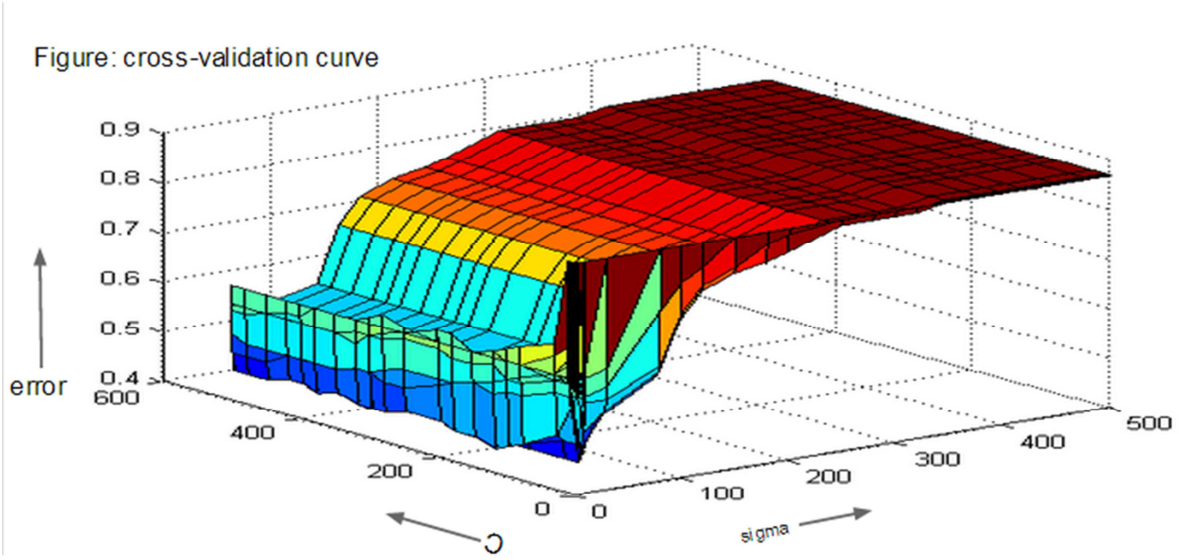


Figure 17: Cross-validation of sigma, and C on SVM

4.3 Accuracy of NN

The accuracy of NN is roughly around 55%. The accuracy also depends on the regularization parameter lambda that was optimized using cross validation with the minimum error.

The accuracy also depends on the regularization parameter lambda that was optimized using cross validation. The value of lambda is also not a constant.

NN was able to classify only ⅗ classes. Figure 18 shows the frequency of the predicted output by NN.

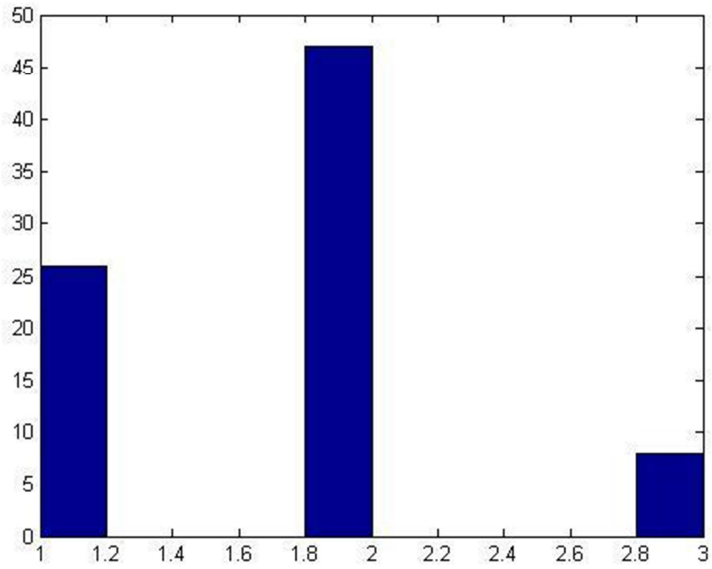


Figure 18: histogram showing the prediction result by NN

4.4 Accuracy of SVM

The accuracy of SVM is roughly around 60%.

SVM was able to classify all the classes. Figure 19 shows the frequency of the predicted output by SVM.

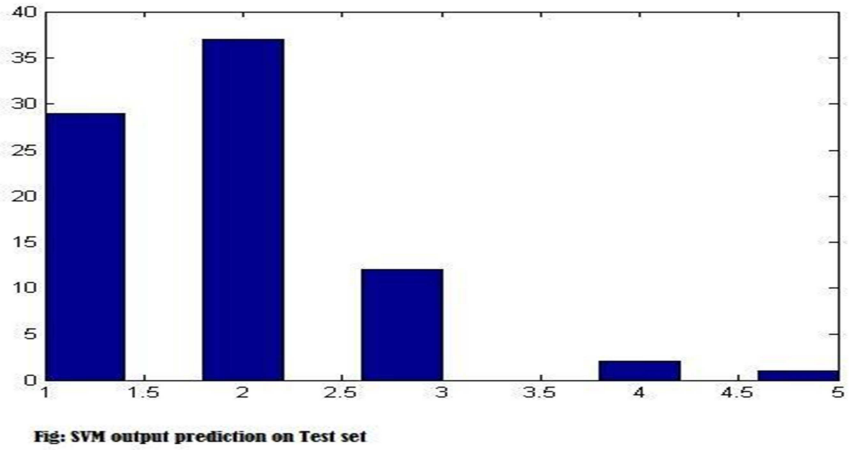


Figure 19: histogram showing the prediction result by NN

The accuracy shows that SVM does better in but this is not the only factor that decides whether an algorithm is doing well. Error analysis needs to be done to decipher if the algorithm is actually performing well.

4.5 Error Analysis

4.5.1 Precision and Recall

Precision is the ratio of true positives over the number of predicted positives. In other words, of all the classes that have been predicted to have a value of 1, what fraction actually has a value of one [25]. Below shows the formula for Precision.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive + False positive}}$$

Recall is the ratio of true positives over the actual positives. In other words, of all the classes that have a value of 1, what fraction was correctly labeled with a value of one by the algorithm. Below shows the formula for Recall.

$$\text{Recall} = \frac{\text{True Positive}}{\text{Actual Positive}} = \frac{\text{True Positives}}{\text{True Positive + False Negative}}$$

The following table shows the result of Precision & Recall from our research.

Algorithm	**Precision(P)**	**Recall(R)**
NN Backpropagation	0.34	0.37
SVM	0.46	0.43

4.5.2 F1 score

In order to evaluate an algorithm, a single real number evaluation is the best option. Although precision and recall provide real numbers, but both of these represent a different result. A combination of both of its value shows the true evaluation of how well an algorithm is doing.

F1 score solves this problem. It combines both the precision and recall values and produces a single metric that evaluates the performance of an algorithm. Higher the F1 score, the better the performance of an algorithm [26]. Below shows the formula of F1 score.

$$F_1 \text{ Score} = 2 \frac{PR}{P+R}$$

The following table shows the result of F1 score from our research.

Algorithm	**Precision(p)**	**Recall(R)**	**F_1 Score**
NN Backpropagation	0.34	0.37	0.35
SVM	0.46	0.43	0.44

Chapter 5

Discussion

5.1 Our findings

All the results indicate clearly that for classification of arsenic contamination given the current database, SVM outperforms NN backpropagation. Since the dataset is small, SVM predicts with a low accuracy.

Furthermore, SVM is also able to classify all the classes, which NN cannot do. Finding the correct combination of values of C and gamma also play an important part in the accuracy of SVM. Initially C-test values were taken to be within the range of 0-100. This produced an accuracy of 55% but when this value was changed to 0-200, the accuracy improved to 60%.

5.2 Future Work

Due to the constraint of the dataset, the algorithms SVM and NN were not able to classify very well. If more data of arsenic contaminated water can be collected, the algorithms will be tested again to see how the performance changes. Other algorithms like C4 and ID3 will be applied on the same dataset to see how they perform.

More study on data preprocessing will be done so that a better feature set can be produced. Model Selection and other methods will be applied on it. Better feature values might lead to a better performance of the algorithms.

Chapter 6

Bibliography

[1] http://water.epa.gov/drink/contaminants/basicinformation/arsenic.cfm Retrieved 2014-01-12

[2] Wikipedia. http://en.wikipedia.org/wiki/Arsenic_contamination_of_groundwater

[3] http://www.globalhealingcenter.com/natural-health/health-dangers-arsenic-toxicity/
Published on 2013-3-6. Dr. Edward F.
Retrieved 2014-01-12

[4] http://www.ncbi.nlm.nih.gov/pubmed/22833016 Retrieved 2014-01-12

[5] http://www.ncbi.nlm.nih.gov/pubmed/22138666 Retrieved 2014-01-12

[6] http://www.ncbi.nlm.nih.gov/pubmed/22776062 Retrieved 2014-01-12

[7] Wikepedia http://en.wikipedia.org/wiki/Machine_learning

[8] Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7., McGraw-Hill, Inc. New York, NY, USA. Published on March 1, 1997

[9] Wikipedia. http://en.wikipedia.org/wiki/Supervised_learning#cite_note-1

[10] Wikipedia. http://en.wikipedia.org/wiki/Unsupervised_learning

[11] Wikipedia. http://en.wikipedia.org/wiki/Artificial_neural_network

[12] Burges.C. J. C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition" *Data Mining and Knowledge Discovery*, Volume 2, Issue 2, June 1998 Kluwer Academic Publishers, Boston 1-43. On Page(s): 121-167.

[13] Wikipedia http://en.wikipedia.org/wiki/Feedforward_neural_network

[14] Stergiou.C, Siganos.D. "Neural Networks."
http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.
Retrieved 2014-1-12

[15] Boser B. E, Guyon I. M. , Vapnik V. N. (1992). “A training algorithm for optimal margin classiers”.*Proceedings of the 5th Annual Workshop on Computational Learning Theory COLT'92*, 152 Pittsburgh, PA, USA. ACM Press, July 1992. On Page(s): 144-152

[16] Wikipedia http://en.wikipedia.org/wiki/Support_vector_machine

[17] Ben-Hur.A, Weston.J (2009) ."A User's Guide to Support Vector Machines". *Data Mining Techniques for the Life Science.*Humana Press. On Page(s): 223-239

[18]Kiyan T., Yildirim T. (2004)"Breast cancer diagnosis using statistical Neural Network" *Istanbul University – Journal of Electrical & Electronics Engineering*.2004. Volume 4, No. 2, June 2004, On Page(s):1149-1153.

[19]Kaur A., Jindal S., Jindal R. (2012) "License plate recognition using Support Vector Machine". *International Journal of Advanced Research in Computer Science and Software Engineering* 2012. Volume 2, Issue 7, July 2012. On Page(s): 403-406.

[20] Magaji A., Isah A., Waziri V., Adeboye K.R.(2013) "A Conceptual Nigeria Stock Exchange Prediction: Implementation Using Support Vector MachinesSMO Model". *World of Computer Science and Information Technology Journal (WCSIT)*, Volume 3, Issue 4, No. 4, On Page(s):85-90.

[21] Chakrabarty H., Murthy C.A., Bhattacharya S., Gupta A. (2013) "Application of Artificial Neural Network to predict Squall-Thunderstorms using RAWIND data". *International Journal of Scientific & Engineering Research*, Volume 4, Issue 5, May-2013, On page(s):1313-1318

[22]Bouamar M., Ladjal M. (2007). "Multisensor system using support vector machines for water quality classification". *Signal Processing and Its Application, 2007. ISSPA 2007. 9th International Symposium on*, 12-15 February 2007, Sharjah. Page(s):1-4

[23]Camejo J, Pacheco O, Guevara M. "Classifier for Drinking Water Quality in Real Time". *Computer Applications Technology (ICCAT), International Conference on*, 20-22 January, 2013, Sousse. Page(s):1-5

[24]Simon P. (March 18, 2013). "Too Big to Ignore: The Business Case for Big Data". Wiley. ISBN 978-1118638170.

[25] Wikipedia http://en.wikipedia.org/wiki/Precision_and_recall. Retrieved 2014-01-12

[26] Wikipedia http://en.wikipedia.org/wiki/F1_score. Retrieved 2014-01-12