Experiments on Conformal Predictions

Nimmy George

Submitted for the Degree of Master of Science in

MSc Data Science and Analytics

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

Dec 6, 2021

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 12483

Student Name: **Nimmy George**

Date of Submission: **6 December 2021**

Signature: Nimmy George

Abstract

This paper focuses on the advances in Conformal Predictions, which is a machine learning method for producing predictions with guaranteed confidence level of new samples with the support of previous experience i.e., training dataset. Its suitability for online learning environments, where outcomes are projected one by one, is one of its most enticing features. Nowadays, conformal prediction has acquired expanding consideration in the different application fields such as anomaly discovery, pharmaceutical domain etc due to its prediction errors and validity conformal predictors guaranteed in both classification and regression tasks.

With the given training dataset, significance level and non-conformity measures we could make the predictions of labels. This method can generate the single predictions as well as the prediction set. Any machine learning algorithm including classical algorithms such as nearest-neighbour method, a support-vector machine, ridge regression, random forest and so on can be applied for generating y with the help of conformal prediction. CP are generally studied in the classical methods such as Classification, Regression and Clustering.

This paper explores various versions of the conformal prediction. It is designed in both batch and online settings. And used different algorithm methods such as nearest neighbour and support vector machine in both transductive as well as inductive conformal prediction. I had done the experience over the USPS handwritten dataset and Wisconsin breast cancer dataset. Here, USPS dataset has higher accuracy than the Wisconsin breast cancer dataset in both methods. The validity of forecasts is assured in online mode, which implies that the error rate of region predictions never surpasses the user-defined significance level. This study is completely based on the "Algorithmic Learning in a Random World", by Vladimir Vovk, Alex Gammerman, and Glenn Shafer (Springer, 2005).

Keywords: conformal prediction, inductive conformal prediction, transductive conformal prediction, k-nearest neighbours, svm

Acknowledgements

I'm extremely thankful to Prof. Alex Gammerman and Dr Ilia Nouretdinov for their supervision over my studies, especially advice and teaching the topic as well as broadening my academic interests.

I'd like to express my gratitude to all of the MSc students and staff in the Computer Science Department for their support and kindness in my studies.

I am also grateful to my family for their love, care, support and encouragement throughout the years.

I'd like to express my gratitude to everyone.

Contents

1. Introduction

In the modern era, the importance of data is increasing day by day and which provides a new dimension to the system. Today, businesses are data-driven and make important decisions based on the data analysis to predict the future by providing more confidence in their actions. At the beginning of 21st century the three scientist Alex Gammerman, Glenn Shafer and Vladimir Vovk introduced the new concept of predicting data with confidence, Conformal Prediction. It is an interesting application of data science. The estimation of predicting confidence is an important challenge faced in both machine learning and statistics, and this challenging issue is solved by the new concepts Conformal Prediction.

The complete definition of Conformal Prediction is defined by Alex Gammerman, Glenn Shafer and Vladimir Vovk as *"Conformal Prediction uses past experience to determine precise levels of confidence in new predictions."* From the definition it mentioned that practical experiences help to provide confidence in prediction. The conformal predictions are applied into many real-world problems in medicines, engineering and other applications such as change detection, anomaly detection, feature selection, fault tolerance and so on. There are many conformal predictors to solve one problem. Efficient performance of conformal prediction results in the increased use of algorithms around the world over the last few years. The CP framework has primarily been used to investigate "traditional" scenarios like classification, regression, and clustering. The goal of this thesis is to investigate performance of Conformal Prediction.

Let's have a look at basic concepts of conformal prediction. The major area in machine learning is the development of algorithms to compute the problems. The machine learning algorithms can be classified as supervised machine learning algorithm and unsupervised machine learning algorithm. The Supervised machine learning algorithm is the model for estimating an output based on one or more inputs with a known desired output. It has two problems: regression and classification. If the set of possible labels in supervised learning is finite, the problem is called classification, here the goal is to predict class labels. If the set of possible labels is infinite, the problem is called regression and the goal is to predict continuous number or floating-point number. Whereas an unsupervised machine learning algorithm receives inputs such as $x_1, x_2, \ldots x_n$ and obtain target output without any known desired output.

In the conformal prediction setup, we are using training data and test data. Training data is used to train the model and test data is used to measure the performance. Based on the availability of labels in the dataset, conformal prediction uses various methods such as supervised, semi-supervised and unsupervised methods. Some of these methods provide general solutions and some are specific to particular domains. In the supervised method, the availability of the training dataset has been labelled normal. Here, the dataset combined, the algorithms reveal how to predict the output. The semi-supervised method is used to build a model for the class corresponding to normal behaviour and use the model to identify confidence in the test data. The unsupervised method is opposite to the supervised method which is mentioned as learning what normal behaviour is from a large volume of data and detecting the deviations from this behaviour. Generally, clustering methods are used in unsupervised learning.

Conformal Prediction provides how confident we are that $y = y^{\wedge}$. The inspiration behind the conformal prediction is that we could expect predictions from past experience and the results are valid and efficient and also provide confidence for the predictions of new sample. In regression it evaluates prediction intervals and in classification a set of classes, which are true values. Conformal Prediction can be implemented in different types of parameters which are non-conformity measure which also include the classical methods and significance level. The non-conformity measure is the output of prediction algorithms such as random forest, neural network, decision trees, support vector machines, nearest neighbours etc.

This paper also describes the batch also known as offline and online conformal prediction. Batch learning or offline learning represents the training of models in a batch manner. One of the disadvantages of batch learning is the process takes more time and resources for re-training models. The main aim of the anomaly detection algorithm is to reduce the false negatives and positives in the dataset. However, generally algorithms can trade-off by adjusting the parameters i.e, when we increase the sensitivity of the algorithm false positive also increases. Anomaly detection also reduces the number of false alarms. Later, the scientists introduced the new concept called online learning which receives data as a continuous flow that is fast and cheap. An anomaly detector detects the anomalies promptly and we are able to know when the anomaly happens right away.

This report explains the implementation of conformal prediction algorithm in both batch and online processes. It also explains the transductive and inductive

methods using support vector machine and nearest neighbour algorithms. In the next section, we formalize the conformal prediction framework and related works. Section 3 explains the dataset that is used in this study. Section 4 describes the results of my experiments with a different dataset as well as algorithms such as nearest neighbours and support vector machines. Section 5 summarizes my study and provides directions for future work.

2. Background Research

2.1 Conformal Prediction

In this section we introduce the theoretical concepts of Conformal Predictions, types of conformal predictions in offline as well as in online methods, how to implement etc.

Conformal Prediction was introduced by Alex Gammerman, Glenn Shafer and Vladimir Vovk in their book *"Algorithmic Learning in a Random Word"* in 2007 where they describe it as *"Conformal Prediction uses past experience to determine the precise levels of confidence in new predictions"*, i.e., it helps us to produce the confidence values for prediction by itself. In the basic setting, $z_1, z_2, z_3, \ldots \in Z$ where $z_i = (x_i, y_i)$ which is the sample and Z is the measurable space. Here, we could find that Z contains more than one element and each is measurable. Here, our goal is to predict the value of a new sample. In regression, conformal prediction evaluates a prediction interval i.e., a floating-point number while in classification it sets the classes that are a point prediction. However, both prediction intervals and sets result in true values with high probability.

Typically, we are given a set of training set m= $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_i is a vector of attributes or object space and $y_i \in \{y_1, y_2 \ldots y_n\}$ is a label or label space given to the instance x_i. And we need to predict the label y_{n+1} for a new instance x_{n+1}. In order to make a prediction, we assume all classes $y \in \{y_1, y_2 \ldots y_c\}$ for the new instance, and we test for each one how likely the prediction is going to be correct. For the assumption we append the new instance x_{n+1}in our training set together with the assumed class y_n and then we train the machine learning algorithm with extend training set i.e.

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \qquad (1)$$

Then we need to calculate the non-conformity scores for each sample in (1) using non-conformity measures. Let us consider a bag of the size of n belongs to N is the collection of n elements $(z_1, z_2 \ldots z_n)$ which is unordered and some elements may be identical, then ordered the elements based on our notation. The Z(n) is the measurable space with the 'n' size of elements. A non-conformity measures is a function that assigns a score to a data sample based on how well it conforms to other data samples i.e., non-conformity measure is a measurable mapping to each possible bag of old samples and each possible new sample. There are different nonconformity

measures each one defines a conformal predictor as well as a smoothed conformal predictor.

Conformal Predictor

Conformal Predictor is a technique which provides predictions with a valid measure of predictions. Suppose we have a training set $(x_i, y_i)$and we need to predict the labels (y_i) and non-conformity measure (A_n), which is a function that maps finite sequences of values $(z_1, z_2 \ldots z_n)$ i.e, a bag and corresponding non-conformity scores are $\propto_1, \propto_2, \propto_3, \ldots \propto_n$. Here, the conformal predictor assumes all pair originate from the same probability distribution and are independent of previous output which is called independently and identically distributed (i.i.d) data assumption.

To calculate whether the new object x belongs to the label y, we must compare this object to all the previous samples belonging to the bag. To achieve this we need to calculate the p-value. The p-value gives an upper estimate of the probability of x_{n+1} appearing, and which assumes it belongs to y, this is called a conditional probability. To compute p-value for each label y where the training set is $z_1, z_2 \ldots z_n$ and test sample x*:

$$\frac{|\{i = 1, \ldots, n+1 : \alpha_i \geq \alpha_{n+1}\}|}{(n+1)} > \epsilon$$

P-value is:

$$p^y = \frac{|\{i = 1, \ldots, n+1 : \alpha_i \geq \alpha_{n+1}\}|}{(n+1)}$$

Where α is the non-conformity score of a sample and n is the size of initial dataset, n+1 is the size of the dataset with the new sample.

Based on the selection of non-conformity measure the performance is evaluated. The non-conformity measure is denoted by A, which calculates the distance between the sample and a set of the same type of sample where the samples are taken together with their labels which results it have the property of equivalent i.e., nonconformity score for a particular sample should have the same value but no matter its position in the sequence. If α_i is small, the z_i is non- conforming or strange.

Let us consider an example for the concept of k-Nearest Neighbours, where the non-conformity score is the distance to the nearest neighbour of the same class. Here, if the distance to the nearest sample of the same class is high then we get a high nonconformity score. By calculating non -conformity scores of every sample with new sample as well as calculating the ranking. If the new sample has low

nonconformity, it fits the dataset and has a high rank within the dataset and if the new sample have high non-conformity, it fits the dataset and have a low rank within the dataset. Then let's calculate implausibility using p-value notation as:

$$p^y = \frac{|\{i = 1, \dots, n+1 : \alpha_i \geq \alpha_{n+1}\}|}{(n+1)}$$

Where $\propto_i$ is the non-conformity score of a sample and n is the size of initial dataset, $n+1$ is the size of the dataset with the new sample

The p-value is the fraction that ranges between 0 and 1 as well as the rank cannot be greater than the number of samples in the datasets. A high p-value indicates that the sample conforms to the data set. If the p-value is 0 or near to 0 it shows that the sample would be most strange and p-value is 1 then the sample is least strange. The level of significance $\in$ determines the amount of conformity.

The other parameter which is used in the conformal prediction is significance level for outputting a prediction set. The significance level denoted by ε.The significance level determines the confidence. The confidence of the prediction set is $1-\varepsilon$.

Smoothed conformal predictor

In order to simplify and strengthen the scientist introduce modified version of conformal predictor, called smoothed conformal predictor which determined by the nonconformity measure (A_n) is the following randomized confidence predictor Γ: the set:

$$\frac{|\{i = 1, \dots, n+1 : \alpha_i > \alpha_{n+1}\}| + \tau|\{i = 1, \dots, n+1 : \alpha_i = \alpha_{n+1}\}|}{(n+1)} > \in \quad (3)$$

Here, smoothed p-value is:

$$p^y = \frac{|\{i = 1, \dots, n+1 : \alpha_i > \alpha_{n+1}\}| + \tau|\{i = 1, \dots, n+1 : \alpha_i = \alpha_{n+1}\}|}{(n+1)}$$

Where $\propto_i$is the non-conformity score of a sample and n is the size of initial dataset, n+1 is the size of the dataset with the new sample

The main difference between the (2) and (3) is the smoothed p-value takes more care in breaking the ties between non-conformity scores. In p-value, it will have the same rank when multiple samples have the same non-conformity score, which results in a great change in the results. But in smoothed p-value, dividing the rank into two parts to solve the issue:

1. The number of samples that have greater non-conformity score
2. The number of ties multiplied by τ

The number τ is generated randomly from a uniform distribution between [0,1], which provides stronger validity than regular p-values. For example, the non-conformity scores [1,2,2,2,2,3,4,5]. The rank is [1,2,3,4,5,6,7,8]. The rank of all samples with nonconformity score 2 is 5 and the p-value is 0.70. If one score is changed then the p-value is reduced; this slight change reflects in the p-value. In certain cases, we get new data without knowing the labels of the samples and here we need to predict the label. It is difficult to compute a non-conformity score and p-value without knowing the label. So, the p-value is calculated for every potential label of a new sample.

Point Predictions

At the point when we perform point prediction, the expected results are credibility and confidence. Confidence implies that if the p-value of a label Y is modest than or equivalent to the significance level, then Y is not the genuine label. If the significance level is equivalent to the second-largest p-value for a given input with a set of all possible labels which means that the probability of any label other than Y. The confidence value for each prediction is one minus the probability of any one of the other labels. The higher the confidence value for a prediction the less likely it is not to be the true label. Credibility is the highest p-value of the prediction, which indicates how good the training set is, i.e., if the credibility is low then the training set is not random.

Credibility = Maximum p-value of the prediction
Confidence = 1 - second largest p-value

Prediction sets

Prediction sets are predictions that have multiple labels. These predictions can be changed based on the significance level. The significance level is seen as one minus confidence level i.e. The confidence of the prediction set is $1-\varepsilon$. By using it we control the rate of error of the predictions where the significance level of ε, error rate do not exceed ε. If the significance level is lower, we get multiple predictions, so the accuracy will be higher. But for a higher significance level, no prediction set due to no p-values greater than the significance level, so the accuracy is lower.

Induction and Transduction

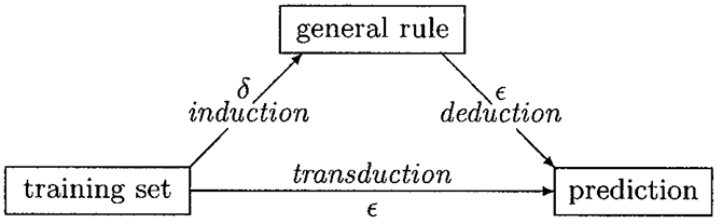

Figure1*:Inductive and transductive prediction [Figure taken from [47]]*

The above figure describes the difference between inductive and transductive prediction. In inductive prediction, from examples, we get some general rules or predictions or a model which describe how we use a training dataset to generate a model about the data, then we apply this model to the test dataset to obtain the prediction. In the deductive step, we apply this to a new sample and derive a prediction from the general rule. But in transductive, we skip the step of generating the rule, so no processing is applied to the training dataset and all are computed on each individual test dataset.

Batch and Online Learning

Both transduction and induction can be performed either by a batch process or an online process. The batch or offline learning represent the training of models in a batch manner i.e., the models get trained with the accumulated data from time to time in a batch way. The models moved into production only at regular intervals based on the performance of models trained with new data. The main cons of this learning are it takes a lot of time and resources to re-training the model i.e., this algorithm will perform optimization steps to reduce the cost function until it reaches specific stopping criteria. On large datasets, common batch learning techniques scale between quadratically and cubically in the number of samples, resulting in quite long training durations.

Online learning, training occurs by continuously feeding data in an incremental way. In this learning method, there are optimization approaches that use an online learning strategy to approximate the solution to a fixed batch problem. Here, each time we observe the sample and predict the label and then add it to a training set that will be used to predict the label for future samples. So, the training set is constantly growing as well as predictions also improved. For example, First,

we observe the sample x_1 and predict its label y_1. Then we observe the sample x_2 and predict the label y_2 and the process is going on till it reaches the nth step, so in each step, it becomes faster and cheaper. The learning rate is an important concept which is defined as the rate at which the model adapts to a new data set. System with a high learning rate will tend to forget the learning quickly compared to a system with a low learning rate similar to batch learning. One of the disadvantages is that if it is fed with bad data, the system will have bad performance. The main advantage of this learning is limited to less computing resources and lesser time consumed.

Validity and Efficiency

Validity and efficiency are the two factors to evaluate the conformal predictor as good or bad. Conformal predictors are valid consistently given that the irregularity or exchangeability principles hold. A Conformal Predictor is valid in the case of regression if the confidence level matches the portion of instances whose true value falls inside the predicted confidence zone. Conformal Predictors are valid in classification tasks since the set of predicted classes for new instances will contain the true label. The validity is described that the p-values are valid if they met the requirement. If the significance level $\alpha \in [0,1]$ then the set predictor is exactly valid and the probability of making an error is α, under any probability distribution on Z^{n+1}. In the same condition, the probability does not exceed ε, then the set predictor is conservatively valid. The validity provides the correctness of the predictions if the object belongs to a particular class but there are no guarantees if the object truly belongs to a class. This can be achieved by reducing the number of output classes of an object which is called efficiency. If the non-conformity measure is efficient then it leads to the best performance. The average size of the predicted confidence intervals is referred to as the effectiveness of a CP in regression. A conformal predictor is more effective when the intervals are smaller. Efficiency in classification refers to the percentage of single-class predictions that are correct.

2.1.1 Transductive Conformal Prediction

Conformal Predictors are classified into two: transductive conformal predictors (TCP) and inductive conformal predictors (ICP), both predictors are based on the same principles and place the same exchangeability requirement on the data. The main difference is in the usage of the training dataset. In the transductive conformal prediction the whole dataset is split into a training dataset and test dataset and the transductive predictors are determined by transductive nonconformity measures.

2.1.1.1 Offline Transductive Conformal Prediction

In the offline or batch prediction, the predictor is given a training series of instances such as $\big((x_1, y_1), \dots, (x_{n,}y_n)\big)$and its aim is to predict labels for a sequence of new samples with the help of a training series of data. Let us begin with a general algorithm for transductive conformal prediction:

Input
—----
Training Set: $\big((x_1, y_1), \dots, (x_{n,}y_n)\big)$,
Test Set: $\big((x_1, y_1), \dots, (x_l y_l)\big)$
Transductive non-conformity measures: A
Significance Level: $\varepsilon \in (0,1)$
Step 1:
Initialize the variables
Step 2:
for in each sample in the test set do
Step 3:
for possible labels do
Step 4:
Add training set with a new sample and postulated label
Step 5:
Calculate nonconformity score with nonconformity measure for every sample in the new training set
Step 6:
Calculate the p-value or smoothed p-value of test sample based on nonconformity measure:

Regular p-value:

$$p^y = \frac{|\{i = 1, \dots, n+1 : \alpha_i \geq \alpha_{n+1}\}|}{(n+1)}$$

or

smoothed p-value:

$$p^y = \frac{|\{i = 1, \dots, n+1 : \alpha_i > \alpha_{n+1}\}| + \tau|\{i = 1, \dots, n+1 : \alpha_i = \alpha_{n+1}\}|}{(n+1)}$$

end for

Step 7:

Calculate the Prediction using p-value i.e., Label which has maximum p-value.

Step 8:

Calculate the Credibility using p-value i.e. Maximum p-value.

Step 9:

Calculate the Confidence using p-value i.e., 1 - second largest p-value

Step 10:

Calculate the total sum of the p-value

End for

Step 11:

Calculate the false_p_value i.e.,

$$\frac{\text{total sum of p} - \text{value}}{2 * length(X\_test)}$$

Output

—-------

Prediction, Confidence, Credibility, and false p-value

The above algorithm describes a batch process. We can measure transductive non-conformity measure it depends on ordering or not ordering the elements in it. The transductive conformal predictor (TCP) corresponding to A finds the prediction region for the test set $x_{l+1}, \dots x_{l+k}$ $\varepsilon \in (0,1)$. Validity and efficiency are the two factors which are desired in confidence predictors. Here, validity means the error rate does not exceed for the selected confidence level and efficiency is the prediction set which is as small as possible. There are a lot of non-conformity measures in TCP, commonly used NCM are nearest-neighbours and support vector machines. The TCP starts all computations from scratch which makes it inefficient especially in a large dataset, which means for every test sample it has to apply the underlying algorithm and compute all nonconformity scores.

2.1.1.2 Online Transductive Conformal Prediction

In the online transductive conformal prediction, the strongest notion of validity for conformal predictors can be provided. The only difference in Online Transductive Conformal Prediction is the training occurs by continuously feeding data in an incremental way. Here, each time we observe the sample and predict the label and then add it to a training set that will be used to predict the label for future samples. So, the training set is constantly growing as well as predictions also improved. For example, First, we observe the sample x_1 and predict its label y_1. Then we observe the sample x_2 and predict the label y_2 and the process is going on till it reaches the nth step, so in each step, it becomes faster and cheaper.

Let's consider a sequence of positive elements k_1, k_2 ...and some inputs such as $z_1, z_2 \ldots z_n$ and test as n= 1, 2, … in the online transductive conformal prediction and predictor as the k_n labels $y_{ln+1}, \ldots y_{ln+kn}$ for the samples $z_1, z_2 \ldots z_{ln}$. Here, the predictor is the subset Γ_n of Y^{kn} . It can be greater than 1 i.e., $(|\Gamma_n| > 1)$ and when it equal to 1 it will become singleton i.e., $(|\Gamma_n| = 1)$ and empty at $(|\Gamma_n| = 0)$. The smoothed transductive conformal prediction, makes error with significance level at various trials in the online mode. And this issue is solved by a series of independent Bernoulli trials with progress probability equivalent to the significance level. A transductive confidence predictor is a quantifiable procedure for predictor in the online prediction which is based on the significance level (0,1) so that the prediction at a larger significance level is a subset of the prediction at a smaller significance level for each training set and each test set. If the series of errors made by the transductive confidence predictor at any significance level is controlled by a succession of independent Bernoulli trials.

2.1.2 Inductive Conformal Prediction

Inductive Conformal Prediction is the improved or modified version of transductive conformal prediction which solves the issue of poor computational efficiency. The main idea behind the ICP is to estimate how well a new test sample fits into the training set. The training dataset $(z_1, z_2 \dots z_l)$i.e., $\big((x_1, y_1), (x_2, y_2), \dots, (x_{l,} y_l)\big)$ split into two parts proper training dataset $\big((x_1, y_1), (x_2, y_2), \dots, (x_{m,} y_m)\big)$ of size $m < 1$ and the calibration dataset $\big((x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \dots, (x_{l,} y_l)\big)$ of size l-m. Here, the total datasets are 3 proper training datasets, calibration datasets and test datasets. Using a proper training set to generate the model for new sample and then assign the non-conformity score to each sample using the calibration set which results the sequence of $\alpha_{m+1} \dots \alpha_{m+n}$ Then compute the p-value. We could perform inductive conformal prediction in two ways: batch/Offline and online process.

2.1.2.1 Offline Inductive Conformal Prediction

As we discuss earlier, consider the series of samples $z_1, z_2 \dots z_n$ were each sample contains two elements such as $\big((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\big)$ where x is object and y is label. Then, Let's look at the general algorithm of an inductive conformal prediction:

Input

———

1.Training dataset: It split into 2:

i. Proper Training dataset: $\big((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\big)$

ii. Calibration dataset: $\big((x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \dots, (x_n, y_n)\big)$

2.Test dataset: $\big((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\big)$

3.Inductive non-conformity measure: A

4. Significance Level: $\alpha \in (0,1) \in$

Step 1:

Initialize the variables

Step 2:

Calculate the non-conformity scores of the calibration set based on non-conformity measure

Step 3:

for each sample in test set do

Step 4:

for all possible labels in training set do

Step 5:

Calculate non-conformity score of test sample based on proper training set.

Step 6:

Calculate the p-value or smoothed p-value of test sample based on the rank of its non-conformity score within the calibration set.

Regular p-value:

$$p^y := \frac{|\{i = m+1, \ldots, l \mid \alpha_i \leq \alpha^y\}| + 1}{(l - m + 1)}$$

or

smoothed p-value:

$$p^y := \frac{|\{i = m+1, \ldots, l | \alpha_i < \alpha^y\}| + \tau|\{i = m+1, \ldots, l | \alpha_i = \alpha^y\}|}{(n+1)}$$

end for

Step 7:

Calculate the Prediction using calculated p-value or smoothed p-value i.e., Label which has maximum p-value.

Step 8:

Calculate the Credibility using calculated p-value or smoothed p-value i.e. Maximum p-value.

Step 9:

Calculate the Confidence using calculated p-value or smoothed p-value i.e., 1 - second largest p-value.

End for

Step 11:

Calculate the total sum of the p-value.

Step 12:

calculate the false-p-value i.e.,

$$\frac{\text{total sum of p} - \text{value}}{2 * length(X\_test)}$$

Output

—--------

prediction, confidence, credibility and false p-value.

This model is trained first on the appropriate training set and then used to predict the activity label for the occurrences in the calibration set. The calibration set instances true and anticipated class labels are used to compute non-conformity scores using the non-conformity measure of choice. The non-conformity score is

used to quantify a new instance's strangeness in comparison to previous instances. And then applied to the test set to calculate the p-value. The non-conformity scores for a new instance are ranked against the non-conformity score list generated for the calibration defined by the p-value and then compared to significance level ε. Here, all inductive conformal predictions are valid and smoothed conformal predictions are exactly valid.

The concept behind the calibration set is it enables us to calibrate test example conformity scores by converting them to a probability-type scale. The ICP corresponding to A is defined as set predictor

$\Gamma^{\epsilon}(z_1, \ldots, z_l,\ x) := \{y | p^y > \epsilon\}$ where $\varepsilon \in [0,1]$ is the chosen significance level ($1 - \epsilon$ is known as the confidence level). Here, the given training set and a new object x in the ICP predicts its label y and it makes an error if $y \notin \Gamma^{\epsilon}(z_1, \ldots, z_l,\ x)$. This is proved by Vovk(2005) for inductive conformal prediction.

In TCP, we need to recalculate the non-conformity score of all training for every test sample which results in more time consumed. But ICP reduces the number of non-conformity scores that needs to be calculated, here nonconformity scores are calculated relative to the training set.

The important drawback of the ICP is, it doesn't give proper accuracy, this is because TCP uses all training samples underlying the algorithm whereas ICP uses only proper training set for computing non conformity scores. TCP uses all training samples whereas ICP uses small portion of training sample and the calibration set.

2.1.2.2 Online Inductive Conformal Prediction

The validity is stronger in online mode at each significance level. The online learning environment for inductive conformal prediction has its own set of difficulties, as there is no single training set to which new examples may be added after generating predictions. Adding additional samples to the training set proper is an option, but it negates the purpose of inductive conformal prediction because the calibration set non-conformity scores would have to be recalculated every time a new sample is added to the training set proper, reducing computing efficiency. Based on the following the prediction, more test samples could be added to the calibration set, resulting in better forecasts as the calibration set expands. Another option is for the calibration set to become part of the training set proper once it reaches a specific size, and the calibration set to start over with the new test sample. Because the recalculation of calibration set scores will be infrequent, it is unlikely to have a significant impact on computational efficiency.

2.1.3 Non- Conformity Measures

The non-conformity measures(A) are an important factor that are used in conformal prediction to estimate the p-value. The non- conformity measure defined as real-valued function A(B,(z)) where z = (x,y) which results how dissimilar a new sample (z) is from the sample in the bag or model (B) i.e. non-conformity score. The algorithm produces prediction regions with the help of real valued functions as non-conformity measures. Generally, we are calculating the distance between two samples which is pointed out in the text by Gammerman and Vovk in 2005. The results of prediction set is dependent upon the non-conformity measure chosen that in the algorithm. So, accuracy and confidence depend upon the chosen non-conformity measure. There are many non-conformity measures used in conformal prediction including support vector machine, k-nearest neighbour, random forest, decision trees, neural network etc. Here, I concentrate on two non-conformity measures k-nearest neighbour and support vector machine. To avoid excessive computing costs, it seems more logical to focus on top-k-distances of the symmetric group, as a nonconformity measure requires a notion of distance on the underlying data space.

2.1.3.1 K-Nearest Neighbour

The next two paragraphs discuss about the understanding of K-nearest neighbour algorithm which includes KNN classifier, KNN regressor and unsupervised nearest neighbour. K-Nearest Neighbour algorithms are classical algorithms in machine learning. These are simple and easy-to-implement algorithms which can be used in any statistical learning such as supervised and unsupervised machine learning. Whilst Supervised machine learning algorithm is the model for estimating an output based on one or more inputs. Unsupervised machine learning the machine receives inputs such as X_1, X_2 …..... X_n and obtain target output without any desired output. The supervised machine learning has two problems regression and classification. If the set of possible labels in supervised learning is finite, the problem is called classification. The goal is to predict class labels. If the set of possible labels in the supervised machine learning is infinite, the problem is called regression and the goal is to predict continuous number or floating-point number. In nearest neighbour algorithm, make prediction for new data point which is closest data points in the training dataset.

This section describes the general algorithm of kNN algorithm as:

Step 1:

Load the dataset and initialize the value of 'k' to choose the nearest neighbours.

Step 2:

For each sample in the data do:

Step 3:

Calculate the distance between test sample and nearest training sample based on the 'k'. To calculate the distance generally using Euclidean distance, Manhattan distance or Minkowski distance.

Step 4:

Add the calculated distance to the index of the sample to an ordered collection.

Step 5:

Sort the ordered collection in ascending order

Step 6:

Select the first 'k' sample from the sorted collection

Step 7:

Get the label of selected samples

end for

Step 8:

If we are choosing classification, return the mode of 'k' labels or if we are choosing regression then return the mean of the 'k' labels.

The equation to calculate Minkowski distance:

$$\boldsymbol{d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^c \right)^{1/c}}$$

The equation to calculate Euclidean distance:

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$$

The equation to calculate Manhattan distance:

$$\boldsymbol{d(x,y) = \sum_{i=1}^{n} |x_i - y_i|}$$

Given a positive integer 'k' and a test sample x_0, the KNN first identify the "k" points in the training data that are closest to x_0 and represented by N_0. Then estimates $k= f(x_0)$ using the average response of all the training sets. When k=1, KNN fit perfectly interpolate the training observations and consequently takes hence form of the step function. Let's consider when k=10, KNN fit step function and averaging all the observations which results constant prediction and smooth fit. The general optimal value for k depends on the bias-variance trade-off which means small value of k provides most flexible fit with low bias and high variance. Large k value provides smoother and less fit. As we discussed earlier unsupervised learning does not have known output, which has only input data and needs to extract the knowledge which are generally using in clustering dataset.

This section discusses, about the conformal prediction on nearest neighbour classification. The main idea behind is samples of same class, group together and away from the different class. The conformal prediction can use in nearest neighbour classification in two: Calculating the distance to the nearest sample of different class and calculating the distance to the nearest sample of same class. The non conformity score is defined as:

$$\propto_i = \frac{\text{distance to nearest sample of same class}}{\text{distance to the nearest sample of different class}}$$

In some situations, the certain sample is far away from the same class and near to the sample of a different class, the non-conformity score is large, which results the sample does not belongs to particular dataset. Then, calculate non-conformity scores for the sequence of 'k' samples nearest neighbours of same class and different class.

$$\propto_i := \frac{\sum_{j=1}^{k} d_{ij}^{+}}{\sum_{j=1}^{k} d_{ij}^{-}} \; i = 1, \ldots, n,$$

where d^{+}_{ij} is the shortest distance from x_i to another sample which is in the same class and d^{-}_{ij} is the shortest distance from x_i to another sample which is in a different class.

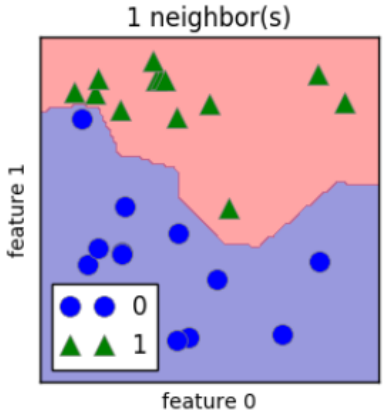

***Figure 2**:Non-conformity measure in nearest neighbour[fig from:6]*

2.1.3.2 Support Vector Machine

This section provides the basic idea of Support Vector Machine, which is an important algorithm in machine learning. In SVM, the initial data is transformed into high-dimensional space and tries to fit in the best line. It can solve both linear and non-linear problems. In linear, the hypothesis is defined as linear function on input space. The simplest way to define SVM is the algorithm creates a hyperplane which helps to separate the data into classes. The below picture shows the basic concept of SVM:

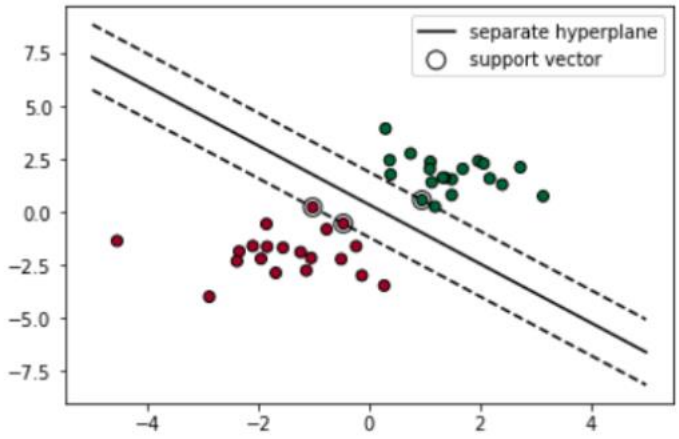

Figure3 :SVM [figure from 26]

The SVM takes the data as input and outputs a line that separates the classes, let's consider the above example we could find the two sets of classes and separate them by a hyperplane. The hyperplane has one less dimension than that of the data. There are also two parallel planes that passes one of the nearest samples. The mathematical representation of hyperplane is:

$$\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_n * x_n = 0$$

Then we can look on to margin, as per the above figure the distance between either side of the dashed line is the margin, the distance between two margins is called marginal distance. During the training section, it helps to find the hyperplane that maximises the margin between the closest samples and the hyperplane.

These hyperplanes are also performed as non-linear. Kernels are a versatile and effective tool for simply modifying linear algorithms to produce non-linear decision functions in the input space. SVM kernel is the technique that is used to convert not separable problem to separable problem which means that in the existing characters additional characters are added and create a new one. There are different types of kernel techniques such as linear, poly, RBF, sigmoid. The kernel

polynomial is a processor, where a polynomial feature is added to existing features and generates a new one.

RBF or Radial Basis Function kernel is the processor which calculates the distance between a specific dot with all other dots to generate a new feature. The common RBF kernel is Gaussian Radial Basis Function where:

Φ(x, center) = exp(-γ ||x-center||2) here new feature in decision boundary is controlled by gamma, if gamma is small influence of new feature on the decision boundary is small and if gamma is big then the influence of new feature on the decision boundary get wiggled.

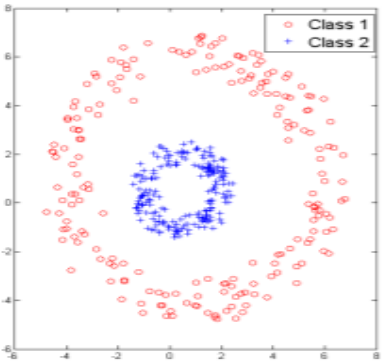

Figure 4:*Example for RBF kernel*

A polynomial kernel is a generalized form of the linear kernel and it can distinguish curved or non-linear input space.

SVM are commonly used to make predictions on new sample and it also produce Lagrange multipliers as the output. The Lagrange multipliers in the binary case roughly correlate to the distance to the separating hyperplane and the side of the hyperplane it lies on. As a result, a high multiplier indicates that the sample is far from the hyperplane and conforms to its class. As a result, you multiply the multiplier by -1 to make it a non-conformity metric. The Lagrange multipliers in the binary situation, where the labels are {-1,1} which is positive or negative and this depends on the side of the hyperplane. The sign of the Lagrange multiplier must also be multiplied on its label. The non-conformity measure is i.e.$\alpha i = -sign(yi)\beta i$ and β is Lagrange multipliers

The SVM can also perform in multi-class, it consists of more than hyperplane. By training a sequence of binary SVMs or solving a single optimization problem, SVM can be extended to handle such instances. The one-versus-one (OVO) approach and the one-versus-all (OVA) method are two common methods for training a series of binary SVMs. However, because they are relatively similar, we

will only examine the OVA technique. The OVA technique combines numerous binary SVM classifiers that have been trained separately into a single classifier. This technique has the benefit of being straightforward to implement on top of a binary SVM solver that already exists. It does, however, have the drawback that even if all binary classifiers are consistent, the resulting OVA classifier is not an all-in-one technique. In this scenario, multiplying by the class label is unnecessary because a Lagrange multiplier is only negative if it sits on the incorrect side of the one-vs-rest hyperplane, which simply indicates that it does not conform, not that it belongs to Class -1. Non-conformity measure is $\propto_i = -\beta_i$.

Datasets:

This section briefly describes the dataset used in this study. Here I, used two datasets USPS dataset and Wisconsin breast cancer dataset:

3.1 USPS Dataset:

The USPS dataset or US Postal Dataset is the series of normalized handwritten digits taken from the zip codes on mail. The digits were impulsively scanned from envelopes by the U.S Postal Service. At the time of scanning, the original digits have different orientations and different sizes they are then deslanted and size normalized. And each resulting sample is made up to 16 * 16 grayscale image of the digit. The brightness ranges between (-1, 1), while the label is a decimal digit from 0 to 9. The data available in two zipped files as training and test. Each line consists of the image of digit 0 to 9 and followed by the 256-grayscale values example of the data are shown in the below figure. The total number of data samples are 9298 and this data samples divided into two as training dataset and test dataset. The training dataset consists of 7291 samples and test dataset consists of 2007 samples. Here, no pre-processing steps are applied, we need to apply pre-processing steps before performing our experiments. The different distribution for training and testing dataset.

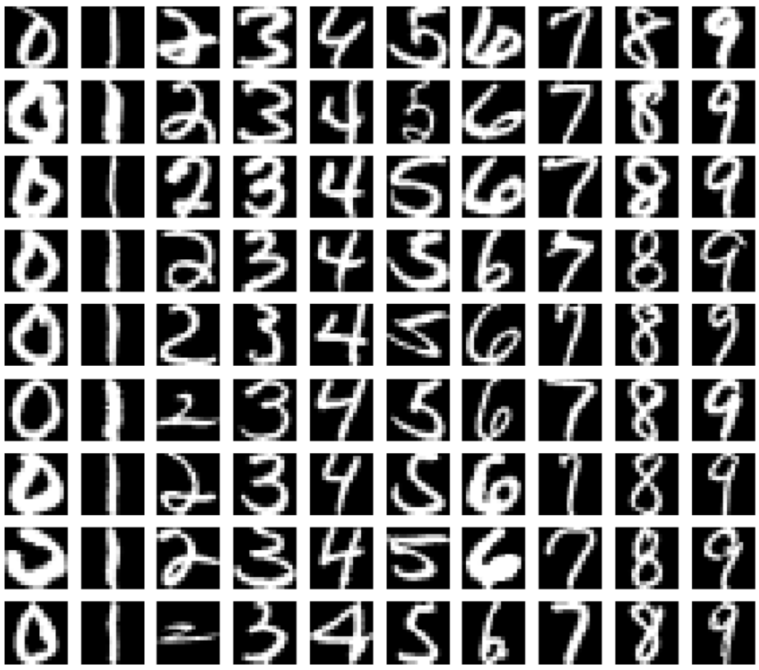

Figure 5:*Example of the data samples in the USPS handwritten digit dataset[Fig]*

3.2 Wisconsin breast cancer dataset:

Breast cancer is one of the most widespread cancer diagnosed around the world. According to a survey in the US, nearly 1 in 3 cancers diagnosed and which increases the death rate among women is breast cancer. Breast cancer is the abnormal growth of the cells in breast tissues generally known as Tumour. When the tumour is detected in the early stage can treat it effectively – this tumour can be benign(not cancerous), pre-malignant(pre-cancerous) and malignant(cancerous). The tests that are used to identify breast cancer are MRI, ultrasound, biopsy etc.

This dataset was created by Dr William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, USA. Dr Wolberg used fluid samples from patient's solid breast masses and also a graphical computer program called Xcyt to create the Wisconsin breast cancer dataset. Xcyt is the program that is able to analyse the cytological features based on the digital scan, which uses curve-fitting algorithms, that calculates mean value, extreme value and standard error of each feature from the image. The below image shows malignant tumour where the visible cell is outlined by the curve fitting algorithm:

Figure7 : *Image of malignant where visible cell nuclei is outlined by curve fitting algorithm* [Figure from [56]]

The main aim of the analysis is to predict malignant or benign cancer and we convert this to numerical representation as 1 for malignant and 0 for benign. The number of instances is 569 and the number of attributes is 30 and the attributes that are used to calculate the prediction are: mean radius, mean-texture, mean-perimeter, mean-area, mean smoothness, mean-compactness, mean

concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error, fractal dimension error, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points,

worst symmetry, worst fractal dimension. Here, no pre-processing steps are applied, we need to apply pre-processing steps before performing our experiments.

4 Results:

4.1 Conformal Prediction with USPS

4.1.1 Transductive Conformal Prediction

The USPS dataset is the handwritten database which is split into two training datasets and test dataset were training dataset consist of 7291 samples and test dataset consist of 2007 samples. As part of my study, I implement the transductive conformal prediction. The used non-conformity measure are nearest neighbour algorithm and support vector machine(SVM). The more difficulties I found in the transductive conformal prediction due to the huge executing time. The nearest neighbour algorithm took ~3632.5699 secs runtime to test 10 test samples and it has 2007 test sample and accuracy for 10 test sample is 0.9,average false p-value is 0.111, confidence results 0.999 and credibility is 0.477. In SVM, the runtime for 10 test sample is ~556.72secs and the accuracy is 0.9, false p-value is 0.282, confidence is 0.971 and credibility is 0.469. While compare both NCM, svm is faster and efficient than the nearest neighbours.

4.1.2 Inductive Conformal Prediction

Here, also used two non-conformity measures NN and SVM . Firstly, load the dat aset and split into 3 as mentioned earlier Training Proper, calibration and test data set. As per algorithm, calculate the non-conformity score and regular p-value. Base d on this, the results is obtained as the accuracy is ~0.944 , false p-value is ~0.013,cr edibility is ~ 0.478 and confidence measure as ~0.9942 and the total run time is ~61 23.51 seconds .And the results of smoothed p-value are accuracy is ~0.945, false p-value is ~0.011,credibility is ~ 0.477 and confidence measure as 0.9943 and the tot al run time is ~5514.935 seconds. From these results we could conclude that the more run time in regular p-value .And there is no difference in the validity cu rve of regular p-value and smoothed p-value. And it is important to look the v alidity of the algorithm. Validity curve is used to show the validity of conforma l prediction. Here, I used different significance level ranging from [0,1] and the graph is given below is the plot of significance level to error rate .When predict ion set does not contain true value of new sample, this condition causes the err or. If the predictor is diagonal then we can conclude that it is valid.Figure 8 rep resents NN validity curve.

Then we can look at the implementation of the support vector machine in the inductive conformal prediction. Here, load the USPS dataset and split the dataset into 3 as mentioned earlier as proper training dataset, calibration dataset and test dataset. And its size as :

Training Proper Size	:	7291
Calibration Set Size	:	1823
Test Set Size	:	2007

Then, by using scikit learning, implement the SVM algorithm and fit a proper training set and the resultant accuracy is 0.9207772795216741.

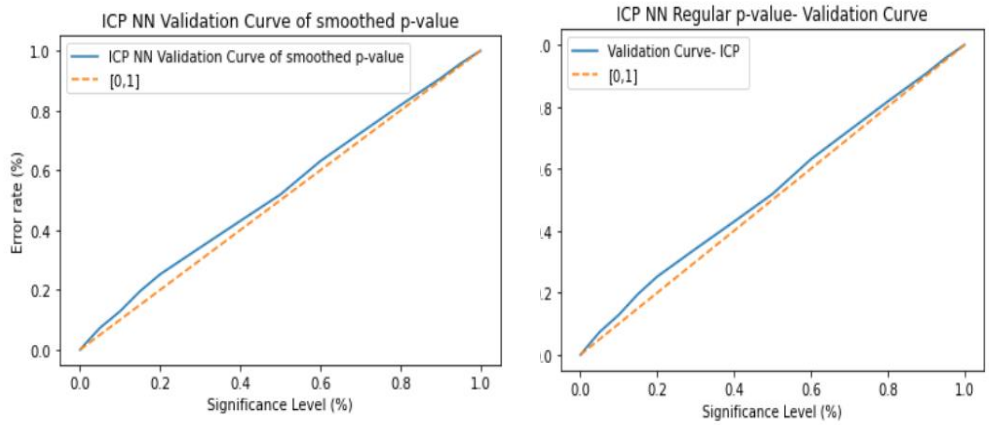

Figure 8: *The validity curve representation of inductive NN with regular p-value(left) and smoothed regular p-value(right)*

Then apply the SVM algorithm in the inductive conformal prediction algorithm in three different kernels in both p-value and smoothed p-value. From this study, When the kernel is "linear" with p-value and smoothed p-value: The total run time for smoothed p-value is more we could find the results as below:

kernel = linear	p-value	smoothed p-value
Total run time	3.6315sec	6.57149 sec
Accuracy	0.9357249626307922	0.9217737917289487
Average False p-value	0.03784301545468041	0.03659720843620257
Average Credibility	0.5221499969405328	0.5218783787203363
Average Confidence	0.9693733664629935	0.9696469953982276

From the results, smoothed p-value takes more time to get results and confidence is also more in smoothed p-value. The graphical representation given below is the validity of the conformal prediction using the validity curve and the curve is a plot of significance level against the error rate. The significance level ranges from 0 to 1. And here error occurs when the prediction set does not contain new sample's true value. Here, when the diagonal line along with square [0,1] then the given predictor is valid. It also shows the validity curve with smoothed p-value which results in the difference between regular p-value and smoothed p-value in the particular kernel. Figure 9 represents validity curve for linear kernel.

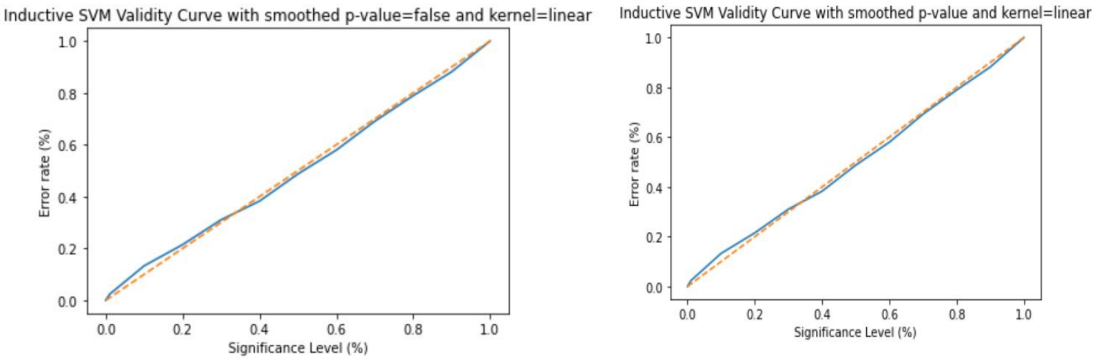

Figure 9:*The validity curve representation of inductive SVM, when kernel = linear with regular p-value(left) and when kernel = linear with smoothed regular p-value(right)*

When, kernel = rbf.
The accuracy of SVM when the kernel is "rbf" = 0.9471848530144494
When the kernel is "rbf" with p-value and smoothed p-value:
The total run time for smoothed p-value is more i.e, results we could see below:

kernel = rbf	p-value	smoothed p-value
Total run time	4.2445 sec	4.4986
Accuracy	0.9481813652217239	0.9481813652217239
Average False p-value	0.005561401323438179	0.004323825586524834
Average Credibility	0.4593590743800208	0.45908275661185144
Average Confidence	0.9977985493754317	0.9979476063191467

From the results, smoothed p-value takes more time to get results but compared to linear kernel RBF is little bit efficient and faster and confidence is also more in smoothed p-value. The significance level ranges from 0 to 1. And here error occurs when the new sample's true value is not in the prediction set. Here, when the diagonal line along with square [0,1] then the given predictor is valid. The graphical representation error rate versus significance level is given below: Figure 10 represents validity curve for linear kernel.

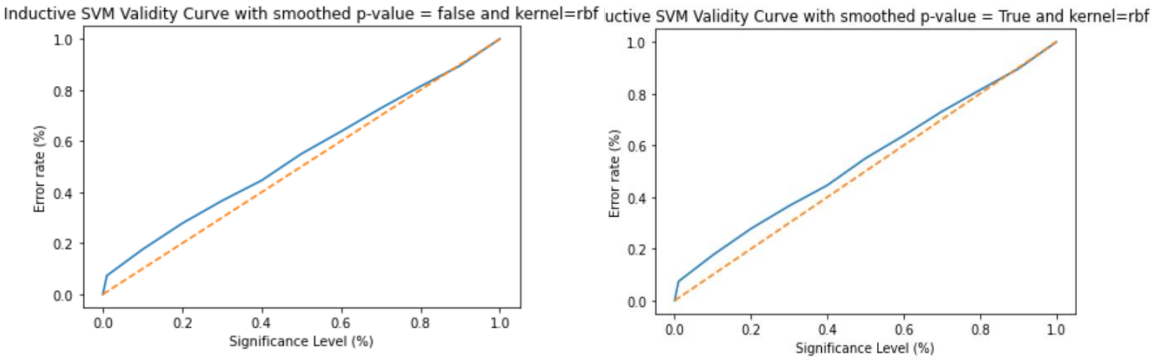

Figure 10 : *The validity curve representation of inductive SVM, when kernel = RBF with regular p-value(left) and when kernel = RBF with smoothed regular p-value(right)*

A look on the status of kernel = poly.
The accuracy of SVM when the kernel is "poly" =0.9347284504235177
When the kernel is "poly" with p-value and smoothed p-value:
The total run time for smoothed p-value is more, results are as seen below:

kernel = poly	p-value	smoothed p-value
Total run time	4.2445 sec	4.0279
Accuracy	0.9357249626307922	0.9362232187344295
Average False p-value	0.013920849395536726	0.012684709534856793
Average Credibility	0.5176848136784412	0.5174098651246708
Average Confidence	0.995693799770977	0.9959660250700785

From the results, smoothed p-value takes more time but there is only a small difference compared to other kernels. The graphical representation is Figure 11.

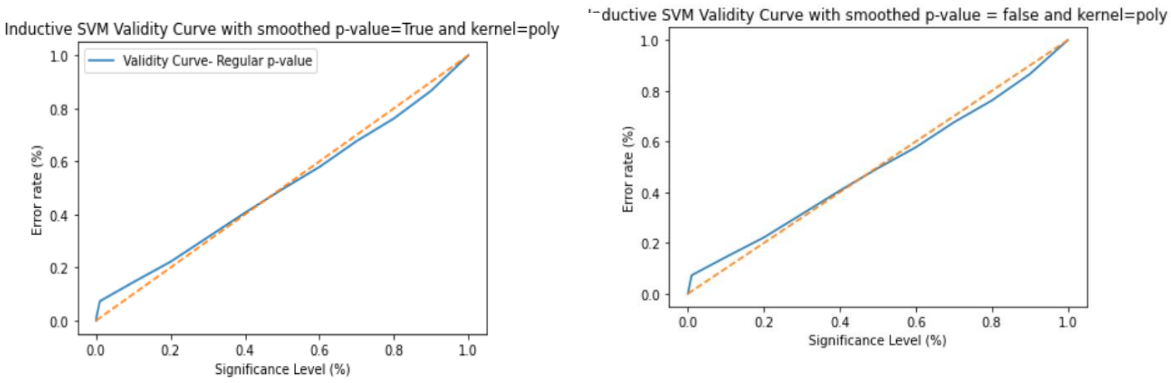

Figure 11: *The validity curve representation of inductive SVM, when kernel = poly with regular p-value(left) and when kernel = poly with smoothed regular p-value(right)*

From the above all results, we could find that the kernel "RBF" is more efficient, high validity as well as faster. And linear is the least one in all cases. While we compare p-value and smoothed p-value, the smoothed p-value has better performance in all kernels.

We could find that the validity curve of both, non-conformity measures are not exact diagonal. From this we can conclude that batch learning mode has weaker validity, but it also provides a smaller relation between significance level and error rate. While looking on to total run time, there is a great difference between the nearest neighbours and SVM, SVM takes much lesser than the NN. This shows efficiency is more in SVM algorithm. But in case of accuracy, all are similar which is ~0.94. There is difference in kernel as mentioned earlier. In case of smoothed and regular p-value we could find not much difference in the validity curve.

Then we can look at the online SVM conformal prediction: As we discussed earlier, the online method consumes significantly less time and computation and comparable accuracy to the batch process. At the same time, it has strong validity compared to batch. The sample is predicted one by one. Here, the implementation of the online conformal predictions in 3 different phases. Initially, the dataset split into 2; training and test dataset. Training data set consists of 20% of data. Let's see with method 1, here add the new samples into the calibration set and make predictions on it. And implement the cumulative multiple predictions, empty predictions and errors . Predictions with more than one label is multiple prediction, prediction set with no labels are empty predictions and prediction set which doesn't contain label of samples are errors. The graphical representation of multiple

predictions, empty prediction and errors with significance level 0.20 and 0.20represent in the Fig:12.

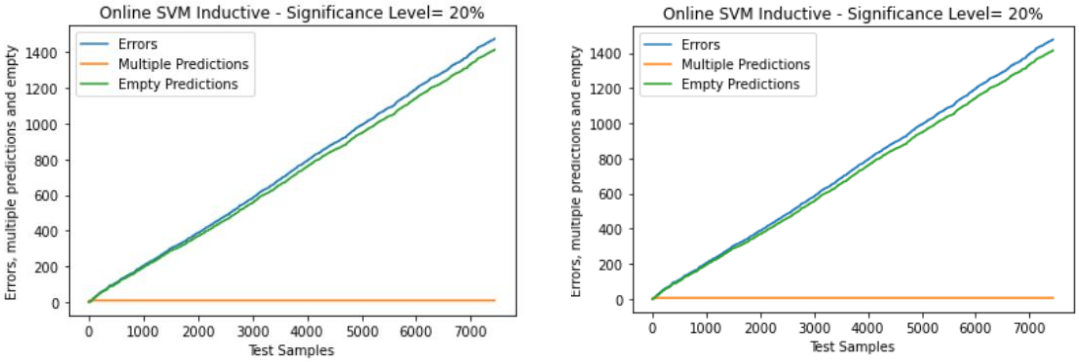

Figure 12: *The graphical representation of multiple predictions, empty prediction and errors with significance level 0.20 and regular p-value(left) and smoothed p-value(right)*

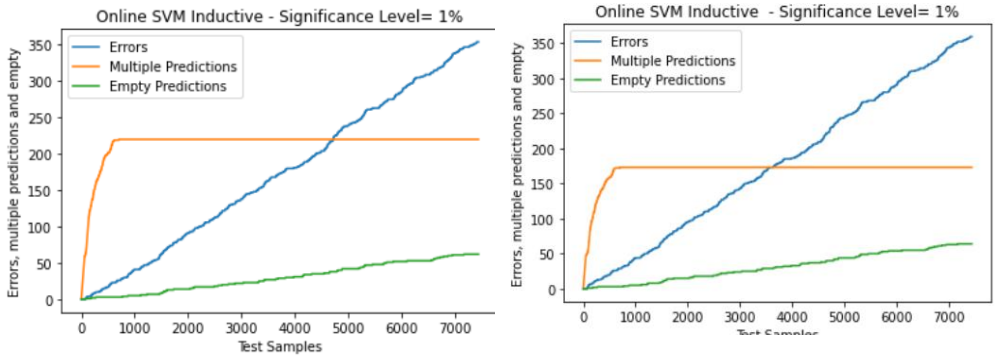

Figure13: *The graphical representation of multiple predictions, empty prediction and errors with significance level 0.05 and regular p-value(left) and smoothed p-value(right)*

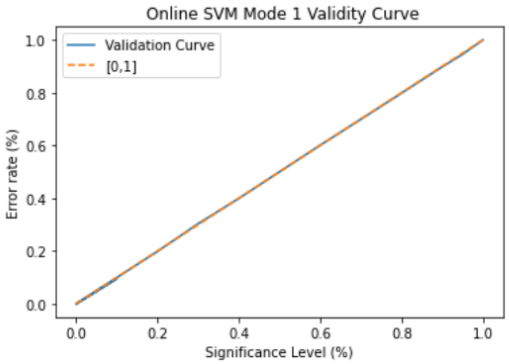

Figure14: *Validity cure of online inductive conformal prediction with svm*

There is no difference in the plotted graph [fig12 and fig13]from the regular p-value and smoothed p-value which results the values are same. When the significance value change, there is a great change in the results. When significance value is 0.05 t initially rapid increase in the number of multiple predictors then it is slower and stable at ~228. And this initial increase due to the lack of samples in calibration set. When the significance value is 0.20, the errors and empty predictions increases together , both are reaching ~1500 and number of multiple predictions barely grows at the beginning. As we discuss earlier, the validity curve of online method have stronger validity than batch process because validity curve is exactly same as the diagonal of the square [Fig:14].

The second phase is to add the calibration set of the training set at particular interval. Here, the particular interval is 1000 samples i.e.all the1000 samples were added to the training set, whenever the calibration set reached a size of 1000 samples.

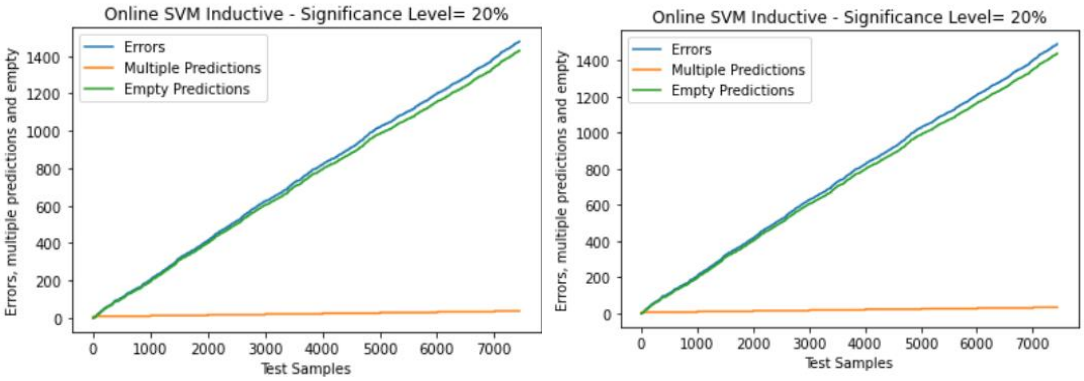

Figure 15 :*The graphical representation of phase 2 multiple predictions, empty prediction and errors with significance level 0.20 and regular p-value(left) and smoothed p-value(right)*

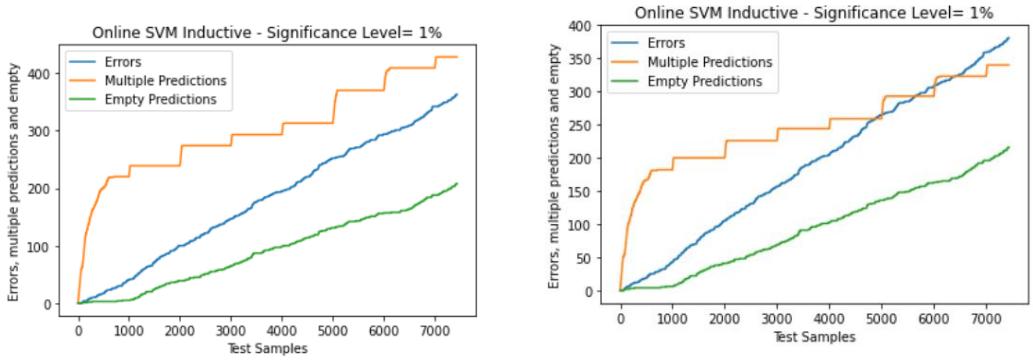

Figure 16: *The graphical representation of phase 2 multiple predictions, empty prediction and errors with significance level 0.05 and regular p-value(left) and smoothed p-value(right)*

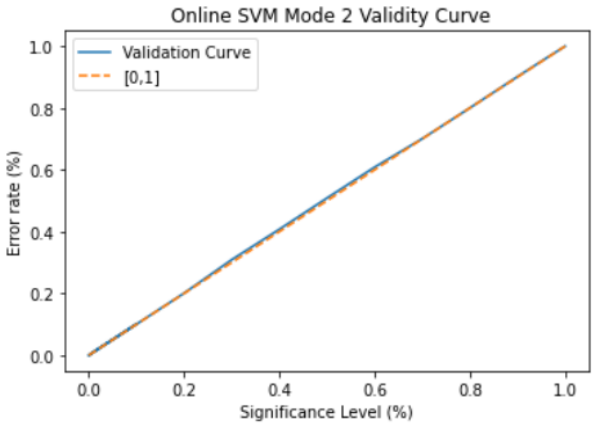

Figure 17:*Validity cure of phase 2 online inductive conformal prediction with svm*

The fig 15 and fig 16 are the phase 2 graphical representation. When significance level 0.20, there is no differences from phase 1 i.e. the errors and empty predictions increases together , both are reaching ~1500 and number of multiple predictions barely grows at the beginning. But the clearest difference between phase 1 and phase 2 with 0.05 significance level. Initially there is a rapid increase in the number of multiple prediction. Then there is no increase, at this time calibration set have been added to the training set with interval of 1000 samples. Then,we could see the increase in multiple prediction. And we get the graph with step like pattern. The final total is near to 350. To regain this we implement phase 3. Validity curve of phase 2 is shown in the fig 17 which is similar to phase 1.

The phase 3 is retaining the samples in the calibration set and move other samples to the training set. Here, we use interval as 1000 samples and 200 samples were retained in the calibration set i.e. once the calibration set reach 1000 samples, the first 800 samples are added to the training set and other 200 samples retained in the calibration set. This results, there is no less sample when the training set is updates. When look on the graphical representation the plot with 0.20 is same as phase 1 with small change in the number of multiple prediction and phase 2 but when the significance level is 0.05 there is difference with phase 2 and same as phase 1. i.e. in phase 3 it retain the cumulative set.

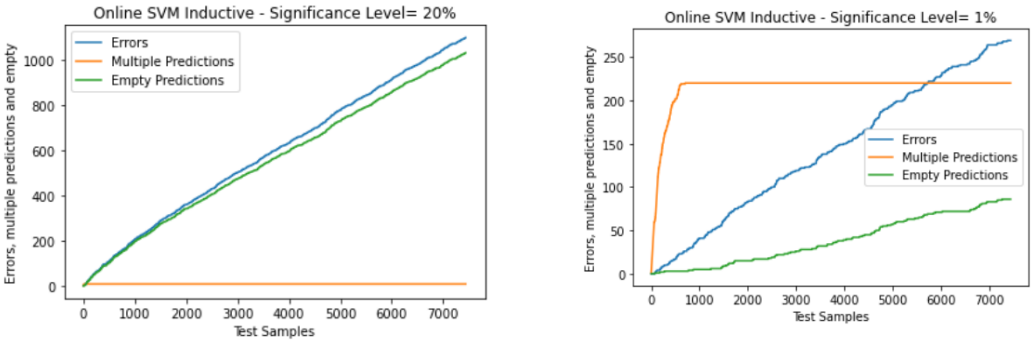

Figure 18: *The graphical representation of phase 3 cumulative total of multiple predictions, empty prediction and errors with significance level 0.20(left) and significance level 0.05(right)*

4.2 Conformal Prediction with Wisconsin Breast cancer dataset

As discuss earlier, breast cancer is an important cancerous disease. Let's look up the dataset, this dataset is used for the research purpose. As part of data mining application, we need to analyse and understand the data. The important dataset preparation are: data understanding, feature selection, eliminate the missing values and redundant records and data normalization.

In this dataset consist of 569 rows and attributes is 30 numeric attributes,predictive attribute and class. The attributes that are used to calculate the prediction are: mean radius, mean-texture, mean-perimeter, mean-area, mean smoothness, mean-compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error, fractal dimension error, worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension. These attributes are computed from the cell nucleus.

The experiment start with load data, then we manipulate into data frames for the easy manipulation. The attribute diagnosis is target and which results the tumour is either benign or malignant . Benign refers to a condition, tumour, or growth that is not cancerous and Malignant refers to cancer cells that can invade and kill nearby tissue and spread to other parts of your body[56]. And we convert this to numerical value for easy calculation. As part of pre-processing removed the missing and redundant values. Then focused to correlation. The statistical relationship between two variables are correlation[54]. For better analysis and prediction lets have a look

on the correlations. Fig 19 represents correlation between the variables. strong correlation between mean radius and mean perimeter, as well as mean area and mean perimeter. i

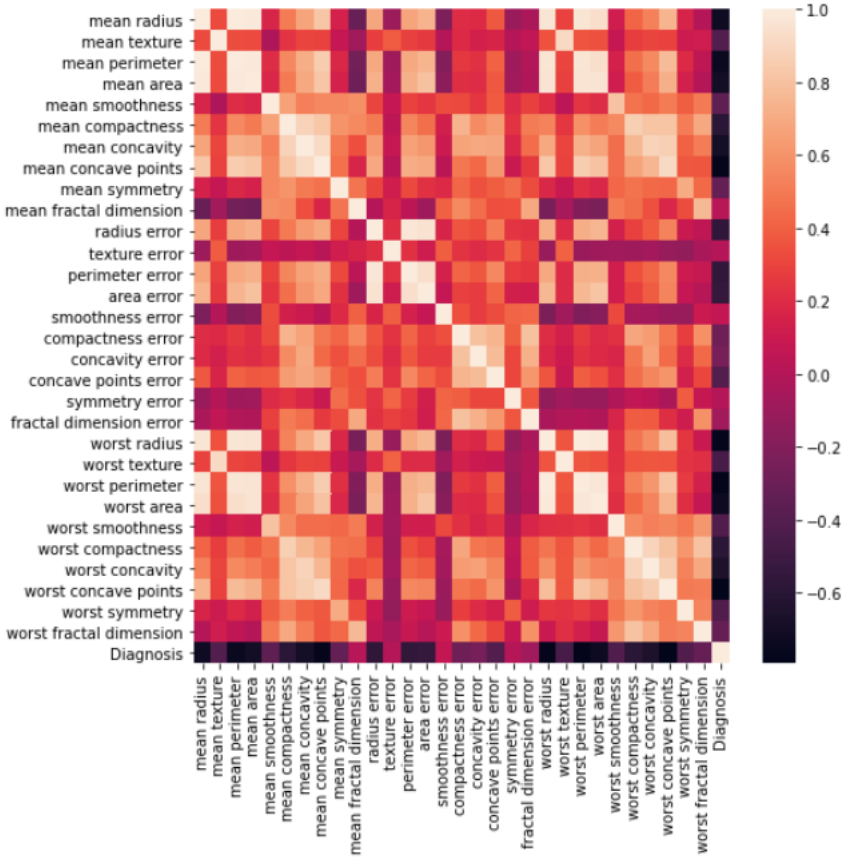

Figure 19: *correlation between the variables*

After that,I had done data normalization using standardscaler and normalizer.

4.2.1 Transductive Conformal Prediction

The data split into training set and test set for the transductive conformal prediction. The training data size is 455 and test data size is 114. While implementing nearest neighbour algorithm in conformal prediction, first create the nearest neighbour model and fit the value into model , then find the conformity score and p-value. Here, I didn't implement smoothed p-value because already identify that there is no much difference in regular and smoothed p-value. The total running time is ~2.379, the accuracy is ~0.939, the average false-p-value is 0.0256, the confidence results ~0.989 and the credibility is 0.573. Then the implementation of SVM using breast can dataset

4.2.2 Inductive Conformal Prediction

The inductive conformal prediction, is also following the same pre-processing and normalizing method as mentioned earlier. Then the data split into three as mentioned the algorithm as proper training dataset, calibration dataset and test dataset. The proper training dataset consist of 341 data, calibration dataset consist of 114 and test dataset consist of 114. Then implement the nearest neighbour and conformal prediction algorithm with regular p-value and the results are the total running time is ~0.467, the accuracy is ~0.885, the average false-p-value is 0.0197, the confidence results ~0.984 and the credibility is 0.437. Then implement the nearest neighbour and conformal prediction algorithm with smoothed p-value and the results are the total running time is ~0.506, the accuracy is ~0.877, the average false-p-value is 0.0193, the confidence results ~0.988 and the credibility is 0.433.

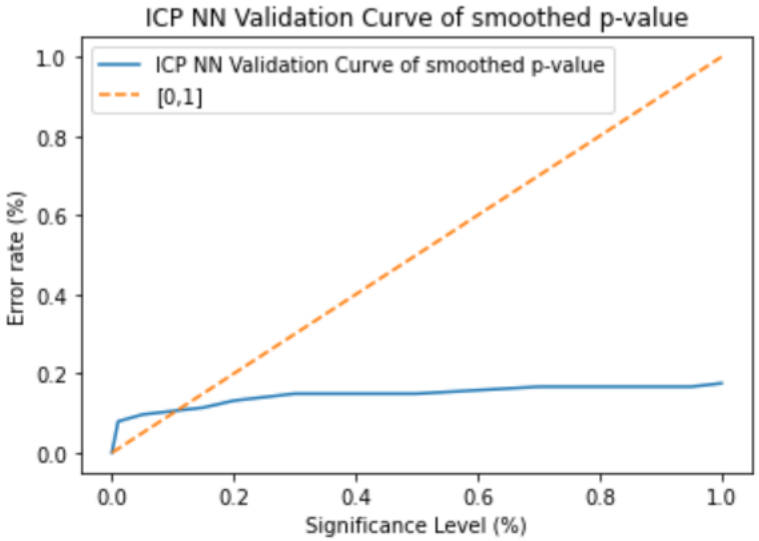

Figure 20: *The validity curve of inductive cp*

Then we can look to the online svm inductive conformal prediction on breast cancer dataset. : As we discussed earlier, the online method consumes significantly less time and computation and comparable accuracy to the batch process. At the same time, it has strong validity compared to batch. The sample is predicted one by one. Here, the implementation of the online conformal predictions in 3 different phases.

Initially, the dataset split into 2; training and test dataset. Training data set consists of 20% of data. Let's see with method 1, here add the new samples into the calibration set and make predictions on it. And implement the cumulative multiple predictions, empty predictions and errors . Predictions with more than one label is multiple prediction, prediction set with no labels are empty predictions and prediction set which doesn't contain label of samples are errors. The graphical representation of multiple predictions, empty prediction and errors with significance level 0.20 . The graphical representation fig:21 When the significance value is 0.20, the errors and empty predictions increases together , both are reaching ~1500 and number of multiple predictions barely grows at the beginning.

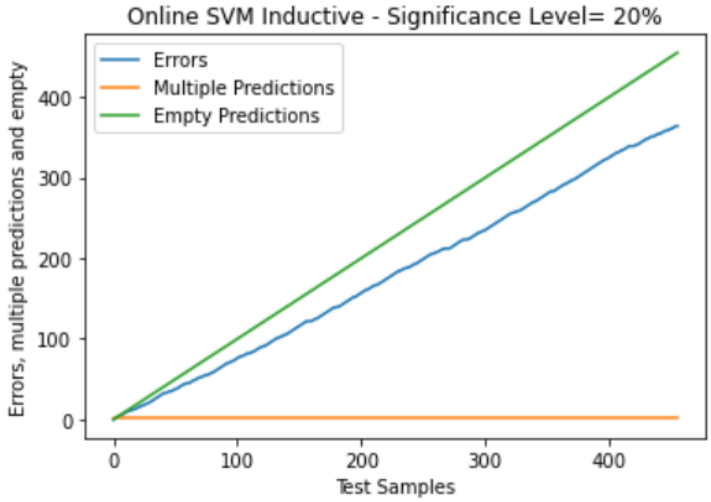

Figure 21: *The graphical representation of method 1 multiple predictions, empty prediction and errors with significance level 0.20 and regular p-value*

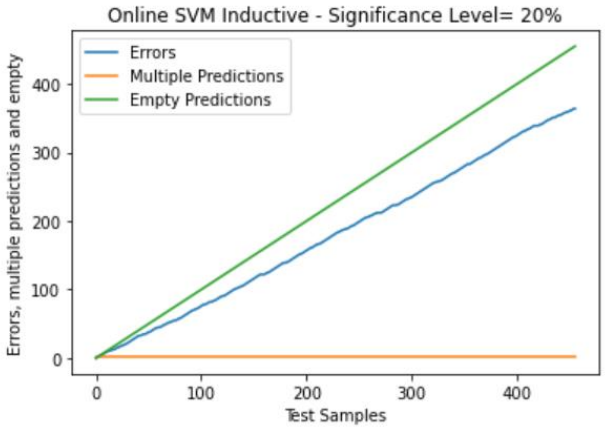

Figure 21: *The graphical representation of phase 2 multiple predictions, empty prediction and errors with significance level 0.20 and regular p-value*

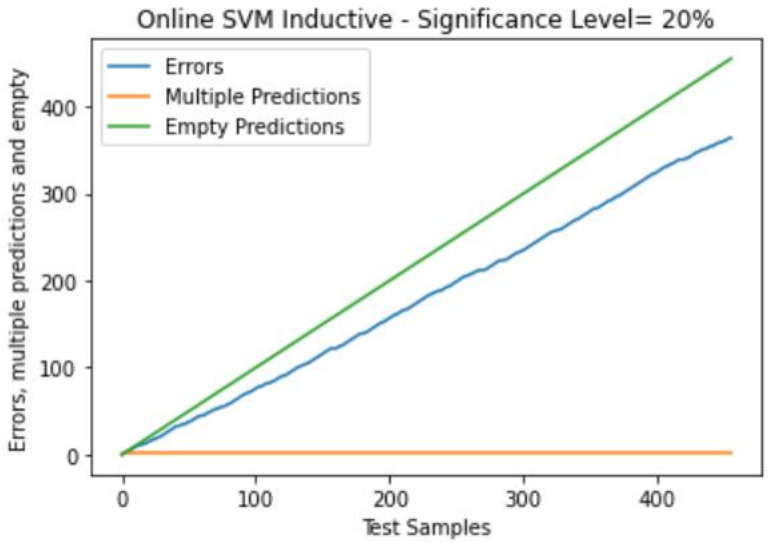

Figure 21: *The graphical representation of phase 3 multiple predictions, empty prediction and errors with significance level 0.20 and regular p-value*

Here, online we could same for all the significance leve =0.20. And the three phases are same method as we discuss in usps dataset.Conformal prediction provide confidence to usps dataset and Wisconsin dataset.

5. Discussion

This paper gives the brief introduction of conformal prediction including various methods, steps, formulas etc. In machine learning, we have datasets which can be split into training dataset and test dataset. The training dataset is to learn and understand the data. And we need to make the prediction of a new sample in the test data with confidence Conformal Prediction. Here, I implement transductive and inductive conformal prediction in both batch and online approach. From the various experience we can conclude that online learning method has high validity and efficiency compared to batch process. And, also, I come to conclusion as inductive conformal prediction is efficient method, we get the results within minutes, but in the case of transductive, it results ours, which is depends upon the size of data. From my experience, usps performs well with this method with high accuracy and confidence. Conformal prediction provide confidence to usps dataset and Wisconsin dataset.

6 Conclusions and Further Work

6.1 Conclusions

This paper has studied Conformal Prediction, which was introduced by Alex Gammerman, Glenn Shafer and Vladimir Vovk, through the paper *"Algorithmic Learning in a Random World"*, (Springer, 2005). Conformal Predictions is the statistical approach and a machine learning method for producing predictions with a guaranteed confidence level of new samples with the support of previous experience i.e. training dataset. This concept can be implemented in any machine learning algorithm including classical methods. Here, I implement transductive and inductive conformal prediction based on nearest neighbours and support vector machine. The TCP takes much more time to run the data than ICP but ICP is more efficient than TCP. I applied these theoretical concepts to different datasets such as the USPS handwritten dataset and the Wisconsin Breast Cancer dataset to obtain experimental results.

6.2 Further Work

There are different methods to extend the study. In this paper, I had just implemented the selected models and some techniques in certain datasets. After my study, I think that the study can be extended to different datasets. It could be interesting to see how changing the underlying algorithm to kNN or another non-conformity method affects the results. Other methods such as neural networks, gradient boosting, random forest etc can also be explored. Normalized non-conformity measure is another interesting area. Moreover, an important future aim is the cross-conformal predictions and applications to different fields especially in the medical field to get a better guarantee for predicting new samples.

7 Professional Issues

In this section, I plan to present the issue which I frequently uncover in my project. My project topic is Experiments in Conformal Prediction. I faced some professional issues during this time period. While performing TCP, which takes much loading time makes an issue .An important issue is connected to the quality of code. In the beginning stage, due to this issue, the code's run time was nearly 10 hours and sorted with help of the professors. I also faced the issue of choosing a dataset from a different source. Time limitation is an important factor to completing this project on time. Initially, we planned to implement conformal anomaly discovery, but due to this factor my project topic was changed to "Experiments in Conformal Prediction". I believe that this is due to difficulty in understanding the concepts, the less knowledge in practical side. But from this project I gained lot of skills and practical experience to implement

8 Self-Assessment

As per my point of view, the project was not going well due to lack of time. It took me month and a half to gain knowledge of the concept behind conformal predictions. Then I implemented it in python using Jupiter notebook. First of all, I faced difficulty in selecting the dataset. When I implemented the transductive conformal prediction in USPS which ran for more than 10 hours only with the help of professors, I was able to sort out the issue, later the concepts became more complicated and for a short period, it was a challenge to fully understand and implement them. However, by the end of this project, I believe that I have a much stronger understanding of various conformal predictions methods and their implementation. Here, I used 2 sets of datasets USPS and Wisconsin breast cancer datasets, which produced a range of results with the algorithms. Actually, we planned to implement a conformal prediction anomaly due to lack of time. The topic was changed into Experiments on Conformal Prediction. I believe the study might have been much better if we were able to meet our initial objectives, I'd like to try methods on some other datasets. Other underlying algorithms as non-conformity measures would also be interesting to test. Finally, this project was a new experience for me I thoroughly loved, and believe that it has improved my research, coding, and dissertation preparation skills. My objective is to work in any sector of machine learning, data science, deep learning, or natural language processing, and the skills I got will assist me in growing in it.

9 References

[1]. Alvarsson, J. *et al.* (2021) 'Predicting With Confidence: Using Conformal Prediction in Drug Discovery', *Journal of Pharmaceutical Sciences*. Elsevier, 110(1), pp. 42–49. doi: 10.1016/j.xphs.2020.09.055.

[2]. Balasubramanian, V. N. *et al.* (2013) 'PyCP: An Open-Source Conformal Predictions Toolkit', in Papadopoulos, H. et al. (eds) *Artificial Intelligence Applications and Innovations*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 361–370. doi: 10.1007/978-3-642-41142-7_37.

[3]. Balasubramanian, V., Vovk, V. and Ho, S.-S. (2014) *Conformal Prediction for Reliable Machine Learning - 1st Edition*. Available at: https://www.elsevier.com/books/conformal-prediction-for-reliable-machine-learning/balasubramanian/978-0-12-398537-8 (Accessed: 10 November 2021).

[4]. Burruss, M. P. (2020) *Detecting Weird Data: Conformal Anomaly Detection*, *Medium*. Available at: https://towardsdatascience.com/detecting-weird-data-conformal-anomaly-detection-20afb36c7bcd (Accessed: 3 December 2021).

[5]. Félix, E. (2020) *cbl-nonconformist* https://thedatascientist.com/anomaly-detection-why-you-need-it/. Available at: https://github.com/eloyfelix/cbl-nonconformist/blob/ad21c3f9d27780162d5f7503892a000cbe807711/README.ipynb (Accessed: 3 December 2021).

[6]. Fiori, L. (2020) 'Distance metrics and K-Nearest Neighbor (KNN)', *Medium*, 22 May. Available at: https://medium.com/@luigi.fiori.lf0303/distance-metrics-and-k-nearest-neighbor-knn-1b840969c0f4 (Accessed: 3 December 2021).

[7]. Gammerman, A. (2006) *Transductive Learning*. Available at: https://www.jstage.jst.go.jp/article/softscis/2006/0/2006_0_878/_pdf/-char/ja (Accessed: 3 December 2021).

[8]. Gammerman, A. and Vovk, V. (2007) 'Hedging Predictions in Machine Learning: The Second Computer Journal Lecture', *The Computer Journal*, 50(2), pp. 151–163. doi: 10.1093/comjnl/bxl065.

[9]. Harrison, O. (2018) *Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science*. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (Accessed: 3 December 2021).

[10]. Holloway, R. (2016) 'The efficiency of conformal predictors for anomaly detection', p. 90.

[11]. Johnson, J. (2020) *Anomaly Detection with Machine Learning: An Introduction*, *BMC Blogs*. Available at: https://www.bmc.com/blogs/machine-learning-anomaly-detection/ (Accessed: 3 December 2021).

[12]. Kampakis, D. S. (2020) 'What is anomaly detection, and why you need it.', *The Data Scientist*, 14 February. Available at: (Accessed: 3 December 2021).

[13].vKulkarni, A. (2021) *Atita05/Data-analysis-with-Python*. Available at: https://github.com/Atita05/Data-analysis-with-Python/blob/d21e6fd30ba5acb049b0b08e40754c7e37954737/NN.ipynb (Accessed: 3 December 2021).

[14]. 'Laxhammar - Anomaly Detection in Trajectory Data for Surveilla.pdf' (no date). Available at: http://www.diva-portal.org/smash/get/diva2:440646/FULLTEXT02 (Accessed: 3 December 2021).

[15]. Laxhammar, R. (2011) 'Anomaly Detection in Trajectory Data for Surveillance Applications', p. 156.

[16]. Laxhammar, R. and Falkman, G. (2015) 'Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories', *Annals of Mathematics and Artificial Intelligence*, 74(1), pp. 67–94. doi: 10.1007/s10472-013-9381-7.

[17]. Liu, F. T., Ting, K. M. and Zhou, Z.-H. (2008) 'Isolation Forest', in *2008 Eighth IEEE International Conference on Data Mining*. *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[18]. markpurtle (2020) *Conformal-Anomaly-Detection*. Available at: https://github.com/markpurtle/Conformal-Anomaly-Detection/blob/cdff79a1532934286da496792424e68e099b132d/README.md (Accessed: 3 December 2021).

[19]. 'Papadopoulos et al. - 2011 - Regression Conformal Prediction with Nearest Neigh.pdf' (no date). Available at: https://arxiv.org/ftp/arxiv/papers/1401/1401.3880.pdf (Accessed: 3 December 2021).

[20]. Papadopoulos, H. (2008) 'Inductive Conformal Prediction: Theory and Application to Neural Networks', in. doi: 10.5772/6078.

[21]. Papadopoulos, H., Vovk, V. and Gammerman, A. (2011) 'Regression Conformal Prediction with Nearest Neighbours', *Journal of Artificial Intelligence Research*, 40, pp. 815–840. doi: 10.1613/jair.3198.

[22]. *sklearn.neighbors.NearestNeighbors* (no date) *scikit-learn*. Available at: https://scikit-

learn/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html (Accessed: 3 December 2021).

[23]. 'Vovk et al. - 2005 - Algorithmic learning in a random world.pdf' (no date). Available at: https://download.e-bookshelf.de/download/0000/0004/78/L-G-0000000478-0002330810.pdf (Accessed: 3 December 2021).

[24]. Alvarsson, J. *et al.* (2021) *Predicting With Confidence: Using Conformal Prediction in Drug Discovery - ScienceDirect*. Available at: https://www.sciencedirect.com/science/article/pii/S002235492030589X (Accessed: 5 December 2021).

[25]. Borges, L. (2015) 'Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection', in.

[26]. Chen, L. (2019) *Support Vector Machine — Simply Explained*, *Medium*. Available at: https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496 (Accessed: 5 December 2021).

[27]. Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, 20(3), pp. 273–297. doi: 10.1007/BF00994018.

[28]. Khedkar, M. (2020) *Importance & Impact of Data Analytics in the Modern World. | Medium*. Available at: https://medium.com/@mahi.khedkar/importance-impact-of-data-analytics-in-the-modern-world-21c3f92c1955 (Accessed: 5 December 2021).

[29]. Kularathne, S. (2020) *Prediction and Data Visualization of Breast Cancer using K-Nearest Neighbor (KNN)Classifier Algorithm.| Analytics Vidhya | Medium*. Available at: https://medium.com/analytics-vidhya/prediction-and-data-visualization-of-breast-cancer-using-k-nearest-neighbor-knn-classifier-df7adadc4872 (Accessed: 5 December 2021).

[30]. Laxhammar, R. and Falkman, G. (2010) 'Conformal prediction for distribution-independent anomaly detection in streaming vessel data', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi: 10.1145/1833280.1833287.

[31]. Linusson et al. - 2014 - Efficiency Comparison of Unstable Transductive and.pdf' (no date). Available at: https://www.diva-portal.org/smash/get/diva2:888036/FULLTEXT01.pdf (Accessed: 5 December 2021).

[32]. Linusson, H. *et al.* (2014) 'Efficiency Comparison of Unstable Transductive and Inductive Conformal Classifiers', in Bayro-Corrochano, E. and Hancock, E. (eds) *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Cham: Springer International Publishing, pp. 261–270. doi: 10.1007/978-

3-662-44722-2_28.

[33]. Navlani, A. (2019) *Sklearn SVM (Support Vector Machines) with Python*, *DataCamp Community*. Available at: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python (Accessed: 5 December 2021).

[34]. Papadopoulos et al. - 2011 - Regression Conformal Prediction with Nearest Neigh.pdf' (no date). Available at: https://arxiv.org/ftp/arxiv/papers/1401/1401.3880.pdf (Accessed: 5 December 2021).

[35]. Papadopoulos, H. (2008) 'Inductive Conformal Prediction: Theory and application to neural networks.', in *Tools in Artificial Intelligence*, pp. 315–330.

[36]. Papadopoulos, H., Vovk, V. and Gammerman, A. (2007) 'Conformal Prediction with Neural Networks', *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*. doi: 10.1109/ICTAI.2007.47.

[37]. Papadopoulos, H., Vovk, V. and Gammerman, A. (2011) 'Regression Conformal Prediction with Nearest Neighbours', *Journal of Artificial Intelligence Research*, 40, pp. 815–840. doi: 10.1613/jair.3198.

[38]. Pupale, R. (2019) *Support Vector Machines(SVM) — An Overview*, *Medium*. Available at: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 (Accessed: 5 December 2021).

[39]. Sandeep, V. (2020) *A Guideline to Conformal Prediction*. Available at: https://medium.com/analytics-vidhya/a-guideline-to-conformal-prediction-7a392fc29bc1 (Accessed: 5 December 2021).

[40]. Shafer and Vovk - A Tutorial on Conformal Prediction.pdf' (no date). Available at: https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf (Accessed: 5 December 2021).

[41]. Shafer, G. and Vovk, V. (2008) 'A Tutorial on Conformal Prediction', p. 51.

[42]. Singh, M. (2021) *Breast Cancer Prediction*. Available at: https://github.com/mani24singh/Breast-Cancer-Prediction/blob/a5f4c2fc7603845b7f5d50435bbf3d827c60d97c/Machine%20Learning-%20Breast%20Cancer%20Prediction%20(Project).ipynb (Accessed: 5 December 2021).

[43]. Smith, J. (2016) 'The efficiency of conformal predictors for anomaly detection', p. 90.

[44]. Trinh, X. T. (2012) *Online learning of multi-class Support Vector Machines*. Available at: http://uu.diva-portal.org/smash/get/diva2:570643/FULLTEXT02.pdf (Accessed: 5 December 2021).

[45]. US Post Office Zip Code Data (no date). Available at: https://hastie.su.domains/StatLearnSparsity_files/DATA/zipcode.html (Accessed: 5 December 2021).

[46]. Vajda, S. and Santosh, K. (2017) 'A Fast k-Nearest Neighbor Classifier Using Unsupervised Clustering', in, pp. 185–193. doi: 10.1007/978-981-10-4859-3_17.

[47]. Vovk, V., Gammerman, A. and Shafer, G., 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

[48]. Vovk, V. (2012) 'Conditional Validity of Inductive Conformal Predictors', in *Proceedings of the Asian Conference on Machine Learning*. *Asian Conference on Machine Learning*, PMLR, pp. 475–490. Available at: https://proceedings.mlr.press/v25/vovk12.html (Accessed: 5 December 2021).

[49].Vovk, V. *et al.* (2017) 'Criteria of efficiency for set-valued classification', *Annals of Mathematics and Artificial Intelligence*, 81(1–2), pp. 21–46. doi: 10.1007/s10472-017-9540-3.

[50]. Wingate, J. (2019) *Breast Cancer Prediction using k-Nearest Neighbors Algorithm in Python, R-ALGO Engineering Big Data*. Available at: https://www.engineeringbigdata.com/breast-cancer-prediction-using-k-nearest-neighbors-algorithm-in-python/ (Accessed: 1December 2021).

[51]. Yael, K. (2017) *Wisconsin Breast Cancer (Diagnostic) DataSet Analysis*. Available at: https://rstudio-pubs-static.s3.amazonaws.com/344010_1f4d6691092d4544bfbddb092e7223d2.html (Accessed: 5 December 2021).

[52]. Zeni et al. - Conformal Prediction a Unified Review of Theory a.pdf' (no date). Available at: https://www.mate.polimi.it/biblioteca/add/qmox/22-2020.pdf (Accessed: 5 December 2021).

[53]. Zeni, G., Fontana, M. and Vantini, S. (2020a) 'Conformal Prediction: a Unified Review of Theory and New Challenges', *arXiv:2005.07972 [cs, econ, stat]*. Available at: http://arxiv.org/abs/2005.07972 (Accessed: 5 December 2021).

[54]. Zeni, G., Fontana, M. and Vantini, S. (2020b) 'Conformal Prediction: a Unified Review of Theory and New Challenges', p. 36.

[55]. D., U. and Ramachandra, B. (2016) 'Big Data Analytics to Predict Breast

Cancer Recurrence on SEER Dataset using MapReduce Approach', *International Journal of Computer Applications*, 150(7), pp. 7–11. doi: 10.5120/ijca2016911549.

[56]. Islam, Md. M. *et al.* (2020) 'Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques', *SN Computer Science*, 1(5), p. 290. doi: 10.1007/s42979-020-00305-w.

[57]. *Prediction of Breast Cancer Using SVM Algorithm* (2021). Available at: https://www.ripublication.com/ijaer21/ijaerv16n4_11.pdf (Accessed: 5 December 2021).

10. Notations

NN:	Nearest Neighbours
SVM:	Support Vector Machines
USPS:	United States Postal Service
Bc :	Breast Cancer Dataset
Iid:	Independent and Identically Distributed
NCM:	Nonconformity Measures
NCS:	Nonconformity Score
Knn :	K-Nearest Neighbours
ICP	Inductive Conformal Prediction
TCP	Transductive Conformal Prediction

11. How to use my project

My project is done in the python code and developed in the Jupyter notebook.

Step 1:

Download and install the jupyter notebook.

Step 2:

Start the Jupyter notebook, and we could find the root directory open at browser

Step 3:

Download the zip file of my project.

Step 4:

Unzip the file to the root folder

Step 5:

Below are the program files run the files using: The dataset are USPS and Wisconsin breast cancer dataset and

usps_nn_transd_prediction.ipynb --- Transductive conformal Prediction in USPS

usps_svm_tcp.ipynb --- Transductive conformal Prediction in USPS using SVM

usps_nn_inductive.ipynb –Inductive conformal Prediction in USPS using NN

usps_svm_ind.ipynb ---- Inductive conformal Prediction in USPS using SVM

usps_online.ipynb ---Online Inductive conformal Prediction in USPS using SVM

bc_svm_transductive.ipynb--- Transductive conformal Prediction in bc using SVM

bc_nn_tcp.ipynb --- Transductive conformal Prediction in bc dataset

bc_nn_ind.ipynb –Inductive conformal Prediction in bc using NN

bc_online.ipynb -----Online Inductive conformal Prediction in bc using SVM

12. Appendix A : My Code

The code is developed and presented in a series of Jupyter notebooks. The detail which notebooks correspond to the dataset files they used. To get accurate results the dataset and Jupyter notebook should be in the same directory.

usps_nn_transd_prediction.ipynb --- Transductive conformal Prediction in USPS

usps_svm_tcp.ipynb --- Transductive conformal Prediction in USPS using SVM

usps_nn_inductive.ipynb –Inductive conformal Prediction in USPS using NN

usps_svm_ind.ipynb ---- Inductive conformal Prediction in USPS using SVM

usps_online.ipynb ---Online Inductive conformal Prediction in USPS using SVM

bc_svm_transductive ipynb--- Transductive conformal Prediction in bc using SVM

bc_nn_tcp.ipynb --- Transductive conformal Prediction in bc dataset

bc_nn_ind.ipynb –Inductive conformal Prediction in bc using NN

bc_online.ipynb -----Online Inductive conformal Prediction in bc using SVM