NCC-EM: A HYBRID FRAMEWORK FOR DECISION MAKING WITH MISSING INFORMATION

A THESIS IN
Computer Science

Presented to the Faculty of the University
Of Missouri-Kansas City in partial fulfillment
Of the requirements for the degree

MASTER OF SCIENCE

By
Varun Chavakula

B. Tech, SRM University,
Chennai, Tamil Nadu, India, 603203

Kansas City, Missouri
2017

NCC-EM: A HYBRID FRAMEWORK FOR DECISION MAKING WITH MISSING INFORMATION

Varun Chavakula, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2017

ABSTRACT

Accounting for uncertainty is important in any data driven decision making. The popular treatment of uncertainties is to employ classical probability theory by expressing variables as random variables or processes in terms of random distributions. This precise approach encounters difficulty and leads to deceptive predictions when the sources of uncertainty are epistemic in terms of incomplete (missing), conflicting, or erroneous information due to the lack of knowledge. There have been many frameworks developed against the precise probability formalism, and one of such frameworks is the Imprecise Probability (IP) based modeling.

In this thesis, we develop and provide a novel hybrid framework, Naïve Credal Classifier with Expectation-Maximization data imputation, for decision making with missing information. The IP-based Credal Set concept is first introduced to model uncertainties for data with missing information. Then the Naïve Credal Classifier (NCC) is employed in this work, which is provided by the latest JNCC2 package. The key idea and research findings in this research is to model missing data using advanced imputation techniques to minimize the performance (accuracy) loss in NCC. The resulting NCC-EM framework is hybrid where the EM imputation technique is used as a preprocessing step. To verify and validate this hybrid framework, the NCC-EM is extensively tested on open machine learning datasets by simulating missing values, and it is shown that NCC-EM outperforms the existing NCC framework and traditional supervised classification methods.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering, have examined a thesis titled "NCC-EM: A HYBRID FRAMEWORK FOR DECISION MAKING WITH MISSING INFORMATION" presented by Varun Chavakula, candidate for the Master of Science degree, and hereby certify that in their opinion, it is worthy of acceptance.

Supervisory Committee

Chen ZhiQiang, Ph.D., Committee Chair
Department of Civil and Mechanical Engineering

Yugyung Lee, Ph.D.,
Department of Computer Science Electrical Engineering

Rao Praveen, Ph.D.,
Department of Computer Science Electrical Engineering

TABLE OF CONTENTS

Chapter Page

ILLUSTRATIONS

Figure Page

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Chen ZhiQiang for all the innovative ideas, insights, advice and challenging deadlines that have helped me achieve this thesis. He has been a constant source of motivation and zeal during my thesis. He was always welcoming for all the help I needed throughout my work, it has always amazed me for the kind of support and inspiring suggestions he has given me for the development of this thesis.

Secondly, I would like to thank the University of Missouri-Kansas City, without which this research would not be possible. The school provided me with good opportunities to support myself and a Lab for my research.

Finally, I would like to express my heartful gratitude to my family and friends for providing me with constant support and encouragement. This accomplishment would not be possible without them.

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

Representing uncertainty has always been a subject of debate. Although there have been successful applications of the precise probability, it's inherently high precision can often lead to inaccurate predictions. There have been many frameworks developed against the uncertainty representations of the precise probability and one such significant framework is the Imprecise Probability(IP). The goal of this research is to enhance decision making on small, incomplete datasets where accounting for uncertainty becomes much more complex. This research work presents one such framework that can help in making objective decision with missing information.

1.2 Proposed Solution

In this thesis, we provide with a java implementation of a novel hybrid framework called NCC-EM. This framework is developed in Java extending the existing JNCC2 software. NCC-EM is developed using a hybrid approach where the Expectation-Maximization(EM) imputation technique is used as a preprocessing step. After the imputation of the missing values, the data is then analyzed using the Naïve Credal Classifier which is an extension of Naïve Bayes Classifier to Imprecise Probability. NCC-EM has been rigorously tested on different datasets containing missing values and shows to outperform NCC. The key idea presented in this research is to model missing data using advanced imputation techniques to minimize the performance(accuracy) loss in NCC. This work is against a single framework approach by testing and proving the potential of mixed framework approaches like NCC-EM as a holistic way to treat uncertainty that occurs due to missing data.

1.3 Summary

The rest of the thesis is organized as follows. Chapter 2 provides a background and literature review of the Imprecise Probability framework. In Chapter 3, we review the Naive Bayes Classifier and its extension towards Imprecise Probability Theory. In Chapter 4, we present NCC-EM, a hybrid framework for decision making with missing information. This chapter presents a section of evaluation of NCC-EM. Finally, Chapter 5 concludes this dissertation.

CHAPTER 2

INTRODUCTION TO IMPRECISE PROBABILITY

This chapter introduces the concepts of the Imprecise Probability (IP) Theory and reviews its applications in the field of machine learning. It also focuses on exploring the IP approach for uncertainty and risk modelling.

2.1 Imprecise Probability

According to the studies by Miranda & de Cooman (2014), Imprecise Probability is considered as a subjective probabilistic approach that deals with the behavioral interpretation of a subject (the person involved in betting). The behavioral interpretation of a subject on any event represents the subject's willingness to take action whose outcomes depends on the occurrence of the event, such as accepting bets on or against the event at a certain betting rate. In 1975, Williams made a first attempt to study imprecise subjective probability based on de Finetti's (1937) work which was addressed in more details by Walley (1991) by developing an arguably more mature behavioral theory of coherent lower previsions (Miranda & de Cooman, 2014). The IP theory is widely applied in the field of Economics and Finance where decision making involves taking high risks. The work of Shafer and Vovk (2005) show the applications of IP theory towards finance thereby extending Shafer's most significant work i.e. Dempster-Shafer Theory (DST).

2.2 Dempster-Shafer Theory

In an article by Shafer (1992), he explains DST as the theory of belief functions and a generalization of Bayesian theory towards subjective probability. As mentioned in this article, the Dempster-Shafer theory owes its name to the work of A.P Dempster (1968) and Glenn Shafer (1976). It introduces DST as a framework consisting of two main steps, first step is obtaining belief function of a question by analyzing a similar question. The second step includes applying Dempster's rule which combines such belief function when they are independent evidences. This has been illustrated by the following example. Suppose the subject has the subjective probability of his friend Betty. The probability that she is reliable is 0.9 and the probability that she is unreliable is 0.1. Suppose she makes a statement that the limb fell on the subject's car. Her statement justifies a 0.9 degree of belief that the limb fell on the subject's car but only a zero degree of belief (not 0.1 degree of belief) that no limb fell on his car. The zero does not mean that the subject isn't sure about the limb falling on the car. It only means that Betty's testimonial gives the subject no reason to believe that the limb fell on his car. The 0.9 and the zero together form the belief function. To illustrate Dempster's rule for combining the degrees of belief, another piece of evidence is taken into consideration. Suppose the subject has another friend Sally who also has the subjective probabilities for reliability as 0.9 and unreliability as 0.1 and she also testifies independently of betty that the limb fell on the subject's car. The event that Betty is reliable is independent of the event of Sally being reliable and we may multiply the probabilities of these events. The probability that both are reliable is 0.9 × 0.9 = 0.81, the probability that neither of them are reliable is 0.1 × 0.1 = 0.01, probability that at least one is reliable is 1 – 0.01 = 0.99. Since they both said that the limb fell on the car, at least one of the being reliable means that the limb fell on the car and we can assign the degree of belief as 0.99 for this event. Suppose they both contradict and Sally say's limb no limb fell on the car. In this case, since both cannot be right we have only three cases: only Sally is reliable, only Betty is reliable and neither of them are reliable. The prior probability for the cases only Sally is reliable and only Betty is reliable is 0.9 × 0.1= 0.09 as it is the product of one being reliable which is 0.9 and other being unreliable which is 0.1. The probability for the case neither being reliable is 0.1 × 0.1 = 0.01. Therefore, the prior for the above three cases are 0.09, 0.09 and 0.01 respectively. The posteriors probabilities for these cases given that both cannot be reliable is as follows. Only Sally being reliable: 0.09 / (0.09 + 0.09 + 0.01) = 9 / 19, Similarly for only Betty being reliable is 9 / 19 and neither being reliable 1 / 19. Thus, the subject assigns a 9 / 19 degree of belief for the limb falling on his car because Betty is reliable and 9 / 19 degree of belief for the limb not falling on his car because Sally is reliable. (Shafer G. , 1992)

As illustrated above implementing DST on a problem generally involves solving two related problems for example in the above case we answer two questions, did the limb fall on the car? Is the witness reliable? (Shafer G. , 1992). Dempster's rule begins with an assumption that the questions to which we have probabilities are independent items. This framework enables to develop probabilities for answering question whose answers are unknown based on the question whose answers are known through belief functions. The kind of reasoning behind DST has its roots from the 17th Century however it came to attention in the 1980's when the AI researchers were trying to use the probability theory into building expert systems (Shafer G. , 1976).

The DST is a general framework for reasoning with uncertainties with connections to other frameworks like IP theory. A unification of the IP theories was proposed by Walley (1991) in an attempt to formalize imprecise probabilities. Walley's theory extends the subjective probabilities through a gambling interpretation via buying and selling prices.

2.3 Uncertainty Modelling

Representing uncertainty has been a subject of heated debate and there is only one arguably appropriate model for numeric uncertainty and that is probability however it may not always be the best way to represent uncertainty (Halpern, 2017). From the examples in Halpern's (2017) work, if Alice has two coins, one is fair and the other is biased and has the probability of 2/3 for heads. Suppose Bob chooses the coin and Alice will flip it and Bob knows which coin is fair and which is biased. He also gets 1$ for heads and loses 1$ for tails. Clearly, in this case he would choose the coin that is biased as it has the probability of 2/3. In this case reasoning uncertainty through probability works however if the bias of the coin is unknown, what should Bob choose then? If he to model the uncertainty of this unknown biased coin through probability the best value to give would be 1/2. Is this the best approach? Is using a single probability measure appropriate here? (Halpern, 2017) IP theory introduces the betting interpretation framework to model uncertainty in such cases that is missed by the probability framework. There are two kinds of uncertainty: Aleatory and Epistemic and the need of this categorizations is crucial in refining the models and addressing the uncertainty as a whole (Ditlevsen & Kiureghian, 2009).

Referring to Troffaes's (2016) notes at the SIPTA workshop. An event in probability theory is defined as a set of outcomes for an experiment to which probability is assigned. Suppose the probability of an event that it will snow tomorrow is 0.4. In probability theory through frequency interpretation, it means that on a day like this (today) it generally snows 4 out of 10 times the next day. In probability theory, there is always a need of the reference class to determine the probability of the event where 4 out of 10 times determines the frequency of the event based on the observation of the reference class. Such a frequency interpretation is possible only for repeatable events and to model the uncertainty in it random variables are used. This kind of uncertainty is called the aleatory uncertainty whose modelling needs plenty of data as it is dependent on frequency.

The events for which there is no reference class (also called as one-shot events by the author) and leads to an uncertainty due lack of knowledge about the event is called the Epistemic uncertainty. The epistemic uncertainties are modelled using belief functions that was introduced under DST. Such probabilistic approaches where the belief is modelled is called as the subjective probabilistic approach and IP theory is a part of the subjective probabilistic approach with a behavioral interpretation. These approaches need a lot of elicitation and rules to maintain the coherence in the logic of the belief functions. Though the DST framework is more general and subsumes the IP theory framework, it is important to have IP theory as it formalizes the belief functions though the use betting interpretation (analogous to behavioral interpretation) and the rules of coherence over the betting rates.

In the betting interpretation, the events are interpreted as gambles and the probability values (analogous to the betting rates) given to these events are based on belief. For example, I would utmost pay 0.1$ now if I am paid 1$ if it snows tomorrow or I would pay 1$ tomorrow if I am paid at least 0.4$ now (Troffaes, 2016). These betting rates bound our confidence. Based on work of (Williams, 2007), the basic concepts of uncertainties representation using IP theory are introduced in the following sections, and simple examples are designed to facilitate the understanding.

2.4 Lower and Upper Previsions

As explained in the work of Miranda & de Cooman(2014), behavioral theory of imprecise probabilities provides a mathematical framework for representing the subject's belief in terms of accepting transactions(gambles) whose outcome depend on that of the experiment. The reward of the gamble f whose value is f(x) represents an uncertain value, that depends on the outcome of the experiment which is x. The theory of lower previsions accepts two types of transaction involving the gamble f: accepting to buy f for a price μ, which amounts to accepting the gamble f-μ; and accepting to sell f for a price λ, which amounts to accepting the gamble λ-f. (Miranda & de Cooman, 2014)

Lower prevision $\underline{P}(f)$ for a gamble is the supremum acceptable buying price that the subject is willing to pay to purchase the gamble. In other words, the subject will purchase the gamble for all the prices lower than μ. It is mathematically given as follows.

$$\underline{P}(f) \coloneqq \sup\{\mu \in \mathbb{R} : f - \mu \in D\}$$

Here D is the set of the desirable gambles. The logic of decision making in the gambling (to the subject who buys) is to make maximum net reward through paying no more than his maximum buying price (i.e. the supremum) given the reward, such that the gamble is reasonable (to the buyer). In other words, he is willing to pay $\underline{P}(f)$ - ϵ for the uncertain reward f or the transaction f - $\underline{P}(f)$ + ϵ is desirable to him for every ϵ > 0.

The Upper prevision $\overline{P}(f)$ on the other hand is defined as the infimum selling price for a gamble; the subject is willing to sell the gamble for all the prices higher than λ. It is mathematically given as follows.

$$\overline{P}(f) \coloneqq \inf\{\lambda \in \mathbb{R} : \lambda - f \in D\}$$

where f again is the uncertain reward of the gamble, λ is the selling price of the gamble, and λ-f belongs to the desirable gambles. The logic of decision making in the gambling (to the subject who sells) is to make is to make maximum net reward through selling at the gamble at no less than greatest minimum selling price (i.e. the infimum), such that the gamble is reasonable (to the seller). This implies that the subject accepts to get a price $\overline{P}(f)$ + ϵ for selling f or in other words the transaction $\overline{P}(f)$ + ϵ - f is desirable for every $\epsilon > 0$.

Let us now understand the lower and upper prevision with a simple decision-making example where the subject needs to buy the gamble f which is as follows.

f(rain) = 10$, f(cloudy) = 5$, f(sunny) = 0$

where the set of possible outcomes are {rain, cloudy, sunny}, and the reward of the gamble (f) is that to bet on 'rain', if correct, one wins $10; bet on cloudy, if correct, wins $5; and bet on sunny, if correct, win $0.

If he should pay a certain fixed price x to buy the gamble to obtain this uncertain reward. In that case his increase in wealth would be (10-x) $ if it rains, (5-x) if it is cloudy and (-x) $ i.e. he loses (x) $ if it is sunny. The maximum amount of price x he is willing to pay to buy this gamble is called the lower prevision for the gamble. If the subject is more confident that it will rain tomorrow he would tend to pay more and if he feels less confident he might pay less so that he doesn't encounter any losses.

On the other hand, if he must sell the gamble f to others for a fixed price y, his increase in wealth would be (y-10) $ if it rains, (y-5) $ if it is cloudy and (y) $ if it is sunny. The minimum price y, for which the subject is willing to sell his gamble is called the upper prevision. If the subject knows for sure that it would rain or be cloudy tomorrow he would be cautious and not sell it for less than 10 $ however if he feels it could be sunny or at the most cloudy he might drop down his price to 5$.

As we try to model the confidence of the subject through the gambles, it is important to have some guidelines and rules while deciding the price of the gamble to maintain the consistency in the betting behavior. These rules are essential in building up a model which we shall discuss under the theory of coherent lower prevision. Another important point to note is, most of these theories are discussed in terms of lower prevision for simplicity as we can justify that selling a gamble f for a price μ is the same thing as buying the gamble -f for the price -μ. This can be mathematically written by the equation $\overline{P}(f) = -\underline{P}(-f)$ (Miranda & de Cooman, 2014).

2.5 Coherent Lower Previsions

As described in Walley's (1991) work, a subject's lower previsions represent his willingness to pay for a desirable gamble. Such betting rates need to be consistent with respect to all the gambles which is ensured by the rules of coherent lower prevision which have been derived from the rationality requirements of desirability. In Walley's (1991) work the rules are summarized as follows:

1. If a transaction makes the subject loses money irrespective of the outcome, he should never accept such transactions. (avoids sure loss)
2. If the subject considers a transaction acceptable he should consider all other transactions that at least gives him the same reward as acceptable.
3. A positive linear combination of acceptable transaction should also be acceptable.
4. The subject must be consistent with the supremum acceptable buying price and cannot raise it by considering a positive linear combination of other acceptable transaction (coherence) (Walley, 1991).

Let us now understand these rules with examples as explained in the work of Miranda & de Cooman(2014). Suppose a ball is to be drawn from an urn containing green, red and black balls, and a subject is offered a reward depending on the color of the ball drawn: he gets 10 euros for a green ball, 5 euros for a red ball and nothing if it is a black ball. The gamble f representing these rewards are given by:

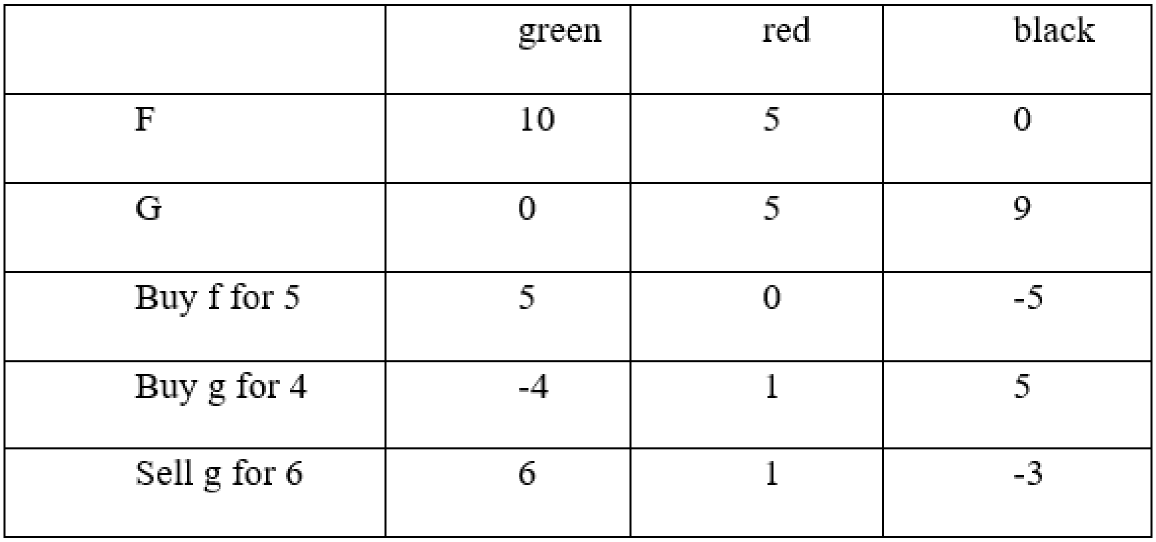
f(green) = 10, f(red) = 5, f(black) = 0

It should be desirable to our subject as it is not reducing his wealth through the reward and now he would buy the gamble for a certain price i.e. lower prevision based on his confidence. Let us suppose he is confident that there are no black balls in the urn then his supremum buying price i.e. lower prevision(x) is 5, then his increase in money would f-x. It would be 5 for a green ball, 0 for a red ball and -5 for a black ball. Now if our subject considers another gamble g which gives him a reward of 9 for a black ball, 5 for a red ball and nothing for a green ball.

g(green) = 0, g(red) = 5, g(black) = 9

Suppose the subject decides to pay a supremum amount of 6 i.e. lower prevision for the gamble g. Then total amount he spends for both the gambles is $\underline{P}(f)$ +$\underline{P}(g)$ = 5 + 6 = 11, then the subject gets a total reward (f+g) i.e. 10 for a green or a red ball and 9 for a black ball however as he spends 11 for both the gambles which is more than any of his rewards, he will incur a sure loss of at least 1. Thus, such combination of transactions is restricted according to the rule of avoiding sure loss. Now after some reflection our subject decides to pay up to 5 for the first gamble f and up to 4 for the second gamble g. He also decides that he would sell g for a price higher than 6, but not less. In the language of lower and upper previsions $\underline{P}(f)$ =

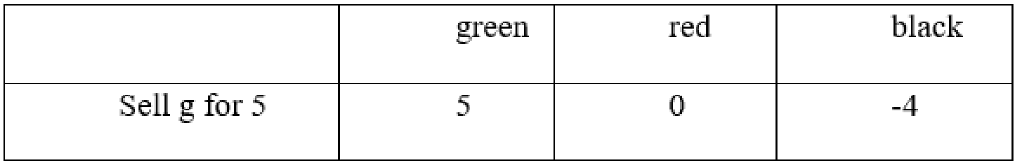
5, $\underline{P}(g) = 4$, $\overline{P}(g) = 6$, the corresponding changes to his wealth are as shown in the figure1 below.



	green	red	black
F	10	5	0
G	0	5	9
Buy f for 5	5	0	-5
Buy g for 4	-4	1	5
Sell g for 6	6	1	-3

Figure 1: Rewards of gamble f and g

These assessments avoid sure loss however since our subject accepts to buy f for up to 5, he should also be willing to sell g for any price higher than 5 which give him the rewards as follows.



	green	red	black
Sell g for 5	5	0	-4

Figure 2: Reward of gamble g

The total reward on it is more than buying f for 5, hence this should also be a desirable gamble if f is a desirable gamble thereby correcting the upper prevision $\overline{P}(g)$ = 5 instead of 6 as it is the infimum selling price. The above betting value of $\overline{P}(g)$ which was 6 earlier was corrected 5 by comparing the reward of gamble g with the gamble f thereby explaining the rule of coherence. The above rules bring consistency in the betting behavior. In the next section, we shall extend the coherent lower prevision theory towards machine learning and see how probabilities can be formed by using these concepts. (Miranda & de Cooman, 2014)

2.6 Lower and Upper Probabilities

As summarized in the work of Walley (1991), The usual approach in constructing the theory of probability is to start with axioms of probability and then define expectations with the mathematical properties of linear previsions however previsions are regarded as more fundamental than probabilities. The upper and lower probability are regarded as special case of upper and lower prevision. A lower probability is just a lower prevision defined on special type of domain that contains only indicator functions of events. If A is an event which is subset of Ω which consists of all the events, we use the same symbol A to denote its indicator function on the domain Ω. Indicator function is defined by A(ω) = 1 if ω ϵ A and A(ω) = 0 if ω ϵ A^c(compliment of A). The lower prevision defined on such a class of events like A is called lower probability. As interpreted before $\underline{P}(A)$ is interpreted as the supremum buying price you are willing to pay for the gamble A, whose reward is 1 if event A occurs (and nothing otherwise).

$$A(\omega) = 1 \text{ if } \omega \in A \text{ and } A(\omega) = 0 \text{ if } \omega \in A^c$$

Thus $\underline{P}(A)$ is the supremum betting price you are disposed to bet on A. If $\underline{P}(A)$= μ then the increase in wealth would be 1- μ if A occurs or -μ if A doesn't occur. We can extend the same reasoning for upper probability as one minus the supremum betting price against the event A. Such representation as given below restricts our focus to only events.

$$\overline{P}(A) = 1 - \underline{P}(A^c)$$

We should note that an event and gamble are exchangeable terms in behavioral interpretation where we see events as gamble whose outcome determines an uncertain reward (Walley, 1991).

2.7 Credal Set

Credal set M of $\underline{P}$ is defined as the set of all the probability measures P for which $\underline{P}(A) \leq$ P(A) $\leq \overline{P}(A)$ for all A $\subset \Omega$. Credal sets are convex sets of probability distributions. It is intended to express uncertainty by relaxing the requirement of probability to be a precise value. A credal set is defined to be the convex hull of a non-empty and finite family of probability measures. Credal sets have great expressive power, encompassing other models of uncertainty, and they are equivalent to coherent lower prevision as defined in Walley (1991). The axioms of probability are maintained for every distribution in the set.

The theory of lower and upper previsions is used to build models in machine learning to solve various classification problems. There are a few machine learning algorithms that work using the IP theory, one such algorithm includes the Naïve Credal Classifiers(NCC) in which the credal set is used for modelling. NCC is used and tested extensively in order to use its property of modelling uncertainty to our research problem. These classifiers are used solve the problem of uncertainty in data while modelling in small incomplete datasets. In this chapter, only the basic concepts of the IP theory were introduced that supports the understanding of the objective modelling framework. It is strongly advised to look at the references for more content on this theory and its applications.

CHAPTER 3

NAÏVE BAYES CLASSIFIER AND NAÏVE CREDAL CLASSIFIER

This chapter introduces the Naïve Bayes Classifiers(NBC) and the Naïve Credal Classifier(NCC), the latter is an extension of NBC towards the IP theory. NCC uses the concepts of credal sets to deliver robust classification even on small incomplete datasets (Corani & Zaffalon, 2008). The working of the NBC and NCC has been reviewed and compared in this chapter.

3.1 Naïve Bayes Classifier

NBC has always been a competitive machine learning algorithm (Wu, et al., 2008). Its success lies on its ability to probabilistically model the prior knowledge of the events. In its core lies the Bayes Rule which has extended its application to various inference models in recent years. In this chapter, the concepts the concepts of Bayesian Inference models have been reviewed briefly and for more complete and rigorous treatment of the topic, readers may refer to (Lavrenko & Sutton, 2011) and (McCallum & Nigam, 1998).

Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A & B are events and P(A) & P(B) are the probabilities of the respective events such that P(B) ≠ 0.

P(A|B) is conditional probability: the likelihood of event A occurring given B is true.

P(B|A) is conditional probability: the likelihood of event B occurring given A is true.

Derivation of Bayes Rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Since $P(A \cap B) = P(B \cap A)$ by the Commutative property of intersections

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

NBC is an implementation of the Bayes' rule, which is interpreted in terms of data and classes. This classifier which seems to be reasonably simple has been proved to be very effective in data based prediction. It is given by the following equation.

$$P(Class|Data) = \frac{P(Data|Class)P(Class)}{P(Data)}$$

$P(Class|Data)$, also called as Posterior probability determines the probability of a class conditioned on the given data. In NBC, the probability of all the classes is computed and the most probable class is the output of the classification.

$P(Data|Class)$, also called as the Likelihood, is the probability which determines how likely this data will fit into the given class. The NBC gets its name from it's naïve assumption that the features of the data are conditionally independent. This makes this classifier computationally easy. Although such an assumption may be wrong, it is quite reasonable as in most cases our data points have high dimensions and it is impossible to compute the probability without such an assumption. These classifiers tend to fail drastically when we have redundant data or proportional data in the dataset as it makes the attribute completely dependent contradicting it's assumption. Hence it is advisable to compute and check the covariance matrix of the dataset before applying the NBC.

$P(Class)$, also called as the Prior, is the probability of the class before assessing the data. Modelling priors sets NBC apart from the Maximum Likelihood Estimation algorithm. The prior modelling becomes important as it determines how probable a class is even before analyzing the data. For Example, to determine if a person is suffering from Ebola in Europe, irrespective of the symptoms, we would need to know how frequent are the cases of Ebola in Europe. Such an estimate brings out the frequency of the classes with respect to the dataset analyzed, in other words finding the relevance of Ebola in a particular continent.

$P(Data)$, also called as the evidence or the normalizer, is the probability that the data belongs to the data set. It is basically used to understand if the data belongs to the dataset of either one class and is not an outlier. In most classification problems evidence is ignored as it has no affect in the classification process or prediction of a class.

NBC can be written as follows in plain English interpretation.

$$Posterior = \frac{(Prior)(Likelihood)}{(evidence)}$$

Or if evidence is ignored

$$Posterior \propto (Prior)(Likelihood)$$

3.2 Bayesian Inference

Bayesian Inference is a method of statistical inference in which Bayes' theorem is used to update probability of a hypothesis as more evidence or information becomes available. In Bayesian Inference models, a probability distribution is assumed, that suitably represents the prior, likelihood and posterior. Gaussian Distribution is one of the most widely used distribution for the NBC models. The Bayesian Inference models also helps in modelling the prior with your belief (e.g. what do you think the probability of Ebola in

Europe can be). Such approaches which involve modelling belief are called subjective probabilistic approach. The posterior which is obtained after assessing the data is the product of our prior belief functions and the likelihood. Likelihood refines the model after looking at the data. In Bayesian Inference, prior is of utmost importance as it is based on one's belief and can result in either a very reliable model or unreliable model.

Such approaches that involve human belief in the modelling have both advantages and disadvantages. Few of the advantages include better inference via modelling which helps to answer classification-related questions. The disadvantage is that by assuming a prior for modelling, we can include a wrong logical reasoning via modelling and can lead to deceptive predictions. Another disadvantage is that human logic based computation can be intractable. Despite the drawbacks, Bayesian Inference is considered admissible as its performance is consistent and reliable; hence it is still among the most popular Machine Learning methods.

3.3 The Conjugacy Models

The Bayesian Conjugate models are the Bayesian Inference models where in the prior and posterior belong to the same probability distribution, which makes the distribution of the likelihood to be the conjugate of the prior and posterior distribution. A few examples of Bayesian Conjugate Models include the pairs of (Beta, Bernoulli), (Beta, Binomial), (Gamma, Poisson), (Dirichlet, Categorical), (Dirichlet, Multinomial), (Gaussian, Gaussian) model etc. The first distribution in the above conjugate pairs represents the distribution of the prior and posterior, and the later distribution represents the distribution of the likelihood. For example, in the Beta-Bernoulli model, the prior and the posterior belongs to the Beta distribution and the likelihood belongs to the Bernoulli distribution. The Conjugate models are very popular in the Bayesian Decision Theory as they make the models easily computable and make it more inferential by understanding the effect of the conjugate likelihood on the priors. Since the prior and the posterior belong to the same distribution (with different parameters) it becomes easy to understand the effects of data on your prior belief by understanding how the parameters change after assessing the data through the likelihood function.

3.4 Derivation of Naïve Bayes Dirichlet-Categorical Model

The Dirichlet-Categorical model is a popular conjugate model due to its simplicity in application and its capability in modelling uncertainty. As mentioned earlier in Dirichlet Categorical model we use a Dirichlet prior and a Categorical likelihood and the posterior also is a Dirichlet distribution. In this section, we shall review the derivation of the Naïve Bayes Dirichlet-Categorical model and NCC as shown in the work by (Corani, Abellan, Masegosa, & Zaffalon, 2014). Let us denote C the classification variable (taking values in *C*) and $\mathrm{A}_{1,\ldots\ldots,}\mathrm{A}_{\mathrm{k}}$ as the *k* discrete feature variables (taking values from the sets $A_{1,\ldots\ldots,}A_k$). The values of the class and features are denoted by lower case letters as c and **a** = $(\mathrm{a}_{1,\ldots\ldots,}\mathrm{a}_{\mathrm{k}})$. For a variable X, P(X) denotes probability mass function over all states of x ϵ *X,* while P(x) denotes the probability of X=x. Both P(X) and P(x) are subjective probabilities. The physical probability of the actual data generation mechanism is called chance; it is unknown and we aim at making inference about it. $\theta_{c,\boldsymbol{a}}$ denote the chance that (C, $A_{1,\ldots\ldots,}A_k$) = (c, **a**), $\theta_{a_i|c}$ denotes the chance that A_i = a_i conditional on C = c, $\theta_{a|c}$ the chance that $A_{1,\ldots\ldots,}A_k$= $(\mathrm{a}_{1,\ldots\ldots,}\mathrm{a}_{\mathrm{k}})$ conditional on C = c.

The Naïve assumption of conditional independence of features can be expressed as follows:

$$\theta_{a|c} = \prod_{i=1}^{k} \theta_{a_i|c}$$

We denote n as the total number of instances; n(c) is the count of class c and similarly n (a_i, c) is the count of instances where A_i = a_i and C = c; **n** is the vector of all counts of type n(c) and n (a_i, c). The multinomial likelihood can be expressed as:

$$L(\theta|\mathbf{n}) \propto \prod_{c \,\epsilon C} [\, \theta_c^{n(c)} \prod_{i=1}^{k} \prod_{a_i \epsilon A_i} \theta_{a_i|c}^{n(a_i,c)} \,]$$

The prior conjugate to the likelihood is a Dirichlet distribution where count n(.) is replaced everywhere by st(.) – 1. The parameter s is equivalent to the sample size and t(.) is interpreted as proportion of hidden instances, for instance t (c_1) is the expected proportion of hidden instances C = c_1.

For a non-informative prior:

$$\mathrm{t(c)} = \frac{1}{C};\ t(a_i, c) = \frac{1}{A_i\ C}$$

which is also called BDe prior.

By multiplying the joint Dirichlet prior and the multinomial likelihood, we get the posterior density $\theta_{c,\boldsymbol{a}}$, which is a Dirichlet density as well.

$$\mathrm{P}\big(\theta_{c,a}|\boldsymbol{n}, \boldsymbol{t}, s\big) \propto \prod_{c \,\epsilon C} [\, \theta_c^{n(c)+st(c)-1} \prod_{i=1}^{k} \prod_{a_i \epsilon A_i} \theta_{a_i|c}^{n(a_i,c)+st(a_i,c)-1} \,]$$

$\mathrm{P}\big(\theta_{c,a}|\boldsymbol{n}, \boldsymbol{t}, s\big)$ is estimated in the learning step and NBC can classify instances, on the new assignment **a** = $(a_{1,\ldots\ldots,}a_k)$. Considering that $\mathrm{P}(c|\boldsymbol{a}, \boldsymbol{n}, \boldsymbol{t}) \propto \mathrm{P}(c, \boldsymbol{a}|\boldsymbol{n}, \boldsymbol{t})$

$$\mathrm{P}(c, \boldsymbol{a}|\boldsymbol{n}, \boldsymbol{t}) = \mathrm{P}(c|\boldsymbol{n}, \boldsymbol{t}) \prod_{i=1}^{k} \mathrm{P}(a_i|c, \boldsymbol{n}, \boldsymbol{t}),$$

and by taking expectation we get:

$$\mathrm{P}(c|\boldsymbol{n},\boldsymbol{t}) = \frac{n(c)+st(c)}{n+s}$$

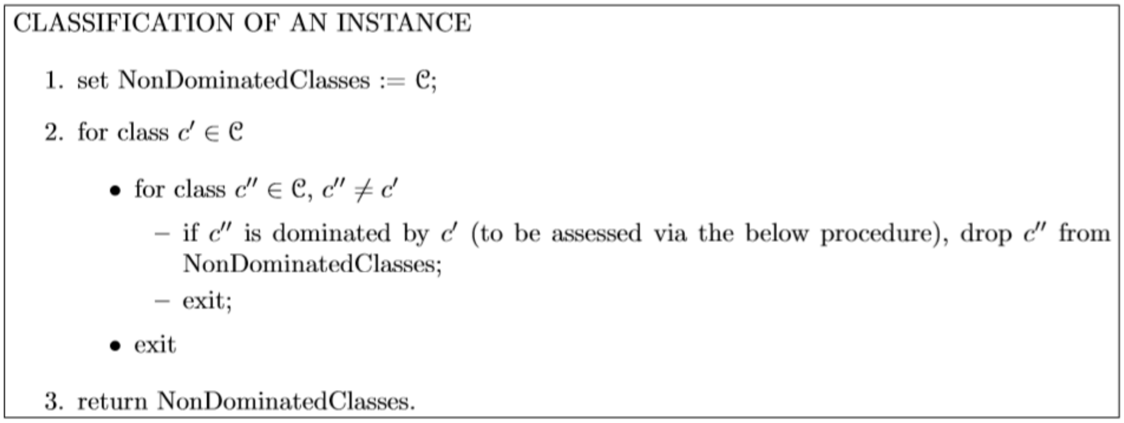
$$\mathrm{P}(a_i|c,\boldsymbol{n},\boldsymbol{t}) = \frac{n(a_i,c)+st(a_i,c)}{n(c)+st(c)}$$

The major problem in NBC is that sometimes the classification is prior dependent and this can be acceptable if the prior is carefully elicited and thus models' domain knowledge; yet this situation is uncommon. Taking the uniform prior solves the above problem by not making the model prior dependent however it only makes the model prior indifferent. NCC extends NBC by substituting the non-informative prior by the Imprecise Dirichlet Model(IDM).

3.5 Naïve Credal Classifier

The Naïve Credal Classifier(NCC) is an extension of NBC to the Imprecise Probability. NCC adopts a joint credal set which is modelled by IDM to represent prior where **t** ranges within the polytype **T,** rather than being fixed. The polytype consists all the densities for which **t** satisfy the constraints, ∀ (i, c): $\sum_c \mathrm{t}(c) = 1$, $\sum_{a_i} \mathrm{t}(a_i|c)$ = t(c), 0 < $\mathrm{t}(a_i|c)$< t(c), 0 < t(c) < 1.

The Credal- dominance is checked i.e. checking which classes are non-dominated by the others. These non-dominated classes is returned as output by NCC. This checking is done using the criteria of maximality: c_1 dominates c_2 if $\mathrm{P}(c_1, \boldsymbol{a}|\boldsymbol{n}, \boldsymbol{t})$ > $\mathrm{P}(c_2, \boldsymbol{a}|\boldsymbol{n}, \boldsymbol{t})$ ∀ **t ϵ T**. The NCC identifies the non-dominated classes as shown in the figure 3 below.



```
CLASSIFICATION OF AN INSTANCE
1. set NonDominatedClasses := $\mathcal{C}$;
2. for class $c' \in \mathcal{C}$
    • for class $c'' \in \mathcal{C}$, $c'' \neq c'$
        – if $c''$ is dominated by $c'$ (to be assessed via the below procedure), drop $c''$ from NonDominatedClasses;
        – exit;
    • exit
3. return NonDominatedClasses.
```

Figure 3: Pseudo code for NCC

It is formulated as follows:

c'dominates c'' in the above algorithm iff:

$$inf_{\boldsymbol{t \epsilon T}} \frac{\mathrm{P}(c',\boldsymbol{a}|\boldsymbol{n},\boldsymbol{t})}{\mathrm{P}(c'',\boldsymbol{a}|\boldsymbol{n},\boldsymbol{t})} > 1$$

The non-dominant classes are those classes which cannot be precisely ranked by the NCC. The NCC is indeterminate in classifying those instances precisely, hence it returns multiple classes. The instances that are determinately classified by NCC are the ones where the NCC returns a single valued output.

The NCC uses the Imprecise Dirichlet Model to model uncertainty, which makes it a more reliable classifier. It performs well especially in small and incomplete data sets which lead to epistemic uncertainties. Assuming a specific prior distribution for such datasets like how NBC does, may lead to strong assumptions and misrepresentation of the data. The NCC instead uses Imprecise Dirichlet Model for modelling prior where a set of probability distributions are considered also called as prior credal set. The NBC's alternative approach to deal with prior ignorance is to assume a uniform prior. This only makes the prior non-informative, thereby resulting in a prior indifferent model as discussed in the earlier section.

Another type of ignorance that may arise from small incomplete data set is the missing data ignorance. Both NBC and NCC ignore the missing attributes and are incapable of modelling such ignorance. In the Chapter 4 we shall introduce a Novel Framework which is an extension to NCC. It deals with missing data through advanced imputation techniques thereby expanding the ability of NCC to handle missing data ignorance.

3.6 Extensions of NCC

This section very briefly introduces a few extensions of NCC that can handle missing data inherently. These extensions shall be compared with our novel framework in chapter 4. For more detailed study on the extensions of NCC, readers may study the material referred.

- Lazy Naïve Credal Classifier(LNCC): The LNCC combines NCC and lazy learning to improve the high bias error and high indeterminacy rate in NCC. It does this by ranking the instances of the training set according to the distance from the query. It then trains a local classifier on the k instances nearest to the query and returns the classification using local classifiers. Please refer to the works of (Giorgio & Zaffalon, 2009) for a detailed treatment of this topic.
- Credal Model Averaging(CMA): CMA is an extension of Bayesian Model Averaging(BMA) to Imprecise Probability. The BMA addresses model uncertainty by averaging over the set of candidate models by assigning a weight to a model proportional to its prior. CMA models substitutes the prior of BMA models with a credal set to address the drawbacks of uniform priors. The disadvantages of assuming a uniform prior has been reviewed in the earlier section. Thus, CMA is a more robust

model than BMA. For more rigorous treatment of this topic, the readers are advised to refer to (Corani & Zafflon, Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities, 2008).

CHAPTER 4

NCC-EM: A NOVEL CREDAL BASED FRAMEWORK

In this chapter, a novel framework approach is introduced to improve the performance of the Naïve Credal Classifier(NCC). We shall exploit the existing advanced techniques and use it along with NCC. This Chapter covers the aspects of missing data, their types and how our framework deals and accounts for the epistemic uncertainties due to missing data.

4.1 Missing Data

As reviewed in Little & Rubin's work in 2014, missing data arises in almost all the real time statistical analysis. Ignoring missing data can lead to highly biased estimates thereby reducing the accuracy of the models. Before we tackle missing data, it is very important to understand the types of missing data and how they are generated (Little & Rubin, 2014).

Let us introduce some notations from Little & Rubin's work. Data matrix Y, missing data matrix M = $\{M_{i\,j}\}$:

$$M_{i\,j} = \begin{cases} 1\; if\; Y_{i\,j}\; is\; missing \\ 0\; if\; Y_{i\,j}\; is\; observed \end{cases}$$

$$Y = (Y_{obs}, Y_{mis})$$

where Y_{mis} denote the data that is not observed.

Missing data is generated by the following mechanisms and are classified as follows:

1. Missing Completely at Random (MCAR): as the name suggests is a type of missing data that happens totally at random and is not dependent on any factor. For example, if a few survey papers are lost from the top of the box, we can say that the data is missing

completely at random. Such type of missing data generation can reduce the data power and thereby the efficiency of the Predictive algorithm also decreases with it.

$$f(M|Y,\theta) = f(M|\theta), i.e. M \amalg Y|\, \theta.$$

2. Missing at Random (MAR): This is the most common type where the generation is due to other attributes in the data however, it is still at random. For example, a patient might not take an expensive medical test because his or her insurance type doesn’t cover it. This is the type of missing data, we frequently find in datasets and is the only type that can be handled using various imputation techniques.

$$f(M|Y,\theta) = f(M|Y_{obs},\theta), i.e. M \amalg Y_{mis}|\, (Y_{obs},\theta).$$

3. Missing Not at Random(MNAR): This type of missing data generation occurs due to the nature of the attribute itself. For example, in a survey, people might not feel comfortable filling details about their psychiatric treatment as a result most of the data under that column goes missing. It is important to differentiate between MAR and MNAR carefully before using the missing data imputation techniques as they are applicable only for MAR type. All the cases that are not MAR and MCAR fall under MNAR.

4.2 Missing Data Imputation Methods

The process of replacing missing data with substituted values is called Imputation. The different techniques to handle missing data is reviewed briefly in this section. For more rigorous treatment of this topic, readers may refer to Little & Rubin’s work from 2014.

1. Complete case analysis: In this method only the complete cases of the data, where all the variables are observed is used for analysis. Can be adequate if most of the cases are present. The disadvantage of this method is that it reduces the effective sample size and is a bad technique on small datasets.

2. Mean Imputation: In this method, we substitute the missing value with the mean value of the attribute. It is too simple and can still lead to highly biased estimates however it reduces the bias error as compared to the complete case analysis.
3. K- Nearest Neighbors: In this method, the missing values are imputed by averaging the values of the attributes of its K nearest neighbors. It is an extension of mean imputation and it reduces the bias error further as compared to mean imputation.
4. Hot deck: In this method, the missing values are substituted with the values from a random instance with same attribute values. This is also called as last observation carried forward.
5. Regression: This is a model based imputation technique where a regression model is estimated to predict observed values of a variable based on other variables. This regression model is used to impute missing values of the variable.
6. Predictive Mean Matching (PMM): This method is a combination of regression and hot deck imputation technique. It initially estimates a regression model and then substitutes the variable with the predicted value only if it appears previously.
7. Expectation Maximization (EM): The expectation maximization is an alternative to Newton Raphson method where the missing data is imputed maximizing the likelihood function. It basically consists of two steps. The Expectation step and the Maximization step. In the expectation step the mean and variance of the gaussian distribution is calculated from the complete cases and in the maximization step we try to maximize the likelihood of the data. This is an iterative process until it reaches a convergence point for a certain mean and variance value. A more detailed review of EM imputation is done in the later section.

8. Multiple Imputation: To deal with the noise due to imputation technique, Rubin (1987) developed a method for averaging across multiple imputed datasets. It consists of three phases namely: Imputation phase, Analysis phase and Pooling phase. In the imputation phase, multiple data sets are created based on one of the above models. In the Analysis phase, the statistical analysis is carried out on all the imputed datasets. Pooling phase finally combines the result of all the datasets based on a combining rule. For more detailed understanding of this techniques, readers are suggested to refer to (Little & Rubin, 2014).

4.3 Missing Data Handling (NBC vs NCC)

The Naïve Bayes Classifiers as discussed earlier ignores missing data while calculating the likelihood. It is equivalent to replacing the probability of the missing event with 1 while computing the likelihood. It is reasonable to ignore the missing attributes in the Naïve Bayes Classifier as the model takes care of it on its own whereas most of the other classifiers cannot handle the missing values at all and would need a preprocessing step to get them working. Most software's that implement the Machine Learning algorithms do a complete case analysis as a preprocessing step when they encounter missing data which can lead to highly biased estimates as it reduces the sample size. On the other hand, the NCC which models epistemic uncertainties handles missing data through prior modelling by including a set of probabilities in the prior credal set. It is noted that while computing the likelihood it does exactly what the NBC does. The only reason that NCC has a better performance is due to its Imprecise Dirichlet prior that allows modelling of epistemic uncertainties in data. In the next section, we shall share our idea of a novel framework that can improve the performance of NBC and NCC by adopting a hybrid framework.

4.4 Hybrid Credal Based Framework

In this section, we shall present our key research idea of using a hybrid framework to improve the performance of NBC and NCC. The framework includes both precise and imprecise probabilistic methods in combination, hence it's been termed as a hybrid framework. This framework is more robust in handling missing data as it tackles the bottlenecks of the existing precise and imprecise probabilistic frameworks by complimenting each other. It tackles bias error that occurs due to missing data by using the advanced imputation techniques and later uses imprecise probabilistic framework to tackle the overfitting that occurs due to the imputation techniques. The block diagram shown in figure 4 below explains the key features of the framework and shows how it improves the existing methods.

Key Research idea:

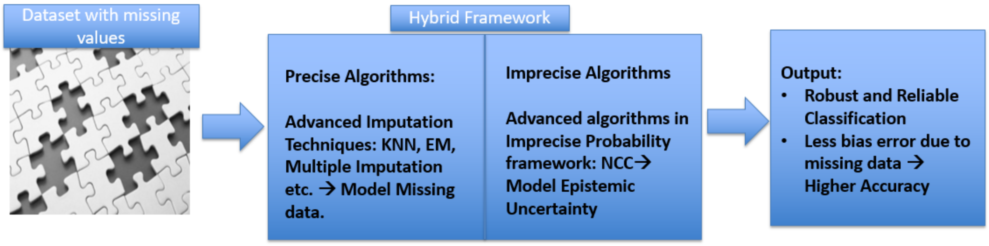

Dataset with missing values
Hybrid Framework
Precise Algorithms:
Advanced Imputation Techniques: KNN, EM, Multiple Imputation etc. → Model Missing data.
Imprecise Algorithms
Advanced algorithms in Imprecise Probability framework: NCC→ Model Epistemic Uncertainty
Output:
• Robust and Reliable Classification
• Less bias error due to missing data → Higher Accuracy

Figure 4: Block diagram of Hybrid Framework

The missing data imputation can be done through various methods which include simple methods like mean imputation to complex methods like multiple imputation using chained equation however in our experiments we shall stick to the model based imputation techniques like KNN, Regression, Expectation Maximization which strike a balance between the complexity, accuracy and the computation time. The more complex algorithms like Multiple Imputation take more computational time and tend to over fit data on small datasets. It is advised to choose the imputation techniques based on the dataset we are trying to impute.

We have test and tried various techniques and have stuck to EM imputation for modelling as it combines well with generative models like NBC and NCC. The experiment results in the figure 5 shows that EM imputation combines well with NCC. Discounted Accuracy is the metric used for comparing the accuracy of the credal classifiers in the figure below. Its formulation is discussed under the evaluation section.

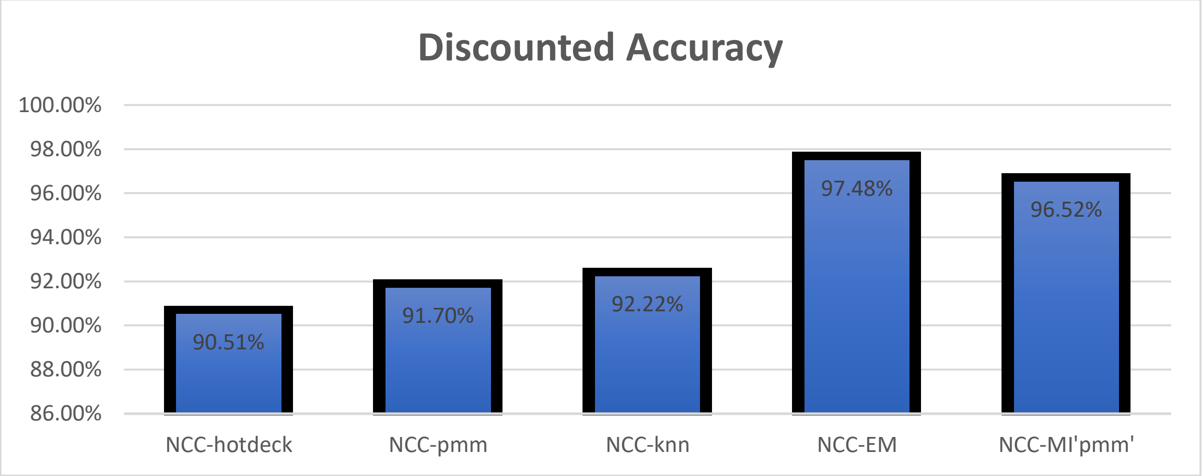


Figure 5: Comparison of EM with other imputation techniques

4.5 NCC-EM

NCC- EM is an implementation of the proposed hybrid framework, it is an extension of NCC with EM imputation as a preprocessing step. It is implemented by using the EM imputation algorithm over the existing Java based NCC library.

The EM Imputation algorithms works as follows.

1. Calculate log-likelihood of the observed data.
2. Estimate the parameters (mean and variance).
3. Replace the missing values by estimated values conditional on observed data.
4. Re-estimate the parameters.
5. Iterate until convergence or maximum number of iterations.

As illustrated in Rubin's work (2014). Let Data (X, Y) are the complete data whereas only incomplete data Y = y is observed.

The complete data log -likelihood is:

$$l(\theta) = \log L\ (\theta; x, y) = \log f(x, y; \theta).$$

The marginal log -likelihood or incomplete data log -likelihood is based on y alone is equals

$$l_y(\theta) = \log L(\theta; y) = \log f(y; \theta).$$

The EM algorithm maximizes the marginal log-likelihood iteratively and alternates between two steps E-step and the M-step.

1. **E-step**: Calculates the expected completed data log-likelihood ratio q(θ|θ*).

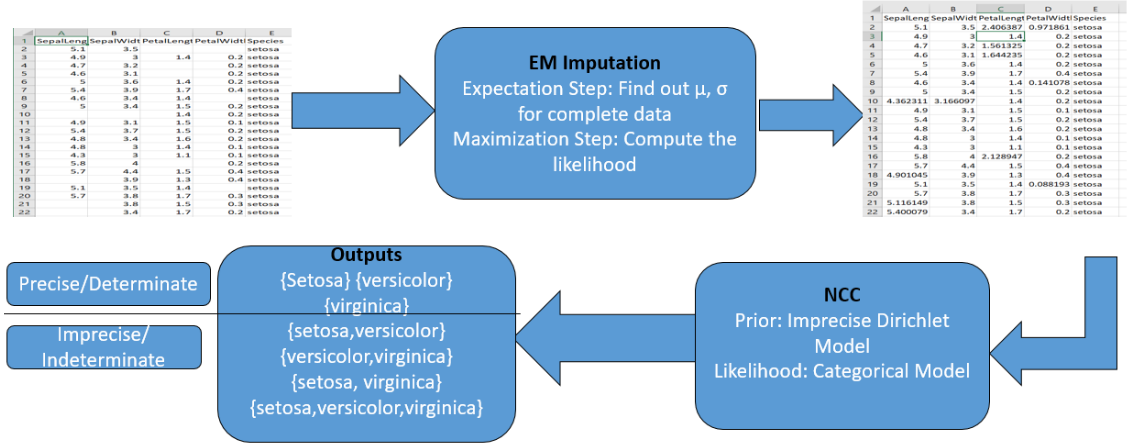
$$\text{q}(\theta|\theta^*) = E_{\theta*}[log\frac{f(X,y;\theta)}{f(X,y;\theta*)}|Y = y\ ]$$

where θ* is the value of θ for the current iteration.

2. **M-step**: Maximizes log likelihood ratio q(θ|θ*) in θ for fixed θ*,

$$\theta^{**} = \text{argmax } q(\theta|\theta^*)$$

In the next step of NCC-EM, the completed data from the EM imputation calls the JNCC2 library which is developed by Corani (2008), for a credal based imprecise classification. The algorithm is discussed in detail in the chapter3. The NCC tackles the overfitting of data due to the EM imputation and outputs multiple classes when it is doubtful. The following figure 6 summarizes the workflow of NCC-EM.



	A	B	C	D	E
1	SepalLeng	SepalWidt	PetalLengt	PetalWidtl	Species
2	5.1	3.5			setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2		0.2	setosa
5	4.6	3.1		0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4		setosa
9	5	3.4	1.5	0.2	setosa
10			1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4		0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18		3.9	1.3	0.4	setosa
19	5.1	3.5	1.4		setosa
20	5.7	3.8	1.7	0.3	setosa
21		3.8	1.5	0.3	setosa
22		3.4	1.7	0.2	setosa

EM Imputation
Expectation Step: Find out μ, σ for complete data
Maximization Step: Compute the likelihood

	A	B	C	D	E
1	SepalLeng	SepalWidt	PetalLengt	PetalWidtl	Species
2	5.1	3.5	2.406387	0.971861	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.561325	0.2	setosa
5	4.6	3.1	1.644235	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.141078	setosa
9	5	3.4	1.5	0.2	setosa
10	4.362311	3.166097	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	2.128947	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	4.901045	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.088193	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.116149	3.8	1.5	0.3	setosa
22	5.400079	3.4	1.7	0.2	setosa

NCC
Prior: Imprecise Dirichlet Model
Likelihood: Categorical Model
Outputs
{Setosa} {versicolor} {virginica}
{setosa,versicolor}
{versicolor,virginica}
{setosa, virginica}
{setosa,versicolor,virginica}
Precise/Determinate
Imprecise/ Indeterminate

Figure 6: Workflow diagram of NCC-EM

4.6 Evaluation and Results

This section reviews the performance of NCC-EM. It is rigorously tested and compared with the state of art IP algorithms introduced in chapter 3. These algorithms are compared across various datasets from the University of California Irvine(UCI) – Machine Learning repository. The performance metrics used for comparison are as follows:

- Determinacy: The percentage of instances for which the algorithm gives a single output.

- Indeterminacy Size: The average number of classes as output when the algorithm is indeterminate or gives multiple/set values as output.
- Single Set Accuracy: The percentage of correct class prediction when the algorithm is determinate or gives single output.
- Multiple Set Accuracy: The percentage of correct class prediction when the algorithm is indeterminate or gives multiple classes as output.
- Discounted Accuracy (DACC): This metric compares accuracy of Credal Classifiers. It takes into account the determinacy and accuracy of the classifier thereby making it easy to compare.

$$\text{DACC}=\frac{1}{N}\sum_{i=1}^{N}((accurate)_i)/|Z_i|$$

Where $(accurate)_i$ is 0 -1 variables, 0 if classifier is inaccurate and 1 if classifier is accurate for the I instance, N is number of instance for the test set, $|Z_i|$ is the number of classes returned (Corani, Abellan, Masegosa, & Zaffalon, 2014).

The missing datasets are generated using an R code by randomly deleting values from the dataset. Suppose the data set has 3 features and 10 instances, the x% of missing data for that dataset is created by deleting [(x/100) ×3×10] values from the dataset.

Case-Study 1: Iris dataset with 10 %missing data using 10-fold cross validation over 10 iterations.

The figure 7 below compares the determinacy percentage of NCC-EM with the state of art machine learning algorithms. As discussed earlier the NCC gives a set valued output in case of doubtful instances. In the results below we can see that NCC-EM is determinate only for 95% of the instances and for the rest 5% of instances it gives a set valued output. The other probabilistic ML algorithms as discussed are always precise, therefore are 100% deterministic.

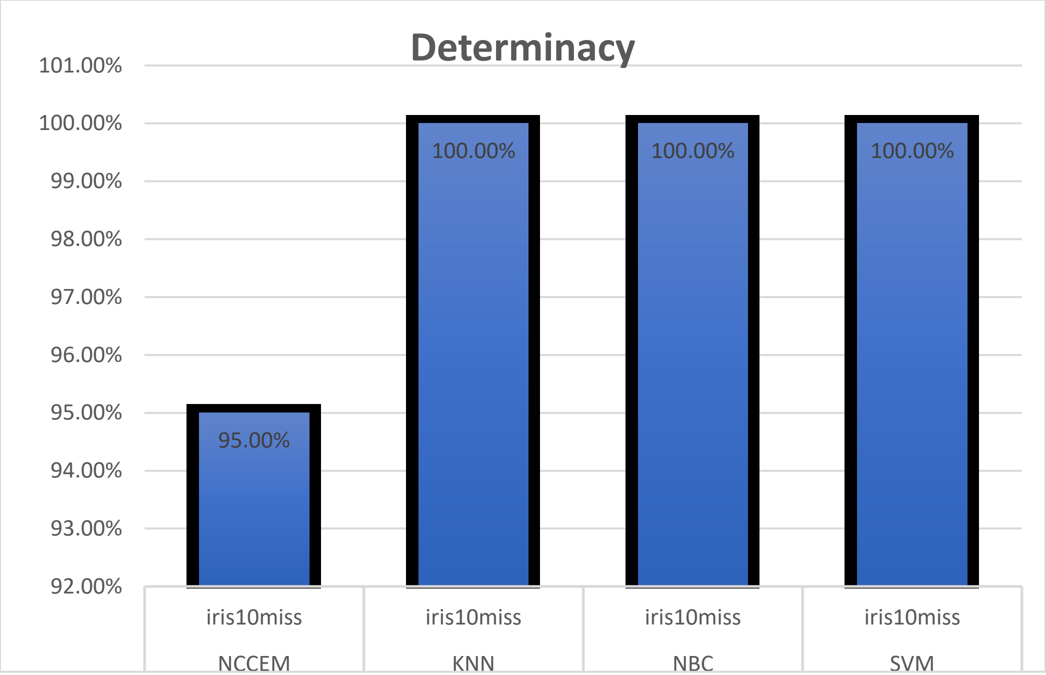


Figure 7: Determinacy comparison of NCC-EM with other Machine Learning Algorithms

The figure 8 below shows that the NCC-EM has an accuracy of 100% on the 95% of the instances where it is determinate. The NCC- EM's ability to distinguish doubtful instance helps it achieve a higher accuracy over the determinate instances which makes NCC-EM more reliable. NCC-EM is noted to be more accurate than most of the precise ML algorithm when it is determinate. We can also note that the accuracy of these precise algorithms reduces significantly in those instances where NCC is indeterminate.

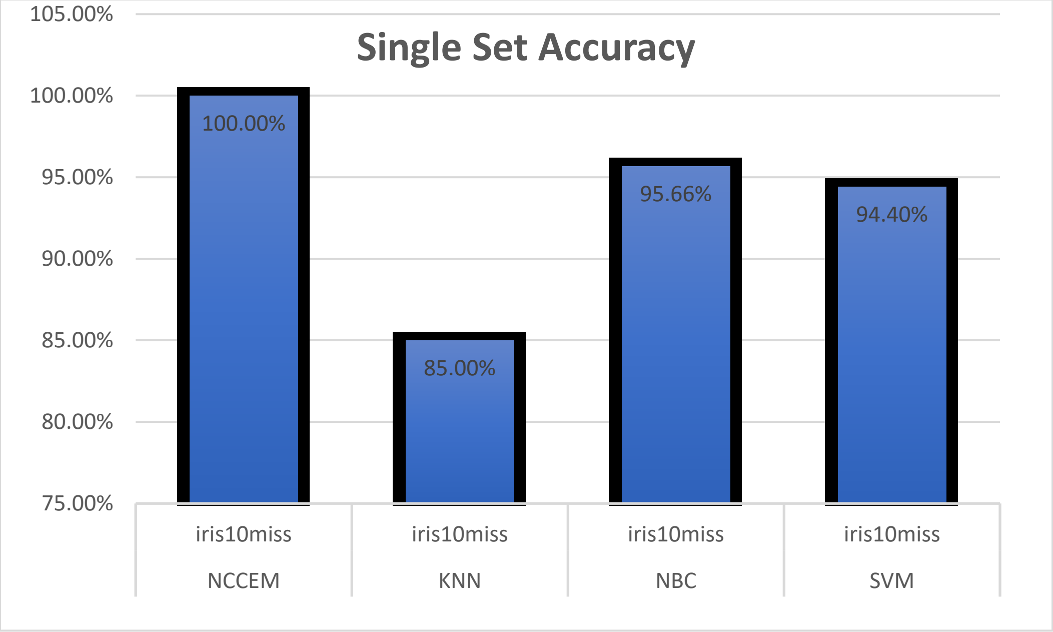


Figure 8: Accuracy comparison of NCC-EM with other Machine Learning Algorithms

The figure 9 below shows that NCC-EM has a set accuracy of 100% on the 5% of the instances where it is indeterminate. This makes NCC-EM a more robust classifier where it can still achieve a higher accuracy over small incomplete datasets. In NCC-EM reliability is given more preference over preciseness. It is reliable as it returns set valued output with a high accuracy rather than a single valued output with lower accuracy. The set valued output may draw weaker conclusions but are more reliable (Corani & Zaffalon, 2008).

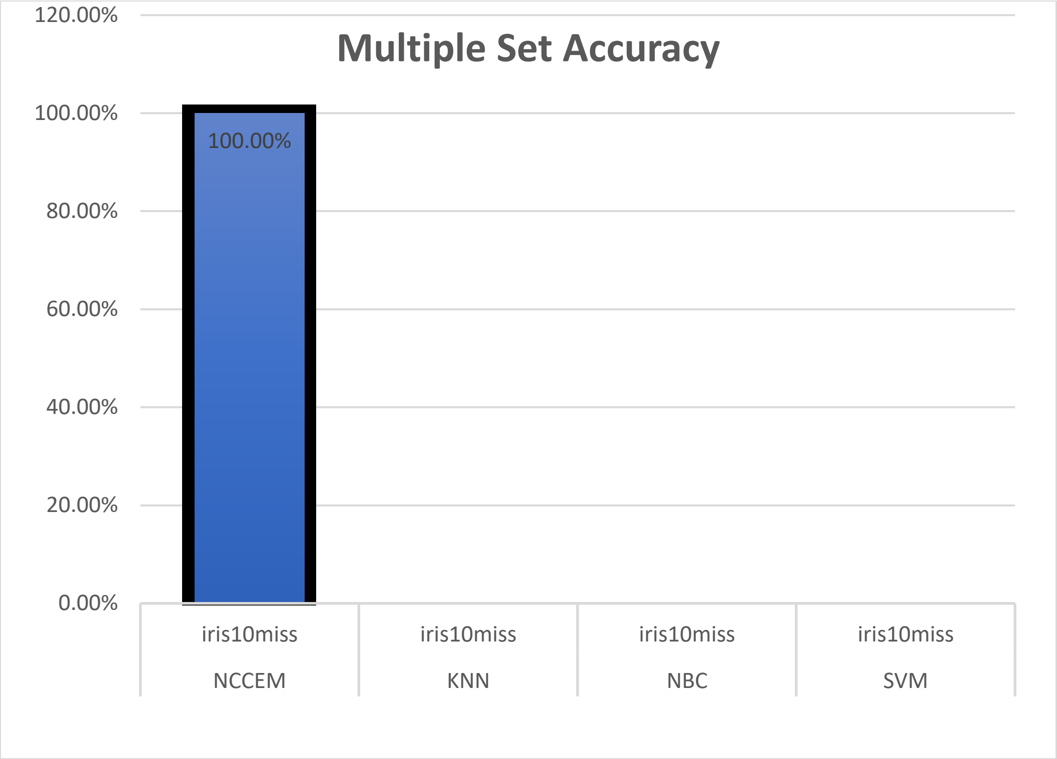


Figure 9: Set Accuracy of NCC-EM

The figure 10 below shows that the discounted accuracy of NCC-EM is higher than the other state of art NCC extensions. As discussed earlier, the discounted accuracy is a measure to compare credal classifiers. It accounts for the indeterminacy, by penalizing the accuracy by the indeterminacy size.

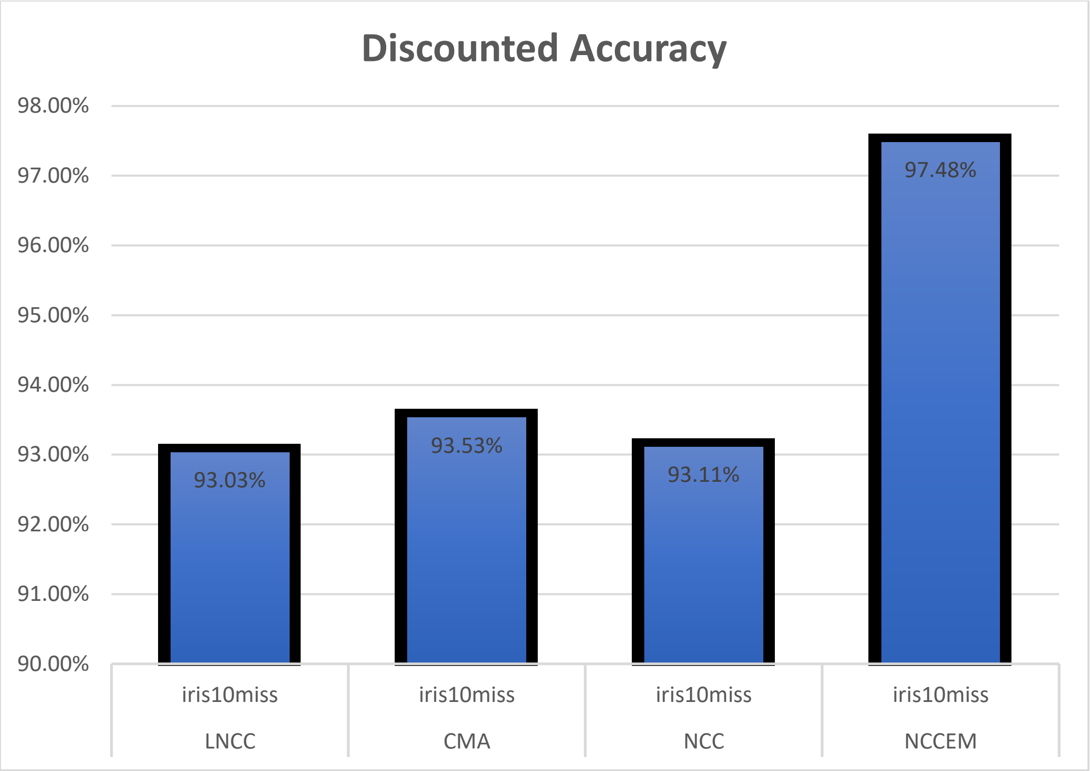
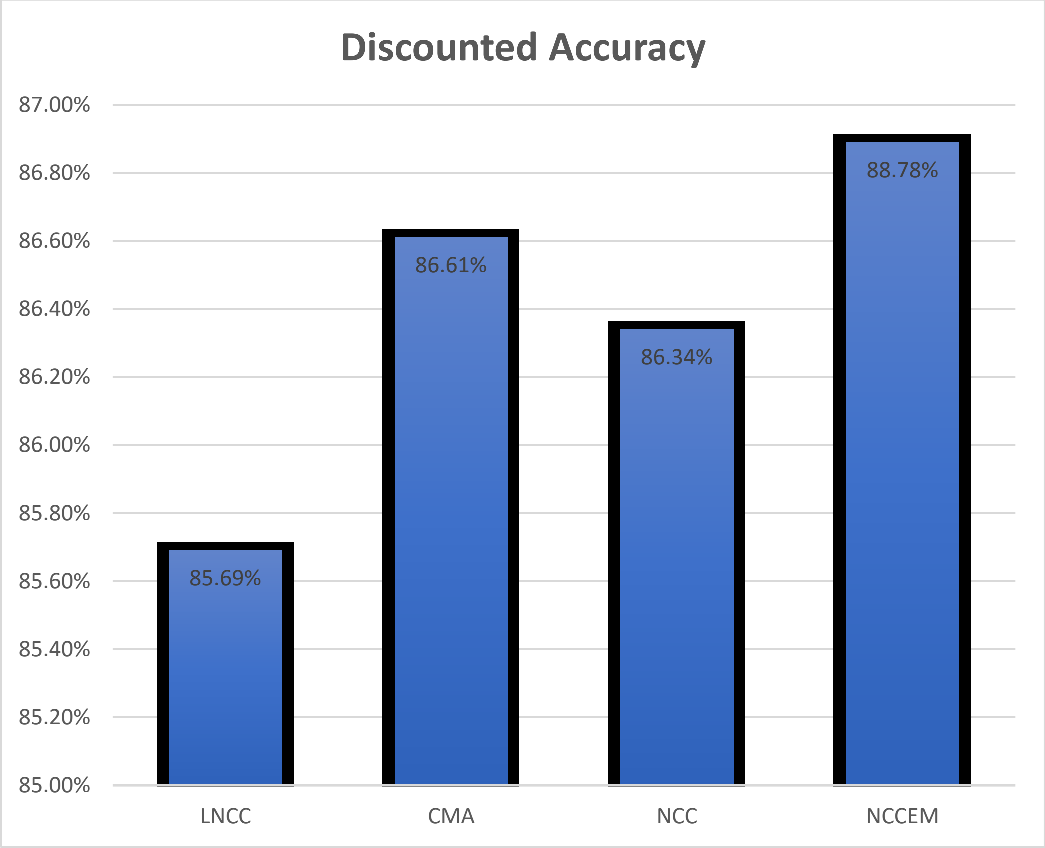


Figure 10: Discounted Accuracy comparison of NCC-EM with other NCC extensions.

Case-Study 2: Average performance over all the 4 datasets mentioned below, using 10 %missing data and 10-fold cross validation method over 10 iterations.

- Iris dataset
- Wine dataset
- Wisconsin breast cancer dataset
- Pima Indians diabetes dataset

The figure11 shows the average performance of the classifiers over 400 calls of classification. We notice that NCC-EM is equally competitive with the other extensions over a variety of datasets and slightly outperforms the others.



.

Figure 11: Discounted Accuracy comparison of NCC-EM with other NCC extensions.

The figure 12 below, shows that the NCC-EM outperforms all the state of art ML algorithms below when it is determinate.

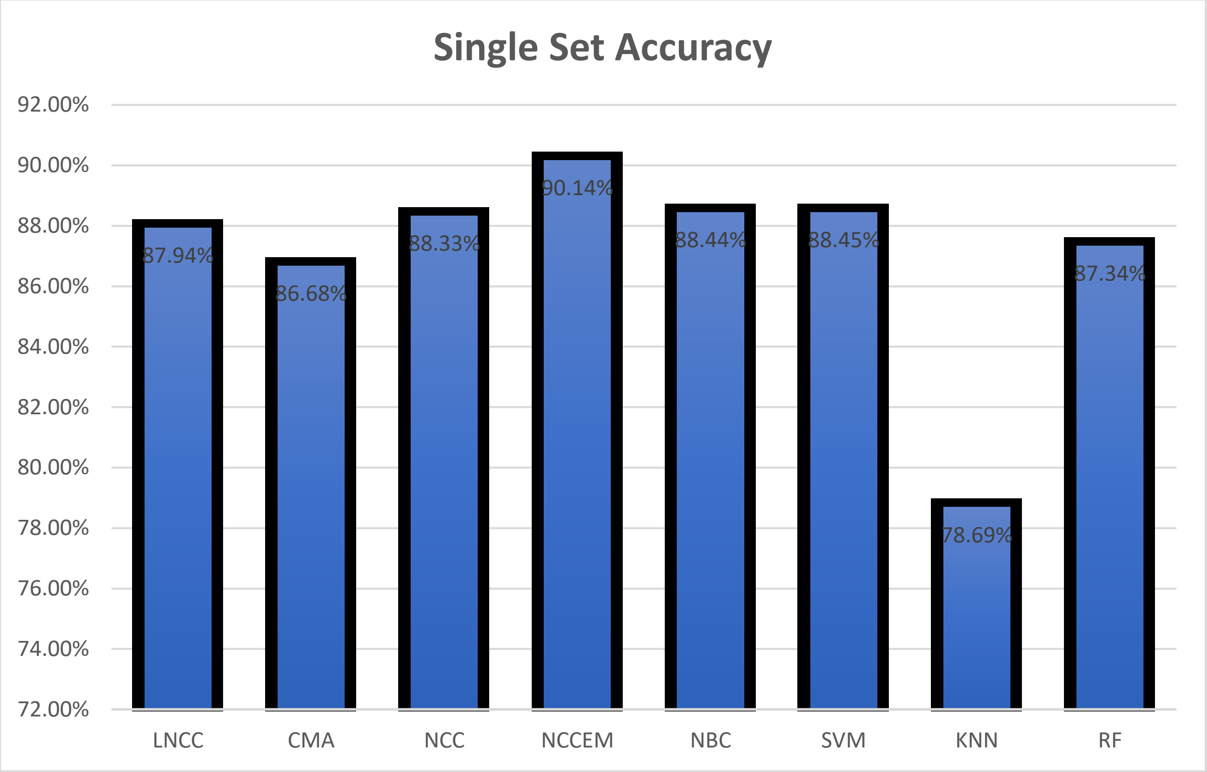


Figure 12: Comparison of Single Set Accuracy over all the classifiers.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this chapter, we shall draw some conclusions about the NCC-EM and explore the one of the many possible application of this framework.

5.1 Conclusion

From the experiments in the previous chapters, we can conclude that the NCC-EM will play an effective role in objective decision making. It is a more accurate and more reliable framework than compared to the other machine learning approaches due to its ability in handling missing data holistically and its ability to handle epistemic uncertainties through the NCC. This makes it an effective tool for creating objective models thereby facilitating real time objective decision making even with small incomplete data.

5.2 Future Work

This research work is a part of the bigger ongoing project for real time community resilience predictions. The future application scenario involves analyzing disaster related data from different sources including satellite, crowd sourced and aerial image data from drones. This real-time image data is passed through a deep learning network for primary labelling of the images, which later is passed through the NCC-EM for objective decision making. The output of the NCC-EM is compared with the true labels once available, thereby updating its network from it. The above process is shown as a workflow in the figure 12 below.

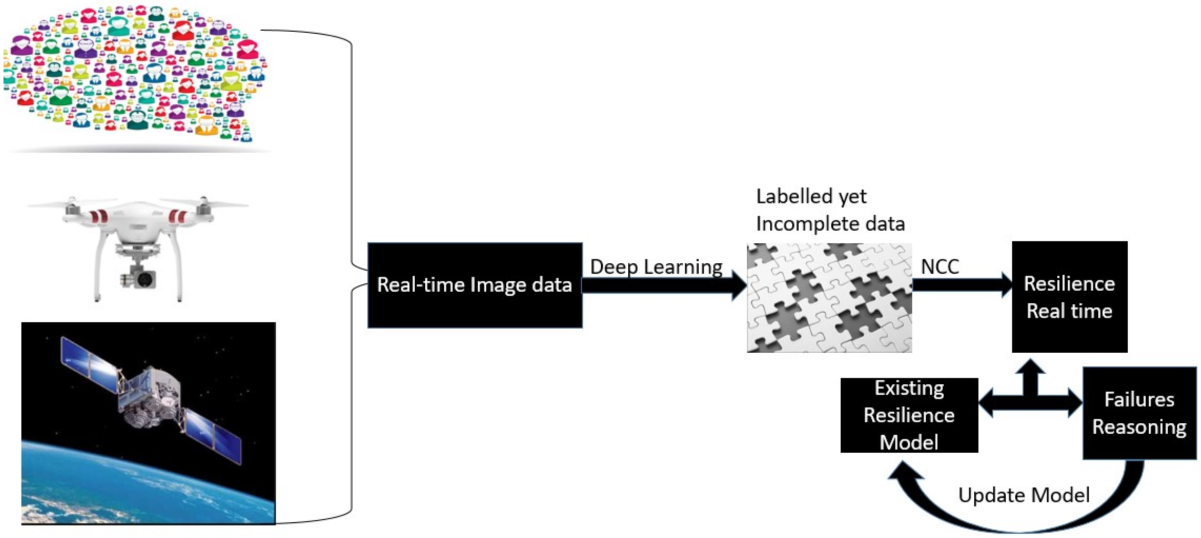


Figure 13: Internet of Things framework for Resilience Modelling

The NCC-EM is developed in Java programming language to make it platform independent and easily deployable on other platforms. As a future work NCC-EM will be more rigorously tested with a variety of datasets and will be extended to big data platforms aswell.

BIBLIOGRAPHY

- Abellán, J., & Mantas, C. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications, 41*(10), 4625-4637.
- Bernard, J. M. (2005). An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning, 39*(2-3), 123-150.
- Cooleen, F. P., Troffaes, M. C., & Augustin, T. (2011). *International Encyclopedia of Statistical Science.* Berlin, Heidelberg: Springer.
- Corani, G., & Benavoli, A. (2010). *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods. IPMU 2010. Communications in Computer and Information Science* (Vol. 80). Berlin, Heidelberg: Springer.
- Corani, G., & de Campos, C. (2010). A tree augmented classifier based on Extreme Imprecise Dirichlet Model. *International Journal of Approximate Reasoning, 51*(9), 1053-1068.
- Corani, G., & Zaffalon, M. (2008). JNCC2, the implementation of naive credal classifier 2. *Journal of Machine Learning Research, 9*, 2695–2698.
- Corani, G., & Zafflon, M. (2008). Credal model averaging: an extension of Bayesian model averaging to imprecise probabilities. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 257-271). Antwerp, Belgium: Springer.

- Corani, G., Abellan, J., Masegosa, A., & Zaffalon, M. (2014). *Introduction to Imprecise Probabilities.* Chichester: John Wiley & Sons Ltd.
- Cozman, F. G. (2000). Credal Networks. *Elsevier, 120*(2), 199-233.
- de Campos, C. (2010, September 1). *Conferences, Schools and Special Sessions: The Society for Imprecise Probability: Theories and Applications.* Retrieved from The Society for Imprecise Probability: Theories and Applications: http://www.sipta.org/ssipta10/material/classification/classification-decampos.pdf
- de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré, 7*, 1-68.
- Ditlevsen, O., & Kiureghian, A. D. (2009). Aleatory or epistemic? Does it matter? *Structural Safety, 31*(2), 105-112.
- *Event (probability theory):Wikipedia*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Event_(probability_theory)
- Giorgio, C., & Zaffalon, M. (2009). Lazy naive credal classifier. *Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data* (pp. 30-37). New York: Association for Computing Machinery.
- Halpern, J. Y. (2017). *Reasoning About Uncertainities* (2nd ed.). Cambridge, MA: The MIT Press.
- Lavrenko, V., & Sutton, C. (2011). *Courses: The University of Edinburgh- School of Informatics*. Retrieved from The University of Edinburgh- School of Informatics: http://www.inf.ed.ac.uk/teaching/courses/iaml/2011/slides/naive.pdf

- Little, R., & Rubin, D. (2014). *Statistical analysis with missing data.* New Jersey: John Wiley & Sons.
- Liu, Z.-g., Pan, Q., Dezert, J., & Martin, A. (2016). Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recognition, 52*, 85-95.
- McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-Worksop on Learning for Text Categorization.* Madison.
- Miranda, E., & de Cooman, G. (2014). *Introduction to Imprecise Probabilities, chapter 2.* (T. Augustin, G. de Cooman, F. Coolen, & M. Troffaes, Eds.) Chichester, West Sussex, United Kingdom: Wiley.
- Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press.
- Shafer, G. (1992). *Encyclopedia of Artificial Intelligence* (2nd ed.). (S. C. Shapiro, Ed.) Wiley. Retrieved from Glenn Shafer.
- Shafer, G., & Vovk, V. (2005). *Probability and finance: it's only a game!* John Wiley & Sons.
- Troffaes, M. (2016, August 29). *Conferences, Schools and Special Sessions: The Society for Imprecise Probability: Theories and Applications.* Retrieved from The Society for Imprecise Probability: Theories and Applications: http://www.maths.dur.ac.uk/users/matthias.troffaes/siptass16/slides/01-monday/01-matthias/matthias-foundations.pdf
- Walley, P. (1991). *Statistical Reasoning With Imprecise Probabilities.* New York, New Jersey, United States of America: Chapman and Hall.

- White, I., Wood, A., & Royston, P. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399.
- Williams, P. M. (2007). Notes on condition previsions. *International Journal of Approximate Reasoning, 44*(3), 366-383.
- Wu, X., Kumar, V., Quinlan, R. J., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1-37.
- Zaffalon, M., & Fagiuoli, E. (2003). Tree-Based Credal Networks for Classification. *Reliable Computing, 9*(6), 487-509.
- Zaffalon, M., Wesnes, K., & Petrini, O. (2003). Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine, 29*(1-2), 61-79.
- Zafflon, M. (2002). The naive credal classifier. *Journal of statistical planning and inference, 105*(1), 5-21.
- Zelterman, D. (2005). Bayesian Artificial Intelligence. *Technometrics, 47*(1), 101-102.

VITA

Varun Chavakula completed his Bachelor's degree in Electrical and Electronics Engineering from SRM University, India and then worked as a Software Engineer in Tata Consultancy Services, India Pvt. Ltd. for 3 years. With an interest in Data-Science, he started his Masters in computer Science at the University of Missouri-Kansas City (UMKC) in August 2015. While studying in UMKC, he has worked as a Graduate Research Assistant under the guidance of Dr. Chen ZhiQiang and worked on a Machine Learning project with emphasis on data science. Upon completion of his requirements for the Master's Program, he plans to work as a Data Scientist.