Data Cleaning Using Batch Reinforcement Learning

[1]Ghalage Prajakta J., [2]Sonawane Nalini N., [3]Tamhane Madhuri S., [4]Deshmukh Nutan S.
Student
Computer, GCOEARA, Pune, India
[1]ghalagepraju@gmail.com, [2]sonawanenalini27@gmail.com, [3]tamhanepradnya@gmail.com,
[4]deshmukhnutan62@gmail.com

***Abstract*- In real world raw data is highly affected by Missing value and uncertainty. This missing and uncertain data leads some distraction in dataset. So that before storing that data in dataset we have to clean that data first. Data cleaning is an important step in data mining [3]. In this paper we introduce some methods to find and remove the missing data and uncertainty. We generate the missing data using Q-Learning Algorithm. In Q-Learning Algorithm the missing data is generate and replaces the Null values with generated one. We use new Discretization Algorithm called UCAIM (Uncertain Class-Attribute Interdependency Maximization) that will find and replace uncertain data. Batch Reinforcement Learning is area of machine learning. By using batch reinforcement learning we can batch the transitions. Removal of Poor Attribute is also part of data mining, especially for high dimensional datasets. We use attribute selection algorithm for the selection and removal of poor attribute.**

Index Terms — Missing data, Uncertain Data, Reinforcement Learning, Q-Learning, Discretization Algorithm, UCAIM, Poor Attribute Removal.

I. INTRODUCTION

It is well known to all of us that success of every data mining is strongly depend on data pre-processing. Data cleaning is an important part in the data pre-processing. Data cleaning is the process of correcting or removing inaccurate record from dataset [1]. After cleaning, the data become consistent with other data in dataset. Before storing that data in database we have to detect that missing values. We present a Q-learning method when some of the data are missing from that dataset.

A. Missing Data

We present a method of learning Q-functions, and demonstrate our methods on real multistage clinical trial data. Our approach uses Bayesian multiple imputation to build Q-functions using all of the observed data.

It is difficult to use traditional classification system, when data arrive sequentially and it is not in batch form [1]. Dirty data (unclean data) can cause confusion for the mining procedure. Data is missing due to following reason:-

1. Not giving so much importance to that field, hence ignore.
2. Because of misunderstanding about that field.
3. Equipment malfunctioning.
4. Incomplete and unlabeled incoming objects.

B. Uncertain Data

In computer science, uncertain data is notion of data that contain specific uncertainty. Uncertain data is the data which is not consistent.

Uncertainty is widely spread in real-world data. Numerous factors lead to data uncertainty includes:-

1. Data acquisition device error,
2. Approximate measurement,
3. sampling discrepancy,
4. Measurement inaccuracy,
5. Data integration error.

In many cases, estimating and modeling the uncertainty for underlying data is available and many classical data mining algorithms have been redesigned or extended to process uncertain data. It is extremely important to consider data uncertainty in the Discretization methods as well. In this paper, we propose a new Discretization algorithm called UCAIM (Uncertain Class-Attribute Interdependency Maximization). Uncertainty can be modeled as either a formula based or sample based probability distribution function (*pdf*). We use probability cardinality to build the quanta matrix of these uncertain attributes, which is then used to evaluate class-attribute interdependency by adopting there designed *claim* criterion. The algorithm selects the optimal Discretization scheme with the highest *claim* value. Experiments show that the usage of uncertain information helps UCAIM perform well on uncertain data. It significantly outperforms the traditional CAIM algorithm, especially when the uncertainty is high.

C. Poor Attribute Removal

Poor attribute Removal is an important topic in data mining, especially for high dimensional datasets. Poor attribute Removal (also known as subset selection) is a process commonly used in machine learning, wherein subsets of the attributes available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions

II. MATHEMATICAL MODULE

Set Theory:

1. U is main set of Users like u1, u2, u3.
   U= u1, u2, u3.
2. TD is main set of temporary database of user td1, td2, td3.
   TD= td1, td2, td3.
3. A is the set of actions like a1,a2,a3
   A= a1, a2, a3,
4. O is the set of observation for observed values for imputation.
   O= o1, o2 , o3,....
5. S is the set of States which are require in q-learning algorithm.
   S= s1, s2, s3,

Q (St, At) <- Q (St, At) + (Rt + MAX Q (St+1) Q (St, At)

Where,

Q(St,At) is the value function of the state-action pair (St,At) at moment t.

(Alpha) and (Gamma) are the learning rate and discount factor respectively.

Rt is the reward value.

Q-Learning Algorithm

```
1. Initiate arbitrarily all values.
2. Repeat (for each episode):
   a) Choose a random (initial) state .
   b) Repeat (for each step in the episode):
      i. Select an action according to the policy;
      ii. Execute the action A , receive immediate reward R, then observe the new state S.
      iii. Q(S,A) -> Q(S,A) + alpha(r +gamma MAX Q(S, A),Q(S,A))
      iv. S -> S
3. Until is one of the goal states.
4. Until the desired number of episodes have been investigated [1].
```

Uncertainty Algorithm

1. Find the maximal and minimal possible values of the uncertain attribute. A Fun, recorded as d0, dn.
2. Create a set B of all potential boundary endpoints. For uncertain attribute modeled in sample based pdf, simply sort all distinct possible values and use them as the set. For uncertain data modeled as formula based pdf, we use the mean of each distribution to build the set.
3. Set the initial Discretization scheme as D: [d0, dn], set GlobalUCAIM = 0.
4. Initialize k=1;
5. Tentatively add an inner boundary, which is not already in D, from B and calculate corresponding UCAIM value
6. After all the tentative additions have been tested, accept the one with the highest value of UCAIM.

7. If UCAIM > GlobalUCAIM or k<S, update D with the accepted boundary and set GlobalUCAIM = UCAIM, else terminate.
   Set k=k+1 and go to step 5 [4].

Poor Attribute Removal Algorithm

The basic Poor attribute removal algorithm is shown in the following:

Input:

S - Data sample with attributes X,|X| = n

J - Evaluation measure to be maximized

GS – successor generation operator

Output:

Solution – (weighted) attribute subset

L: = Start Point(X);

Solution: = {best of L according to J };

Procedure:

Repeat

- L := Search Strategy (L,GS(J),X);
- X' := {best of L according to J };
- if J(X')≥J(Solution) or (J(X')=J(Solution) and |X'| < |Solution|) then

```
Solution: =X';
Until Stop (J, L).
```

With the help of that algorithm, find the missing attribute and poor attribute.

III. EXISTING SYSTEM

In some systems because of unprocessed and unclean data, mining process is not efficiently worked. It consume lots of time [3]. And also data retrieval method becomes complicated. In some cases of missing value it take garbage value. Its create lots of confusion in data mining process [2].

Increasingly, reinforcement learning algorithms are being applied to problems that in the past would have been analyzed in the framework of classical statistics. Take clinical randomized trials for example: traditional randomized clinical trials compare two groups of patients (treatment and control) at a single point in time. Analyses of these trials use standard statistical approaches to deal with missing data (which is always a problem), and they provide scientists with well-established measures of confidence.

IV. PROPOSED SYSTEM

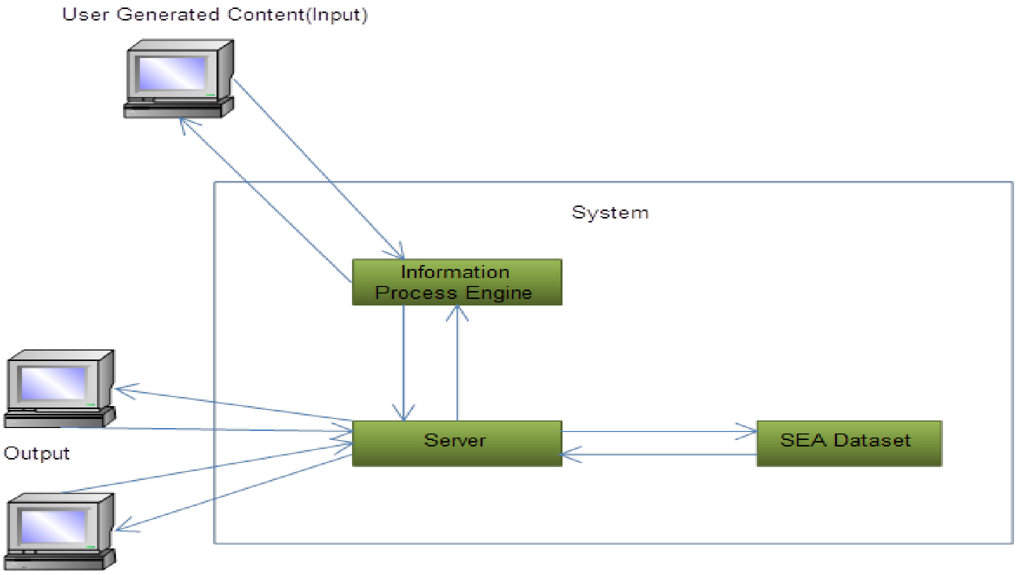

Figure 1: System Architecture

For efficient data cleaning following steps are to be implemented:

- Input Dataset: In this class we are taking the particular dataset as input from user.
- Load Data: In this class we are going to store whatever data we are taking from user as input.
- Save Observation: In this class previously present data, or history related to that particular application is maintain.
- Detection of missing value: In this we compare the user's given input with the saved observation of detecting missing values. We use Q-learning algorithm to find missing values.
- Uncertainty Checking: Under this function check that is there any uncertainty in dataset. We use UCAIM algorithm for finding the uncertain data[4].

In Implementation Detail we will follow the following four states to classify the problem:

State 1:- Incomplete object: the incoming object, xj, has one missing attribute in j.
State 2:- Complete object: the missing attribute of xj has been filled with either imputation approach.
State 3:- Classified complete object: x is classified with a confidence p higher or equal than a minimum predefined threshold, thmin (more on this later).
State 4:- Classified incomplete object: the incomplete object xj is classified, using the projection method, with confidence p thmin.

Actions defined in the algorithm are applied to the incoming object. They are related basically with the methods of dealing with incomplete data, classification, whether to incorporate or not the incoming object into the training set and returning to the initial state. Nine actions were included in the algorithm:

Action 1:- Projection: it classifies the incomplete object using the projection approach, which objects from the training set, T, are projected to one dimension less.
Action 2:- SVR: the missing attribute j is imputed using support vector regression (SVR). To this end, objects from T are used to build a new space using SVR. Then, xj is mapped to this space to estimate the value for the missing attribute.
Action 3:- 1NN: the missing attribute of xj is filled with the attribute j of its nearest neighbor (1NN).

Action 4:- Random: the missing attribute of xj is imputed with a random value estimated in the range between the minimum and maximum values of the attribute j from the training set objects.
Action 5:- Mean: the missing attribute of xj is filled with the mean value obtained from attributes j of objects from T.
Action 6:- Expert intervention: the expert provides the class label of the incoming object when the algorithm asks for it. The aid of the expert may be requested either for complete or incomplete objects, but it is only for the true class label and never for the missing attribute.
Action 7:- Classification: it classifies x using the confidence based on its k-NN.
Action 8:- Insert x into T: objects classified with thmin p thmax, where thmax is a maximum predefined threshold, are incorporated into T with the aim to insert only objects which may provide useful information.
Action 9:- Return to the initial state: if in the classification p < thmin the algorithm returns to the initial state to repeat the process using a different method of dealing with missing attributes.

These are steps and actions involving in the finding missing data and uncertain data.

V. CONCLUSION

We have shown how methods for missing data can be applied to batch reinforcement learning. We have explained why reasoning about missing data is important in theory and shown its effect in practice on the clinical dataset.

We also observe that the proper use of data uncertainty information can significantly improve the quality of data mining results. A lot a methods have been developed but data preprocessing still an active area of research.

VI. REFERENCES

[1] Missing Data and Uncertainty in Batch Reinforcement Learning (Daniel J. Lizotte Lacey Gunter Eric Laber Susan A. Murphy)
[2] Sutton, R.S., Barto, and A.G.: Reinforcement Learning: An Introduction. MIT Press (1998)
[3] Data Pre-Processing and Intelligent Data Analysis (Famili, Fazel; Shen, W.-M.; Weber, R.; Simoudis, E.)
[4] A Discretization Algorithm for Uncertain Data (Jiaqi Ge, Yuni Xia Department of Computer and Information Science)