Uncertain Ontology Model for Knowledge Representation and Information Retrieval Using Decision Rules

Sanjay Kumar Anand and Suresh Kumar

December 11, 2022

Uncertain Ontology Model for Knowledge Representation and Information Retrieval Using Decision Rules

1st Sanjay Kumar Anand
CSE, Netaji Subhas University of Technology, East Campus
Guru Gobind Singh Indraprastha University
New Delhi, India
anandsk19@gmail.com

2nd Suresh Kumar
CSE, Netaji Subhas University of Technology, East Campus
Netaji Subhas University of Technology
New Delhi, India
suresh.kumar@nsut.ac.in

***Abstract*—Knowledge plays a vital role for an effective operation and decision making. Relevance of Information Retrieval (IR) depends on an efficient Knowledge Representation (KR). Ontology constitutes rich-set of knowledge formalism for KR but inconsistency, vagueness, incompleteness etc., are major limitations shows uncertainty. In this paper, we have presented Rough Bayesian (RB) approach for uncertain ontology. Under this work, we have presented decision rules to determine attributes reduction, estimate the outcomes of a set of queries and decision class for PIMA Indians Diabetes Ontology. The model identifies a group of relational rules, reduct calculations, minimal rules and utilizes it in inferences and query retrieval with the help of probabilistic BN. The model captures the ontology's knowledge as a whole, and accurately predicts the average of belief with 91% accuracy for four queries in terms of precision, recall and accuracy and 98% accuracy of decision class on approximation.**

***Index Terms*—Information Retrieval, Knowledge Representation, Ontology, Decision rule.**

I. Introduction

Modeling the knowledge and retrieval of information depends on the proper representation of knowledge in a domain. Information retrieval (IR) [1] [2] deals with overall management of Information in knowledge base (KB). Since ontology provides a background knowledge to Semantic Web (SW) based applications with its strict rules and formal specification. But it is not able to handle uncertainty (i.e. incomplete, inconsistent, vague, missing or ambiguous knowledge) [3] [4] [5]. Ontology itself (at component level) or association of two or more ontology (i.e. Ontology mapping, matching, alignment etc.) are affected by any form of uncertainty [31], [6], [7] in knowledge representation (KR) and retrieval process of any SW based applications. Thus, ontology should be modeled and allow reasoning by its uncertain nature.

Many researchers focused their work to handle uncertainty and applied different methods to find better decision and produced solutions of their model used. Classic Logistic Regression [11] [12] [13], Decision Trees [14] [15] [16], and, more recently used techniques such as - Neural Networks (NN) [17] [18] and Support Vector Machines (SVM) [19] [20] are some of the techniques, they applied to perform various tasks such as inferences, classification, rules generation to evaluate the model accuracy.

Rough set [8] is an intelligent approximation method and provides a solutions to complex real-life problems. It mathematically processes the uncertain information. Bayesian model is a graphical knowledge modeling approach with a collection of nodes and directed edges that depends on dependent and independent relationships between random variable (nodes) [9]. In this paper, we have ensembled the strength of RS [10] and BN approach to model the uncertainty in PIMA Indian Diabetics ontology. The combined model identifies a group of decision rules and utilizes it in inferences and query retrieval from BN.

This paper is categorised in six sections. Section one is about Introduction. Section two explores related literature under Related work. Section three deals with Reviews of preliminary concepts. Section four presents the Proposed solution. Section five deals with Experimental result. Finally, section six summarises with conclusion and future work.

II. Related Work

There are several models available for modelling the knowledge, inferences and classification tasks using decision rules. Muchlinski et al (2016) [13] modelled the Civil War Onset Data and compared Random Forest (RF) with LR. A Classic LR is well described in [11]. de Souza, et al (2011) [12] proposed LR method to classify different pattern for interval data and evaluate the model for its usefulness. Song, Y. Y., & Ying, L. U. (2015) [14] proposed Decision Tree (DT) model to classify the attributes using decision rules. Brijain et al (2014) [15] conducted a survey of existing DT algorithms to classify and discover new patterns from large data sets. They applied various algorithms of Decision tree in terms of characteristic, challenges, advantage and disadvantage. Similarly, Charbuty, B., & Abdulazeez, A. (2021) [16] presented a comparative analysis and evaluated various DT algorithms in terms of advantages, disadvantages and data set used. More recently used technique such as- Neural Networks [18] are also focused. Gurney, K. (2018) [17] presented the detail overview on NN with mathematical approach. Similarly, Meyer, D., & Wien, F. T. (2015) [19] describe SVM method. García et al (2011) [33] proposed SVM and decision rules to discriminate the fuel classes. They identified overall accuracy (92.8%) and kappa coefficient of 0.9 to classify Multispectral and LiDAR data.

All the above models work well with crisp data, but they are less efficient to model uncertainty (inconsistent or ambiguous data) in a domain. The RS approach efficiently deals with uncertain knowledge. Due to such capability, it has attracted the interest among many researchers and practitioners. It plays a vital role in machine learning, intelligent system, expert system, knowledge discovery etc., to find hidden patterns in data, induction of learning approximations of concepts, constitution of knowledge discovery, attributes selection, reduction and extraction, rules generation and extraction and many more [10]. Thus, we have ensembled RS approach and BN model.

III. Reviews of Preliminary Concepts

A. Ontology

Ontology is one of the knowledge formalism technique to represent knowledge in term of RDF triples (subject, predicate, object) and modeling approach in SW technology, providing backbone to SW application/s [22] [23]. W3C defines an ontology as "the term used to describe and represent the area of knowledge" [24]. The core ontology includes knowledge, logical mapping among knowledge, lexicon and knowledge base.

B. Rough Set

Rough sets theory is a powerful method, algorithm, and mathematical instrument for detecting hidden patterns in data. The primary purpose of rough set analysis is the induction of learning approximations of concepts. It allows us to use ontology-based flexible information systems. Information system can also be referred as decision system or a knowledge system. An Information system of given ontology is defines as below -

$$S = (U, A) \tag{1}$$

where U and A are non-empty finite set of objects termed the universe and non-empty finite set of attributes respectively. In an information system, there is a function that computes f : U $\times$ A $\rightarrow$ V for every a $\in$ A. V_a is referred to as A's value set. The union of C and D refers to A, whereas the intersection of C and D is empty. C are known as conditional attributes, and D are known as decision attributes.

The Rough set produces equivalence classes from the given Information system [21]. All of the data tuples in an equivalence class are indistinguishable, which means that the samples are identical in terms of the attributes describing the data. Finally, lower and upper approximation of a data set are generated by indiscernibility relation.

C. Bayesian Network Model

A Bayesian network (BN) [35] is a probabilistic graphical model to describe knowledge in an uncertain domain, where each node and edge represents a random variable and conditional probability for the connected random variables respectively [29]. A BN model [30] can be considered as a directed acyclic graph (DAG) G = (V, E), where x_i (random variable) for each node $i \in V$ [29], holds the conditional probability distribution (CPD) P(x_i | x A_i), and specify the probability of x_i conditioned on its parents' values. The nodes are assigned discrete random variables V = X_1, X_2, ..., X_n, while the degree E represents the causal probabilistic relationships among the nodes. The conditional independence, joint probability, marginalization rules are the foundation of BN model's inferences.

IV. Proposed Solution

Rough set is commonly used mathematical method for removing unnecessary attributes by rule base, classification, and approximation of knowledge system for model diagnosis [27]. In this work, ontology is used as knowledge system. Using probability theory, BN processes the information and deal with uncertainty between rule bases. The probability computations speed up the search result for matching rules. As a result, in this paper, Rough Sets and Bayesian Networks are combined. The model is further described as follows:

A. Model Architecture

Model architecture is presented in figure 1. At initial stage, domain ontology of PIMA Indian Diabetes[1] is converted and constructed to ontology knowledge system. The dataset includes 768 observations and 8 attributes names as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI (Body Mass Index), Diabetes Pedigree Function, Age, and Outcome (Decision class). This knowledge system is treated as training data set, is fetched from PIMA Indian Diabetes ontology. Decision rules are generated based on attributes of reduct set. Using final decision table, we measure conditional probability and perform query retrieval. The step-by-step processes and its descriptions are mentioned in subsequent section.

B. Steps used in proposed model

We have first used PIMA Indian Diabetes ontology as knowledge base to initiate and evaluate the model. In preprocessing step, the model is initially used to deal with missing values and inconsistencies in Information system. The intermediate steps of model are described below.

- **Step-1: Construction of Ontology knowledge base/system.** In this step, we have made decision table for PIMA Indian Diabetes based on patient/s records in knowledge system and assign YES for the presence value of each patient's attribute. If no attribute value presence denotes it NO. Similarly, Normal, High and Very High are assigned to attributes. For example, if patient has glucose, then YES value is assigned and if patient has insulin level, then we assign Normal, High,or Very High as according to the patient's record.

[1] https://data.world/data-society/pima-indians-diabetes-database

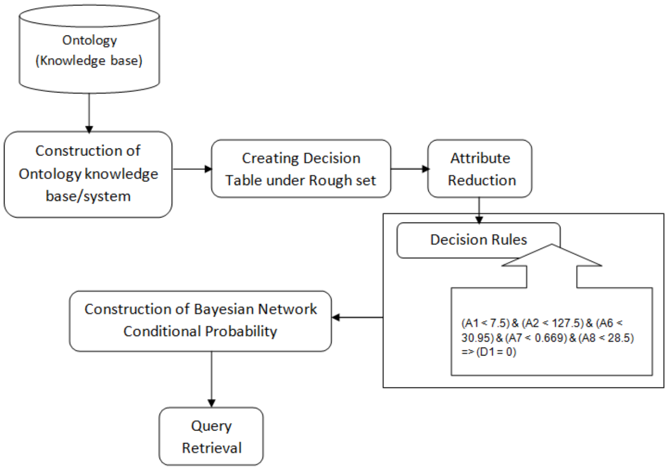
Ontology
(Knowledge base)
Construction of Ontology knowledge base/system
Creating Decision Table under Rough set
Attribute Reduction
Decision Rules
(A1 < 7.5) & (A2 < 127.5) & (A6 < 30.95) & (A7 < 0.669) & (A8 < 28.5) => (D1 = 0)
Construction of Bayesian Network Conditional Probability
Query Retrieval

Fig. 1. Model Architecture.

- **Step-2: Creating Decision Table under Rough set.** Based on Rough set theory, we assigned each non-numeric value given in step 1 to numeric form. For example, we assigned - Yes → 1, No → 0, Normal → 0, High → 1, Very High→ 2.
- **Step-3: Attribute Reduction.** In this part, we calculated the equivalence classes and relative reduct. After performing equivalence and relative reduct [25], we have found the reduction of attributes (C_1, C$_3$). During the attribute calculation of all observation and objects, we have found the No of Reduct (3) in Table I and attributes statistics in Table II.
- **Step-4: Rules Generation and extraction.** We have obtained following rules, based on final decision table. $R_1 : C_1 = 1\cap C_3 = 1\rightarrow D = 1$, $R_2 : C_1 = 1\cap C_3 = 2\rightarrow D = 1$, $R_3 : C_1 = 0\cap C_3 = 1\rightarrow D = 1$, $R_4 : C_1 = 0\cap C_3 = 2\rightarrow D = 1$

  Rule (R$_1$) can be understood as if a patient has glucose and insulin, then he or she is diabetic patient. Decision (D) to 1 tells the presence of disease in a patient. In Pima Indian Ontology, there exist 768 objects, 9 attributes including decision attribute (D). Based on the combination of objects and attributes, decision rules are created in two decision class [0,1]. Different combination of attributes and objects, 444 rules are generated where from rules 1 to 355 exists in No decision class (0) whereas from rules 356 to 444 i.e., 88 rules are in Yes decision class (1). In addition to it, we have also measured the minimal rules for PIMA indian Diabetes in terms of decision rules and total objects classified are listed in Table III.
- **Step-5: Construction of BN probability.** From the final decision table, we also calculated the conditional probability mentioned in Table IV and BN diagram is generated, depicted in Fig. 2.

TABLE I
NO OF REDUCT

SNo	Reduct	Length
1	A1, A2, A7	3
2	A2, A3, A7	3
3	A1, A3, A7	3

TABLE II
ATTRIBUTES STATISTICS

SNo	Attribute	Frequency	Frequency (In %)
1	A1	2	66.67
2	A2	2	66.67
3	A3	2	66.67
4	A7	3	100

V. EXPERIMENTAL RESULT

This section shows experimental result to evaluate the model accuracy in terms of query retrieval and approximation.

A. Query Retrieval and Implication

In this section, we have evaluated the model's accuracy in terms of precision and recall for IR to check the effectiveness of the model [36]. Recall (R) is the proportion of relevant documents retrieved, and precision (P) refers to the proportion of retrieved documents that are relevant for a given query. Based on the Recall and Precision, the accuracy of the model is calculated using the following equation (2), (3) and (4).

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)} \tag{2}$$

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN} \tag{3}$$

$$Accuracy(F - Measure) = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{4}$$

TABLE III
MINIMAL RULES FOR PIMA INDIAN DIABETES

Rules	Total Objects Classified
(A1== 1) && (A2== 66)	6
(A2== 110)	6
(A1== 5) && (A8== 20)	5
(A2== 99)	17
(A1== 92) && (A5== 0)	6

TABLE IV
BAYESIAN NETWORK CONDITIONAL PROBABILITY.

S.No.	Event	Number of Transaction	Probability
1	G	$\lvert G\rvert$=3, n=8	p(G) = 3/8 =0.375
2	I	$\lvert I\rvert$= 6, n=8	p(I)=6/8= 0.75
3	D—G,I	$\lvert D, G, I\rvert$ = 2, —G,I— = 2	p(D—G,I) = 2/2 =1
4	D—G, $\neg I$	$\lvert D, G, \neg I\rvert$ =0, —G, $\neg I\rvert = 1$	p(D—G, $\neg I) = 0$
5	D—$\neg G, I$	$\lvert D, \neg G, I\rvert$= 2 , $\lvert\neg G, I\rvert$ = 4	p(D— ,I) =2/4 = 0.5
6	D	$\lvert D\rvert$ = 4, n = 8	p(D) = 4/8=0.5

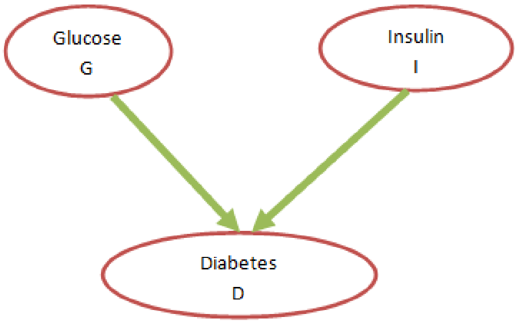

Fig. 2. Constructing Bayesian Network.

TABLE V
MODEL ACCURACY BASED ON PATIENT'S CASES.

Query	Precision (%)	Recall (%)	Accuracy (%)
Q_1	90.32	87.62	88.95
Q_2	92.71	91.89	92.30
Q_3	92.79	90.63	91.11
Q_4	93.74	92.35	93.04
Average	92.39	90.62	91.35

For query retrieval, we have applied probabilistic model to get more refine result in response to the query. The probabilistic model estimates the likelihood that a given document d_n will be relevant with respect to given query (q), indicated as $P(R \mid q, d_n)$, where R is a relevant decision. The query retrieval process initiates with a query by user enters into the model. The probabilistic model $P(R \mid q, d_n)$ can mathematically be calculated as mentioned in equation (5).

$$p(R \mid q, d_n) = \sum_{t_i \in d_n} W_{rc}\ p(t_i \mid Q) \tag{5}$$

The notions t$_i$ and W$_{rc}$ indicate the *Term* and value (weight) at each row and column respectively.

B. Model Accuracy

For testing the accuracy of model, we have taken case-based queries, and presented the outcomes of each query, mentioned in Table V. The Table VI shows the overall accuracy of model using approximation method and classified all observations and objects in two decision classes No (0) and Yes (1).

VI. CONCLUSION AND FUTURE WORK

In this study, we have presented a Rough Bayesian model to predict diabetes among patients records using the PIMA Indian Diabetes Ontology. This approach significantly generates decision rules based on attributes and observations as Reduct. The model determines the decision class using approximation method. Further, Bayesian model encapsulates domain ontology with Information Retrieval to perform better user response and query optimization. From Table V, our ensemble approach of rough set and Bayesian model give better accuracy (91%) and optimizing performance in query results. Also from Table VI, the model outperforms in classifying the objects with its boundary regions. Since model gives empirical result but we need to more improvement with maximum queries and different ontological dataset. For this purpose, We will implement clustering approach [34] in existing model in future work.

TABLE VI
MODEL ACCURACY BASED ON APPROXIMATION.

Decision Class	No. of Objects	Lower Approx-imation	Upper Approx-imation	Accuracy (In %)
0	500	500	506	98
1	268	262	268	97

REFERENCES

[1] Kobayashi, Mei and Takeda, Koichi, Information retrieval on the web, ACM Computing Surveys (CSUR), vol. 32, num. 2, pp. 144–173 (2000), publisher ACM New York, NY, USA.
[2] Manning et al., Introduction to information retrieval, Natural Language Engineering, vol.16, num. 1, pp. 100–103 (2010), publisher Cambridge university press.
[3] Smith, Barry, Ontology, The furniture of the world, pp. 47–68 (2012), Brill
[4] Guarino, et al. What is an ontology?. In Handbook on ontologies (pp. 1-17). Springer, Berlin, Heidelberg. (2009).
[5] Gangemi et al. Modelling ontology evaluation and validation. In European Semantic Web Conference (pp. 140-154). Springer, Berlin, Heidelberg (2006, June).
[6] Amrouch, S., & Mostefai, S. Survey on the literature of ontology mapping, alignment and merging. In 2012 International Conference on Information Technology and e-Services (pp. 1-5). IEEE. (2012, March).
[7] Haase, P., & Völker, J. Ontology learning and reasoning—dealing with uncertainty and inconsistency. In Uncertainty reasoning for the semantic web I (pp. 366-384). Springer, Berlin, Heidelberg. (2006).
[8] Yao et al. A review of rough set models. Rough sets and data mining, 47-75. (1997).
[9] Margaritis, D. Learning Bayesian network model structure from data. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science (2003).
[10] Pawlak, Zdzisaw, Rough set theory and its applications, Journal of Telecommunications and information technology, 7–10, (2002).
[11] Monroe, W. Logistic regression. Recall, 1(1) (2017).
[12] de Souza et al. Logistic regression-based pattern classifiers for symbolic interval data. Pattern Analysis and Applications, 14(3), 273-282 (2011).
[13] Muchlinski et al. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Analysis, 24(1), 87-103 (2016).
[14] Song et al. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130 (2015).
[15] Brijain et al. A survey on decision tree algorithm for classification, (2014).
[16] Charbuty, B., & Abdulazeez, A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28 (2021).
[17] Gurney, K. An introduction to neural networks. CRC press (2018).
[18] Aggarwal, C. C. (2018). Neural networks and deep learning. Springer, 10, 978-3.
[19] Meyer, D., & Wien, F. T. Support vector machines. The Interface to libsvm in package e1071, 28 (2015).
[20] Zhang, X. D. Support vector machines. In A Matrix Algebra Approach to Artificial Intelligence (pp. 617-679). Springer, Singapore (2020).
[21] Słowiński et al. Rough-set-based decision support. In Search Methodologies (pp. 557-609). Springer, Boston, MA (2014).
[22] Anand, S., & Verma, A. Development of Ontology for Smart Hospital and Implementation using UML and RDF. International Journal of Computer Science Issues (IJCSI), 7(5), 206 (2010).
[23] Flynn et al. The Knowledge Object Reference Ontology (KORO): a formalism to support management and sharing of computable biomedical knowledge for learning health systems. Vol. 2, No. 2, p. e10054 (2018).

[24] Antoniou, G., & Harmelen, F. V. . Web ontology language: Owl. In Handbook on ontologies (pp. 67-92). Springer, Berlin, Heidelberg, (2004).

[25] Meng, Z., & Shi, Z. Extended rough set-based attribute reduction in inconsistent incomplete decision systems. Information Sciences, 204, 44-69 (2012).

[26] Pawlak, Z. Rough set theory and its applications to data analysis. Cybernetics & Systems, 29(7), 661-688 (1998).

[27] Wang, J., & Miao, D. (1998). Analysis on attribute reduction strategies of rough set. Journal of computer science and technology, 13(2), 189-192.

[28] Friedman et al. Bayesian network classifiers. Machine learning, 29(2), 131-163 (1997).

[29] Cobb et al. Bayesian network models with discrete and continuous variables. In Advances in probabilistic graphical models (pp. 81-102). Springer, Berlin, Heidelberg (2007).

[30] Friedmanet et al. Bayesian network classifiers. Machine learning, 29(2), 131-163 (1997).

[31] Anand, S. K. & Kumar, S. Uncertainty Analysis in Ontology-Based Knowledge Representation. *New Generation Computing*. pp. 1-38 (2022).

[32] Robinson, L. & Jewell, N. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale De Statistique*. pp. 227-240 (1991).

[33] Garcıa, M., Riaño, D., Chuvieco, E., Salas, J. & Danson, F. Multispectral and LiDAR data fusion for fuel type mapping using Support Vector Machine and decision rules. *Remote Sensing Of Environment*. **115**, 1369-1379 (2011).

[34] Anand, S. K. & Kumar, S. Experimental Comparisons of Clustering Approaches for Data Representation. *ACM Computing Surveys (CSUR)*. **55**, 1-33 (2022).

[35] Sharma, A. & Kumar, S. Bayesian rough set based information retrieval. *Journal Of Statistics And Management Systems*. **23**, 1147-1158 (2020).

[36] Kumar, N. & Kumar, S. Querying RDF and OWL data source using SPARQL. *2013 Fourth International Conference On Computing, Communications And Networking Technologies (ICCCNT)*. pp. 1-6 (2013).