Integration of Healthcare domain Ontologies using Bayesian Networks

Monika P, G T Raju

Abstract*: Semantic Web (SW) was created with the vision of knowledge sharing. Knowledge from the past and present help predict the future with the use of Machine Learning (ML) algorithms. SW powered with ontologies help in realizing machine interactions supporting automated knowledge extraction. Healthcare as a field of medical domain gives lot of importance for timely accurate decisions with the available features. Representing existing information in terms of ontologies, retrieving the decisions upon establishing interaction between the relevant ontologies within the same domain, knowledge sharing & reusing the existing facts are of great benefit to the medical practitioners and researchers which has lot of open challenges to be resolved in order to realize the same. To address the stated issues, an algorithmic approach – Ontologies Integration algorithm using Bayesian Networks (OIBN) based on Bayesian Belief Networks (BBN) working on Naïve beliefs has been proposed which works on symptoms through the attributes of related ontologies within the same domain exploring the symptom dependencies and their probability of occurrences in combination. Selection of features for integration will follow the steps proposed in Sequential Forward Feature Selection algorithm (SFFS). The observation on the correctness of the presented method over diabetic datasets represented in ontological form with integration of relevant features reveals that the knowledge graphs have been efficiently explored discovering the facts based on the probability theory. The experimental results conclude that the proposed technique is showing enhanced prediction accuracy of 80.95% which is better compared to accuracies of the individual ontologies prior to integration and existing state-of-art technique.

***Keywords* : Semantic web, Ontologies, Ontology agents, Ontologies Integration, Health care, Diabetology, Domains.**

I. INTRODUCTION

Today's web is loaded with enormous information which is being generated every second in various formats pertaining to several domains. Semantic web manages actualities and significance associated inside the web meddling with the facts related to terms of available information. Healthcare in the sector of human services is one such space which contributes towards tremendous volume of information generation on daily basis. Knowledge graphs are extensively used in the recent years for representing data in machine readable form for automating the decision processes. Converting the existing facts, which are presently accessible in various formats like Electronic Medical Record (EMR), clinical diagnosis, pathology observations is a tedious job as it requires human intervention as part of semi-automated conversions. Lot of diagnostic conclusions solely depends on the symptoms and related close observations.

Construction of ontologies can be automated with compromise in the representation quality as semantic extraction of contents will be limited as per the technique employed. Semi-automated approaches with medical or domain experts' intervention guarantees the semantic relevance to certain extent. Generalizing and reusing the existing or constructed onto-graphs is a complex task which needs sophisticated ML algorithms for knowledge extraction, which has remained as a complicated challenge yet to be addressed.

Bayesian Belief Networks is an acyclic graphical representation simplifying mathematical model into graph of dependency between variables. With the advantage of easy to follow, and understanding inter-relationships of various attributes of the given dataset as features on which ontologies have been built; it has been extensively used to provide desired complexity by representing uncertainty of the predicted results of a model. Based on the evidence, the Naïve method estimates the degree of belief for the possible outcomes. Predicting future based on past observations lends itself naturally to applications requiring predictive analysis along with updating the belief with new evidence. Hence an algorithmic approach - Ontologies Integration algorithm using Bayesian Networks (OIBN) has been proposed which works on the concepts of Naïve beliefs on the symptoms treated as attributes of different ontologies in same domain for exploration of facts inter related between onto-graphs.

The rest of the paper is organized as follows: section 2 describes applications and techniques employed by other researchers towards medical knowledge representation, ontologies usage for knowledge extraction and diagnosis of diseases with ML concept applications. In section 3, the proposed algorithmic approach working on Naïve Bayesian Belief concepts and Sequential Forward Feature Selection algorithm for related feature extraction from various ontologies within domain followed with integration of ontologies for efficient knowledge extraction are briefed in detail. Section 4 summarizes the Observations and Results. Section 5 concludes, followed with references.

Revised Manuscript Received on December 12, 2019.
***** Correspondence Author
Monika P*, Research Scholar, R&D Centre, CSE Dept., RNS Institute of Technology, Assistant Professor, Dept. of CSE, Dayananda Sagar College of Engineering, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka. Email: monikamanjunath@gmail.com
G T Raju, Professor, Dept. of CSE, RNS Institute of Technology, Bengaluru, Visvesvaraya Technological University, Belagavi, Karnataka. Email: gtraju1990@yahoo.com

Retrieval Number: B10281292S19/2019©BEIESP
DOI: 10.35940/ijitee.B1028.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. RELATED WORK

Knowledge graph is termed as ontology in Semantic Web. Knowledge graphs enable the information to be read and understood by machines with the support of machine learning algorithms upon presenting the proof of automation by extracting the facts as human brains derive to certain level. Several applications working on ontologies have been noted in recent years particularly in the domain of medicine [1] for automated disease classification and conclusion of the diagnosis [2] based on the clinical observations. Automation through inference learning [3] in medicine has huge demand as it enables knowledge sharing across the world coupling the experience of experts over the practical fields. For a patient, diagnosis of disease can be done through various ways like clinical examination, physical examination, instrumental observations etc. Prior to determining the effective measures towards treatment, the experts should have clear knowledge of symptoms and the related diseases [4]. Learning of the relationships between the various symptoms as attributes and the diseases as conclusions from the ontological representation can be done easily at entity level discovery with the use of triplet representation [5], discovery of diseases through the symptoms in the graphical path [6] and through the mining algorithms [7].

From the survey, it is observed that the ontologies as triple store have been extensively applied with Bayes classifiers through various approaches in the general fields for classification of documents [8] and mapping of ontologies for information extraction [9-10]. In the literature, Naïve Bayes classifier has been used frequently in many clinical decision support tasks including curing of mammographic mass lesions [11] supporting the optimization of observations and treatments which has been considered on the truth of independence between symptoms. But independence amongst the symptoms is not always favorable for efficient diagnosis conclusion as there will be strong co-relation between the symptoms and related diseases [12] most of the time. In-order to ensure the completion of information in the available ontologies, either they should be enriched with additional knowledge through automated techniques [13] or to be reconstructed with probability details [14] along with a check of special constraints [15]. The data conversion from MySQL format to onto-graph can be done through the ready tool - Owlready [16] easily. Further many ontology development tools are available for machine supported construction of ontologies with care of correctness of attributes representation [17].

III. PROPOSED ALGORITHMIC APPROACH

Naïve Bayesian classifier is based on Bayes theorem (1) which computes probability of occurrence of one event under the circumstance of occurrence of the other event using prior knowledge. It is one of the strong probabilistic classifier which imposes assumptions of independence amongst the features of the dataset selected for experimentation. Prediction using ontologies work on logical reasoning, whereas prediction using Bayesian concepts work on probabilistic reasoning.

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} \quad (1)$$

Where *A* and *B* are 2 different events.

The combination of Naïve Bayesian classifier and Onto-graphs for disease diagnosis in medical domain remains unexplored to major extent. Bayesian Belief Networks has been widely used in medical domain as it handles uncertainty to maximum extent [18]. Hence an attempt is made to propose an approach for detection of diabetes by combining both the concepts with logical and probabilistic reasoning for effective diagnosis and knowledge sharing in the upcoming years. The architecture of the suggested idea is as shown in fig. 1. The algorithmic steps followed in the presented architecture are depicted in fig. 2.

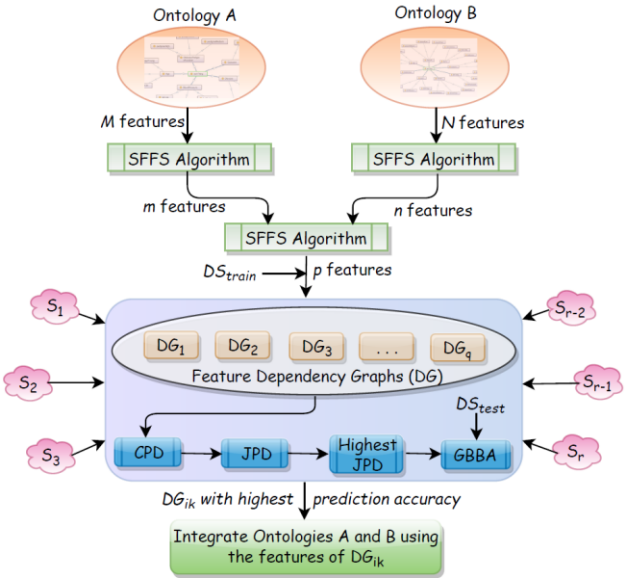


Fig. 1. Proposed Architecture for ontologies integration within domain

The approach of using BBN concept by combining 2 different ontologies in the same domain to enhance the prediction accuracy as stated in Ontologies Integration algorithm using Bayesian Networks (OIBN) follows the sequence of feature selection using Sequential Forward Feature Selection (SFFS) procedure as a first step. It is a famous wrapper method for selecting eligible features termed as *m* and *n* from both the ontologies. The eligible candidate features *p* across ontologies from the subsets *m* and *n* will be selected by reapplying SFFS method. Sample datasets for experimentation which is combination of inputs related to both the ontologies will be segregated into training ($DS_{training}$) and test (DS_{test}) sets. For a set of unique scenarios (S_1, S_2, S_3, …, S_r) considered individually one at a time, Feature Dependency Graphs $DG = \{DG_1, DG_2, DG_3, \dots, DG_q\}$ of all possible feature combinations to solve the scenario will be identified based on the learning from $DS_{training}$. For each and every *DG*, Conditional Probability Density (CPD) (5) function and Joint Probability Density (JPD) (6) function

based on the historical data and features selected as parameter for processing will be recorded. The features of particular DG_k of each scenario S_i with highest JPD will be forwarded for prediction accuracy evaluation using Gaussian Naïve Bayes technique (7). The features of DG_{ik} with best prediction accuracy will be finalized as best candidate features for integrating ontologies based on the appropriate ontological relationships.

The Sequential Forward Feature Selection algorithm (SFFS) in fig. 3 sequences the steps involved in selecting the eligible features *F'* from the given list of features *F* of a particular ontology. The python library package implementation: *Boruta* has set of functions to select the features using the concept of SFFS. *Boruta* receives the features list as input. The processing steps to filter the eligible features are as follows: Shadow Features (*SF*) will be added and dataset will be shuffled. Random forest classifier works on the *F* and *SF*. Importance of the features will be evaluated with the help of Z-score. Highly rated features will be retained and least rated features will be eliminated. The process repeats until all the features are processed or the classifier threshold is met. Finally the selected candidate features set *F'* will be returned to the OIBN for integration of ontologies to enhance the prediction accuracy.

Algorithm: Ontologies Integration algorithm using Bayesian Networks (OIBN)

Input: Ontology-1 (O_1), Ontology-2 (O_2), Sample dataset (DS)
Output: Integrated Ontology with better prediction accuracy

1. Let *M* be number of features of O_1 and *N* be the number of features of O_2
2. Let $m \subset M$ be the subset of features of O_1 selected for integration using SFFS
3. Let $n \subset N$ be the subset of features of O_2 selected for integration using SFFS
4. Let $p \subset \{m,n\}$ be the subset of features selected as candidates for integration using SFFS
5. Categorize the sample datasets into training (DS_{train}) and test (DS_{test}) sets
6. **for** $i = S_1$ *to* $S_r = |S|$ **do** /* $S_1, S_2, \ldots, S_r$ are set of unique scenarios */
7. Under the usage of DS_{train}
8. Compute possible Feature Dependency Graphs (DG) = $\{DG_1, DG_2, \ldots, DG_q\}$ using *p*
9. $\forall DG_k \in DG$ where *k=1 to q*, compute Conditional Probability Density (CPD) scores
10. $\forall DG_k \in DG$ where *k=1 to q*, compute Joint Probability Density (JPD) scores
11. Select the DG_k with highest JPD score
12. Using DS_{test} and Gaussian Bayesian Belief Algorithm (GBBA), tabulate the Prediction Accuracy of DG_k.
13. **end for**
14. Select the highest Prediction Accuracy score with the corresponding DG_{ik}
15. Integrate the ontologies A and B using the features of DG_{ik} with appropriate ontological relationships.
16. Return the integrated ontology

Fig. 2. Algorithm for integrating ontologies using Bayesian Networks

Algorithm: Sequential Forward Feature Selection Algorithm (SFFS)

Input: Set of features (F = {$f_1, f_2, f_3, \ldots, f_n$}) of a particular dataset, number of features *n*, objective function *boruta()*
Output: Best subset of features producing good prediction accuracy (F')

1. Let $|n| \in F$ be number of features of a particular Ontology *O*
2. *F' = Call boruta(n,F)*
3. Return the best set of features stored in F'

F' boruta(int n, features F)
1. $\forall m \in M$, Add randomness to the dataset by creating shadow features (SF) and shuffling
2. Apply Random forest classifier on F and SF
3. Compute importance of the available features
4. Retain features with high Z-score
5. Repeat steps 2-4 until all features are processed or Random forest classifier limit is reached
6. Return F'

Fig. 3. Sequential Forward Feature Selection Algorithm (SFFS)

Given 2 symptoms, S1 and S2 as in fig. 4, the CPD function for (a), (b) and (c) can be computed as in (2), (3), and (4) concluding that the conditional probability function products jointly form the JPD scores.

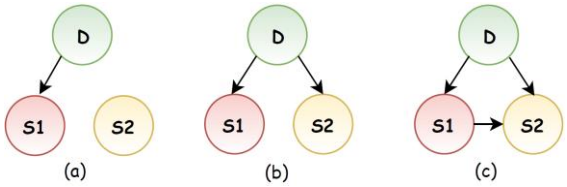


Fig. 4. Conditional Independence among the features

$$P(S_1, S_2, D) = P(D/S_1)P(S_1) \quad (2)$$

$$P(S_1, S_2, D) = P(D/S_1, S_2)P(S_1)P(S_2) \quad (3)$$

$$P(S_1, S_2, D) = P(D/S_1, S_2)P(S_1/S_2)P(S_1) \quad (4)$$

Summarizing the CPD, for any given Bayesian network, which is a directed acyclic graph, CPD for each vertex s_i in the given set of vertices S is defined as

$$P(S_1, S_2, S_3, \ldots, S_n/D) = \frac{P(S_1/D)P(S_2/D)\ldots P(S_n/D)P(D)}{P(S_1)P(S_2)\ldots P(S_n)} \quad (5)$$

Where $S_1, S_2, \ldots, S_n$ are the states of a node S in a DAG. As denominator in (5) will be 1 remaining as constant for any given inputs, the JPD of the states can be concluded as product of the CPD as shown in (6).

$$D = P(S_1, S_2, S_3, \ldots, S_n) = argmax_D\, P(D) \prod_{i=1}^{n} P(S_i/D) \quad (6)$$

Where D is set of parent nodes which decides the features set combination which has maximum probability of prediction as the candidates for ontology merging.

$$P(x_i/y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (7)$$

Where, $x=\{x_1, x_2, \ldots, x_n\}$ are probable features for measuring the quality of the output predictions *y*, σ measures standard deviation and μ measures mean of all possible predictions present in *y*.

IV. OBSERVATIONS AND RESULTS

As proof of concept, couple of sample computations has been done on the ontologies built using two different diabetic datasets: diabetes prediction during pregnancy and diabetic type prediction in common man, obtained from University of California, Irvine (UCI) machine learning repository [19]. Dataset – 1 is a Pima Indian Diabetic dataset with 768 records, 8 attributes namely Number of times pregnant, Plasma glucose concentration at exact 2 hours in an oral glucose tolerance set, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2- Hour serum insulin

(mm U/ml), Body mass index (weight in kg, height in m), Diabetic pedigree function, Age (years) and a predictor class (0 – non diabetic, 1 – diabetic). Dataset – 2 is a diabetic type predictor set with 1010 records, 8 attributes namely Age (years), Blood Sugar Fasting(mmol/L), Blood Sugar Post Prandial (mmol/L), Plasma R (mmol/L), Plasma F (mmol/L), HbA1c (average 3 months blood sugar mg/dL), Diabetes Type (Normal, Type 1, Type 2) and a predictor class (0 – non diabetic, 1 – diabetic).

Ontologies were constructed using protégé [20] ontology editor which is open source software from Stanford University to analyze the nature of the features and their dependencies through onto-graphs. Analyzing the datasets which have numerical values, it's understood that the samples are continuous in nature. Hence Gaussian Naïve Bayesian network method (7) has been used for analyzing the prediction accuracies. Ontologies before integration are as shown in figures 5 and 6.

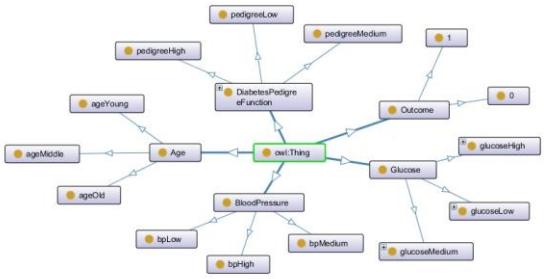


Fig. 5. Onto-graph of Diabetic dataset-1

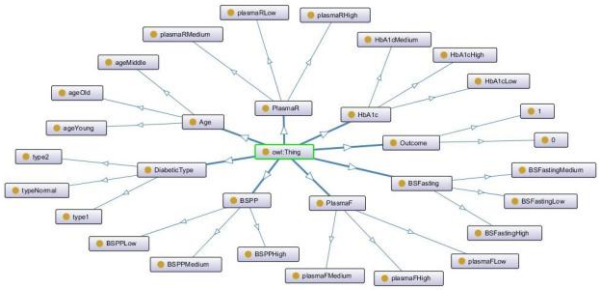


Fig. 6. Onto-graph of Diabetic dataset-2

The output of Sequential Forward Feature Selection algorithm applied on the features of both the ontologies in the form of feature list ratings ordered as per the importance of feature towards good prediction accuracy of individual ontologies are depicted in table I. The eligible candidate features from both the ontologies chosen for integration are tabulated in table II. Rest of the attributes has been neglected as they were contributing towards deprecating the prediction accuracy or remained as no change in prediction accuracy.

Among the list of features in table I, Diabetic Pedigree Function (DPF) feature from dataset-1 and Blood Sugar Fasting (BSF) & HbA1c (HbA1c) features from dataset-2 were selected as candidate features for integration of ontologies based on highest JPD and Gaussian Belief Networks accuracy score observations as the computed results were satisfactory. The 4th feature BSPP from table II has been stepped away from the conditional independence consideration as inclusion of the same resulted no change in the performance for the current dataset.

Table-I: Feature ranking of both the ontologies based on SFFS algorithm

Rank	Features of Dataset -1 (m)	Features of Dataset -2 (n)
1	Plasma Glucose Concentration (PGC)	Blood Sugar Fasting (BSF)
2	Age (A1)	Blood Sugar Post Prandial (BSPP)
3	Diabetic Pedigree Function (DPF)	HbA1c (HbA1c)
4	Diastolic Blood Pressure (DBP)	Age (A2)

The Conditional Independence among the candidate features BSF, HbA1c and DPF from both the ontologies is represented in fig. 7. Feature DPF and Feature BSF are considered as conditionally independent though selected from different ontologies. The features DPF and BSF causes variation in the values of feature HbA1c resulting in the conclusion of diabetes presence or absence in the samples.

Table-II: Candidate Feature ranking based on SFFS algorithm

Rank	Candidate Features for Ontologies Integration(p)
1	Blood Sugar Fasting (BSF)
2	HbA1c (HbA1c)
3	Diabetic Pedigree Function (DPF)
4	Blood Sugar Post Prandial (BSPP)

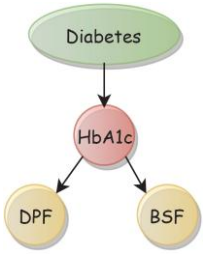


Fig. 7. Conditional Independence of candidate features belonging to 2 diabetic Ontologies

The possible CPD and JPD computations for the test dataset of size 400 are as shown in table III and IV. Ontologies were combined with the concept of added features and relationships like *hasSymptom*, *causes*, *leadsTo*. Sample scenario is formulated as: "*An aged patient having complaints of high Blood Sugar Fasting, average Diabetic Pedigree Function and border line high average blood sugar level (HbA1c)*".

Sample ontology triplets are as follows:
(Patient, *hasSymptom*, DPFAverage)
(Patient, *hasSymptom*, BSFHigh)
(DPFAverage, *causes*, HbA1cHigh)
(BSFHigh, *causes*, HbA1cHigh)
(HbA1cHigh, *leadsTo*, DiabetesYes)

Similarly, rest of the triple stores was finalized upon the possible scenarios. The performance of the proposed technique has been evaluated based on the observations made on dataset of sizes 50, 100, 200 and 400 by applying GBBA. Table V depicts the performance readings of individual ontologies prior to integration. Performance of integrated ontologies was documented as in table VI for analysis. The performance accuracy of the integrated ontologies were measured using the metrics: Precision (8), Recall (9) and Accuracy (10), where TP is count of True Positive, TN is count of True Negative, FP is count of False Positive and FN is count of False Negative predictions.

Table-III: Conditional Probability Distribution observations

Features		C		
A	B	L	M	H
A=L	B=L	0.895	0.05	0.055
	B=M	0.14	0.78	0.08
	B=H	0.0375	0.1425	0.82
A=M	B=L	0.8725	0.125	0.0025
	B=M	0.44	0.4725	0.0875
	B=H	0.0625	0.17	0.7675
A=H	B=L	0.4675	0.4825	0.05
	B=M	0.395	0.5025	0.1025
	B=H	0.005	0.2225	0.7725

Table-IV: Joint Probability Distribution observations

FEATURES		D=YES			D=NO			
		C=L	C=M	C=H	C=L	C=M	C=H	TOTAL
A	B							
A=L	B=L	4.9659E-05	2.8E-05	2.1E-05	0.00432	0.00072	0.00011	0.00525
	B=M	5.2323E-05	0.00155	0.00186	0.00138	0.00943	0.00018	0.01445
	B=H	0.00025062	0.00197	0.03856	0.00052	0.00187	0.00078	0.04395
A=M	B=L	0.00024061	0.00059	2.6E-05	0.00818	0.00136	3E-06	0.01039
	B=M	0.00054487	0.00458	0.00177	0.00336	0.01746	0.0006	0.02831
	B=H	0.00168061	0.00692	0.06313	4.9E-05	0.00087	0.00182	0.07448
A=H	B=L	0.00034379	0.00124	0.00162	0.00593	0.01047	0.00017	0.01978
	B=M	0.00163049	0.0056	0.0063	0.00293	0.03879	0.00038	0.05564
	B=H	8.3042E-05	0.01513	0.16373	7.9E-05	0.00088	0	0.17991

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (10)$$

Table-V: Accuracies before integrating Ontologies by applying GBBA

Dataset size	Accuracy Comparison (GBBA)					
	Ontology 1			Ontology 2		
	Precision	Recall	Accuracy (%)	Precision	Recall	Accuracy (%)
50	0.60	0.60	73.33	0.50	0.58	66.66
100	0.67	0.50	78.94	0.71	0.54	78.77
200	0.78	0.48	77.03	0.72	0.44	81.22
400	0.78	0.45	73.00	0.75	0.40	78.25
Average →	**0.71**	**0.51**	**75.58**	**0.67**	**0.49**	**76.23**

The graphical representations of the precision, recall and accuracy scores of the individual ontologies and integrated ontology for datasets of size 50, 100, 200 and 400 during prediction are presented in fig. 8. The average precision and recall scores have been depicted in fig. 9(a). It is observed that the integrated ontology is demonstrating high precision and low recall scores compared to individual ontologies witnessed on various sized datasets and on average readings. Increase in precision scores is concluded as correctness of classification of cases as predicted. Decrease in recall scores is concluded as reduction in misclassification rate as the test samples might have cases more into negative conclusion of the disease being diagnosed

Table-VI: Accuracies after integrating Ontologies with best features by applying GBBA

Dataset size	Accuracy Comparison (GBBA)		
	Integrated Ontology		
	Precision	Recall	Accuracy (%)
50	0.68	0.45	74.33
100	0.75	0.33	80.95
200	0.79	0.33	82.25
400	0.80	0.30	86.25
Average →	**0.76**	**0.35**	**80.95**

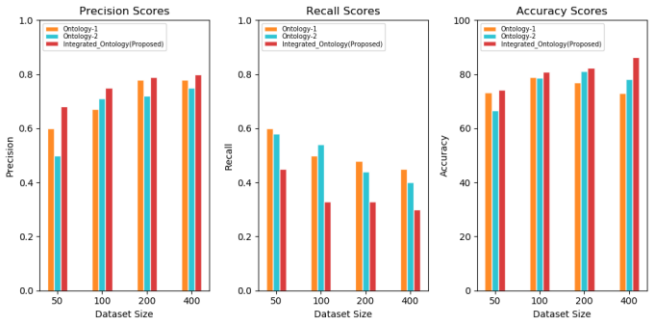


Fig. 8. Precision, Recall and Accuracy scores of individual and integrated ontologies

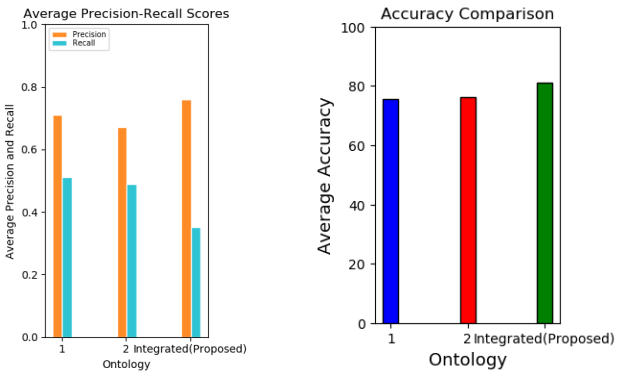


Fig. 9. (a) Average Precision & Recall scores of individual and integrated ontologies. (b) Average Accuracy scores of individual and integrated ontologies

Retrieval Number: B10281292S19/2019©BEIESP
DOI: 10.35940/ijitee.B1028.1292S19

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table-VII: Accuracies after integrating Ontologies with best features by applying C4.5

Dataset size	Accuracy Comparison (C4.5)		
	Integrated Ontology		
	Precision	Recall	Accuracy (%)
50	0.62	0.46	72.31
100	0.73	0.35	78.90
200	0.75	0.34	80.66
400	0.78	0.32	83.58
Average →	**0.72**	**0.37**	**78.86**

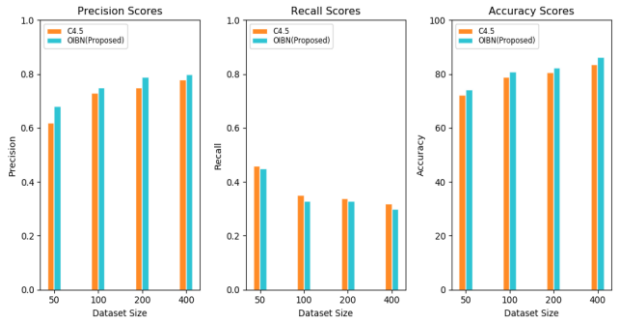
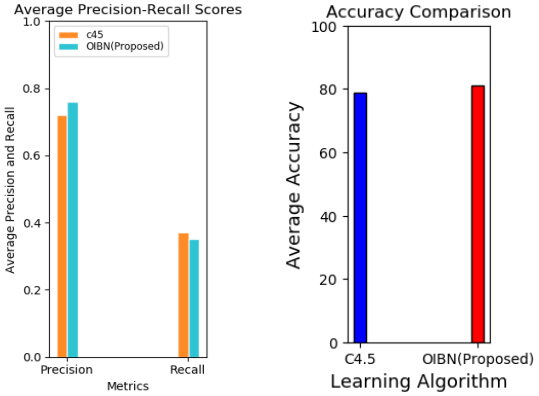


Fig. 10. Performance comparison between C4.5 and OIBN after integration of ontologies



(a) **(b)**

Fig. 11. (a) Mean Precision & Recall scores comparison of C4.5 and OIBN after integration of ontologies. (b) Mean Accuracy scores of C4.5 and OIBN after integration of ontologies

The average accuracy scores comparison of ontologies prior and after integration are as illustrated in fig 9(b). Performance of integrated ontology was evaluated using the well-known C4.5 decision tree algorithm and the observations were tabulated as in table VII. The plot of existing state-of-art and proposed technique prediction observations on various sized datasets in terms of precision, recall and accuracy of models' performance is as shown in fig. 10. Fig. 11(a) and fig. 11(b) depicts the mean precision, recall and accuracy readings comparison between C4.5 and OIBN (proposed) algorithms. From the documented results it is evident that the accuracy scores of the integrated ontology is better with the average score of 80.95% compared to accuracies of the existing state-of-art technique and independent ontologies.

V. CONCLUSION

Today's web functions with semantic applications using knowledge graphs rather than schema tables. Understanding the knowledge graph requires intelligence in the form of machine learning algorithms. ML algorithms perform well in understanding the past and predicting the future in the presence of ontologies. Ontologies are the triple store representation of information which provides easy access for ML techniques resulting in better prediction accuracy. In the field of medicine, lot of challenges has been observed for representing the existing knowledge in machine readable form and reusing the same when needed. Ontologies are the best choice to represent the knowledge in the form of .rdf, .owl, .ttl etc., formats which ease the reading task by machines. The challenge of extracting information from different ontologies related to the same domain has been addressed in the proposed algorithm using the concept of Naïve Bayesian Belief Networks.

The proposed steps (OIBN) were experimented on two different diabetic datasets for relevant feature selection using Sequential Forward Feature Selection algorithm applied individually on ontologies chosen for integration. The selected subsets of features were again meddled with SFFS algorithm for combined candidate features selection. Using the training datasets, candidate features and different scenarios, DG_s were designed followed with computation of CPD and JPD functions. Highest JPD combinations were tested with Gaussian Naïve Bayes algorithm and test datasets to predict the accuracies. The features combination with highest prediction accuracy was chosen for ontology integration based on relevant onto – relationships.

The experimental results reveal that the proposed technique is serving as reliable proof with enhanced accuracy of prediction promised if the relevant ontologies within the same domain are integrated based on the candidate features selected as additional features. The average enhanced prediction accuracy of 80.95% is obtained after integration, which is better than the prediction accuracies of the individual ontologies and existing state-of-art technique thereby concluding that the knowledge graphs have been proficiently explored discovering the facts based on the probability theory.

REFERENCES

1. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, Smolinski RJ, Wind WM, "Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain", American Journal of Sports Medicine, 2014, PP. 42(10):2371–6, DOI: 10.1177/0363546514541654.
2. Power D, Sharda R, Burstein F, "Decision support systems", New Jersey: John Wiley & Sons; 2015.
3. Zhu J, Fung GPC, Lei Z, Yang M, Shen Y, "An in-depth study of similarity predicate committee", Journal of Information Processing and Management (Elsevier), 2019; PP. 56(3):381–393.
4. Seidenberg J, Rector A, "Web ontology segmentation: analysis, classification and use", 15th International Conference on World Wide Web, May 22–26, Edinburgh: ACM, 2006. PP. 13–22, DOI: 10.1145/1135777.1135785.

5. Yin X, Tan W, "Semi-supervised truth discovery", 20th International Conference on World Wide Web. ACM, 2011. PP. 217–226, DOI: 10.1145/1963405.1963439.
6. Cheng Li, Santu Rana, Dinh Phung, Svetha Venkatesh, "Hierarchical Bayesian Nonparametric Models for Knowledge Discovery from Electronic Medical Records", Journal of Knowledge-Based Systems (Elsevier), 2016, PP. 168–82, DOI: 10.1016/j.knosys.2016.02.005.
7. Julien Tourille, Olivier Ferret, Aurelie Neveol, Xavier Tannier, "Neural Architecture for Temporal Relation Extraction: A Bi-LSTM Approach for Detecting Narrative Containers:, Proceedings of the 55th annual meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2017. PP. 224–30, DOI: 10.18653/v1/P17-2035.
8. Chang YH, Huang HY, "An automatic document classifier system based on naive bayes classifier and ontology", IEEE international conference on Machine learning and cybernetics, 2008; PP. 3144–3149.
9. Kim H, Chen SS, "Associative Naive Bayes classifier: Automated Linking of gene ontology to medline documents", Journal of Pattern Recognition (Elsevier), 2009, PP. 1777–1785, DOI: 10.1016/j.patcog.2009.01.020.
10. Choi N, Song IY, Han H, "A survey on ontology mapping", ACM SIGMOD Record, 2006, PP. 34–41, DOI: 10.1145/1168092.1168097.
11. Benndorf M, Kotter E, Langer M, Herda C, Wu Y, Burnside E, "Development of an online, publicly accessible naive Bayesian decision support tool for mammographic mass lesions based on the American College of Radiology (ACR) BI-RADS lexicon", European Society of Radiology (Springer), 2015, PP. 1768–1775, DOI: 10.1007/s00330-014-3570-6
12. Wu J, Cai Z, Pan S, Zhu X, Zhang C, "Attribute weighting: how and when does it work for Bayesian network classification", IEEE International Joint Conference on Neural Networks (IJCNN), 2014, July 06–11, Beijing (China), New York, PP. 4076–4083.
13. Moon C, Jones P, Samatova NF, "Learning Entity Type Embeddings for Knowledge Graph Completion", ACM Conference on Information and Knowledge Management, 2017, PP. 2215–2218, DOI: 10.1145/3132847.3133095.
14. Jiang Jingchi, Li Xueli, Zhao Chao, Guan Yi, Yu Qiubin, "Learning and inference in knowledge-based probabilistic model for medical diagnosis", Journal of Knowledge-Based Systems (Elsevier), 2017, Volume 138, PP. 58–68, DOI: 10.1016/j.knosys.2017.09.030.
15. Chekol MW, Pirrò G, Schoenfisch J, Stuckenschmidt H, "Marrying Uncertainty and Time in Knowledge Graphs", AAAI Conference on Artificial Intelligence, 2017, PP. 88–94.
16. Lamy JB, "Owlready: ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies", Artificial Intelligence in Medicine (Elsevier), 2017, volume 80, PP. 11–28, DOI: 10.1016/j.artmed.2017.07.002.
17. Shen Y, Wen D, Li Y, Du N, Zheng HT, Yang M, "Path-based attribute-aware representation learning for relation prediction", SIAM International Conference on Data Mining: Society for Industrial and Applied Mathematics, 2019, PP. 639–647, DOI: 10.1137/1.9781611975673.72.
18. Tore Bruland, Agnar Aamodt, Helge Langseth, "Architectures Integrating Case – Based Reasoning and Bayesian Networks for Clinical Decision Support", international Conference on Intelligent Information Processing V (Springer), 2010, PP: 82-91, DOI: 10.1007/978-3-642-16327-2_13.
19. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
20. Musen, M.A, "The Protégé project: A look back and a look forward", AI Matters, Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.

AUTHORS PROFILE

Monika P has received M. Tech. (CSE), Degree from Visvesvaraya Technological University (VTU), Belagavi, Karnataka in 2011. Currently pursuing research at Department of Computer Science & Engineering, RNS Institute of Technology, Bengaluru, Karnataka – 560 098, affiliated to VTU and working as Assistant Professor in Department of Computer Science & Engineering at Dayananda Sagar College of Engineering, Bengaluru, Karnataka – 560 078. She has published research papers in reputed International Journals and conferences. She has 10.9 years of teaching experience and 1.6 years of Industry experience. Her areas of research interests include Web Mining, Semantic Web, Artificial Intelligence and Machine Learning.

Dr. G T Raju has received M.E. (CSE), Degree from Bangalore University in 1995 and Ph. D (CSE) from Visvesvaraya Technological University (VTU), Belagavi, Karnataka in 2008. Currently working as Vice-Principal, Professor & Head in the Department of Computer Science & Engineering , RNS Institute of Technology, Bengaluru, Karnataka – 560 098. He has 24 years of teaching and research experience. His areas of research interests include Web Mining, Semantic Web, Artificial Intelligence, Machine Learning, Knowledge Data Discovery, Internet of Things, Image Processing and Pattern Recognition. He has published 100+ research papers in reputed International Journals and conferences. He has authored 5 technical text books. He has completed two funded projects. 10+ Research Scholars have been awarded Ph. D degree under his supervision.