Data Pre-Processing and Customized Onto-Graph Construction for Knowledge Extraction in Healthcare Domain of Semantic Web

Monika P, G T Raju

Abstract*: Present electronic world produces enormous amount of data every second in various formats, especially in healthcare units. To efficiently utilize the available data by representing it in the machine readable form, the concept of Semantic web stepped in progressing towards automated knowledge discovery process. In this paper, comprehensive pre-processing techniques have been proposed for preparing the raw data to be presentable in structured format so as to construct the onto-graph for selected features in a health care domain. Cluster based Missing Value Imputation Algorithm (CMVI) has been proposed to enhance the quality of the imputed data which is the most important step during data pre-processing. Missing values were randomly induced into the Pima Indian Diabetic dataset with the missing ratio of 1%, 3% and 5% for each attribute up to 50% of the attributes in the original diabetic dataset. The experimental observations reveal that the quality of the pre-processed data is better compared to raw, unprocessed data in terms of imputation accuracy measured against coefficient of determination (R^2), Index of agreement (d_2) and Root Mean Square Error (RMSE).Documented results proved that the proposed techniques are comparatively superior than the traditional approaches with increased R^2 & d_2 and decreased RMSE scores. Further, importance of knowledge graph and various ontological representation types are discussed in short as construction of .owl file is the first step towards automation in semantic web.

Keywords* : Semantic web, Ontologies, Ontology agents, Onto-graph, Knowledge graph, Knowledge extraction, Data pre-processing, Handling missing values, Missing Value Imputation, Health care, Diabetology.

I. INTRODUCTION

In order to enhance the search and response capabilities of Google semantically, technology started advancing from various sources in the direction of usage of knowledge base called Knowledge graph during May, 2012. Since then research and experimentations grew rapidly exploring the technology of Knowledge graph enhancing search quality and experience in various disciplines heading towards automated knowledge extraction process connecting the fields of Artificial Intelligence and Machine Learning. However due to the limited exposure about the technology constructs and confidentiality of data, the knowledge graph construction procedure is quite intricate [1].

Knowledge graphs also known as Onto-graphs or Ontologies in general can be developed through top down approach focusing on schema & domain ontologies or through bottom up approach focusing on instances of knowledge like Linked Open Data datasets storing in Resource Description Framework [RDF] or graph database formats [2]. In the field of data mining, data representation in the graph format benefits knowledge acquisition through interaction between the ontologies to a large extent. Lot of challenges has been thrown in the field of medicine for automating the knowledge discovery, there by benefiting the efficient utilization of the available medical data as means of diagnosis and treatment all over the world.

Publicly available ontologies including Yet Another Great Ontology (YAGO), WordNet, Google Knowledge Vault, YAGO2, Friend Of A Friend (FOAF) etc. can be integrated with the customized ontologies during construction process for standardizing the knowledge graphs quality. Most frequently used development tools include Protégé, Open Sesame Framework, Neon toolkit, Java ontology editor, OntoStudio, Swoop and Allegrograph. Zotero and Mendeley are client appplications consuming machine readable metadata for semantic conclusions. GraphDB, MongoDB and Neo4j are the widely used knowledge graph Storage formats for representing and querying the knowledge semantically.

In this paper, an attempt is made to pre-process the diabetic related data collected from UCI machine learning repository for construction of onto-graphs. Right from data collection & cleaning, relevance analysis, integration, missing values imputation, transformation and construction of Ontologies till storage of knowledge graph, a comprehensive approach has been employed for pre-processing so as to make information retrieval process effective with the help of query engines benefitting the medical researchers and practitioners. Figure 1 illustrates the flow diagram of data pre-processing and customized ontology construction steps for knowledge extraction in diabetic healthcare domain of semantic web.

The work includes collection of syntactic data from the web containing structured, semi-structured and un-structured data in the domain of Healthcare specifically with the subdomain of Diabetology.

Revised Manuscript Received on September 2, 2019.

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

Data collected will be pre-processed further by cleaning, handling missing data with the proposed Cluster based Missing Value Imputation (CMVI) algorithm and verified for data redundancy, inconsistency, reduction & transformation.

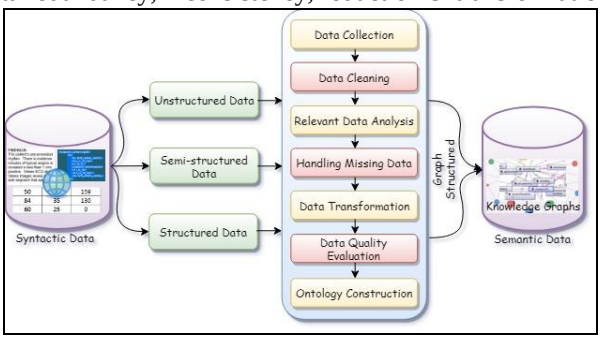

Fig. 1 Flow diagram of Data Pre-processing

The data quality is evaluated with the standard performance evaluation metrics such as coefficient of determination (R^2), Index of agreement (d_2) and Root Mean Square Error (RMSE). Finally the Onto-graph is constructed and stored in respective format. With the enhanced data quality, automated information retrieval process can then be initiated towards extraction of knowledge.

The rest of the paper is organized as follows: section 2 describes the algorithms and techniques employed by other researchers towards pre-processing the data prior employing Machine Learning techniques for knowledge extraction. In section 3, the data pre-processing steps, suggested missing value imputation algorithm, other data handling techniques, data quality check and ontology construction process are briefed in detail. Section 4 concludes the paper with summary of the results and conclusions.

II. RELATED WORK

Constructing Knowledge graphs from scratch is a challenging task as technologies to be involved are not completely detailed due to complexity of data availability in heterogeneous and multilingual formats, throwing lot of challenges [3-4]. General sequence to be followed includes knowledge acquisition, fusion, storage, query and visual display [2]. During the data construction process, handling missing data plays a major role as quality of the data gets enhanced to the maximum extent. Various techniques employed to handle missing data include deletion of the records leading towards information loss, averaging the values may induce irregularities in the prediction sequences with lot of systematic differences [5]. However the probability of missing values for an attribute may depend on that particular attribute's values to a major extent [6].

Several imputation techniques have been used in practice. The list includes mean, median, most common occurrence, most frequently used etc. Anomaly detection with missing values considers all features during learning and prediction time [7]. Hotdeck imputation is a technique which replaces the missing data by the attribute value from the input set closest to both the patterns [5]. Fuzzy k-means clustering technique works on the degree of clustering. Every unreferenced attribute will be replaced on the basis of attribute membership degree and centroid of the cluster which may result in high computation time [8-10].

Support Vector Machine (SVM) imputation takes the condition and decision attributes for predicting the missing values with a good memory management skill but may result in poor performance if number of features are more than the sample set [11]. Expectation Maximization (EM) Algorithm is one of the imputation techniques which keep trying with the replacement of imputed values in a scaled incremental manner until expected accuracy is reached. This method is lot of time consuming and costly due to repeated testing.

Md. Geaur Rahman et. al. [12] have innovated two novel techniques for categorical and numerical missing values imputation based on decision trees and forests for segmenting the dataset with highest similarity and attribute correlations. Further authors have experimented on merging couple of segments to maximize the quality of the imputed values. Results conclude that the proposed method is far superior in achieving the filling of missing values. Decision tree based missing value imputation coupled with EM algorithm promises better imputation accuracies compared to standard imputation techniques [13]. However the execution time of the EM may deteriorate the model as the code runs into n number of iterations until completion.

Deep learning technique can be employed for imputing non numerical data in tabular format including unstructured texts resulting in better quality of the imputed data compared to simple methods [14]. Irregularly sequenced multivariate time series missing values can be imputed using recurrent neural networks automating the machine interactions with humans, trustable to certain extent [15]. Imputing missing values using Mean computation, Decision tree, K Nearest Neighbors and self - organizing maps and being evaluated using Bayesian Rule learning, it's evident that the learning model demonstrate robustness towards learning as quality of data is enhanced [16].

III. DATA PRE-PROCESSING

Data mining as a knowledge discovery process involves data gathering, cleaning, relevant analysis, integration, selection & transformation, pattern discovery and evaluation followed by knowledge representation. Diversity of underlying databases with various complexities and data representations offer great challenges during the mining process. Data pre-processing is one of the important step in the process of data mining which involves transforming raw data into machine understandable format thereby eliminating program inconsistencies and erroneous behaviors. The detailed pre-processing steps followed in the proposed work are elaborated in the subsequent subsections.

A. Data Collection

Gathering datasets with variety of attribute types consisting of nominal, binary, ordinal or numeric values initiates the pre-processing procedure. Data can be collected from internet sources like webpages, wiki source, readily available ontologies, medical hubs like NCBI, UCI machine learning repositories etc.

Pima Indian Diabetic dataset collected from UCI machine learning repository [17] has been used for experimentation and drawing the observed conclusions in the presented work. The dataset contains around 768 samples consists of 8 attributes with a diabetic predictor class. The attributes include Number of times pregnant, Plasma glucose concentration at exact 2 hours in an oral glucose tolerance set, Diastolic Blood Pressure (BP) (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 Class Variable (0 – non diabetic, 1 – diabetic).

B. Data Cleaning

Data collected from real world will be generally incomplete and noisy in majority of the cases. Incomplete data are termed as missing values and may be caused due to human errors, type mismatches, for security reasons, equipment malfunctioning etc. Imputing missing values in the dataset is one of the major pre-processing steps as it directly impacts on the quality of the data. The attributes in dataset used in this work are mostly numerical, hence quantitative in nature. This reduces the sensitivity of the attribute towards the extremes as outliers to certain extent. Outliers are out of box data which can be identified by plotting the scatter plots with the combination of the relevant attributes as in figures 2-3. Relevant attributes can be initially chosen based on the expert's opinion. Later they can be reframed by verifying with the machine learning algorithmic measures like entropy, information gain, gain ratio etc.

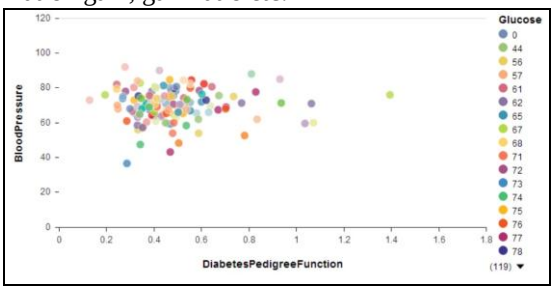

Fig. 2 Scatter plot of attributes Blood Pressure and Diabetes Pedigree Function by Glucose level

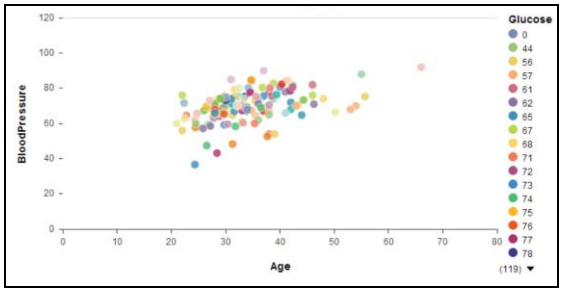

Fig. 3 Scatter plot of attributes Blood Pressure and Age by Glucose level

The attribute Glucose Level has been chosen as the most prominent attribute contributing majorly for the diabetic prediction. Scatter plot in the figure 2 has been plotted with the attributes blood pressure and diabetic pedigree function by glucose level. Scatter plot in the figure 3 has been plotted with the attributes blood pressure and Age by glucose level. From the above observations, it is evident that the outliers are scattered from the cluster signifying no much contribution towards automated prediction as the occurrences are sparse. Such observations were documented for all other combination of attributes by Glucose Level in the given dataset. These values if retained may mislead the learning algorithms working on functions like standard deviation, aggregate and other statistical measures. Hence such entries which were less than 1% of original data size are considered as outliers and are eliminated from the dataset prior to handling imputation of missing values.

C. Relevant Data Analysis

Attributes and their dependencies give way for redundancy measures. The subset of diabetic dataset with no missing values were selected and applied with C4.5 decision tree algorithm for decision rules generation in order to verify the prominent attributes majorly contributing for the decision process. It was observed that the decision tree was getting pruned after processing the 4 most important attributes: Glucose level, Age, BP, Diabetic Pedigree function with satisfactory accuracy scores marking the best features selection. If the tree intentionally left un-pruned, the decision rules generated were inconsistent resulting in poor accuracy scores. Redundant data may also confuse the knowledge discovery process to certain extent. In order to address these issues of redundancy and inconsistency, the least prominent attributes: skin thickness, Insulin level, Body Mass Index, count of Pregnancies were eliminated from the dataset resulting in dataset narrow downing up to 50% of the original size.

From this relevant data analysis step, it is evident that feature selection contributes majorly for checking the quality and predictive power of the attributes. Selected features should be consistent enough to demonstrate their support for prediction with very less or no imputed values. Features with too many missing values are scored to be inconsistent as bulk of their values are imputed declining the quality of data. Hence those attributes can also be deleted from the dataset.

D. Handling Missing Data

Irrespective of type of values, handling missing data is one among the most important data pre-processing steps. If not handled carefully, the entire model may be deteriorated with decreased prediction accuracy. The dependency of missing values can be categorized into Completely Missing At Random (CMAR) in which missing values will be independent of all the features, Missing At Random (MAR) in which values depend on other features and No Missing At Random (NMAR) in which missing values depend on other missing features. In the present dataset, missing values and their dependencies can be categorized into MAR type.

The statistical imputation methods like Hot Deck, central tendency measures (mean, median, mode) etc. impute values irrespective of the neighboring attribute dependencies. But in reality, particularly in healthcare domain, the attribute values depend on the other features to certain extent as stated earlier. So, Machine learning methods and imputation techniques using algorithms like SVM, k-means, K – Nearest Neighbors (KNN), Ensemble methods, Gradient boosting etc. can be employed for missing values imputation to enhance the data

quality. Expectation Maximization is a model based imputation algorithm which imputes values based on the statistical measures supported with step increments until local optima reached. However, the vertical and horizontal features dependencies also to be considered to get better data quality after imputation. The above listed algorithms consider vertical dependencies to maximum extent neglecting the horizontal features dependencies.

To address the above problem, *Cluster based Missing Value Imputation (CMVI)* procedure depicted in figure 4 has been proposed in this work. The proposed idea is constructed on the concepts of k-means and SVM clustering techniques for imputing the missing values. SVM based imputation considers conditional and decision attributes for linear separation of the dataset on the linear regression kernel (1) by computing dot product of given input(x) and each support vector (x_i). SVM has been chosen for the initial computation as it's the only linear model which can classify data which is non-linearly separable.

$$f(x) = B(0) + sum\big(a_i * (x, x_i)\big) \quad (1)$$

Where x_i includes all support vectors of the training set, x is the input vector, the co-efficients B(0) and a_i (for each and every input) will be assessed using the training set by the learning algorithm.

K-means being one of the simplest, fastest and robust clustering based algorithm with reliable results on a linear fashioned data, it best suites when cluster numbers are specified clearly based on the data sequence. The algorithm uses (2) to recalculate the new cluster center and aims to minimize the squared error function (3) giving best results on datasets which are well separated and distinct in nature.

$$v_i = (1|c_i) \sum_{i=1}^{c_i} x_i \quad (2)$$

$$J(V) = d(i,j) \quad (3)$$

Where d(i,j) is computed as in (4), c_i is the number of data points in i^{th} cluster, x_i is attribute value in that particular cluster.

K-means concept computes the mean value within the separated clusters based on Euclidean distance (4) and imputes the missing values. '*k*' value depends upon the count of the categorical range of the numerical attributes.

$$d(i,j) = \sqrt{(x_{i1} - O_{j1})^2 + (x_{i2} - O_{j2})^2 + \cdots + (x_{ip} - O_{jp})^2} \quad (4)$$

Where i = (x_{i1}, x_{i2}, … , x_{ip}) and j = (x_{j1}, x_{j2}, … , x_{jp}) are two objects described by p numeric attributes.

Glucose level has been selected as the most prominent attribute followed with age, blood pressure, diabetic pedigree as important attributes for diabetic prediction. The prominent features have been chosen as per the opinion of domain experts and upon ranking the feature importance based on the C4.5 decision tree classification algorithm. The proposed CMVI algorithm reads training dataset (D_T) with *n* tuples and its corresponding *m* attributes (A) as input. The dataset with more than |n|/2 missing values, where *n* is the total number of attributes being considered for decision rule generation; may lead to incorrect decisions by reducing the entropy and increasing the information gain of the misleading attribute. Rules generated with such dataset results in decreased accuracy of the model prediction. Such entries are considered as outliers and will be dropped from the dataset to reduce the noise in the resultant model.

Algorithm: Cluster based Missing Value Imputation (CMVI)

Input: Training dataset (D_T), set of Attributes (A)
Output: Training dataset (D_T) with Imputed Missing Values

1. Let 'n' be number of tuples in the training dataset (D_T)
2. Let 'm' be the number of Attributes ($A_1, A_2, A_3, \ldots, A_m$)
3. ∀ t ∈ D_T, if ∃ t with |A| < (|n|/λ), then (where n is the total number of attributes being considered for decision rule generation, λ is a Greek symbol with values 1, 2, ..)
   delete t
   /* Assumption is if the number of missing attributes are more than the present attributes, * then the prediction accuracy may get reduced with all imputed values */
4. Split D_T into D_{NM} and D_{WM} clusters
   /* D_{NM} is Dataset with No Missing values and D_{WM} is Dataset with Missing Values */
5. Split D_{NM} into D_{NT} and D_{NF} sub-clusters
   /* D_{NT} is No missing value Dataset with True classification and D_{NF} is No missing value Data * -set with False classification */
6. Split D_{WM} into D_{WT} and D_{WF} sub-clusters
   /* D_{WT} is With missing value Dataset with True classification and D_{WF} is With missing value * Dataset with False classification */
7. Identify the most prominent attribute 'a_{p1}' & second most prominent attribute 'a_{p2}' from A.
8. ∀ t ∈ D_T, if ∃ t with a_{pm} ∈ a_{p1}, then /* a_{pm} is missing value of the a_{p1} */
   if a_{pm} in D_{WT} and is numerical, then impute a_{pm} using SVM(a_{p1}, a_{p2}) from D_{NT}
   if a_{pm} in D_{WT} and is categorical, then impute a_{pm} using mode(a_{p1}) from D_{NT}
   if a_{pm} in D_{WF} and is numerical, then impute a_{pm} using SVM(a_{p1}, a_{p2}) from D_{NF}
   if a_{pm} in D_{WF} and is categorical, then impute a_{pm} with mode(a_{p1}) from D_{NF}
9. Split D_{NT}, D_{NF}, D_{WT} and D_{WF} based on various ranges of most prominent attribute selected to form additional sub-clusters
10. Match the additional sub-clusters between D_{NT} & D_{WT} and D_{NF} & D_{WF} based on prediction class
11. ∀ a ∈ A_m with missing values in the D_{WT} and D_{WF} /*
    if *a* is numerical, then
    - Within the matched additional sub-clusters, impute the missing values of 'a' using k-means applied on No missing additional sub-cluster values
    - *k* – value ranges between 1 and count(categorical range of A_m)
    - Substitute the imputed values in the records of D_{WT} and D_{WF}

    else impute the values of 'a' using mode imputation method
12. Compute $D_T = D_{NT} \cup D_{NF} \cup D_{WT} \cup D_{WF}$

Fig. 4 Cluster based Missing Value Imputation (CMVI) Algorithm

Dataset is clustered based on With Missing values (D_{WM}) and No Missing values (D_{NM}) of the attributes. Depending on the prediction class, Dataset with No missing True classification (D_{NT}), Dataset with No missing False classification (D_{NF}), Dataset With missing True classification (D_{WT}) and Dataset With missing False classification (D_{WF}) sub-clusters are formed from the above clusters. Before imputation of missing values of all the features, it is important to impute the most prominent attribute missing values based on the so formed sub-clusters.

In the current dataset, *Age* has been chosen as second most prominent attribute supporting missing value imputation of most prominent attribute *Glucose Level*. Missing values in the sub-clusters D_{WT} and D_{WF} are imputed based on the respective imputation method and values in the sub-clusters D_{NT} and D_{NF}. For categorical values, mode (most common values) imputation method has been used and for numerical values, SVM imputation method has been employed to impute the missing values.

Additional sub-clusters of D_{NT}, D_{NF}, D_{WT} and D_{WF} are built on various ranges of the selected prominent feature. Match between the additional sub-clusters of D_{NT}, D_{NF}, D_{WT} and D_{WF} is done based on the prediction class.

Missing values are imputed for numerical values of features

using k-means technique having *k* value as count of the categorical range of the present numerical values of the features and for categorical values of features, the mode imputation method within the additional sub-clusters have been applied accordingly. Finally all the clusters are integrated recursively until a single dataset is obtained. Sample missing value imputation analysis of the feature - *Glucose Level* is done with the help of scatter plot as in figure 5. Calculated values are shown in yellow and blue color. It is evident that the missing value computed falls within the median of each prediction class promising not much deviation in the prediction accuracy in future.

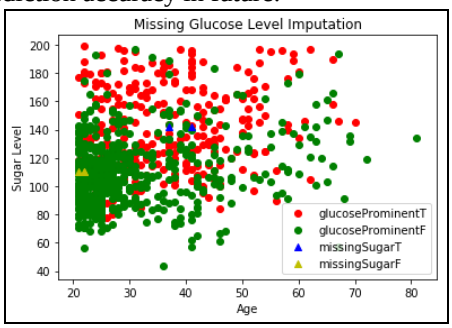

Fig. 5 scatter plot of Missing Value Imputation of Glucose Level

E. Data Transformation

As part of data transformation, normalizing input variables is intent to make the underlying relationship between input and output variables easier to model. Min-Max normalization (5) is applied on attributes with large boundaries by performing linear transformation on original data to preserve relationship among the original values. It maps a value v_i of Attribute (A) to v_i' in the range [new_min$_A$, New_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}\ (new\_max_A - new\_min_A) + new\_min_A \qquad (5)$$

To analyze the relationship between the diabetic pedigree function, glucose and age, Min-Max normalization has been employed.

Discretization, a concept of hierarchy generation which helps reduce the data from lower level (numeric attributes) to higher level (continuous attributes) making data more meaningful by resulting in ease of mining. The bottom up merging strategy is being followed here by clustering the data based on ranges and replacing the numerical values by categorical values. Table I depicts the selected diabetic dataset attributes and their clustering ranges.

Table I Pima Indian Diabetic Dataset and their clustering ranges

Sl. No.	Attribute Name	Attribute Range		
		Low	Medium	High
1	Plasma Glucose Concentration	<70	71-140	>140
2	Diastolic Blood Pressure	<80	81-120	>120
3	Diabetic pedigree function	<0.42	0.43-0.82	>0.82
4	Age (young, Middle, Old)	<40	41-60	>60

F. Data Quality Evaluation

Accuracy, completeness and consistency are the three measures commonly used to access the superiority of the input dataset. The percentile counts of missing values also contribute for assessing the quality of data to maximum extent. Survey concludes that more than 15% missing values results in poor quality though quantity is high. In-order to assess the data quality preprocessed with the proposed CMVI method, Missing values were randomly induced into the dataset with the missing ratio of 1%, 3% and 5% for each attribute. The techniques of SVM Imputation (with linear regression kernel), k-means missing value imputation (k = 3), Expectation Maximization (EM) based missing value imputation (25 iterations) working on Gaussian mixture model and the proposed CMVI method were used to simulate the missing values of various attributes over the missing ratios.

Couple of well-known performance indicators namely coefficient of determination (R^2) and Index of agreement (d_k) have been used for measuring the imputation accuracy with Root Mean Square Error (RMSE) test conducted for error measure. The performance scores are documented for maximum of 50% of attributes with missing values starting from single attribute.

Coefficient of Determination (R^2) (6) determines the degree of correlations between the actual and imputed values. The range varies from 0 to 1. Higher the value towards 1, better the fit of imputed values.

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}[(P_i - \overline{P})(O_i - \overline{O})]}{\sigma_P \sigma_O}\right]^2 \qquad (6)$$

Where,

N is the count of artificially created missing values

O_i is the actual value of the i^{th} artificially created missing value ($1 \le i \le N$)

P_i is the imputed value of the i^{th} missing value

$\overline{O}$ is the mean of the actual values O_i $\forall$ $i \epsilon N$

$\overline{P}$ is the mean of the imputed values

σ_P is the standard deviation of the imputed values

σ_O is the standard deviation of the actual values

Index of agreement (d_k) (7) determines the degree of agreement between the actual and imputed values. The range varies from 0 to 1. Higher value indicates a better fit.

$$d_k = 1 - \left[\frac{\sum_{i=1}^{N}(|P_i - O_i|)^k}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^k}\right] \qquad (7)$$

Where,

N is the count of artificially created missing values

O_i is the actual value of the i^{th} artificially created missing value ($1 \le i \le N$)

P_i is the imputed value of the i^{th} missing value

$\overline{O}$ is the mean of the actual values O_i $\forall$ $i \epsilon N$

k is a constant with values either 1 or 2. In the presented work k value has been chosen to be 2

Root Mean Square Error (RMSE) (8) is to explore the average difference of actual values with the imputed values. The range varies from 0 to ∞. Lower value indicates better matching.

$$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{1/2} \qquad (8)$$

Where,

N is the count of artificially created missing values, O_i is the actual value of the i^{th} artificially created missing value ($1 \leq i \leq N$), P_i is the imputed value of the i^{th} missing value

The average performance of SVM, k-means, EM, CMVI (proposed) techniques upon imputing the artificially induced missing values into the Pima Indian Diabetic dataset are documented in Table II.

Table II Average Performance of SVM, k-means, CMVI on Diabetic dataset

Num of Attributes with Missing Values	Missing Rate	R^2				d_2				RMSE			
		SVM	k-means	EM	CMVI (proposed)	SVM	k-means	EM	CMVI (proposed)	SVM	k-means	EM	CMVI (proposed)
1	1%	0.954	0.886	0.948	0.962	0.987	0.967	0.986	0.989	8.124	12.816	8.616	7.356
	3%	0.897	0.826	0.890	0.927	0.974	0.955	0.972	0.981	8.304	10.795	8.554	6.992
	5%	0.875	0.839	0.861	0.961	0.966	0.958	0.961	0.990	10.330	11.829	11.010	5.799
2	1%	0.901	0.838	0.889	0.935	0.974	0.958	0.972	0.983	6.191	8.902	6.551	5.327
	3%	0.799	0.741	0.826	0.907	0.957	0.946	0.960	0.977	6.561	7.984	6.427	4.976
	5%	0.834	0.815	0.844	0.942	0.961	0.958	0.962	0.986	7.835	8.599	7.942	4.519
3	1%	0.931	0.886	0.922	0.955	0.982	0.970	0.980	0.988	4.141	5.954	4.383	3.562
	3%	0.864	0.826	0.882	0.937	0.971	0.964	0.973	0.984	4.384	5.328	4.294	3.321
	5%	0.883	0.871	0.891	0.961	0.972	0.970	0.973	0.990	5.237	5.747	5.309	3.016
4	1%	0.912	0.906	0.895	0.951	0.978	0.958	0.974	0.987	4.078	6.037	4.398	3.302
	3%	0.887	0.833	0.898	0.949	0.975	0.964	0.976	0.987	3.915	5.135	3.917	2.888
	5%	0.904	0.875	0.903	0.967	0.977	0.970	0.976	0.992	4.539	5.434	4.786	2.663

Figures 6-8 shows the plots of the performance indicators R^2, d_2, RMSE scores plotted against Missing ratio patterns by count of attributes with missing values ranging from 1 to 4 which is 50% of the attributes in the original diabetic dataset.

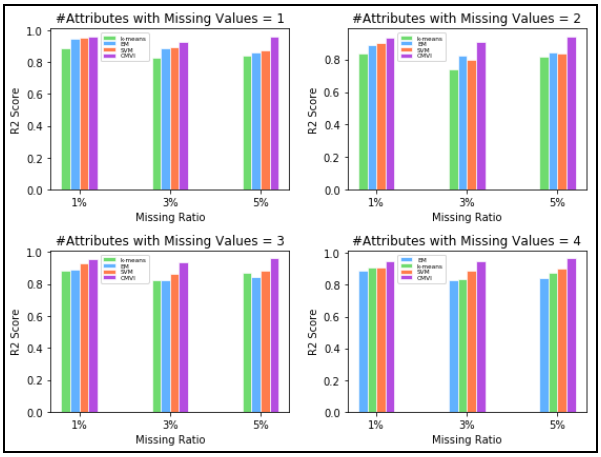

Fig. 6 Average performance of SVM, k-means, EM and CMVI based on R2 score against missing ratios

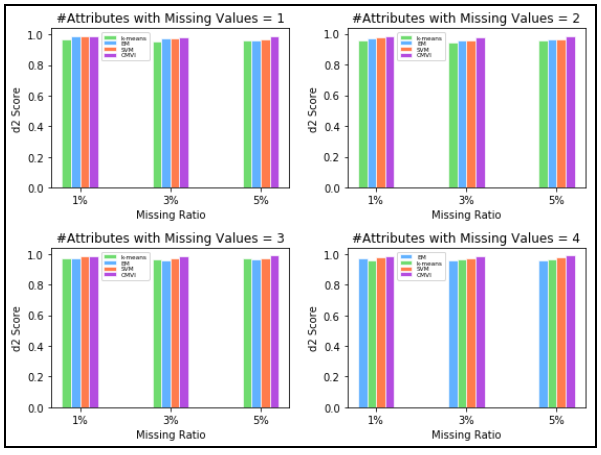

Fig. 7 Average performance of SVM, k-means, EM and CMVI based on d_2 score against missing ratios

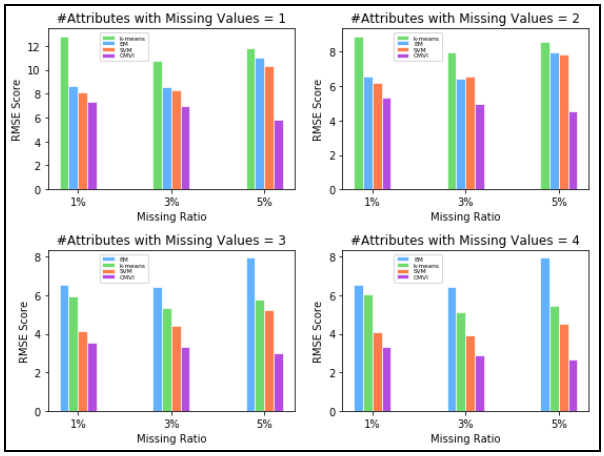

Fig. 8 Average performance of SVM, k-means, EM and CMVI based on RMSE score against missing ratios

Observations state that as the number of attributes with missing values increase, drop in overall metrics performance is being noticed. In majority of the cases, performance can be ordered approximately in ascending sequence of k-means, SVM, EM and CMVI concluding that the performance of the proposed CMVI technique outscores the other individual techniques employed here for comparison. Similar approach has been followed for rest of the datasets.

G. Ontology Construction

Data once completely pre-processed, to be transformed to the usable form suitable for semantic querying and automated knowledge extraction. Research on visualization of the knowledge graph is in boom as to make them human readable [18]. The free, open source ontology editor: *protégé* [19] a tool by Stanford research center is used to construct the knowledge graph called Onto-graph from the pre-processed data. A plug-in called *cellfie* (figure 9) is used to build the ontology from the .xls input file. Cellfie loaded with .json rules, defining the triple store formats and dependencies with limitations of the attributes, entities, domain, range etc., constructs a knowledge graph. Figure 10 depicts a sample diabetic Onto-graph constructed from the .xls input file.

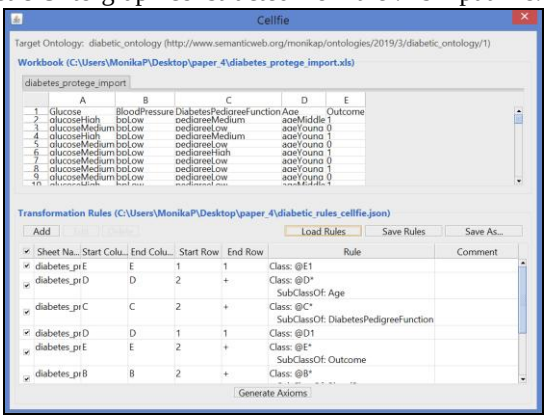

Fig. 9 Protégé plugin - Cellfie for knowledge graph construction from .xls input file

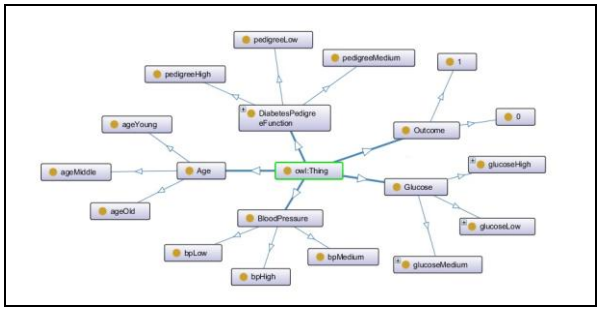

Fig. 10 Diabetic Onto-graph

Knowledge graph can be used for extracting knowledge in various fields like Data mining, Natural Language Processing [NLP], Deep Learning etc., Ontologies can be stored in .rdf, .owl, .n3, .turtle, .xml/rdf, .tz, formats. These days NoSQL, SPARQL, combination of Jena, HBase and Sesame are used for large scale storages [20]. RDF is the most widely used format as merging and integration of ontologies are well supported here. Rule based engines can be constructed for knowledge extraction. Supervised and Unsupervised learning algorithms with classification algorithms like Marcov Chains & Models [21], SVM, KNN, Decision trees etc. can be applied on the knowledge graph for automatic knowledge extraction process realization.

IV. CONCLUSION

Web consists of enormous amount of data pertaining to several domains. Healthcare domain in particular generates loads of data every second in various formats. These data should be handled effectively as the information within is very much important for the medical practitioners and researchers for the benefit of the society. Data availability in various formats includes unstructured, semi-structured and structured representations. Such data should be pre-processed before supporting machine interactions towards automation of knowledge discovery.

Cluster Based Missing Value Imputation (CMVI) algorithm has been presented in this paper to enhance the imputed data quality. The quality of the imputed data values using the proposed CMVI method has been rigorously tested by inducing the missing values into the Pima Indian Diabetic dataset with missing ratios of 1%, 3% and 5% for each attribute. Simulated missing values were evaluated using the data quality evaluation metrics namely R^2, d_2, RMSE. The experimental results revealed that the proposed method imputes better data values compared to the standard imputation algorithms namely SVM, EM and k-means.

Other pre-processing procedures followed for preparing the data for semantic usage which are briefed in the presented work include data collection & cleaning, relevance analysis, integration, transformation and construction of Ontologies for storing the information in the format of knowledge graph. Further work progresses upon proposing a comprehensive approach for efficient information retrieval from the semantic web with the help of soft computing approaches in the form of machine learning algorithms, query languages and engines establishing various ontologies interaction and integration for the benefit the medical researchers and practitioners.

REFERENCES

1. Piskorski J, Yangarber R, "Information Extraction: Past, Present and Future, multilingual information extraction and summarization", Springer, Berlin, Heidelberg, 2013, pp. 23-49, DOI: 10.1007/978-3-642-28569-1_2.
2. Zhanfang Zhao, Sung-Kook Han, In-Mi So, "Architecture of Knowledge Graph Construction Techniques", International Journal of Pure and Applied Mathematics, Volume 118, No.19, 2018, pp. 1869-1883, ISSN: 1314-3395.
3. Han Xianpei, Zhao Jun, "Named Entity Disambiguation by leveraging Wikipedia Semantic Knowledge", Proceedings of the 18th ACM conference on Information and Knowledge Management, NewYork, ACM 2009, pp. 215-224.
4. Mendes P N, Muhleisen H, Bizer C Sieve, " Linked data quality assessment and fusion", Proceedings of the 2nd International workshop on Linked Web Data Management at Extending Database Technology, NewYork, ACM 2012, pp. 116-123.
5. S Thirukumaran, A Sumathi, "Missing Value Imputation Techniques Depth Survey and an Imputation Algorithm to Improve the efficiency of the Imputation", IEEE 4th International conference on Advanced Computing (ICoAC), 2012, DOI: 10.1109/ICoAC.2012.6416805.
6. T V Rajinikanth, "An Enhanced Approach on Handling Missing Values Using Bagging KNN Imputation", International Conference on Computer Communication and Informatics (ICCCI), 2013.
7. Thomas G. Dietterich and Tadesse Zemicheal, "Anomaly Detection in the Presence of Missing Values", Proceedings of ACM SIGKDD 2018, Workshop August 20th, London UK (ODD'18). ACM, New York, NY, USA, Article 4-8 pages, DOI: https://doi.org/10.475/123_4.
8. Sandeep Kumar Singh, Archana Purwar, "Empirical Evaluation of Algorithms to Impute Missing Values for Financial Data Set", IEEE International Conference 2014.
9. Jerry W Grzymala Busse, Ming Hu, "A Comparison of Seven Approaches to Missing Attribute Values in Data Mining", Springer-Verlag Berlin, Heidelberg, 2001, pp. 378-385.
10. Dai Li, jitender Deogun, William Spaulding, Bill Shuart, "Towards Missing Data Imputation: A study of Fuzzy K-means clustering method", Springer-Verlang Berlin, Heidelberg, 2004.
11. Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, Chen Yumei, "A SVM Regression based Approach to Filling in Missing Values", Springer-Verlang Berlin, Heidelberg, 2005.
12. Md. Geaur Rahman, Md. Zahidul Islam, " Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel Techniques", Elsevier Journal of Knowledge based systems, 2013, pp. 51-65.
13. Md. Geaur Rahman, Md. Zahidul Islam, " A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing", 9th Australasian Data Mining conference (AusDM'11), Ballarat, Australia, Vol. 121.
14. Felix Biessmann, David Salinas, Sebastian Schelter, Philipp Schmidt, Dustin Lange, "Deep Learning for Missing Value Imputation in Tables with Non-Numerical Data", 27th ACM International Conference on Information and Knowledge Management (CIKM'18), Oct. 22–26, 2018, Torino, Italy. ACM, New York, USA, 9 pages. DOI: https://doi.org/10.1145/3269206.3272005
15. Zachary C Lipton, David C Kale, Randall Wetzel, "Modeling Missing Data in Clinical Time Series with RNNs", Proceeding of Machine Learning for Healthcare, Journal of Machine Learning Research W& C track volume 56, 2016.
16. Yuzhe Liu, Vanathi Gopalakrishnan, "An Overview and Evaluation of Recent Machine Learning Imputation methods using Cardiac Imaging Data", Journal of MDPI, Basel, Switzerland, 2017, DOI: https://doi.org/10.3390/data2010008.
17. 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.
18. Kai Sun, Yuhua Liu, Zongchao Guo, Changbo Wang, " Visualization for Knowledge Graph based on Education Data", International Journal of Software and Informatics, 2016, Volume 10, Issue 3, ISSN: 1673-7288.
19. 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.
20. Philippe Cudr-Mauroux, Iliya Enchev, Sever Fundatureanu, Paul Groth, Albert Haque, Andreas Harth, Felix Leif Keppmann, Daniel Miranker, Juan Sequeda, Marcin wylot "NoSQL

databases for RDF: an Empirical Evaluation", International Semantic Web Conference, Springer, Berlin, HeidelBerg, 2013, pp. 310-325.
21. Scheffer T, Dccomain C, Wrobel S, "Active Hidden Marcov Models for Information Extraction", International Symposium on Intelligent Data Analysis, Springer, Berlin, Heidelberg, 2001, pp. 309-318.

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 10.7 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.