# AN ONTOLOGY-BASED HIERARCHICAL BAYESIAN NETWORK CLASSIFICATION MODEL TO PREDICT THE EFFECT OF DNA REPAIRS GENES IN HUMAN AGEING PROCESS

Hasanein Alharbi

Department of Computer Engineering Techniques,
Al-Mustaqbal University College, Babylon, Iraq

***ABSTRACT***

*Conventional Data Mining (DM) algorithms treated data simply as numbers ignoring the semantic relationships among them. Consequently, recent researches claimed that ontology is the best option to represent the domain knowledge for data mining use because of its structural format. Additionally, it is reported that ontology can facilitate different steps in the Bayesian Network (BN) construction task. To this end, this paper investigates the advantages of consolidating the Gene Ontology (GO) and the Hierarchical Bayesian Network (HBN) classifier in a flexible framework, which preserves the advantages of both, ontology and Bayesian theory. The proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) is tested using data set in the biomedical domain. DNA repair genes are classified as either ageing-related or non-ageing-related based on their GO biological process terms. Furthermore, the performance of SAHBN was compared against eight conventional classification algorithms. Overall, SAHBN has outperformed existing algorithms in eight experiments out of eleven.*

***KEYWORDS***

*Semantic Data Mining, Hierarchical Bayesian Network, Gene Ontology, DNA Repair Gene, Human Ageing Process*

## 1. INTRODUCTION

The ultimate aim of Data Mining (DM) algorithms is to extract useful knowledge from data. Fayyad et al. have defined these methods as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in databases [1],[2]. However, existing mining algorithms treated data as meaningless numbers disregarding their semantic context [3],[4], Hence, the data mining philosophy has faced a paradigm shift from being a data-centered process to knowledge-centered process that aims to cater for domain knowledge and its integration in the mining process. The process of integrating domain knowledge with DM task is known as Semantic Data Mining [4] ,[5],[6].

Domain knowledge can be represented using various techniques. However, recent researches indicated that ontology are playing significant role in the process of knowledge acquisition and representation [7], [8]. In fact, the formal structure of ontology makes it a strong candidate for knowledge integration in the DM algorithms. Ontology could be intertwined with DM algorithms

Dhinaharan Nagamalai et al. (Eds) : CSEIT, NCS, SPM, NeTCoM - 2018
pp. 15–26, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81802

to bridge the semantic gap, to provide prior knowledge and constraints, and formally represent the mining results [9], [10]. Likewise, ontology can be used to facilitate different steps in the Bayesian Network (BN) construction process. It can assist in the identification of the BN structure and supports the calculation of the Conditional Probability Tables (CPT's) [11].

Hence, the process of developing a framework which systematically consolidates ontology and the Bayesian mining algorithm is investigated in this paper. The aim of this paper is to explore the potential advantage obtained from coupling the domain knowledge in the form of Gene Ontology (GO) and the Hierarchical Bayesian Network (HBN) classifier and then utilizing the developed model to predict the DNA repair gene effect in the human ageing process.

The structure of this paper is organized as follows. Section 2 explains the proposed model. Section 3 presents the experimental results and the evaluation process. Finally, section 4 draws conclusions and discusses possible future research directions.

## 2. PROPOSED MODEL

GO was used in this research because of its high quality and comprehensive nature in biomedical domain [12]. Meanwhile, the structure of the HBN implicitly provides more knowledge about the targeted domain [13]. As a results, the integration of these two concepts, GO and HBN, generate a classification model which seamlessly reflects the domain knowledge.

The proposed SAHBN classification model shares some initial steps with the standard classification algorithms such as data pre-processing and feature selections. However, the essential steps related to BN structure construction and variables probability estimation are designed in such a way that exploits the semantic nature of the GO. Figure1 compares the process sequence of standard classification algorithms and the proposed SAHBN model.

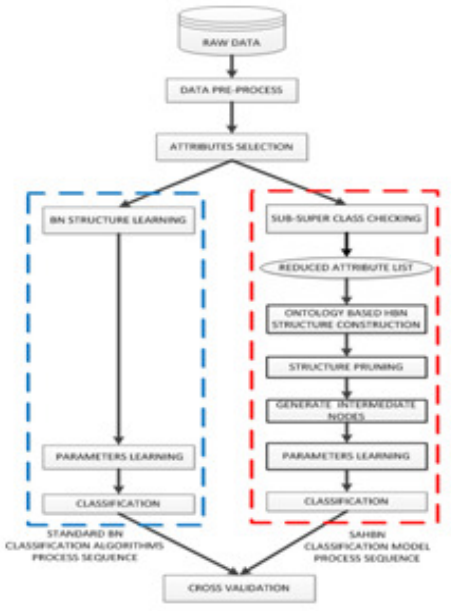


Figure 1 SAHBN Classification Model versus Standard Classification Algorithms Process Sequence

Figure1 shows that the selected prediction attributes have been further processed based on the semantic knowledge extracted from the GO. This can be seen in the steps surrounded by the red dotted line in the same figure. The new steps introduced by SAHBN model are summarized in the following subsections.

## 2.1. SUB-SUPER CLASS CHECKING

The first step which follows the attributes selection task is to check whether there is a semantic relation between the selected attributes. This is done by matching the selected attributes to the GO concepts. The data sets covered in this paper used the GO biological process (GOBP) terms as a prediction attributes. Hence, one-to-one matching between the selected attributes and the GO concepts was implemented. Consequently, the GO structure was exploited to extract the semantic relation between these attributes.

The relation that was targeted in this research is the parent-child ("is-a") class relation. GO used the "is-a" relation to represents the subtype relation between concepts. For example, "Replicative Cell Aging" is a subtype of and less general that the "Cell Aging" process. Likewise, the intermediate nodes in the HBN structure represent an aggregation of simpler nodes. Hence, the "is-a" relation was selected to identify the structure of the HBN.

The "is-a" relation not only facilitates the construction of the HBN structure, but also achieves the following objectives.

Maintain data consistency: The GO "is-a" relation follows the True Path Rule (TPR) which states that if an instance of GO node is proved to be true, so it's ancestors all the way to the root must be true. Otherwise, if an instance found to be false, so all its descendants to the leaf nodes must be false [14], [15], [16]. Thus, any two GO terms connected via the "is-a" relation and used as prediction attributes must follow the TPR. Otherwise, an inconsistent data set can be used to train the classification model, which may lead to inaccurate results.

Table 1 shows some records from the DNA repair gene-PPI data set (discussed in the 3rd section), which highlights the inconsistency in the training data set.

| No. | GO:0007568 | GO:0001302 | Label Class |
|---|---|---|---|
| 1 | TRUE | FALSE | TRUE |
| 2 | FALSE | TRUE | FALSE |
| 3 | FALSE | FALSE | FALSE |
| 4 | FALSE | TRUE | TRUE |
| 5 | TRUE | FALSE | TRUE |
| 6 | FALSE | TRUE | TRUE |
| 7 | FALSE | TRUE | TRUE |
| 8 | FALSE | FALSE | TRUE |
| 9 | FALSE | TRUE | TRUE |
| 10 | FALSE | FALSE | FALSE |

Table 1 Sample of Inconsistent Training Data Set

According to the GO structure, the GO:0001302 attribute is a child class of the GO:0007568. While the former refers to the replicative cell ageing, the latter refers to the ageing biological process, and there is an indirect "is-a" relationship between them. Hence, it can be seen that records 2,4,6,7 and 9 (highlighted in red in Table 1) are inconsistent because the value of the parent class is false, while the value of its child class is true and this violates the TPR.

Thus, this paper proposed the use of the Chi-squared to break the conflict between the contradicted prediction terms and eliminate data inconsistency. This is done in four steps, as follows:

a. Identify the contradictory GO prediction terms, which connected via the "is-a" relationship.

b. Calculate the Chi-squared value between each term and the label class.

c. Delete the GO term, which has the lowest dependency with the label class.

d. Repeat steps 1 throw 3 until all contradiction is removed.

Reduce prediction attributes list dimension: removing the contradicted attributes using GO "is-a" relation not only eliminates the inconsistency in the training data set but also reduces the dimension of the prediction attribute list. High dimensional data poses a serious challenge for data mining techniques, especially in medical domain.

## 2.2. ONTOLOGY-BASED HBN STRUCTURE CONSTRUCTION

The second step, which follows the parent-child class checking, is HBN structure construction. The structure construction task is implemented based on the reduced attributes list and the structure of the GO. The steps involved in this process are summarized in the following points.

a) Match each attribute in the reduced list generated after the parent-child class checking step to node in the GO.

b) Extract the path for each matched node (i.e. attribute node) using the "is-a" relation and the GO structure. The path is extracted from the matched node all the way to the root node. We began by extracting the parent class of the attribute node, and then the extracted parent class was considered as an attribute node and its parent class extracted. This process was repeated until the root node was reached.

c) Combine the extracted path to form a tree-like hierarchical structure.

Figure 2 depicts a sample of ontology-based HBN structure for attributes list consists of five GO terms {GO1, GO2, GO3, GO4, and GO5}. The predication attributes form the terminal nodes in the HBN structure and their parent classes shape the rest of the structure.

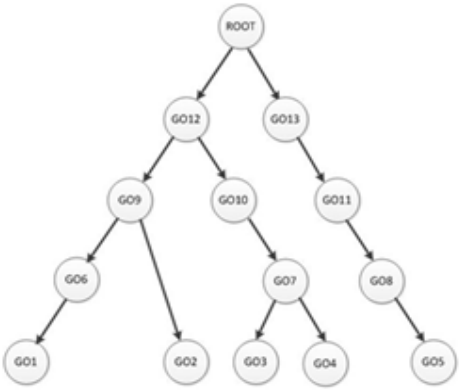


Figure 2 Ontology-based HBN structure

## 2.3. STRUCTURE PRUNING

The structure pruning step exploits the transitive nature of the "is-a" relationship in the GO. The "is-a" relation is transitive which mean that if "A is-a B", and "B is-a C", we can infer that "A is-a C". Hence, it is save to aggregate terms connected by the "is-a" relationship [17]. Figure 3 illustrates the structure pruning process.

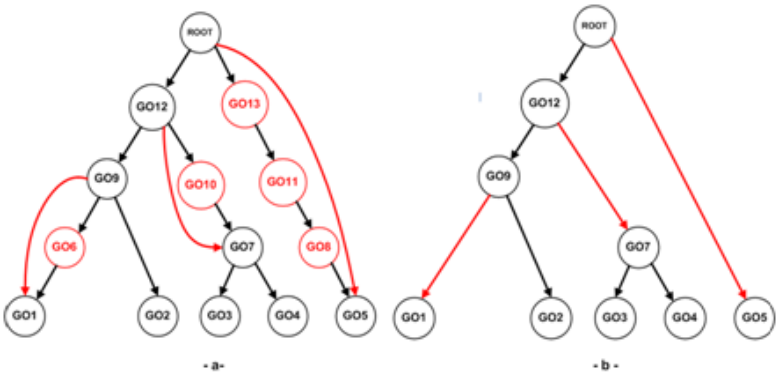


Figure 3 SAHBN Structure pruning process

The aim of this step is to remove redundant nodes that do not affect the principles of the HBN structure and maintain the semantic consistency of the targeted domain. There are two main basic principles underpinning the structure of the HBN. These principles can be summarized in the following points.

- Aggregation: each node in the HBN structure represents an aggregation of simpler nodes.
- Independency: each node in the HBN structure is conditionally independent of its non-descendant node given the value of its direct parent.

  Consequently, and in order to prune the created HBN structure without violating the above principles, the following steps were followed:

1. Delete all intermediate nodes that have only one child class.
2. The child class of the deleted node will be a child class of the deleted node parent class.

To demonstrate the pruning process, the above steps were applied to the structure of the HBN depicted in Figure 3 (a), which was constructed in the previous step. As a result GO6, GO8, GO10, GO11and GO13 terms, and the associated arcs, were deleted. The steps of the pruning process are summarized in Figure 3(b).

## 2.4. GENERATE INTERMEDIATE NODES

Figure 3 (b) shows that three intermediate nodes have been added to the structure of the HBN, namely, GO7, GO9, and GO12. Unlike the observed prediction attributes (i.e., terminal nodes), the value of the added intermediate nodes are unknown. However, as previously explained, the GO "is-a" relation is subject to the TPR. Consequently, the TPR principle was exploited to define the values of the intermediate nodes. This is done by implementing the following rule: "the value of any intermediate node is equal to true if and only if the value of any of its child classes is equal to

true. Otherwise, its value is equal to false". Consequently, semantically consistent and complete training data set was generated.

### 2.5. PARAMETERS LEARNING

Having filled the intermediate nodes with values, the next step is to learn the SAHBN variables probability. There are two main approaches for estimating the probability values in the BN for complete dataset, namely, Maximum Likelihood Estimation (MLE) and Bayesian Estimation [18]–[21].

Despite its various advantages, the MLE method has the following limitation [19].

1. The size of the observed data set has no effect on the estimation process.

2. MLE does not consider the prior knowledge. Therefore, it entirely relies on the observed data set.

Hence, this paper has used a Bayesian-based approach, namely, Maximum a Posterior Estimation (MAP) method to estimate the probability values of the SAHBN variables.

## 3. EXPERIMENTAL RESULTS

This section discusses the experimental implementation and the obtained results. The first subsection gives an introduction on the human ageing case study. The second subsection explains the creation process of the DNA repair gene data set. Finally, the implementation of SAHBN model and the obtained results are analyzed in the third subsections.

### 3.1. HUMAN AGEING CASE STUDY

Human aging is defined as the gradual failure of the physiological function in various cells, tissues and organs in the human body, which ultimately leads to the fragility of body functionalities within the time growth and increases the probability of death [22]–[24].

Human ageing is an extremely complex, mysterious, controversial, and puzzling process that requires more investigation. Furthermore, studying the ageing process has led to challenges, such as ethical factors associated with doing experiments on human data, time form implementing the experiments on human data, and the comprehensive elements that must be considered when analyzing the ageing process. Thus, researchers have alternatively used the gene/protein databases of short living organism models to implement their experiments. Consequently, data mining techniques have been recently applied to analyze large amount of open access gene/protein databases to gain some insight into the human ageing process [25]–[28].

Human genome preserves its integrity by protecting the cellular DNA from both internal and external attacks. Thus, the cellular DNA is steadily monitored by the repair enzymes to correct damages resulting from these attacks. Accordingly, DNA damage is an essential element in the human ageing process, the modification of DNA repair process will result in advance understanding of the cellular ageing phenomena [27][29]. Hence the proposed SAHBN classification model applied to the DNA repair genes database to classify their effect as either an ageing-related or non-ageing-related gene.

### 3.2. DNA REPAIR GENE DATA SET CREATION

The data set used in this research has been created using the DNA repair gene [30], GenAge [31], Human Protein Reference [32], and UniPort [33] databases. The data set creation process can be summarized in the following steps.

1. Download the DNA repair gene database from the Human DNA repair genes website [30].

2. Classify the downloaded DNA repair genes into two categories, ageing-related and non-ageing related. The DNA repair genes appearing in the GenAge [31] database are classified as ageing-related, while the DNA repair genes that do not appear in the GenAge database are classified as non-ageing-related.

3. Represent the downloaded DNA repair genes in the form of proteins. This is done using the following steps.

    a. Download the protein-protein interactions (PPI) from the human protein reference database [32].

    b. Select the interaction that at least one of the proteins is located in the DNA repair gene which generated in step 2.

    c. The type of evidence for the interactions is obtained from either in vitro or in vivo experiments.

4. Since the DNA repair genes are represented in form of PPI. Hence, they can be annotated in form of GO biological process (GO BP) terms using the UniProt [33] database.

Eventually, each DNA repair gene was represented in the form of a sequence of GO BP terms. The selected GO BP terms are associated with proteins that represent the DNA repair genes. The value of the GO BP terms was equal to "T" if it appears in the protein associated the DNA repair gene; otherwise, it was equal to "F". Table 2 presents sample of the DNA Repair Data Set.

| GO:0006281 | GO:0006284 | ..... | ageing-related/non-ageing-related |
|---|---|---|---|
| T | T | Values for other prediction attributes | T |
| T | T | | F |
| F | F | | F |
| T | F | | F |

Table 2 Sample of the DNA Repair Data Set

### 3.3. RESULTS ANALYSIS

This subsection gives a detailed description of the experimental implementation and the obtained results. It discusses the proposed SAHBN model implementation, comparison with classical classification algorithms and results analysis.

Eleven attributes selection methods were used to reduce the dimensions of the created data set. Consequently, the performance of the proposed SAHBN classification model was compared against the performance of different standard classification algorithms, such as Bayesian Network

(BN), Naïve Bayes (NB), Decision Tree (J48), Support Vector Machine (SVM), K Nearest Neighbor (KNN) and Neural Network (NN).

Classification accuracy is the most commonly used criteria to estimate the performance of the classification model. However, it has been argued that the classification accuracy can misjudge the model performance if the tested data set is imbalanced. Hence, this research used the harmonic mean of the average class accuracy as proposed by [34]. Equation (1) presents the formula of the harmonic mean for the average class accuracy.

$$average\ class\ accuracy_{HM} = \frac{1}{\frac{1}{|levels(t)|}\sum_{l \in levels(t)} \frac{1}{recall_l}} \quad (1)$$

The harmonic mean of the average class accuracy was calculated for the proposed SAHBN model and the other 8 classification algorithms against which SAHBN was compared. This process is repeated for all 11 combinations of attribute selection methods. Consequently, the obtained results are summarized in Table 3.

| No. | SAHBN | BN (ICSS) | BN (K2) | BN (TAN) | NB | DT (J48) | SVM | KNN | NN |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.88** | 0.68 | 0.82 | 0.84 | 0.87 | 0.85 | 0.84 | 0.82 | 0.73 |
| 2 | **0.89** | 0.73 | 0.82 | 0.84 | 0.87 | 0.84 | 0.77 | 0.79 | 0.75 |
| 3 | 0.84 | 0.82 | 0.82 | 0.82 | 0.82 | 0.83 | 0.82 | **0.86** | 0.77 |
| 4 | **0.77** | 0.68 | 0.72 | 0.71 | 0.74 | 0.62 | 0.72 | 0.69 | 0.70 |
| 5 | **0.88** | 0.68 | 0.82 | 0.84 | 0.87 | 0.84 | 0.77 | 0.79 | 0.73 |
| 6 | **0.90** | 0.73 | 0.82 | 0.84 | 0.87 | 0.84 | 0.77 | 0.79 | 0.75 |
| 7 | **0.84** | 0.62 | 0.80 | 0.83 | 0.84 | 0.81 | 0.79 | 0.77 | 0.75 |
| 8 | 0.78 | **0.84** | 0.81 | 0.78 | 0.68 | 0.80 | 0.80 | 0.81 | 0.77 |
| 9 | **0.84** | 0.71 | 0.82 | 0.81 | 0.83 | 0.81 | 0.77 | 0.75 | 0.67 |
| 10 | **0.82** | 0.77 | 0.76 | 0.82 | 0.80 | 0.81 | 0.73 | 0.77 | 0.78 |
| 11 | 0.81 | 0.75 | 0.78 | **0.83** | 0.78 | 0.83 | 0.79 | 0.73 | 0.73 |
| Avg. HA | **0.84** | 0.73 | 0.80 | 0.81 | 0.81 | 0.81 | 0.78 | 0.78 | 0.74 |

Table 3 Ten-Folds Cross Validation Test Results in Terms of the Harmonic Mean of the Average Class Accuracy

Table 3 shows that the proposed SAHBN classification model outperformed conventional classification algorithms in eight experiments. Furthermore, SAHBN scored the highest average harmonic class accuracy among all classifiers.

The performance of SAHBN model was further analyzed using more sophisticated non-parametric statistical test. Precisely, Friedman test followed by Nemenyi Post-Hoc evaluation are implemented for all 11 data set combinations. SAHBN was used as a control classifier against which all other 8 conventional classifiers were compared. Consequently the results are presented in Table 4.

| No. | SAHBN | BN (ICSS) | BN (K2) | BN (TAN) | NB | DT (J48) | SVM | KNN | NN |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 9 | 6.5 | 4.5 | 2 | 3 | 4.5 | 6.5 | 8 |
| 2 | 1 | 9 | 5 | 3.5 | 2 | 3.5 | 7 | 6 | 8 |
| 3 | 2 | 5.5 | 5.5 | 5.5 | 8 | 3 | 5.5 | 1 | 9 |
| 4 | 1 | 8 | 3 | 5 | 2 | 9 | 4 | 7 | 6 |
| 5 | 1 | 9 | 5 | 3.5 | 2 | 3.5 | 7 | 6 | 8 |
| 6 | 1 | 9 | 5 | 3.5 | 2 | 3.5 | 7 | 6 | 8 |
| 7 | 1.5 | 9 | 5 | 3 | 1.5 | 4 | 6 | 7 | 8 |
| 8 | 6.5 | 1 | 2 | 6.5 | 9 | 4.5 | 4.5 | 3 | 8 |
| 9 | 1 | 8 | 3 | 4.5 | 2 | 4.5 | 6 | 7 | 9 |
| 10 | 1.5 | 6 | 8 | 1.5 | 4 | 3 | 9 | 7 | 5 |
| 11 | 3 | 7 | 5.5 | 1 | 5.5 | 2 | 4 | 8 | 9 |
| Avg. Rank | **1.86** | 7.32 | 4.86 | 3.82 | 3.64 | 3.95 | 5.86 | 5.86 | 7.82 |
| Nemenyi Test | Control Classifier | **4.67** | 2.57 | 1.67 | 1.52 | 1.79 | **3.43** | **3.43** | **5.10** |

Table 4 Classification algorithms ranking (Friedman Test)

Table 4 reports that SAHBN model was significantly outperformed NN, BN (ICSS), SVM, and KNN. Additionally, it shows that SAHBN perform much better that the K2 search-based Bayesian Network classifier. Furthermore, the performance of DT (J48), TAN search-based Bayesian Network and NB were slightly lower than SAHBN.

Finally, the results are depicted using the Critical Difference (CD) significant diagram proposed by [35]. Figure 4 plots the compared classifiers in Y axis, and their corresponding mean ranks in the X axis. The line to the right of each classifier's mean rank represents the critical difference associated with the classifier. That is, other classifiers mean ranks located to the right of the critical difference line are significantly outperformed by the correspond classifier.

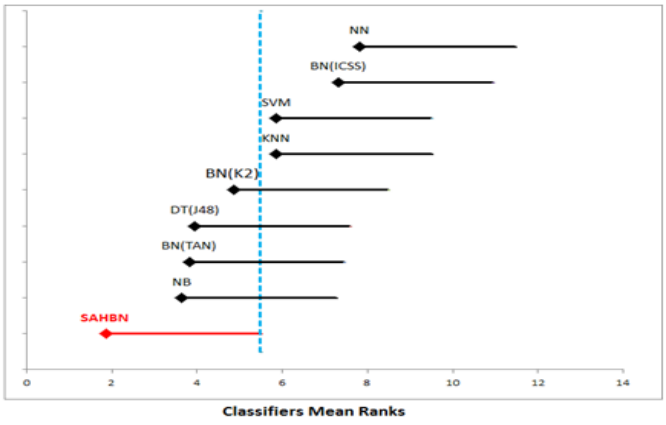


Figure 4 Pairwise comparisions of all classifiers

## 4. DISCUSSION AND CONCLUSIONS

This paper investigated the potential advantages of integrating the domain knowledge in the form of GO and the HBN classifier. Accordingly, the proposed Semantically Aware Hierarchical Bayesian Network (SAHBN) classification model was tested using data set in the biomedical domain. Consequently, the findings extracted from analyzing the obtained results indicated that SAHBN model demonstrated a very competitive performance in comparison with conventional classification algorithms. SAHBN model outperformed existing algorithms in eight experiments

out of eleven. Furthermore, it scored the highest average performance rank among all other classifiers.

SAHBN model exploited the ontological knowledge to construct a consistent training data set and eliminate contradictions between prediction attributes. Additionally, SAHBN structure implicitly reflects the background knowledge of the targeted domain. Hence, it is a self-explanatory structure that can be readily be maintained.

SAHBN model not only highlighted the advantages of integrating ontology with the HBN classifier but also laid out the foundations to consider amore semantic relations between prediction attributes, such as equivalent, disjoint union and intersection. Future works, will investigate the advantages of integrating more ontological knowledge with the SAHBN classification model.

In summary, the SAHBN model consolidated the GO and the HBN classification algorithms in flexible framework that preserves the advantage of ontology and Bayesian theory. Initial results revealed very promising findings that establish a solid foundation for future research.

## REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, And P. Smyth, "From Data Mining To Knowledge Discovery In Databases," Ai Mag., Vol. 17, No. 3, Pp. 37–54, 1996.

[2] C. Zhang And S. Zhang, Association Rule Mining: Models And Algorithms. Springer-Verlag Berlin Heidelberg. Xii, 244., 2002.

[3] P. K. Novak, A. Vavpetic, I. Trajkovski, And N. Lavrac, "Towards Semantic Data Mining With G-Segs," In Proceedings Of The 11th International Multiconference Information Society, Is, 2009.

[4] F. Benites And E. Sapozhnikova, "Using Semantic Data Mining For Classification Improvement And Knowledge Extraction," In Ceur Workshop Proceedings, 2014, Vol. 1226, Pp. 150–155.

[5] L. Cao, "Domain-Driven Data Mining: Challenges And Prospects," Ieee Trans. Knowl. Data Eng., Vol. 22, No. 6, Pp. 755–769, 2010.

[6] C. Antunes And A. Silva, "New Trends In Knowledge Driven Data Mining A Position Paper," Proc. 16th Int. Conf. Enterp. Inf. Syst., Pp. 346–351, 2014.

[7] S. Staab And R. Studer, Hand Book On Ontologies. Springer Science & Business Media, 2013.

[8] G. Mansingh And L. Rao, "The Role Of Ontologies In Developing Knowledge Technologies," Knowl. Manag. Dev. Springer Us, Pp. 145–156, 2014.

[9] H. Liu, "Towards Semantic Data Mining," In Proc. Of The 9th International Semantic Web Conference (Iswc2010). 2010.

[10] D. Dou, H. Wang, And H. Liu, "Semantic Data Mining: A Survey Of Ontology-Based Approaches," In Proceedings Of The 2015 Ieee 9th International Conference On Semantic Computing, Ieee Icsc 2015, 2015, Pp. 244–251.

[11] S. Fenz, "An Ontology-Based Approach For Constructing Bayesian Networks," Data Knowl. Eng., Vol. 73, Pp. 73–88, 2012.

[12] J. A. Blake, "Ten Quick Tips For Using The Gene Ontology," Plos Comput Biol, Vol. 9, No. 11, P. E1003343, 2013.

[13] E. Gyftodimos And P. A Flach, "Hierarchical Bayesian Networks : An Approach To Classification And Learning For Structured Data," Proceedings Of The Ecml/Pkdd - 2003 Workshop On Probablistic Graphical Models For Classification, Vol. 3025. Pp. 291–300, 2004.

[14] J. A. Blake And M. A. Harris, "The Gene Ontology (Go) Project: Structured Vocabularies For Molecular Biology And Their Application To Genome And Expression Analysis," Current Protocols In Bioinformatics, No. Suppl. 23. 2008.

[15] S. Götz And A. Conesa, Visual Gene Ontology Based Knowledge Discovery In Functional Genomics. Intech Open Access Publisher, 2011.

[16] R. P. Huntley, T. Sawford, M. J. Martin, And C. O'donovan, "Understanding How And Why The Gene Ontology And Its Annotations Evolve: The Go Within Uniprot.," Gigascience, Vol. 3, No. 1, P. 4, 2014.

[17] "Gene Ontology Consortium | Gene Ontology Consortium." [Online]. Available: Http://Www.Geneontology.Org/. [Accessed: 21-Dec-2016].

[18] T. D. Nielsen And F. V. Jensen, Bayesian Network And Decision Graph. Springer Science & Business Media, 2009.

[19] D. Koller And N. Friedman, Probabilistic Graphical Models: Principles And Techniques. Mit Press, 2009.

[20] R. G. Almond, R. J. Mislevy, L. S. Steinberg, D. Yan, And D. M. Williamson, "Learning In Models With Fixed Structure," Bayesian Networks Educ. Assessment. Springer New York, Pp. 279–330, 2015.

[21] Z. Ji, Q. Xia, And G. Meng, "A Review Of Parameter Learning Methods In Bayesian Network," In Advanced Intelligent Computing Theories And Applications: 11th International Conference, Icic 2015, Fuzhou, China, August 20-23, 2015. Proceedings, Part Iii, D.-S. Huang And K. Han, Eds. Cham: Springer International Publishing, 2015, Pp. 3–12.

[22] H. E. Wheeler And S. K. Kim, "Genetics And Genomics Of Human Ageing.," Philos. Trans. R. Soc. Lond. B. Biol. Sci., Vol. 366, No. 1561, Pp. 43–50, 2011.

[23] H. Lees, H. Walters, And L. S. Cox, "Animal And Human Models To Understand Ageing," Maturitas, 2016.

[24] T. B. Kirkwood, "The Origins Of Human Ageing.," Philos. Trans. R. Soc. Lond. B. Biol. Sci., Vol. 352, No. 1363, Pp. 1765–72, 1997.

[25] C. Wan, A. A. Freitas, And J. P. De Magalhaes, "Predicting The Pro-Longevity Or Anti-Longevity Effect Of Model Organism Genes With New Hierarchical Feature Selection Methods," Ieee/Acm Trans. Comput. Biol. Bioinforma., Vol. 12, No. 2, Pp. 262–275, 2015.

[26] J. P. De Magalhães Et Al., "The Human Ageing Genomic Resources: Online Databases And Tools For Biogerontologists," Aging Cell, Vol. 8, No. 1. Pp. 65–72, 2009.

[27] A. A Freitas, O. Vasieva, And J. P. De Magalhães, "A Data Mining Approach For Classifying Dna Repair Genes Into Ageing-Related Or Non-Ageing-Related.," Bmc Genomics, Vol. 12, No. 1, P. 27, 2011.

[28] C. Wan And A. Freitas, "Prediction Of The Pro-Longevity Or Anti-Longevity Effect Of Caenorhabditis Elegans Genes Based On Bayesian Classification Methods," In Bioinformatics And Biomedicine (Bibm), 2013 Ieee International Conference On, 2013, Pp. 373–380.

[29] R. D. Wood, M. Mitchell, J. Sgouros, And T. Lindahl, "Human Dna Repair Genes.," Science, Vol. 291, No. 5507, Pp. 1284–9, 2001.

[30] "Human Dna Repair Genes." [Online]. Available: Http://Sciencepark.Mdanderson.Org/Labs/Wood/Dna_Repair_Genes.Html. [Accessed: 08-Dec-2016].

[31] "Genage: The Ageing Gene Database." [Online]. Available: Http://Genomics.Senescence.Info/Genes/. [Accessed: 08-Dec-2016].

[32] "Human Protein Reference Database." [Online]. Available: Http://Www.Hprd.Org/Index_Html. [Accessed: 08-Dec-2016].

[33] "Uniprot." [Online]. Available: Http://Www.Uniprot.Org/. [Accessed: 08-Dec-2016].

[34] J. D. Kelleher, B. Mac Namee, And A. D'arcy, "Fundamentals Of Machine Learning For Predictive Data Analytics." Mit Pr, 2015.

[35] S. Lessmann, B. Baesens, C. Mues, And S. Pietsch, "Benchmarking Classification Models For Software Defect Prediction: A Proposed Framework And Novel Findings," In Ieee Transactions On Software Engineering, 2008, Vol. 34, No. 4, Pp. 485–496.

## AUTHORS

Hasanein Alharbi received his PhD from the School of Computing, Science and Engineering at the University of Salford-Manchester. He holds a master's degree in Computer Application (MCA) from Sardar Patel University in India and a BSc in Computer Science from the University of Babylon, Iraq. Hasanein particular research interests are in the area of semantic data mining, integrating ontology with mining algorithms and mining the linked open data.