Prediction of the Pro-longevity or Anti-longevity Effect of *Caenorhabditis Elegans* Genes Based on Bayesian Classification Methods

Cen Wan
School of Computing
University of Kent
Canterbury, Kent, UK
cw439@kent.ac.uk

Alex Freitas
School of Computing
University of Kent
Canterbury, Kent, UK
A.A.Freitas@kent.ac.uk

***Abstract*—The genetic mechanisms of ageing are mysterious and sophisticated issues that attract biologists' attention. With the help of data mining techniques, some findings relevant to the ageing problem can be revealed. This paper studies the performance of Bayesian network augmented naive Bayes classifier, naive Bayes classifier and proposed feature selection methods for naive Bayes on predicting a *C. elegans* gene's effect on the organism's longevity. The results show that due to the hierarchical structure of predictor attribute values (Gene Ontology terms), the Bayesian network augmented naive Bayes classifier performs better than the naive Bayes classifier, and the proposed feature selection methods for naive Bayes can effectively optimize the predictive performance of naive Bayes.**

***Keywords*-ageing; Gene Ontology; data mining; Bayesian classifiers; feature selection**

I. Introduction

The genetic mechanisms of ageing are mysterious and sophisticated issues that attract biologists' attention on finding the essence of longevity. Although some researches have revealed possible factors relevant to ageing, it is still a puzzle. Caloric restriction has been found to be an approach to extend the longevity of many species [1], probably due to the reduction of metabolic rate, which might be related with the decrease of toxic reactive oxygen species [2]. In addition, environmental factors, such as temperature, oxidative stress, etc., are also related with longevity [3]. Furthermore, mutations in some DNA repair genes lead to accelerated ageing [4]. Given the uncertainty about which types of genes mostly influence the ageing process, the study of ageing should take into account the comprehensive analysis of many types of genes in the genome of organisms.

Data mining (or machine learning) methods have been recently applied to the analysis of ageing-related genes. We focus on the classification task of data mining, where the algorithm builds, from the training dataset, a classification model that predicts the classes of genes in the testing dataset (unseen during training).

There are few works on classification methods for predicting ageing-related gene functions, as follows. Freitas et al. [4] proposed a method to classify ageing-related DNA repair genes by using two datasets. One dataset mainly contains Gene Ontology (GO) terms and protein-protein interaction information as predictor attributes features, and the other one merely contains gene expression data as predictor attributes. Note that, in their research, the hierarchical structure of the GO including generalization/specialization relationships between GO terms was not taken into account by the classifiers. Li et al. [5] also proposed an approach to predict ageing-associated genes based on features of a functional network that was built using information about gene sequence, genetic interactions, physical interactions, etc. In addition, Huang et al. [6] used features derived from a deletion network that was built by merging the information about the deletion effect of genes on longevity and a network derived from STRING database, which was a protein-protein interaction network built to predict whether the deletion of one gene will increase or decrease the lifespan of yeast.

From the perspective of biological data, GO has been considered as an important cornerstone of protein and gene function prediction research because of its contribution to the unification of biological knowledge [7]. There are several works that predict GO term annotations for proteins based on GO term annotations of neighbouring proteins in protein-protein networks, e.g., [8], [9]; or perform hierarchical

predictions of GO terms, e.g., [10]. Note, however, in such works, GO terms have the role of functional classes to be predicted, rather than having the role of predictor attributes as in our work. The relationship between GO terms is a type of generalization/specialization that is associated with some redundancy in the data, from a classification algorithm's perspective. Therefore, it can be speculated that simply and completely ignoring that information will lead to low predictive accuracy.

For example, the naive Bayes (NB) classifier is built on the assumption that its attributes (GO terms in our case) are independent from each other, given the class. However, the redundancy associated with the generalization/specialization relationships between GO terms violates that assumption. Therefore, it is possible that another type of Bayesian classifier which takes into account the hierarchical relationships between GO terms would produce better results.

This work aims to investigate the usefulness of hierarchical relationships between GO terms on improving the predictive performance of Bayesian classifiers, in the context of ageing-related gene classification. We focus on genes from *C. elegans*, a major model organism for ageing research; and we predict the effect of a gene on the organism's longevity, using GO terms as predictor attributes.

We evaluated two well-known types of Bayesian classifiers, Bayesian Network Augmented Naive Bayes (BAN) and Naive Bayes (NB). We propose a new approach to define the network structure (or topology) of a BAN, naturally exploiting the GO's hierarchical relationships, and propose new feature selection methods based on NB.

The remainder of this paper is structured as follows. In section II, background on NB, BAN and feature selection methods is reviewed. Section III describes the classification methods proposed in this work. Section IV presents the experiments' results, followed by discussion in Section V. Finally, a conclusion is presented in Section VI.

II. Background

A. Naive Bayes and Bayesian Network Augmented Naive Bayes

Naive Bayes (NB) and Bayesian Network Augmented Naive Bayes (BAN) are popular classifiers because of their powerful predictive ability and interpretability of their graphical models. The NB classifier uses the inference formula shown in Equation (1):

$$P(y|x_1, x_2, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \qquad (1)$$

where *n* is the number of attributes and the probability of a class attribute value *y* given all attribute values of an instance is estimated by calculating the product of the individual probability of each attribute value x_i given *y* times the prior probability of *y*. However, NB has the limitation of assuming that all the attributes are independent from each other, given the class. In this work, this is a serious limitation, since the attributes are GO terms, and there are strong dependencies between many GO terms, due to hierarchical (generalization/specialization) relationships between GO terms.

On the contrary, BAN does not have that limitation. It represents the dependencies between attributes by a Bayesian network, where nodes are attributes (GO terms, in our case) and edges represent attribute dependencies [11]. The inference formula is shown in Equation (2):

$$P(y|x_1, x_2, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|Pa(x_i), y) \qquad (2)$$

where the probability of a class attribute value *y* given all attribute values of an instance is estimated by calculating the product of the individual probability of each attribute value x_i given its parent attribute(s) value(s) *Pa(x_i)* and class attribute value *y* times the probability of *y*.

B. Feature Selection

Feature selection plays an important role in data mining, since it is often adopted for improving the predictive performance of a classifier based on certain criteria. Feature selection methods can be categorized into two groups, wrapper and filter approaches. The wrapper approach for feature selection directly measures the relevance of a subset of predictor attributes by evaluating the predictive performance of a classifier built using that attribute subset [12]. In contrast, the filter approach indirectly measures the relevance of predictor attributes by using information gain, distance, dependence, consistency or other criteria [12]. Comparing with the wrapper approach, the filter approach has the advantages of low time complexity and better scalability to datasets with large number of attributes, such as the datasets used in this work, which have from 361 to 586 predictor attributes. Therefore, in this work, we use the filter approach for optimizing the predictive performance of NB.

III. Proposed Methods

A. GO-hierarchy-aware BAN

In the data mining literature, the structure (topology) of BANs are usually learned from the data. In this work, we propose instead using the hierarchical relationships between GO terms as the structure of a BAN. This not only avoids the need for a computationally expensive procedure for learning the BAN's structure, but also has the advantage of directly employing the naturally expert-defined hierarchical information of the GO for Bayesian inference. Hence, the proposed classifier can be named as GO-hierarchy-aware BAN classifier. Figure 1 shows the topology of BAN, where each GO term is connected with its parent GO term(s) and class attribute (e.g., GO:0006810 and GO:0044699 are parent GO terms of GO:0044765). In this figure, the solid edges represent dependencies between GO terms (defined by hierarchical relationships specified in the GO documentation), whist the dashed edges represent the fact that each predictor attribute (GO term) depends on the class attribute, as usual in a BAN.

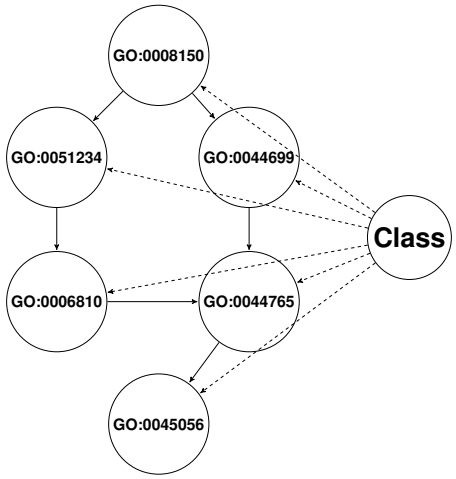


Figure 1. Topology of BAN Classifier Based on Gene Ontology Data

B. New Feature Selection Approaches for Naive Bayes

In order to remove redundant attributes from the dataset, we exploit hierarchical relationships between GO terms. We propose a feature selection method based on two ideas: selecting attributes which are individually good predictors of the class, and removing redundant attributes.

1) Attribute Predictive Power Measure: Intuitively, the predictive performance of naive Bayes is sensitive to the predictive power of individual predictor attributes. Therefore, as a part of our feature selection method, we propose the use of Equation (3) to measure the predictive power of an attribute (GO term).

$$Relevance(GO) = (P(Class = Pro \mid GO = Yes) - P(Class = Pro \mid GO = No))^2 + (P(Class = Anti \mid GO = Yes) - P(Class = Anti \mid GO = No))^2 \quad (3)$$

A general form of Equation (3) was originally used in [13], in the context of an instance-based learning (nearest neighbour) algorithm, but the original formula was adapted to our context of feature selection for naive Bayes. Our feature selection method uses Equation (3) to calculate the relevance of each GO term as a function of the difference in the conditional probabilities of each class given different values (yes or no) of a GO term indicating whether or not a *C. elegans* gene is annotated with that GO term. In addition, in Equation (3), the calculation of each probability uses Laplace correction with the purpose of avoiding extreme probability values (like 0) associated with very few instances [14]. The formula of Laplace correction is shown as follow:

$$P(y \mid x_i) = \frac{C(y|x_i)+1}{C(x_i)+Z} \quad (4)$$

where $C(y \mid x_i)$ denotes the number of training instances belonging to class *y* given attribute value x_i, $C(x_i)$ denotes the number of training instances having attribute value x_i and Z denotes the number of values for the class attribute.

2) Hierarchy-based Redundant Attributes Removal: With the help of hierarchical relationships among GO terms in the dataset, we use Equation (3) to get rid of redundant attributes. The basic principle is that in each path of the DAG built according to GO term relationships, if the value of one GO term attribute GO_i for an instance equals to "0", then the values of all descendant GO terms of GO_i in the instance should equal to "0". This is because, if a descendant GO term of GO_i has the value "1" in an instance, then, due to the "is_a" relationship between GO terms, GO_i would necessarily have the value "1" too, which would be a logical contradiction with the fact that GO_i has value "0". That is, if GO_i has the value "0", this implies that its descendant GO terms have value "0" (in the same instance), characterizing a redundancy of these descendants. In addition, if the value for one GO term GO_i in an instance equals to "1", then the values of all ancestor GO terms of GO_i in the instance

should equal to "1". Due to the "is_a" relationships, so that the ancestors of GO_i represent redundant information.

Based on these principles, we adopt a "lazy" learning method where we build a classifier for each testing instance in turn, rather than building a single classifier for all testing instances. The main advantage of this "lazy" learning approach is that we can select a set of attributes (GO terms) optimized for predicting the class of each testing instance. It is expected that most of redundant attributes in the GO's DAG can be removed from the set of predictor attributes. After that removal process, the remaining predictor attributes are ranked in descending order of their relevance value measured by Equation (3). Then naive Bayes merely adopts the top-k ranked attributes for classification, where k is a user-specified parameter.

As shown in the Pseudocode of Algorithm 1, we firstly create the DAG composed by all GO terms (see Section IV-A), then compute the relevance for every GO term. We use a "lazy" learning method for classifying every testing instance in turn. For each GO term GO_i in the current instance being classified, we select the processing direction according to the value of GO_i in that instance. For example, we compare the relevance of all ancestor GO terms for a GO term GO_i with value "1". If the relevance value of an individual ancestor GO term is equal or lower than the relevance of GO_i, then that ancestor GO term will be removed from the set of predictor attributes. Analogously, in case of "0" being the value of GO term GO_i, descendant GO terms with a value of "0" and with a relevance value equal or lower than the relevance value of GO_i will be removed from the set of predictor attributes. After processing all GO terms in an instance, NB will merely adopt the remaining attributes in the set of predictor attributes.

Concerning the notation used in Algorithm 1, $Dataset_{<Train_1>}$ and $Dataset_{<Test>}$ denote the original training dataset and testing dataset, and they consist of all GO terms used as predictor attributes; $Ancestor_{<GO_i>}$ denotes the set of ancestors for the i^{th} GO term; $Descendant_{<GO_i>}$ denotes the set of descendants for the i^{th} GO term; $Status_{<GO_i>}$ means the selection status of the i^{th} GO term; $Relevance_{<GO_i>}$ denotes the value of relevance for the i^{th} GO term; k means the number of attributes selected to be used as input for naive Bayes; $Instance_{<n>}$ means one instance in $Dataset_{<Test>}$; $Value_{<GO_i>}$ denotes the value of GO_i in that instance;

Algorithm 1 Hierarchy Based Redundant Attribute Removal Naive Bayes Classifier

```
Initialize DAG with all GO terms in Dataset;
Initialize Dataset_<Train_1>;
Initialize Dataset_<Test>;
for each GO_i in DAG do
  Initialize Ancestor_<GO_i> in DAG;
  Initialize Descendant_<GO_i> in DAG;
  Initialize Status_<GO_i> ← "Select";
  Calculate Relevance_<GO_i> in Dataset_<Train_1>;
end for
Initialize k;
for each Instance_<n> ∈ Dataset_<Test> do
  for each GO_i ∈ DAG do
    if Value_<GO_i> ∈ Instance_<n> = 1 then
      for each A_ij ∈ Ancestor_<GO_i> do
        if Relevance_<A_ij> ≤ Relevance_<GO_i> then
          Status_<A_ij> ← "Remove";
        end if
      end for
    else
      for each D_ij ∈ Descendant_<GO_i> do
        if Relevance_<D_ij> ≤ Relevance_<GO_i> then
          Status_<D_ij> ← "Remove";
        end if
      end for
    end if
  end for
  Create Set_<Select_1> with GO_i : Status_<GO_i> =
      "Select";
  Create Set_<Select_2> ← Top-Rank(Set_<Select_1>, k);
  Create Instance_<s> with GO_i ∈ Set_<Select_2>;
  Create Dataset_<Train_2> with GO_i ∈ Set_<Select_2>;
  Classify NaiveBayes(Dataset_<Train_2>, Instance_<s>);
  Re-assign each GO_i : Status_<GO_i> ← "Select";
end for
```

A_{ij} denotes the j^{th} ancestor GO term for the i^{th} GO term; D_{ij} denotes the j^{th} descendant GO term for the i^{th} GO term; $Set_{<Select_1>}$ denotes a set of GO terms whose individual status is "Select" and $Set_{<Select_2>}$ denotes the set of top-k GO terms according to the relevance-based ranking; $Instance_{<s>}$ means the instance that only consists of GO terms in $Set_{<Select_2>}$ from $Instance_{<n>}$; and $Dataset_{<Train_2>}$ means the training dataset that is only composed by GO terms in $Set_{<Select_2>}$ from $Dataset_{<Train_1>}$.

IV. Computational Experiments

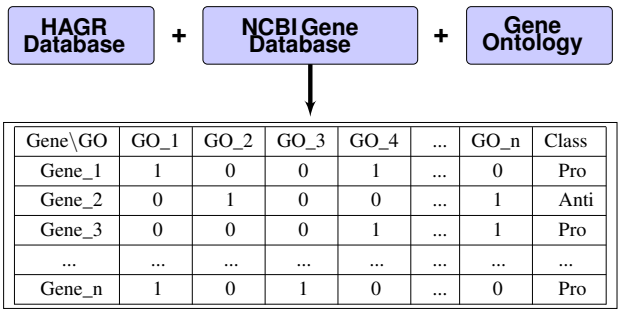
A. Dataset Creation

According to the consideration of reliability of the dataset to be created, the model organism has been chosen as *Caenorhabditis elegans*, which has more completely annotated information about longevity than other types of model organisms in the Human Ageing Genomic Resources (HAGR) database (data-version: Build 16) [15]. The HAGR database provides data about genes and their effects on the organisms' longevity (i.e., pro- or anti- longevity). We also used the Gene Ontology (GO) database (data-version: 2013-01-25), which provides well structured and controlled-vocabulary description of gene functions [7]. The newly created dataset is composed by a set of GO terms associated with each *C. elegans* gene and its effect on the organism's longevity (the class attribute). In addition, it contains information about the hierarchical structure between GO terms, where a GO term can have parent (more general) terms or child (more specific) terms. This hierarchical relationship among GO terms (used as predictor attributes) can be exploited by a classification algorithm.

We firstly match the EntrezID number of each *C. elegans* gene in the HAGR database with the same EntrezID number in the NCBI gene database (data-version: 2012-12-13) [16], which contains the corresponding list of GO terms for each gene. Then the GO terms for matched genes are mapped to data from the GO database, where each GO term has been assigned into one out of three categories of namespaces: biological process, molecular function and cellular component [7]. The biological process GO terms are considered more interpretable in the context of this research, since they refer to biological processes that can be more naturally interpreted as affecting an organism's longevity. Therefore, only the GO terms of the biological process namespace were adopted to build the new dataset.

After retrieving all the biological process GO terms for the *C. elegans* genes, each gene can be described as a set of binary attributes, where each attribute indicates whether or not a given GO term is annotated for a gene. To take into account the hierarchical structure of GO terms, the ancestors for each child GO term are retrieved by following the relationship "is_a" in the GO database. Then each gene is added to the dataset with the information about the gene's effect on longevity (value of the class attribute). Note that the annotations for some genes in the HAGR database contain missing values of longevity, so they were not adopted.

Finally, the structure of the newly created dataset is represented as shown in Figure 2, where the attribute value "1" means the occurrence of a GO term with respect to the corresponding gene. In the column for the class attribute, the values of "Pro" and "Anti" mean pro-longevity and anti-longevity. There are 1507 GO term attributes, plus 1 Class attribute, and 554 gene instances. In addition, the GO term GO:0008150 ("biological process") was removed, since it is the ancestor for all biological process GO terms and not meaningful for classification.



Gene\GO	GO_1	GO_2	GO_3	GO_4	...	GO_n	Class
Gene_1	1	0	0	1	...	0	Pro
Gene_2	0	1	0	0	...	1	Anti
Gene_3	0	0	0	1	...	1	Pro
...	...	...	...	...	...	...	...
Gene_n	1	0	1	0	...	0	Pro

Figure 2. Structure of the Created Dataset

Some GO terms that have few occurrences in the dataset will affect the performance of classifiers, because probabilities calculated from very few instances are not reliable. In particular, the number of GO terms which have only 1 associated gene is 560. In terms of predictive data mining, these GO terms have no generalization ability, and a model that includes these GO terms would be confronted with the overfitting problem and reduction on predictive accuracy. Therefore, it is necessary to investigate what is the most appropriate threshold for the minimum number of occurrences of a GO term, with the purpose of avoiding unreliable probability calculations and mitigating the overfitting problem.

Table 1 shows the number of GO terms remaining in the dataset after adopting different thresholds on the minimum number of occurrences for a GO term, for filtering the dataset. We experimented with thresholds ranging from 4 to 10. The larger the threshold value, the more reliable the probability calculations (used by classifiers) are, but the smaller the number of GO terms available as predictor attributes, possibly leading to a loss of some GO terms with good predictive power for larger thresholds.

Note that, after filtering the dataset by each threshold,

duplicated instances tend to appear [4]. Therefore, duplicated instances are detected and then removed for avoiding interference on the computation of predictive accuracy (i.e., to avoid that an instance occurs in both the training and testing datasets).

Table I
EFFECT OF DIFFERENT THRESHOLDS FOR THE MINIMUM NUMBER OF GENES HAVING A GO TERM ON THE NUMBER OF GO TERMS LEFT IN THE DATASET

Threshold (user-defined parameter) for filtering GO terms	Number of GO terms left in dataset (remove GO term's frequency $<$ threshold)
4	586
5	515
6	465
7	426
8	392
9	373
10	361

B. Evaluation Methodology

The experiments evaluate the predictive performance of 5 types of classification methods, as follows. The first method is the BAN classifier using the GO hierarchical relationships as the structure of the BAN's network. The second method is Hierarchy Based Redundant Attribute Removal Naive Bayes Classifier (HNB). Recall that HNB consists of two phases. First, it executes Algorithm 1 to remove redundant attributes. Second, it selects the top-k remaining attributes based on relevance. The third method is Relevance-based Naive Bayes (RNB), which means naive Bayes merely selects the top-k ranked attributes in descending order of their individual predictive power measured by their relevance (Equation 3). The fourth method is named HNB$_{-s}$, which can be seen as the first phase of HNB. It follows the same approach as HNB, but adopts all remaining attributes after removing redundant attributes (executing Algorithm 1), rather than selecting just the top-k ranked ones. Finally, we included in our experiments NB, as a baseline for measuring the improvement on the predictive performance of our three proposed feature selection methods.

In terms of the k value used to define how many top-quality attributes were selected, we did experiments with the values of 30, 40 and 50, combining them with varying the threshold of minimum number of GO term occurrences ranging from 4 to 10. Hence, our experiments used in total 21 versions of the dataset (3 k values multiplied by 7 threshold values). We report the results with two measures of predictive performance, i.e., accuracy and sensitivity$\times$specificity. Accuracy denotes the proportion of correctly classified instances in the testing dataset. Sensitivity denotes the proportion of correctly classified positive (pro-longevity) genes in the testing dataset, and specificity denotes the proportion of correctly classified negative (anti-longevity) genes in the testing dataset [17]. In all experiments, predictive performance was evaluated by ten-fold cross validation [18].

C. Experiment Results

Table II shows the predictive performance of the algorithms on the 21 dataset versions from the perspective of accuracy (Acc.) and sensitivity$\times$specificity (S.$\times$S.). In the table, "Thr." stands for threshold on minimum number of genes annotated with a GO term and "K" denotes the number of top-ranked GO terms that will be selected as the predictor attributes for the feature selection methods. In each row of the table, the best Acc. and S.$\times$S. are shown in boldface. In addition, the best Acc. and S.$\times$S. results (in each row) out of the three feature selection methods are shown in italic.

For each of the 21 dataset versions, we performed statistical tests of significance (i.e., two-tailed Wilcoxon signed-rank test, at the significance level of 5%) [19] to compare the predictive performance of BAN against the performance of NB, and to compare the predictive performance of the best feature selection method against the performance of NB. The values in underline denote that the corresponding classifier significantly outperforms NB. BAN significantly outperforms NB 13 (out of 21) times, in terms of accuracy, and 8 times in terms of sensitivity$\times$specificity. There is no dataset version where NB significantly outperforms BAN. HNB is overall the best feature selection method, since it significantly outperforms NB 11 times, in terms of accuracy, and 4 times in terms of sensitivity$\times$specificity. There is no dataset version where NB significantly outperforms HNB. RNB is the second best feature selection method, significantly outperforming NB 5 times in terms of accuracy. HNB$_{-s}$ performs worst comparing with the two other feature selection methods, since it never significantly outperforms NB.

In addition, HNB obtained the highest value of predictive accuracy (68.1%) and sensitivity$\times$specificity (41.8%) around the whole experiment, with a sensitivity of 57.5% and a specificity of 72.6%. These are substantial increases against

the baseline values (38.8% and 61.2%) that equal to the relative frequency of instances belonging to the class pro- or anti-longevity, respectively.

Table II
PREDICTIVE PERFORMANCE (%) RESULTS
IN 21 DIFFERENT DATASET VERSIONS

Aliases		**BAN**		**NB**		**RNB**		**HNB$_{-s}$**		**HNB**	
Thr.	**K**	Acc.	S.×S.	Acc.	S.×S.	Acc.	S.×S.	Acc.	S.×S.	Acc.	S.×S.
T4	30	**66.8**	**39.9**	60.0	32.2	*66.4*	26.7	63.4	*33.9*	63.6	33.6
	40	**67.0**	**40.7**	62.5	35.8	63.8	26.1	66.0	*37.7*	*66.4*	35.5
	50	65.5	**39.3**	62.1	35.4	64.2	31.7	63.4	35.2	***68.1***	*37.4*
T5	30	**66.4**	**39.5**	60.8	33.3	63.0	27.5	*63.6*	*35.3*	63.0	34.1
	40	65.1	**38.4**	61.7	34.7	64.9	35.3	64.5	*36.2*	***65.5***	35.8
	50	**67.7**	**41.2**	62.5	35.9	64.9	34.5	64.2	36.5	*65.3*	*36.7*
T6	30	**65.3**	**38.3**	62.1	35.6	62.7	33.4	63.2	*36.1*	*63.4*	35.6
	40	**64.2**	**36.9**	58.0	31.3	62.5	32.8	60.8	32.3	*63.6*	*34.6*
	50	**64.2**	**37.6**	59.3	32.1	63.0	34.5	63.4	35.8	*64.2*	*37.4*
T7	30	**66.3**	**40.0**	59.9	33.0	62.2	32.1	62.9	34.5	*63.9*	*34.6*
	40	63.5	35.5	58.8	31.4	***64.8***	37.7	62.7	35.2	64.4	***39.0***
	50	64.8	**36.3**	59.2	31.1	63.3	30.8	62.0	35.1	***66.1***	*35.4*
T8	30	65.2	37.6	60.1	33.7	63.5	35.3	62.7	36.1	***66.3***	***39.9***
	40	63.3	35.6	58.8	31.6	***63.5***	35.5	60.7	32.2	63.1	***36.4***
	50	65.9	**38.8**	60.7	33.9	61.4	36.7	62.0	34.5	***66.3***	*37.5*
T9	30	**65.7**	**38.9**	59.4	33.0	62.4	*37.9*	59.7	32.2	*63.5*	36.4
	40	65.2	38.5	59.4	32.9	62.2	37.3	60.9	35.0	***66.7***	***41.8***
	50	**65.9**	38.8	59.7	32.2	*65.5*	***39.7***	60.3	32.1	64.4	36.4
T10	30	64.4	36.6	60.1	33.2	61.8	35.7	61.2	33.6	***66.7***	***41.1***
	40	64.6	37.1	58.4	31.6	***65.5***	***40.7***	59.4	32.5	63.9	36.3
	50	**65.9**	**39.1**	59.2	32.5	62.9	*37.0*	58.2	30.3	*65.0*	36.6

V. DISCUSSION

The experiment results reveal that the hierarchical relationships between GO terms are valuable for improving the classification of *C. elegans* genes' effects into pro- or anti-longevity, because BAN, which incorporates hierarchical GO term relationships, significantly outperforms NB in the majority of cases. In addition, HNB, which adopts hierarchical GO term relationships for feature selection, has also successfully improved the predictive performance of NB. The reasons can be explained by further analysis of the three feature selection methods. To begin with, HNB$_{-s}$, which removes redundant attributes but does not select the top-k attributes, performs worse than the other two feature selection methods because the remaining non-redundant attributes do not guarantee a high predictive accuracy. Some of the remaining attributes might still have relatively low predictive power. Hence, selecting the top-k ranked GO terms after processing by HNB$_{-s}$ (i.e., executing the full two-phase HNB method) should improve the predictive accuracy. Indeed, HNB obtained higher accuracy than HNB$_{-s}$ in 20 out of 21 dataset versions, and higher sensitivity×specificity values in 16 dataset versions.

Moreover, the reason why HNB outperforms RNB is that the former removes redundant attributes and selects attributes that have relatively higher predictive power, whereas the latter merely selects a set of relatively higher predictive power attributes that might be redundant. Those results show that the first phase of HNB, where redundant attributes are removed based on GO hierarchical relationships, improves predictive performance in general.

To compare BAN and HNB, we conducted the Wilcoxon signed-rank statistical tests [19]. The results show that there is no significant difference between their predictive performances. Hence, this shows that with the help of a feature selection method based on hierarchical GO term relationships, NB can be improved to the same level of predictive performance as BAN.

In addition, the relevance measure (Equation 3) allows us to rank the GO terms in decreasing order of their individual predictive power. For instance, in the dataset version with minimum GO term occurrence threshold of 4 (which is the dataset version with the largest number of GO terms in our experiments), the most relevant GO term was GO:0009314 ("response to radiation"), reinforcing the association between radiation and ageing suggested decades ago in the literature [20]; and the GO term GO:0031667 ("response to nutrient level") was joint second in the ranking, reinforcing the association between nutrient levels and ageing mentioned in Section I. Note that here we discussed the relevance-based ranking of GO terms in general, rather than using the results of HNB, which combines the relevance measure with GO hierarchy-based attribute removal. This is because, as discussed earlier, HNB performs "lazy" feature selection, where different subsets of attributes (GO terms) are removed for classifying different testing instances (genes). In contrast, although the relevance measure does not cope with redundancy among GO terms, it provides a simple approach to rank individual GO terms across the entire dataset, taking into account all *C. elegans* genes.

VI. CONCLUSIONS

This work reveals that the hierarchical relationships between GO terms are helpful for detecting redundant predictor attribute values, and the naive Bayes classifier's predictive

performance at classifying *C. elegans* genes into pro- or anti-longevity genes can be effectively improved by removing redundant attribute values based on the GO hierarchy. Also, with the help of hierarchical GO term relationships, BAN performs well in the above classification task and is shown to perform statistically as effectively as HNB. In addition, both the proposed BAN and HNB classifiers obtained significantly better results than the baseline naive Bayes classifier, as confirmed by statistical tests of significance.

As future work, we will focus on other approaches for removing redundant attributes and representing dependencies between attributes using the hierarchical structure of the Gene Ontology.

ACKNOWLEDGMENT

We thank Dr. João Pedro de Magalhães for his valuable general advice for this project.

REFERENCES

[1] E. J. Masoro, "Overview of caloric restriction and ageing," *Mechanisms of Ageing and Development*, vol. 126, no. 9, pp. 913-922, Sept. 2005.

[2] L. Guarente and C. Kenyon, "Genetic pathways that regulate ageing in model organisms," *Nature*, vol. 408, no. 6809, pp. 255-262, Nov. 2000.

[3] C. J. Kenyon, "The genetics of ageing," *Nature*, vol. 464, no. 7288, pp. 504-512, Mar. 2010.

[4] 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, Jan. 2011.

[5] Y. H. Li, M. Q. Dong and Z. Guo, "Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*," *Mechanisms of Ageing and Development*, vol. 131, no. 11-12, pp. 700-709, Nov.–Dec. 2010.

[6] T. Huang, J. Zhang, Z. P. Xu, L. L. Hu, L. Chen, J. L. Shao, L. Zhang, X. Y. Kong, Y. D. Cai and K. C. Chou, "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017-1025, Apr. 2012.

[7] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25. no. 1, pp. 25-29, May 2000.

[8] N. Nariai, E. D. Kolaczyk and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS ONE*, vol. 2, no. 3, p. e337, Mar. 2007.

[9] A. Mitrofanova, V. Pavlovic and B. Mishra, "Prediction of protein functions with gene ontology and interspecies protein homology data," *Computational Biology and Bioinformatics, IEEE/ACM Transactions*, vol. 8, no. 3, pp. 775-784, May-June 2011.

[10] A. Sokolov and A. Ben-Hur, "Hierarchical classification of Gene Ontology terms using the GOstruct method," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 2, pp. 357-376, Apr. 2010.

[11] N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131-163, Nov. 1997.

[12] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Norwell, MA: Kluwer Academic Publishers, 1998.

[13] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12. pp. 1213-1228, Dec. 1986.

[14] B. Cestnik, "Estimating probabilities: a crucial task in machine learning," in *Proceedings of the Ninth European Conference on Artificial Intelligence*, 1990, pp. 147-149.

[15] J. P. de Magalhães, A. Budovsky, G. Lehmann, J. Costa, Y. Li, V. Fraifeld and G. M. Church, "The Human Ageing Genomic Resources: online databases and tools for biogerontologists," *Aging Cell*, vol. 8, no. 1, pp. 65-72, Feb. 2009.

[16] National Center for Biotechnology Information. (2012). *gene2go* [Online]. Available FTP: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz

[17] D. G. Altman, and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity," *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, June 1994.

[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1137-1143.

[19] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective.* New York, NY: Cambridge University Press, 2011.

[20] D. Harman, "Aging: a theory based on free radical and radiation chemistry," *Journal of Gerontology,* vol. 2, pp. 298-300, 1957.