Universidade Federal Fluminense
Instituto de Computação
Departamento de Ciência da Computação

Igor Martire de Miranda

Using Protein-Protein Interactions Data to Improve Predictions of the Effect of Aging-Related Genes on the Longevity of Model Organisms

Niterói-RJ
2017

Igor Martire de Miranda

Using Protein-Protein Interactions Data to Improve Predictions of the Effect of Aging-Related Genes on the Longevity of Model Organisms

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Alexandre Plastino de Carvalho - Professor do IC-UFF
Coorientador: Pablo Nascimento da Silva - Doutorando do IC-UFF

Niterói-RJ
2017

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

M672 Miranda, Igor Martire de

Using protein-protein interactions data to improve predictions of the of effect of aging-related genes on the longevity of model organisms / Igor Martire de Miranda. – Niterói, RJ : [s.n.], 2017.
44 f.

Projeto Final (Bacharelado em Ciência da Computação) – Universidade Federal Fluminense, 2017.
Orientadores: Alexandre Plastino de Carvalho, Pablo Nascimento da Silva.

1.Mineração de dados (Computação). 2. Seleção de atributos. 3. Ontologia. 4. Envelhecimento. 5. Longevidade. I. Título.

CDD 005.741

Igor Martire de Miranda

Using Protein-Protein Interactions Data to Improve Predictions of the Effect of Aging-Related Genes on the Longevity of Model Organisms

Trabalho submetido ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Aprovado por:

Prof. Alexandre Plastino de Carvalho, D.Sc. - Orientador

UFF

Pablo N. da Silva

Pablo Nascimento da Silva, M.Sc. - Coorientador

UFF

Prof. Aline Marins Paes Carvalho, D.Sc.

UFF

Prof. Jonnathan dos Santos Carvalho, M.Sc.

IFF

Niterói-RJ

2017

“...inside every old person is a young person wondering what happened.“
— Terry Pratchett, 1990

Acknowledgments

I would like to express my sincerest gratitude to everyone who, directly or indirectly, helped me to reach where I am. From family, friends, professors, and colleagues to the entirety of Brazil, the country that provided me with free, quality education. I hope this paper will be only one of my many contributions to society.

In particular, I would like to give special thanks to my professor Alexandre Plastino, who accompanied me every step of the way and always trusted me much more than I thought I deserved. Also, a special thanks to Pablo Nascimento da Silva for always being accessible to assist me in this research and for working so hard to review my writing. My many thanks to Alex Freitas and Fábio Fabris for the invaluable remote meetings consistently full of knowledge and insights.

I would like to acknowledge the many professors, friends, and colleagues who inspired me in some way to become a better person and professional. Without that inspiration, I would not be anywhere close to where I am.

I would also like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the invaluable opportunity offered through the Science without Borders exchange program by sponsoring my studies at the University of California, Berkeley, which greatly contributed to my academic development.

Finally, my deepest and sincerest gratitude to my family, especially my mother, father and brother. All my achievements are but a reflection of their influence in my life.

Abstract

This work focuses on improving the predictive performance on the task of classifying the effect of aging-related genes on the longevity of model organisms. Usually, datasets for this problem are built using gene ontology features. We show that, by introducing a new set of features based on protein-protein interactions, we can improve the classification in a statistically significant way. Although beneficial, the introduction of these features brings a new problem for this task: performing classification based on uncertain information. This happens because the values of the new features only represent how certain we can be that an interaction exists between two proteins. To cope with this problem, we propose a novel similarity coefficient, based on the Jaccard index, and show that, by using it, we can benefit from the introduced uncertain data without adding too much complexity to the classification solution. Finally, we demonstrate how to achieve even better results on this task by performing a feature selection procedure.

Keywords: aging, classification, feature selection, jaccard similarity, longevity, protein-protein interaction

Resumo

Este trabalho foca em melhorar a performance preditiva na tarefa de classificação do efeito de genes relacionados ao envelhecimento na longevidade de organismos-modelo. Normalmente, bases de dados contruídas para este problema usam atributos da ontologia de genes. Nós demonstramos que, ao introduzirmos um novo conjunto de atributos baseados em interações proteína-proteína, conseguimos melhorar a classificação de forma estatisticamente significativa. Embora benéfica, a introdução desses atributos geram um novo problema: realizar a tarefa de classificação usando dados incertos. Isso ocorre porque os valores desses novos atributos representam apenas um grau de certeza da existência de uma interação entre duas proteínas. Para lidar com esse problema, propomos um novo coeficiente de similaridade, baseado no coeficiente de Jaccard, e mostramos que, ao usá-lo, podemos nos beneficiar desses dados incertos introduzidos nas bases sem aumentar demasiadamente a complexidade da solução de classificação. Por fim, nós demonstramos como obter resultados ainda melhores através da aplicação de uma técnica de seleção de atributos.

Palavras-chave: classificação, interação proteína-proteína, envelhecimento, longevidade, seleção de atributos, similaridade de Jaccard

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

Almost three decades ago, Friedman & Johnson (1988) showed that a single gene mutation could increase the lifespan of the *C. elegans* model organism up to 110%. Since then, biologists have discovered how to extend the lifespan of several other model organisms by single mutations [Ayyadevara et al. 2008].

Also, 20 to 30% of overall variation in human adult lifespan is believed to be due to genetic factors [Iachine et al. 2006], and the positive economic impact of delayed aging in humans is projected in trillions of dollars over 50 years in the United States alone [Goldman et al. 2013].

Therefore, it is clear the potential and importance of finding genes that regulate lifespan. However, due to biological and ethical difficulties associated with conducting aging experiments in humans, model organisms are usually the target of aging research [Wan, Freitas e Magalhães 2015]. Even though it is not possible to simply extrapolate the findings in mutant model organisms to humans, similarities in the phenotypes of many of these mutants indicate that the mutations may have a similar effect on organisms of higher complexity as well [Guarente e Kenyon 2000].

In any case, wet-lab identification of aging genes is a tedious and labor-intensive activity [Feng et al. 2012]. It cannot deal alone with the ever-increasing amount of biological data coming from high-throughput experiments. Thus, it is essential to develop and apply machine learning techniques that can assist aging researchers on reaching meaningful biological conclusions.

1.2 Goals

Our goal is to improve the predictive performance achieved in the task of classifying aging-related genes into either pro- or anti-longevity. As the name suggests, a pro-longevity gene has the effect of extending the longevity of an organism when its expression is increased. On the other hand, an increase in an anti-longevity gene's expression has the opposite effect, reducing an organism's lifespan.

More specifically, we aim at improving the classification performance of the Nearest Neighbor (NN) classifier, since, in previous work, it has been shown effective for this task [Wan, Freitas e Magalhães 2015].

1.3 Proposed Solutions

We propose, for the first time in the literature, the usage of protein-protein interactions (PPIs) data as features in the task of classifying the effect of genes on longevity. We propose the introduction of these features in order to assist the classification task by providing more meaningful data that will then be leveraged by the classifier to improve its predictions.

Also, we propose a novel Jaccard-based similarity coefficient in order to leverage the uncertain information from the PPI features while keeping the desirable Jaccard index properties that make it fit for working in sparse datasets.

Finally, we propose a feature selection procedure in order to improve the predictive performance even further. We propose a Filter technique which uses the F-statistic as a feature evaluation measure.

1.4 Text Organization

The remainder of this work is organized as follows:

Chapter 2 reveals related work in the field of bioinformatics and genetics that had goals aligned with our task or shared similar characteristics, such as the features being used. We also refer to a few of the latest developments on the broader field of data mining and machine learning that relate to the classification problem we face in this work.

Chapter 3 presents our first proposed solution for improving prediction performance by introducing new features. Then, it analyzes the results obtained by this proposed solution in order to verify if there were relevant improvements. In Chapter 4, we define our proposed novel Jaccard-based similarity coefficient and evaluate the statistical significance of its improvements over the results obtained by previous methods. Then, our proposed solution of performing feature selection is described in Chapter 5 along with its results.

Finally, Chapter 6 presents the conclusions of this work as well as future research directions.

Chapter 2

Related Work

There have been many applications of classification techniques in the field of aging research. Freitas, Vasieva & Magalhães (2011) and Jiang & Ching (2011) worked on labeling DNA repair genes as aging-related or non-aging-related. Taking a different direction, Fang et al. (2013) focused on classifying aging-related genes as related or non-related to DNA repair. Huang et al. (2012) proposed a two-stage classification process in order to determine the effect on longevity due to the deletion of a gene. The first stage would predict if the deletion would change yeast's lifespan, while the second would predict if this change in lifespan would be an increase or decrease. Other works [Wan e Freitas 2013] [Wan, Freitas e Magalhães 2015] focused on the same task we study in this work: the classification of aging-related genes into pro-longevity or anti-longevity.

Regarding commonly used features, several studies of classification in the area of aging research make use of the Gene Ontology [Freitas, Vasieva e Magalhães 2011] [Fang et al. 2013] [Fabris e Freitas 2016]. For example, a recent study on this topic [Wan, Freitas e Magalhães 2015] uses Gene Ontology features alone and focuses on hierarchical feature selection methods to improve the predictive performance of the Naive Bayes and NN classifiers.

Also, there have been attempts to use features based on protein-protein interactions (PPIs) in classification tasks [Freitas, Vasieva e Magalhães 2011] [Fang et al. 2013] by using the Human Protein Reference Database (HPRD) [Prasad et al. 2009], from which one can obtain binary features without uncertainty. However, HPRD only provides human PPI data, while, in this work, we are interested in model organisms, which leads us to deal with the uncertain values provided by the STRING database [Szklarczyk et al. 2014].

The classification of uncertain data has been studied extensively in the last two decades. Many different techniques have been adapted to handle uncertain data, such as Bayesian approaches [Ren et al. 2009], Neural Networks [Ge, Xia e Nadungodage 2010], Decision Trees [Tsang et al. 2011], *k*-Nearest Neighbors [Yang et al. 2015] and Support Vector Machines [Yang e Li 2009]. The majority of these techniques focus on uncertain numerical features, not specifically on uncertain binary features. Notwithstanding, very few uncertain data mining studies focus on sparse datasets, and they are usually related to other tasks, such as Frequent Itemset Mining [Xu et al. 2014].

This lack of development in the area of classification with uncertain data in sparse datasets may explain why much of the research done so far in the bioinformatics field has ignored the uncertain information provided by the STRING database about PPIs. This discard of the uncertainty information has been done by applying *ad-hoc* cut-off values (i.e., a feature value greater than a pre-defined threshold is deemed a positive binary feature with value equal to 1) such as 0.3 [Kulmanov, Khan e Hoehndorf 2017], 0.7 [Gao et al. 2017], and 0.9 [Lin et al. 2016].

Finally, even though the introduction of features based on PPIs have assisted in similar tasks [Freitas, Vasieva e Magalhães 2011] [Fang et al. 2013], we could not find any previous work that explored the usage of these features for the task we study here.

Chapter 3

Introducing Protein-Protein Interaction Features

Previous research on classification of genes' effect on longevity [Wan e Freitas 2013] [Wan, Freitas e Magalhães 2015] have used Gene Ontology features, which describe genes functions [Ashburner et al. 2000] and are broadly adopted in aging research.

In this work, in order to improve classification performance, we propose the introduction of features based on protein-protein interactions (PPIs), which indicate proteins that jointly contribute to a shared function. The usage of PPIs are not novel in classification tasks for aging research [Freitas, Vasieva e Magalhães 2011] [Fang et al. 2013], however they have never been explored in the task of predicting genes' effect on longevity.

Also, the ways PPI features were used in previous works were far too different from the one we propose here. Those works were focused on human genes and used PPI features from the Human Protein Reference Database (HPRD) [Prasad et al. 2009]. That database focuses on providing manually curated information on the interactions of human proteins. However, we are interested in the study of model organisms and so we use the popular STRING database [Szklarczyk et al. 2014] as the source of PPI data. The first difference this brings is on the number of PPIs: the STRING database provides us with many more PPIs than the HPRD. This is only possible because it uses several evidence channels, some of which are based on automated processes of determining proteins associations, such as textmining or by prediction methods. While this allows more information to be provided, the data is less precise. Thus, the second major difference to previous works is that we are provided with certainty scores for each interaction, which brings the challenge of dealing with uncertain values in our classification task. But before we discuss how we deal with those introduced uncertain values, we describe the datasets used in this work.

3.1 Datasets

We use 28 datasets of aging-related genes, where instances are genes and the binary class indicates whether or not the genes are related to longevity. These datasets were created by integrating data from the Human Ageing Genomic Resources (HAGR) GenAge database (version: 335 Build 17) [Magalhães et al. 2009] and the Gene Ontology (GO) database (version: 2015-10-10) [Ashburner et al. 2000]. HAGR is a database of aging- and longevity-associated genes in model organisms which provides aging information for genes from four model organisms: *C.elegans* (worm), *D.melanogaster* (fly), *M.musculus* (mouse) and *S.cerevisiae* (yeast). The GO database provides information about three ontology types: biological process (BP), molecular function (MF) and cellular component (CC). Each ontology contains a separate set of GO terms (features). So, for each of the four model organisms, we created seven datasets, with seven combinations of feature types, denoted by BP, CC, MF, BP.CC, BP.MF, CC.MF, and BP.CC.MF.

Hence, each dataset contains instances (genes) from a single model organism. Each instance is formed by a set of binary features indicating whether or not the gene is annotated with each GO term and a binary class variable indicating if the instance is either positive ("pro-longevity" gene) or negative ("anti-longevity" gene) according to the HAGR database. These GO features values are highly sparse and, in order to avoid overfitting, GO terms which occurred in less than three genes were discarded, avoiding the use of rare features with very little statistical support and virtually no generalization power for our set of genes.

Finally, as a contribution to the aging-related genes classification problem, in order to improve the predictive performance achieved in previous works when using only GO terms [Wan, Freitas e Magalhães 2015], we added protein-protein interactions (PPIs) uncertain data from the STRING database (version: 10) [Szklarczyk et al. 2014] to each of the 28 datasets. The data is also highly sparse and, as we did with the GO features, we also filtered out the PPIs that only occurred in less than three genes.

These PPI features values were obtained, in the STRING database, from the *combined_score* field in the *network.node_node_links* table. Their values $s \in [0,1]$ indicate the degree of confidence of their correspondent interactions. We use these values under a probabilistic perspective, where the features can be seen as binary ones (with the value 0 (1) indicating absence (presence) of the correspondent PPI in that instance's set of PPIs) and their values are represented by a probability distribution function f, defined as $f(1) = s$ and $f(0) = 1 - s$.

Table 3.1 shows statistics for each dataset, including information on their sparsity. For each of the four model organisms, each of the seven rows shows information about a specific dataset. The first column identifies the model organism. The second column shows the selected Gene Ontologies type(s) on the dataset. The other columns show, respectively, the number of features, the number (and percentage) of GO features, the number of PPI features, the average percentage of GO features with value 0 in an instance, the average percentage of PPI features with value 0 in an instance, the number of instances, the number (and percentage) of positive-class instances and the number of negative-class instances. For example, for the *C. elegans* dataset with GO terms of the Biological Process (BP) ontology type only (first row), out of the 12,438 features, 991 (7.97%) are GO features and the remaining 11,447 (92.03%) are PPI features. Also, the column "avg. % GO = 0" shows that, on average, an instance of that dataset has 95.48% of its GO features with value 0 and the column "avg. % PPI = 0" shows that, on average, an instance of that dataset has 95.32% of its PPI features with value 0. Finally, the last three columns show that this dataset has 657 instances, from which only 226 (34.40%) are labeled *positive* (Pos) and the remaining 431 (65.60%) are labeled *negative* (Neg).

3.2 Handling Uncertain Values

In this work we use the Nearest-Neighbor (NN) classifier since it has been shown effective [Wan e Freitas 2013] [Wan, Freitas e Magalhães 2015] in the task of predicting genes' effect on longevity. Another advantage of the NN is that it is easy to understand and implement, thus inviting and welcoming a broader range of biologists in the aging field to explore data mining techniques and benefit from them.

Being a distance-based classifier, the NN requires the definition of a distance metric

Dataset		#	# (%)	#	avg. %	avg. %	#	# (%)	#
Organism	GO types	features	GO features	PPI features	GO = 0	PPI = 0	instances	Pos	Neg
C. elegans	BP	12438	991 (7.97)	11447	95.48	95.32	657	226 (34.40)	431
	CC	11163	178 (1.59)	10985	93.35	94.63	484	176 (36.36)	308
	MF	11151	263 (2.36)	10888	94.93	94.58	504	190 (37.70)	314
	BP.CC	12626	1169 (9.26)	11457	95.47	95.35	664	228 (34.34)	436
	BP.MF	12733	1254 (9.85)	11479	95.65	95.35	663	227 (34.24)	436
	CC.MF	11731	441 (3.76)	11290	95.01	94.87	566	205 (36.22)	361
	BP.CC.MF	12912	1432 (11.09)	11480	95.62	95.37	667	229 (34.33)	438
D. melanogaster	BP	7359	800 (10.87)	6559	91.68	91.11	132	95 (71.97)	37
	CC	6549	89 (1.36)	6460	86.98	90.85	122	86 (70.49)	36
	MF	6698	145 (2.16)	6553	92.28	90.92	126	89 (70.63)	37
	BP.CC	7503	889 (11.85)	6614	91.38	91.20	133	95 (71.43)	38
	BP.MF	7559	945 (12.50)	6614	91.89	91.20	133	95 (71.43)	38
	CC.MF	6817	234 (3.43)	6583	90.72	91.17	130	92 (70.77)	38
	BP.CC.MF	7648	1034 (13.52)	6614	91.56	91.20	133	95 (71.43)	38
M. musculus	BP	11513	1332 (11.57)	10181	89.35	90.04	109	75 (68.81)	34
	CC	10236	142 (1.39)	10094	83.20	90.11	107	73 (68.22)	34
	MF	10323	240 (2.32)	10083	90.27	89.86	106	72 (67.92)	34
	BP.CC	11655	1474 (12.65)	10181	88.79	90.04	109	75 (68.81)	34
	BP.MF	11753	1572 (13.38)	10181	89.53	90.04	109	75 (68.81)	34
	CC.MF	10563	382 (3.62)	10181	87.93	90.04	109	75 (68.81)	34
	BP.CC.MF	11895	1714 (14.41)	10181	89.03	90.04	109	75 (68.81)	34
S. cerevisiae	BP	6305	844 (13.39)	5461	94.65	92.25	331	44 (13.29)	287
	CC	5606	145 (2.59)	5461	89.96	92.25	331	44 (13.29)	287
	MF	5682	221 (3.89)	5461	94.27	92.25	331	44 (13.29)	287
	BP.CC	6450	989 (15.33)	5461	93.96	92.25	331	44 (13.29)	287
	BP.MF	6526	1065 (16.32)	5461	94.57	92.25	331	44 (13.29)	287
	CC.MF	5827	366 (6.28)	5461	92.56	92.25	331	44 (13.29)	287
	BP.CC.MF	6671	1210 (18.14)	5461	94.02	92.25	331	44 (13.29)	287

Table 3.1: Statistics for each dataset

to measure how distant (dissimilar) two instances are. We choose the Jaccard distance (defined in Section 4.2) also because of its proven effectiveness in related previous work, but we can see its widespread use in other subfields of bioinformatics as well [Sato et al. 2005] [Schloss et al. 2009] [Prokopenko et al. 2015]. This is because the Jaccard distance is particularly good at dealing with sparse binary values. We can notice how sparse our data is by looking at the sixth and seventh columns of Table 3.1. However, since the PPI data we introduced is uncertain and the Jaccard distance does not handle uncertain values, we are required to somehow convert these values to binary values without uncertainty in order to use this distance metric.

As shown in Chapter 2, a common way in the bioinformatics literature of doing such transformation on PPI data from the STRING database is by applying a confidence cut-off value: values with confidence greater or equal than the cut-off are set to 1 and set to 0 otherwise. The STRING database online search interface suggests four cut-off values: 0.15, 0.40, 0.70 and 0.90, meaning, respectively, low, medium, high and highest confidence, which have also been extensively employed in the related literature [Shi et al. 2017] [Gao et al. 2017]. However, which of these four cut-off values to choose for each dataset?

To solve this problem, for each dataset we can choose the cut-off value that, when applied, allows our classification model to achieve the best performance when evaluated on the training set. It is important to notice that this choice is made without ever accessing the validation set. Because of that, this procedure will not necessarily choose the cut-off value that would allow our classifier to achieve the best predictive performance in the validation set. However, this procedure does its best in choosing the cut-off value given the available information at the time of classification, which is the training set only.

This way, we can still use the Jaccard distance with the NN classifier after removing the uncertainty from the data by using a cut-off value chosen by the explained technique.

As a side note, one could think of using the Euclidean distance with the NN classifier by using the probability values as features values, thus leading to a scenario with "certain" numerical features instead of uncertain binary ones. A preliminary experiment using this strategy has been performed, obtaining very poor results when compared to the other methods explored in this work. These results are somewhat intuitive, since the Euclidean distance is known to be weakly discriminant for multidimensional and sparse data, and also because treating a probability as just a numeric value can lead to wrong assumptions. As an example, think of the case when comparing the distance between two instances with a single uncertain binary feature, and assume this feature's values for both instances are represented by the same probability distribution function f, for which $f(0) = f(1) = 0.5$. The Euclidean distance between these two instances would be zero, even though, if we assume that the (unknown) true value of a feature is binary (an assumption that may or may not be appropriate depending on the application domain), there is a 50% chance that these two instances have the opposite binary values for their single feature, which would result in an Euclidean distance of 1 instead.

Next, we define the experimental methodology we adopt to evaluate the predictive performance obtained by the NN classifier with the Jaccard distance in our datasets. The same methodology is used to obtain the results when the classifier is used on the datasets without the PPIs. This way, we will be able to compare if the introduction of the PPI features indeed improved the classifier's prediction ability or not.

3.3 Experimental Methodology

To evaluate the predictive performance of our classifier we have to be careful due to the class imbalance in our datasets. The last two columns in Table 3.1 show that the distribution of instances belonging to the two classes is imbalanced in all the 28 datasets: usually staying around 70%/30%, but reaching up to 87%/13% on the *S.cerevisiae* organism datasets.

Then, if the simple accuracy measure (the percentage of correctly classified instances) had been used, it would provide us with misleading performance evaluation, since we could trivially obtain a high accuracy (but no useful model) by predicting the majority class for all instances [Japkowicz e Shah 2011]. Hence, we evaluate the predictive performance of the classifiers by using the value of Geometric mean (Gmean), defined as $\mathbf{Gmean} = \sqrt{Sens \times Spec}$, which takes into account the balance of the classifiers's sensitivity (Sens) and specificity (Spec) [Japkowicz e Shah 2011]. Sensitivity (specificity) means the proportion of pro-longevity (anti-longevity) genes that were correctly predicted as pro-longevity (anti-longevity) in the testing dataset [Altman e Bland 1994]. This Gmean is a good choice for our task not only because of the class imbalance issue, but also because it gives the same importance to correctly predicting both classes, which is a desirable characteristic for our task. For illustration purposes, in a fraud detection system, where the impact of a false-negative is much worse than that of a false-positive, the F-measure would be more appropriate than the Gmean for evaluating classifiers.

For model validation, we perform the well-known stratified k-fold cross-validation procedure [Witten et al. 2016] and calculate the Gmean by first aggregating the results of all validation folds into a single confusion matrix, since this way is more unbiased than averaging the Gmeans obtained in each fold [Forman e Scholz 2010], especially because many of our datasets have few instances, leading to small validation folds.

Furthermore, we agree with Zhang & Yang (2015) that the correct choice for the number of folds of the cross-validation technique should take into consideration the goals of the researcher on using the method. Since our goal is model selection, we are more interested in having a low variance than having a low bias when evaluating our classifiers. Also, since our model is unstable, we believe the Leave-one-out cross-validation technique would not meet the requirement of low variance [Zhang e Yang 2015]. Thus, for our specific case, we agree with Breiman & Spector (1992) and Kohavi et al. (1995) and choose $k = 10$ for the external cross-validation and $k = 5$ for the internal cross-validation (to select the cut-off value to be used in the external fold).

Since our model is unstable, we repeat the k-fold cross-validation for a number t of times in order to obtain more stable estimates of performance. In each iteration we generate different folds partitioning by randomizing the instances with a different seed. Then, the final reported result by this procedure is the average of the Gmeans obtained in each of the t cross-validations. Repeating the cross-validation a number of times is important, especially when working with unstable models, to avoid that the variance of the results drastically affect the comparison of different models [Zhang e Yang 2015].

As for the number t of times we repeat each cross-validation procedure, we agree with Breiman & Spector (1992) and Zhang & Yang (2015) that a number between 10 and 20 is already enough for obtaining stable results. For that reason, we choose to repeat the internal cross-validation procedure 10 times. However, we choose to repeat the external cross-validation 30 times to improve even further the reliability of statistical tests.

To summarize, we will be evaluating the classifiers by their average Gmean obtained from 30 repetitions of a stratified 10-fold cross-validation procedure. And, in the case of having to choose a cut-off value to remove the uncertainty of the data, we will be choosing, for each external fold, the best cutoff (in terms of Gmean) from the set of candidates $\{0.15,\ 0.40,\ 0.70,\ 0.90\}$, each of them being evaluated by 10 repetitions of a stratified 5-fold internal cross-validation procedure (on the external training fold only).

3.4 Results

For convenience, we will be referring to the approach of using the NN classifier with Jaccard distance on the datasets with only Gene Ontology features as *Baseline*. And we will refer to the approach of using an Internal Cross-Validation procedure to choose a cut-off value in order to remove the uncertainty of the PPI data as Jaccard-ICV. Table 3.2 compares the classification performance obtained by the Baseline method and

Jaccard-ICV. The first two columns are the same as in Table 3.1, explained in Section 3.1. The third column shows the Gmean values obtained by the Baseline approach on the datasets containing only GO features, while the fourth column shows the results obtained by the Jaccard-ICV approach on the datasets containing both GO and PPI features. Each row represents a different organism and set of GO categories present in the dataset, similar to Table 3.1. The difference is that the fourth column (GO+PPI) shows results for datasets that also contain PPI features. The second to last row, Average Rank, shows the average rank obtained by each method over the 28 datasets. For each dataset, the best method receives the ranking value of 1; conversely, the worst method receives the ranking value of 2. So, the smaller the average rank of a method, the better its overall predictive performance. The last row, #Wins, shows the number of datasets where each method has obtained the best predictive performance. Finally, we highlight the best result obtained in each row by using boldface numbers.

We can see that our proposed Jaccard-ICV solution outperformed the Baseline approach on 20 out of the 28 datasets. Regarding the statistical significance analysis, we use Wilcoxon's Signed-Rank Test for Matched Pairs to compare the two algorithms on multiple domains [Japkowicz e Shah 2011].

Because we are only interested in a positive effect from our new approach, we use a one-tail test, in which the null hypothesis states that the predictive performance of Jaccard-ICV is either equal or lower than that of the Baseline approach. Opposing to it, the alternative hypothesis states that Jaccard-ICV has a better predictive performance than the Baseline method. Finally, we use the well-known significance level threshold of 0.05 to determine if we can reject the null hypothesis.

By performing this statistical test we obtain a p-value of 0.00357, with which we can then reject the null hypothesis and conclude that Jaccard-ICV has a better predictive performance than the Baseline method.

Dataset		GO	GO+PPI
Organism	GO types	Baseline	Jaccard-ICV
C. elegans	BP	55.76	**65.13**
	CC	58.83	**62.66**
	MF	53.01	**63.90**
	BP.CC	60.78	**66.32**
	BP.MF	57.40	**67.81**
	CC.MF	60.15	**63.50**
	BP.CC.MF	57.79	**67.30**
D. melanogaster	BP	**64.87**	59.30
	CC	69.20	**70.81**
	MF	59.80	**66.50**
	BP.CC	**65.42**	62.26
	BP.MF	**66.45**	59.95
	CC.MF	57.79	**62.33**
	BP.CC.MF	66.04	**66.91**
M. musculus	BP	66.96	**68.75**
	CC	53.45	**60.98**
	MF	62.59	**65.15**
	BP.CC	**64.07**	58.20
	BP.MF	66.72	**68.70**
	CC.MF	59.57	**64.34**
	BP.CC.MF	**66.69**	63.00
S. cerevisiae	BP	55.47	**59.78**
	CC	**57.17**	56.93
	MF	47.00	**58.43**
	BP.CC	**61.72**	60.89
	BP.MF	59.63	**61.05**
	CC.MF	51.21	**60.87**
	BP.CC.MF	**60.34**	59.82
Average Rank		1.71	**1.29**
# Wins		8	**20**

Table 3.2: Comparison of predictive performance between Baseline and Jaccard-ICV

Chapter 4

A Novel Jaccard-based Similarity Coefficient

In the previous chapter, we introduced features, based on protein-protein interactions, with uncertain binary values to our dataset. Therefore, to keep using the Jaccard distance with the NN classifier, we had to select a cut-off parameter to remove the uncertainty of the data, since the Jaccard distance only applies to binary values without uncertainty. We believe that the classification performance could be improved if the Jaccard distance could handle uncertain binary values. This belief comes from two reasons: i) we would not need to optimize a cut-off parameter, which can lead to a sub-optimal optimization and ii) the classifier would then be using a richer information, the uncertainty of the interactions, which could lead to better generalization and predictive performance.

In light of this motivation, we propose a novel similarity coefficient and its correspondent distance metric. This novel similarity coefficient is based on the Jaccard index, but it can handle both certain and uncertain binary values. Also, it is equivalent to the original Jaccard coefficient for values without uncertainty.

Next, we present the intuition behind this development and then we formally define it. After the definition, we present and analyze the experimental results, comparing to both the Baseline approach and Jaccard-ICV.

4.1 Intuition

Let us start with an example on how to calculate the Jaccard similarity coefficient for the two instances represented in Table 4.1.

Instance	GO term 1	GO term 2	GO term 3	GO term 4	GO term 5
Gene A	0	1	0	1	1
Gene B	0	0	1	1	0

Table 4.1: Two instances with binary values without uncertainty

For this task, we will be using four counters (C_{00}, C_{01}, C_{10} and C_{11}), each of them with initial value zero. By observing the first feature on both instances, we see that both Genes A and B have value 0 and, thus, we increment counter C_{00} by 1. Then, by analysing the second feature on both instances, we see that Gene A has value 1 and Gene B has value 0, so we increment counter C_{10} by 1. And so on, for each feature with value a in Gene A and value b in Gene B we increment the counter C_{ab} by 1. At the end of this process, we will have $C_{00} = 1$, $C_{01} = 1$, $C_{10} = 2$ and $C_{11} = 1$. Then, we calculate the Jaccard similarity of these two instances as in Equation 4.1.

$$\text{Jaccard similarity} = \frac{C_{11}}{C_{01} + C_{10} + C_{11}} = \frac{1}{4} = 0.25 \tag{4.1}$$

Now, consider the two instances represented in Table 4.2. Notice how the values for the Protein features are uncertain. For example, for the uncertain value (τ_{A1}; 0.3), τ_{A1} represents the true value of feature Protein 1 in Gene A. Since we are only working with binary features, we know that $\tau_{A1} \in \{0, 1\}$, but we do not know which of the two it is. Even though we lack this information, we have the value 0.3 indicating how certain we are that $\tau_{A1} = 1$. In the same way, the uncertain value (τ_{B3}; 0.5) indicates a level of confidence of 0.5 that $\tau_{B3} = 1$. It is worth noting that these values that indicate a degree of certainty are in the range $[0, 1]$.

The Jaccard similarity is not applicable to these instances because it cannot handle the uncertain binary values. For this reason, we propose a novel coefficient, based on the Jaccard index, to deal with them. Next, we examplify its use to show the intuition behind the proposed extension.

Instance	GO term 1	GO term 2	Protein 1	Protein 2	Protein 3
Gene A	0	1	(τ_{A1}; 0.3)	(τ_{A2}; 0.1)	(τ_{A3}; 0.5)
Gene B	0	0	(τ_{B1}; 0.9)	(τ_{B2}; 0.2)	(τ_{B3}; 0.5)

Table 4.2: Two instances with binary values with and without uncertainty

Again, we will be using the four counters as before (C_{00}, C_{01}, C_{10} and C_{11}), each of them with initial value zero. Starting with the first feature, we see that Gene A has value 0 and Gene B has value 0, so we increment C_{00} by 1. Analogously, for the second feature we increment C_{10} by 1, just like we did when calculating the traditional Jaccard similarity. Now, when we reach the third feature, Protein 1, we would like to increment the counter $C_{\tau_{A1}\tau_{B1}}$ but we are not certain about the values of τ_{A1} and τ_{B1}.

We know that $(\tau_{A1}, \tau_{B1}) \in \{(0,0),(0,1),(1,0),(1,1)\}$, but how likely is it that (τ_{A1}, τ_{B1}) is any of these values? Under an evidential interpretation of probability, we can use the certainty values associated with each τ to calculate the probability of (τ_{A1}, τ_{B1}) being any of its four possible values.

First, since the certainty value 0.3 in (τ_{A1}; 0.3) measures how certain we are that $\tau_{A1} = 1$, we can extrapolate that $P(\tau_{A1} = 1) = 0.3$ and $P(\tau_{A1} = 0) = (1 - 0.3) = 0.7$. Then, by assuming indepedence between the values of a feature in different instances, we can calculate the joint probability $P(\tau_{A1} = a \wedge \tau_{B1} = b)$ as $P(\tau_{A1} = a) \times P(\tau_{B1} = b)$. Thus:

- $P(\tau_{A1} = 1 \wedge \tau_{B1} = 1) = P(\tau_{A1} = 1) \times P(\tau_{B1} = 1) = 0.3 \times 0.9$
- $P(\tau_{A1} = 1 \wedge \tau_{B1} = 0) = P(\tau_{A1} = 1) \times P(\tau_{B1} = 0) = 0.3 \times (1 - 0.9)$
- $P(\tau_{A1} = 0 \wedge \tau_{B1} = 1) = P(\tau_{A1} = 0) \times P(\tau_{B1} = 1) = (1 - 0.3) \times 0.9$
- $P(\tau_{A1} = 0 \wedge \tau_{B1} = 0) = P(\tau_{A1} = 0) \times P(\tau_{B1} = 0) = (1 - 0.3) \times (1 - 0.9)$

Finally, we propose that, for each feature, instead of incrementing a single counter by 1, we increment each counter C_{ab} by $P(\tau_{A1} = a \wedge \tau_{B1} = b)$, i.e., the probability of $(\tau_{A1}, \tau_{B1}) = (a, b)$, which can also be interpreted as the probability that counter C_{ab} would be incremented by 1 if we knew the values of τ_{A1} and τ_{B1} all along.

Continuing with the calculation of the extended Jaccard similarity, by analysing the level of uncertainty of Genes A and B for their third feature, Protein 1, we would:

- increment C_{11} by $P(\tau_{A1} = 1 \wedge \tau_{B1} = 1) = 0.3 \times 0.9 = 0.27$;
- increment C_{10} by $P(\tau_{A1} = 1 \wedge \tau_{B1} = 0) = 0.3 \times (1 - 0.9) = 0.03$;
- increment C_{01} by $P(\tau_{A1} = 0 \wedge \tau_{B1} = 1) = (1 - 0.3) \times 0.9 = 0.63$;
- increment C_{00} by $P(\tau_{A1} = 0 \wedge \tau_{B1} = 0) = (1 - 0.3) \times (1 - 0.9) = 0.07$.

Notice how the sum of all these increments $(0.27 + 0.03 + 0.63 + 0.07)$ is 1. This is because we considered *all* of the possible (τ_{A1}, τ_{B1}) values.

By analysing the fourth feature, Protein 2, we increment C_{11} by 0.02, C_{10} by 0.08, C_{01} by 0.18 and C_{00} by 0.72. Notice how intuitive these results are: since we have a low belief that any of the two values is 1, C_{00} gets incremented the most, which corresponds to our belief that the most likely values for τ_{A1} and τ_{B1} are both 0. Finally, by analysing the last feature, Protein 3, we increment all the four counters by 0.25.

Now we can finally calculate the extended Jaccard similarity of these two instances as in Equation 4.2.

$$\text{Extended Jaccard similarity} = \frac{C_{11}}{C_{01} + C_{10} + C_{11}} = \frac{0.54}{2.96} = 0.18243243... \tag{4.2}$$

Notice that Equation 4.2 is the same as Equation 4.1. The only thing that changed was how to increment the counters.

4.2 Definition

For the definition that follows, we assume sparse datasets with binary features that can have either certain or uncertain values. We call *positive value* of a feature its least frequent value, i.e., its sparse value. For the case under study in this work, the *positive value* of the GO features is 1, which indicates the presence of the GO term, and the *positive value* of the protein interaction features is also 1, which indicates the existence of the correspondent interaction.

In the same way as we did in the last section, we start by presenting the definition of the Jaccard similarity in order to facilitate its comparison with the proposed extension.

Let s_j and $s_{j'}$ be the sets of binary features with *positive value* in instances j and j' respectively. The Jaccard index is defined as in Equation 4.3. In the special case when both s_j and $s_{j'}$ are empty, the Jaccard index is defined to be equal to 1.

$$\text{Jaccard}(s_j, s_{j'}) = \frac{|s_j \cap s_{j'}|}{|s_j \cup s_{j'}|} \tag{4.3}$$

And the Jaccard distance between j and j' is simply defined as:

$$\delta_{\text{Jaccard}}(j, j') = 1 - \text{Jaccard}(s_j, s_{j'}). \tag{4.4}$$

Note that Equation 4.3, and consequently Equation 4.4, are limited to scenarios with binary feature values without uncertainty. We then propose an extension of the Jaccard index to take into account the probability $p_i(s_j)$ of a binary feature i (of a total of n features in the dataset) belonging to s_j, i.e., having *positive value* in instance j. Equation 4.5 defines this new similarity coefficient, here called ProbJaccard (Probabilistic Jaccard measure). Again, we define ProbJaccard(s_j,$s_{j'}$) = 1 when the denominator evaluates to zero, which happens when both sets are certainly empty.

$$\text{ProbJaccard}(s_j, s_{j'}) = \frac{\sum_{i=1}^{n}[p_i(s_j) \times p_i(s_{j'})]}{\sum_{i=1}^{n}[p_i(s_j) + p_i(s_{j'}) - p_i(s_j) \times p_i(s_{j'})]} \tag{4.5}$$

Like Equation 4.3, the numerator of Equation 4.5 measures the degree of *intersection* between the two instances, while the denominator measures the degree of *union* between the two instances. Note however, that these degrees of intersection and union are *probabilistic* in Equation 4.5.

Analogously, we define the Probabilistic Jaccard distance between j and j' as:

$$\delta_{\text{ProbJaccard}}(j, j') = 1 - \text{ProbJaccard}(s_j, s_{j'}). \tag{4.6}$$

Note that all these indexes and distances take values in the interval [0,1]. Also note that, when working with certain data, Equations 4.5 and 4.6 become equivalent to Equations 4.3 and 4.4, and, thus, they can be used in datasets with both certain and uncertain binary values.

4.3 Results

We obtain the results for the ProbJaccard approach by the same method we used in Chapter 3 for the Baseline approach: a stratified 10-fold cross-validation repeated 30 times. The classifier we use is still the Nearest-Neighbor and the dataset is the same as the one proposed in Chapter 3, which was used with the Jaccard-ICV method.

We first compare the Baseline results with the ones obtained by our ProbJaccard method in order to evaluate if we obtained a significant improvement of predictive performance, in the same way we did with Jaccard-ICV. Then we compare ProbJaccard with Jaccard-ICV to see how much better our proposed method is in terms of predictive performance.

4.3.1 Baseline x ProbJaccard

In Table 4.3, we show the comparison of the predictive performances obtained by our ProbJaccard approach and the baseline method (which uses the Jaccard distance in a dataset with gene ontology features only).

The organization of this table is identical to Table 3.2, except that the fourth column now shows the Gmean results obtained by the ProbJaccard approach. Again, the boldface numbers define the best result in each row.

The results show that ProbJaccard achieved a better average ranking and number of wins (21 x 7) when compared to the Baseline method. This is similiar to the result obtained from comparing Jaccard-ICV with the Baseline (20 x 8). In the same way as we did in Section 3.4, we apply the Wilcoxon Signed-Rank Test for Matched Pairs to evaluate the statistical significance of these results, yielding a p-value of 0.0008, with which we can conclude that ProbJaccard has a better predictive performance than the Baseline method.

4.3.2 Jaccard-ICV x ProbJaccard

Now we compare our two proposed approaches to try to determine if using ProbJaccard either leads us to better results or if the method fails in obtaining a good classification performance when compared to the traditional internal cross-validation approach. The Gmean results obtained by both approaches in each dataset are shown in Table 4.4. With the same overall organization as the previous table, the difference this time is that we

Dataset		GO	GO+PPI
Group	GO types	Baseline	ProbJaccard
C. elegans	BP	55.76	**64.94**
	CC	58.83	**63.58**
	MF	53.01	**65.00**
	BP.CC	60.78	**66.26**
	BP.MF	57.40	**65.35**
	CC.MF	60.15	**63.61**
	BP.CC.MF	57.79	**65.80**
D. melanogaster	BP	**64.87**	63.63
	CC	**69.20**	64.89
	MF	59.80	**64.68**
	BP.CC	65.42	**65.60**
	BP.MF	**66.45**	65.89
	CC.MF	57.79	**64.45**
	BP.CC.MF	**66.04**	65.52
M. musculus	BP	**66.96**	62.39
	CC	53.45	**63.16**
	MF	62.59	**70.25**
	BP.CC	**64.07**	62.00
	BP.MF	66.72	**66.75**
	CC.MF	59.57	**68.65**
	BP.CC.MF	**66.69**	63.10
S. cerevisiae	BP	55.47	**63.78**
	CC	57.17	**60.87**
	MF	47.00	**58.65**
	BP.CC	61.72	**65.25**
	BP.MF	59.63	**63.19**
	CC.MF	51.21	**59.52**
	BP.CC.MF	60.34	**63.69**
Average Rank		1.75	**1.25**
# Wins		7	**21**

Table 4.3: Comparison of predictive performance between Baseline and ProbJaccard

replace the Baseline results in the third column by the Jaccard-ICV results that were reported in Table 3.2.

The results show that ProbJaccard achieved a better average ranking and number of wins (18 x 10) when compared to the Jaccard-ICV method. Again, in the same way as we did in Section 3.4, we apply the Wilcoxon Signed-Rank Test for Matched Pairs to evaluate the statistical significance of these results, yielding a p-value of 0.03754, with which we can conclude that ProbJaccard has a better predictive performance than the Jaccard-ICV method.

Dataset		GO+PPI	GO+PPI
Group	GO types	Jaccard-ICV	ProbJaccard
C. elegans	BP	**65.13**	64.94
	CC	62.66	**63.58**
	MF	63.90	**65.00**
	BP.CC	**66.32**	66.26
	BP.MF	**67.81**	65.35
	CC.MF	63.50	**63.61**
	BP.CC.MF	**67.30**	65.80
D. melanogaster	BP	59.30	**63.63**
	CC	**70.81**	64.89
	MF	**66.50**	64.68
	BP.CC	62.26	**65.60**
	BP.MF	59.95	**65.89**
	CC.MF	62.33	**64.45**
	BP.CC.MF	**66.91**	65.52
M. musculus	BP	**68.75**	62.39
	CC	60.98	**63.16**
	MF	65.15	**70.25**
	BP.CC	58.20	**62.00**
	BP.MF	**68.70**	66.75
	CC.MF	64.34	**68.65**
	BP.CC.MF	63.00	**63.10**
S. cerevisiae	BP	59.78	**63.78**
	CC	56.93	**60.87**
	MF	58.43	**58.65**
	BP.CC	60.89	**65.25**
	BP.MF	61.05	**63.19**
	CC.MF	**60.87**	59.52
	BP.CC.MF	59.82	**63.69**
Average Rank		1.64	**1.36**
# Wins		10	**18**

Table 4.4: Comparison of predictive performance between Jaccard-ICV and ProbJaccard

Chapter 5

Feature Selection with F-statistic

To improve the predictive performance, we introduced a large number (in between 5000 and 12000 depending on the dataset) of features based on protein-protein interactions. Especially after introducing that many features in our datasets with only hundreds of instances, a question may emerge about their relevance. Are all of these features assisting on the classification task or are some of them redundant or irrelevant? In light of this question, we propose a feature selection procedure in an attempt to filter out the irrelevant features and, ultimately, improving the classification performance.

5.1 Introduction

Given a dataset with W features, there are $2^W - 1$ different subsets of features that could be used for classification. One could evaluate all of them in order to choose the best, but such exhaustive approach is clearly unfeasible even for a small number of features. For this reason, search heuristics are used in an attempt to find good subsets without having to explore all the possibilities [Han, Pei e Kamber 2011]. As for the evaluation of those subsets, there are two main approaches: wrapper and filter. The wrapper method uses the same model for classification to evaluate the subsets, while the filter method evaluates them with measures that are independent of the model being used for classification. For this reason, the wrapper method usually achieves better classification performance, while the filter technique is faster. This difference is aggravated with the increase in the number of features. To balance the pros and cons of both techniques, hybrid approaches have also been proposed [Hsu, Hsieh e Lu 2011]. However, since we are dealing with thousands of features and we want to consider large subsets, we choose to use a filter approach in this work.

We start by describing our proposed feature selection procedure and then we compare the results obtained before and after its application for reducing the dimensionality of the datasets.

5.2 Proposal

We propose an application of the filter technique using the F-statistic as the feature evaluation measure. The F-statistic (also known as F-score) evaluates how well a feature discriminates samples from different classes [Zhao et al. 2010] by considering the ratio:

$$\text{F-statistic} = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{\sum_{i=1}^{c} \frac{n_i}{c-1}(\mu_i - \mu)^2}{\frac{1}{n-c}\sum_{i=1}^{c}(n_i - 1)\sigma_i^2} \tag{5.1}$$

where c is the number of classes, n is the total number of instances, n_i is the number of instances with class i, μ is the mean of the values from all instances, μ_i is the mean of the values from instances with class i and σ_i^2 is the variance of the values from instances with class i.

The higher the F-statistic, the better the feature discriminates the two classes. Thus, once we have calculated this measure for all features, we can select the ones that obtained the highest values. However, how many features should we select? A traditional method in the literature is to select the top k features. Instead of arbitrarily choosing what k value to use for all datasets, we decide to use the same approach we used to select a threshold in Chapter 3 for the Jaccard-ICV method. This way, for each fold of the external cross-validation we will perform an internal cross-validation to choose the k that will be used to select features on that external fold. The set of possibles values for k that we consider in the internal cross-validation procedure is $\{1\%, 5\%, 10\%, 25\%, 50\%, 100\%\}$. For example, if the internal cross-validation chooses $k = 5\%$, then only the top 5% features (in terms of F-statistic, calculated in the current external training fold) will be used for the classification of the current external validation fold. If $k = 100\%$, then no selection is performed and all features are passed to the classifier.

5.3 Results

The classifier we will be using to compare the results obtained before and after feature selection is the NN with the ProbJaccard distance. We choose to use the ProbJaccard approach instead of the Jaccard-ICV because it was the one that presented the best overall results for our datasets. Also, if we choose to use the Jaccard-ICV, we will be increasing the complexity and runtime of the classification process even more, since there would be *two* parameters that need to be chosen internally for each external fold. This is a good practical example of one of the advantages of using ProbJaccard instead of Jaccard-ICV.

To evaluate the predictive performance of the described approach with feature selection (which we will be calling ProbJaccard-FS), we will use the same process used to evaluate the Jaccard-ICV in Chapter 3: 30 repetitions of a stratified 10-fold cross-validation externally and 10 repetitions of a stratified 5-fold cross-validation internally.

First, we compare the results with the Baseline approach to evaluate the improvement we have had so far. Then we compare with ProbJaccard to evaluate the improvement related to the implementation of the described feature selection technique.

5.3.1 Baseline x ProbJaccard-FS

In Table 5.1, we show the comparison of the predictive performances obtained by our ProbJaccard-FS method and the Baseline approach described in Chapter 3.

The organization of this table is identical to Table 3.2, except that the fourth column now shows the ProbJaccard-FS results. As before, the boldface numbers define the best result in each row.

The results show that ProbJaccard-FS achieved a better average ranking and number of wins (25 x 3) than the Baseline method. In the same way as we did in Section 3.4, we apply the Wilcoxon Signed-Rank Test for Matched Pairs to evaluate the statistical significance of the results. This time the test yields a p-value of 0.00000, with which we can conclude that, by performing feature selection with the described approach, we can obtain, with statistical significance, much better results than the baseline.

Dataset		GO	GO+PPI
Group	GO types	Baseline	ProbJaccard-FS
C. elegans	BP	55.76	**66.61**
	CC	58.83	**65.13**
	MF	53.01	**66.87**
	BP.CC	60.78	**66.37**
	BP.MF	57.40	**66.26**
	CC.MF	60.15	**62.99**
	BP.CC.MF	57.79	**67.18**
D. melanogaster	BP	**64.87**	63.54
	CC	**69.20**	65.56
	MF	59.80	**67.69**
	BP.CC	65.42	**65.50**
	BP.MF	**66.45**	63.28
	CC.MF	57.79	**65.62**
	BP.CC.MF	66.04	**66.74**
M. musculus	BP	66.96	**68.25**
	CC	53.45	**64.43**
	MF	62.59	**67.28**
	BP.CC	64.07	**69.55**
	BP.MF	66.72	**70.95**
	CC.MF	59.57	**67.45**
	BP.CC.MF	66.69	**69.42**
S. cerevisiae	BP	55.47	**69.63**
	CC	57.17	**59.58**
	MF	47.00	**63.86**
	BP.CC	61.72	**70.62**
	BP.MF	59.63	**70.22**
	CC.MF	51.21	**59.64**
	BP.CC.MF	60.34	**71.47**
Average Rank		1.89	**1.11**
# Wins		3	**25**

Table 5.1: Comparison of predictive performance between Baseline and ProbJaccard-FS

5.3.2 ProbJaccard x ProbJaccard-FS

Now we compare the ProbJaccard method both with and without the feature selection procedured described in this chapter to analyze how much this feature selection technique assisted in the classification process. The Gmean results obtained by both approaches in each of the experimental datasets are shown in Table 5.2. The organization is the same as the previous table with the difference that, this time, we replace the Baseline results in the third column by the ProbJaccard results previously reported in Table 4.3.

These results show that ProbJaccard-FS achieved a better average ranking and number of wins (21 x 7) than the ProbJaccard method. Once again, in the same way as

Dataset		GO+PPI	GO+PPI
Group	GO types	ProbJaccard	ProbJaccard-FS
C. elegans	BP	64.94	**66.61**
	CC	63.58	**65.13**
	MF	65.00	**66.87**
	BP.CC	66.26	**66.37**
	BP.MF	65.35	**66.26**
	CC.MF	**63.61**	62.99
	BP.CC.MF	65.80	**67.18**
D. melanogaster	BP	**63.63**	63.54
	CC	64.89	**65.56**
	MF	64.68	**67.69**
	BP.CC	**65.60**	65.50
	BP.MF	**65.89**	63.28
	CC.MF	64.45	**65.62**
	BP.CC.MF	65.52	**66.74**
M. musculus	BP	62.39	**68.25**
	CC	63.16	**64.43**
	MF	**70.25**	67.28
	BP.CC	62.00	**69.55**
	BP.MF	66.75	**70.95**
	CC.MF	**68.65**	67.45
	BP.CC.MF	63.10	**69.42**
S. cerevisiae	BP	63.78	**69.63**
	CC	**60.87**	59.58
	MF	58.65	**63.86**
	BP.CC	65.25	**70.62**
	BP.MF	63.19	**70.22**
	CC.MF	59.52	**59.64**
	BP.CC.MF	63.69	**71.47**
Average Rank		1.75	**1.25**
# Wins		7	**21**

Table 5.2: Comparison of predictive performances obtained by ProbJaccard with and without feature selection

we did in the previous analysis, we apply the Wilcoxon Signed-Rank Test for Matched Pairs to evaluate the statistical significance of the results. This time the test yields a p-value of 0.00152, with which we can conclude that, by performing feature selection with the described approach, we can obtain, with statistical significance, even better results than before.

Chapter 6

Conclusions

This work proposed three ways to improve the predictive performance of the NN classifier in the task of classifying aging-related genes into either pro- or anti-longevity.

First, by introducing new features based on protein-protein interactions, we were able to improve the classification performance in a statistically significant way with a confidence interval of 95%.

Then, we proposed a novel Jaccard-based similarity coefficient (ProbJaccard) and its correspondent distance function to handle uncertain values in sparse datasets. By using this novel metric with the NN classifier, we were able to improve the predictive performance in a statistically significant way, with a 95% confidence interval, when compared to both the Baseline and the Jaccard-ICV approaches.

Also, it is worth noting that ProbJaccard presents a simpler and much faster way to handle the uncertain values, since there is no internal cross-validation step. This is an even bigger advantage when there is the need to perform the optimization of other parameters, which can happen, for example, when introducing feature selection procedures. Finally, this novel similarity coefficient can also be useful in tasks other than classification, such as when a researcher wants to calculate the similarity between two genes for other purposes. Therefore, it stands as one of the major contributions of this work.

Finally, we introduced a new methodology to perform feature selection by using the F-statistic to evaluate features individually. The results obtained by the NN classifier (using ProbJaccard) after performing this proposed feature selection procedure represented a statistifically significant improvement, with a 95% confidence interval, in comparison to the results obtained by the same classification approach but without feature selection.

Thus, this work successfully achieved its goal of improving predictions of the effects of aging-related genes on the longevity of model organisms, achieving the best results with the ProbJaccard-FS approach, as seen in Table 5.1. Also, it introduced a novel similarity coefficient that can be used as an extension of the Jaccard index to handle uncertain values. Finally, it is worth noting that part of the results presented in this work was published in [Martire et al. 2017].

We leave as future work the task of analysing how beneficial ProbJaccard can be in other domains. Also, in regards to the central task in this work, we intend to explore the Wrapper method for feature selection. It will require pre-filtering of features as well as optimizations in the implementation to make it feasible for our scenario. We also leave for future work the task of researching and developing a feature evaluation measure more suitable for the features in this task. Another interesting possibility is to explore techniques of dealing with class imbalance. Finally, another research direction consists of exploring other relevant features that can be added to the datasets in order to improve, even more, the predictive performance on this task.

Bibliography

[Aggarwal 2014]AGGARWAL, C. C. *Data classification: algorithms and applications.* [S.l.]: CRC Press, 2014. ISBN 978-1-46-658674-1.

[Aha e Kibler 1991]AHA, D.; KIBLER, D. Instance-based learning algorithms. *Machine Learning*, v. 6, p. 37–66, 1991.

[Altman e Bland 1994]ALTMAN, D. G.; BLAND, J. M. Diagnostic tests 1: sensitivity and specificity. *British Medical Journal*, v. 308, n. 6943, p. 1552, 1994.

[Ashburner et al. 2000]ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, v. 25, n. 1, p. 25–29, 2000.

[Ayyadevara et al. 2008]AYYADEVARA, S. et al. Remarkable longevity and stress resistance of nematode pi3k-null mutants. *Aging cell*, Wiley Online Library, v. 7, n. 1, p. 13–22, 2008.

[Bhaskar, Hoyle e Singh 2006]BHASKAR, H.; HOYLE, D. C.; SINGH, S. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in biology and medicine*, Elsevier, v. 36, n. 10, p. 1104–1125, 2006.

[Breiman e Spector 1992]BREIMAN, L.; SPECTOR, P. Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, JSTOR, p. 291–319, 1992.

[Dornelas et al. 2014]DORNELAS, M. et al. Assemblage time series reveal biodiversity change but not systematic loss. *Science*, American Association for the Advancement of Science, v. 344, n. 6181, p. 296–299, 2014.

[Fabris e Freitas 2016]FABRIS, F.; FREITAS, A. A. New kegg pathway-based interpretable features for classifying ageing-related mouse proteins. *Bioinformatics*, Oxford University Press, v. 32, n. 19, p. 2988–2995, 2016.

[Fabris, Magalhães e Freitas 2017]FABRIS, F.; MAGALHÃES, J. P. D.; FREITAS, A. A. A review of supervised machine learning applied to ageing research. *Biogerontology*, Springer, p. 1–18, 2017.

[Fang et al. 2013]FANG, Y. et al. Classifying aging genes into DNA repair or non-DNA repair-related categories. In: *International Conference on Intelligent Computing*. Nanning, China: [s.n.], 2013. p. 20–29.

[Feng et al. 2012]FENG, K. et al. Topological anaylysis and prediction of aging genes in mus musculus. In: IEEE. *Systems and Informatics (ICSAI), 2012 International Conference on*. [S.l.], 2012. p. 2268–2271.

[Forman e Scholz 2010]FORMAN, G.; SCHOLZ, M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, ACM, v. 12, n. 1, p. 49–57, 2010.

[Freitas, Vasieva e Magalhães 2011]FREITAS, A. A.; VASIEVA, O.; MAGALHÃES, J. P. de. A data mining approach for classifying dna repair genes into ageing-related or non-ageing-related. *BMC genomics*, BioMed Central, v. 12, n. 1, p. 27, 2011.

[Friedman e Johnson 1988]FRIEDMAN, D. B.; JOHNSON, T. E. A mutation in the age-1 gene in caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics*, Genetics Soc America, v. 118, n. 1, p. 75–86, 1988.

[Gao et al. 2017]GAO, Y. et al. The DNA damage response of C. elegans affected by gravity sensing and radiosensitivity during the Shenzhou-8 spaceflight. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, v. 795, n. 1, p. 15–26, 2017.

[Ge, Xia e Nadungodage 2010]GE, J.; XIA, Y.; NADUNGODAGE, C. UNN: a neural network for uncertain data classification. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Hyderabad, India: [s.n.], 2010. p. 449–460.

[Goldman et al. 2013]GOLDMAN, D. P. et al. Substantial health and economic returns from delayed aging may warrant a new focus for medical research. *Health affairs*, Health Affairs, v. 32, n. 10, p. 1698–1705, 2013.

[Guarente e Kenyon 2000]GUARENTE, L.; KENYON, C. Genetic pathways that regulate ageing in model organisms. *Nature*, Nature Publishing Group, v. 408, n. 6809, p. 255–262, 2000.

[Han, Pei e Kamber 2011]HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Morgan Kaufmann, 2011. ISBN 978-0-12-381479-1.

[Hsu, Hsieh e Lu 2011]HSU, H.-H.; HSIEH, C.-W.; LU, M.-D. Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, Elsevier, v. 38, n. 7, p. 8144–8150, 2011.

[Huang 2008]HUANG, A. Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. [S.l.: s.n.], 2008. p. 49–56.

[Huang et al. 2012]HUANG, T. et al. Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. *Biochimie*, Elsevier, v. 94, n. 4, p. 1017–1025, 2012.

[Iachine et al. 2006]IACHINE, I. et al. Genetic influence on human lifespan and longevity. *Human genetics*, Springer, v. 119, n. 3, p. 312, 2006.

[Jaccard 1908]JACCARD, P. Nouvelles researches sur la distribution florale. *Bull Soc Vaud Sci Nat*, v. 44, p. 223–270, 1908.

[Jaccard 1912]JACCARD, P. The distribution of the flora in the alpine zone. *New phytologist*, Wiley Online Library, v. 11, n. 2, p. 37–50, 1912.

[Japkowicz e Shah 2011]JAPKOWICZ, N.; SHAH, M. *Evaluating learning algorithms: a classification perspective*. [S.l.]: Cambridge University Press, 2011. ISBN 978-0-52-119600-0.

[Jiang e Ching 2011]JIANG, H.; CHING, W.-K. Classifying dna repair genes by kernel-based support vector machines. *Bioinformation*, Biomedical Informatics Publishing Group, v. 7, n. 5, p. 257, 2011.

[Kirkwood 2005]KIRKWOOD, T. B. Understanding the odd science of aging. *Cell*, Elsevier, v. 120, n. 4, p. 437–447, 2005.

[Klass 1983]KLASS, M. R. A method for the isolation of longevity mutants in the nematode caenorhabditis elegans and initial results. *Mechanisms of ageing and development*, Elsevier, v. 22, n. 3, p. 279–286, 1983.

[Kohavi et al. 1995]KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

[Kulmanov, Khan e Hoehndorf 2017]KULMANOV, M.; KHAN, M. A.; HOEHNDORF, R. Deepgo: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *arXiv preprint arXiv:1705.05919*, 2017.

[Leydesdorff 2008]LEYDESDORFF, L. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 59, n. 1, p. 77–85, 2008.

[Lin et al. 2016]LIN, D. et al. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, v. 17, n. 1, p. 247, 2016.

[López-Otín et al. 2013]LÓPEZ-OTÍN, C. et al. The hallmarks of aging. *Cell*, Elsevier, v. 153, n. 6, p. 1194–1217, 2013.

[Magalhães et al. 2009]MAGALHÃES, J. P. de et al. The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging cell*, v. 8, n. 1, p. 65–72, 2009.

[Martire et al. 2017]MARTIRE, I. et al. A novel probabilistic jaccard distance measure for classification of sparse and uncertain data. In: *Proceedings of the 5th Symposium on Knowledge Discovery, Mining and Learning*. [S.l.: s.n.], 2017. v. 5, p. 81–88.

[Medvedev 1990]MEDVEDEV, Z. A. An attempt at a rational classification of theories of ageing. *Biological Reviews*, Wiley Online Library, v. 65, n. 3, p. 375–398, 1990.

[Niwattanakul et al. 2013]NIWATTANAKUL, S. et al. Using of jaccard coefficient for keywords similarity. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. [S.l.: s.n.], 2013. v. 1, n. 6.

[Prasad et al. 2009]PRASAD, T. S. K. et al. Human Protein Reference Database - 2009 update. *Nucleic Acids Research*, v. 37, n. 1, p. D767–D772, 2009.

[Prokopenko et al. 2015]PROKOPENKO, D. et al. Utilizing the jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. *Bioinformatics*, Oxford University Press, v. 32, n. 9, p. 1366–1372, 2015.

[Ren et al. 2009]REN, J. et al. Naive Bayes Classification of Uncertain Data. In: *IEEE International Conference on Data Mining*. Miami, United States of America: [s.n.], 2009. p. 944–949.

[Sato et al. 2005]SATO, T. et al. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, Oxford University Press, v. 21, n. 17, p. 3482–3489, 2005.

[Schloss et al. 2009]SCHLOSS, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, Am Soc Microbiol, v. 75, n. 23, p. 7537–7541, 2009.

[Shi et al. 2017]SHI, J. et al. Identification of potential crucial gene network related to seasonal allergic rhinitis using microarray data. *European Archives of Oto-Rhino-Laryngology*, v. 274, n. 1, p. 231–237, 2017.

[Stojanova et al. 2013]STOJANOVA, D. et al. Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC Bioinformatics*, v. 14, n. 1, p. 285, 2013.

[Szklarczyk et al. 2014]SZKLARCZYK, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, v. 43, n. D1, p. D447–D452, 2014.

[Tipping 2001]TIPPING, M. E. Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research*, v. 1, n. Jun, p. 211–244, 2001.

[Tsang et al. 2011]TSANG, S. et al. Decision trees for uncertain data. *IEEE transactions on knowledge and data engineering*, v. 23, n. 1, p. 64–78, 2011.

[Wan e Freitas 2013]WAN, C.; FREITAS, A. Prediction of the pro-longevity or anti-longevity effect of caenorhabditis elegans genes based on bayesian classification methods. In: IEEE. *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*. [S.l.], 2013. p. 373–380.

[Wan e Freitas 2017]WAN, C.; FREITAS, A. A. An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features. *Artificial Intelligence Review*, p. 1–40, 2017.

[Wan, Freitas e Magalhães 2015]WAN, C.; FREITAS, A. A.; MAGALHÃES, J. P. D. Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 12, n. 2, p. 262–275, 2015.

[Witten et al. 2016]WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. ISBN 978-0-12-374856-0.

[Xu et al. 2014]XU, J. et al. Efficient probabilistic frequent itemset mining in big sparse uncertain data. In: *Pacific Rim International Conference on Artificial Intelligence*. Gold Coast, Australia: [s.n.], 2014. p. 235–247.

[Yang e Li 2009]YANG, J.-L.; LI, H.-X. A probabilistic support vector machine for uncertain data. In: *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. Hong Kong, China: [s.n.], 2009. p. 163–168.

[Yang et al. 2015]YANG, L. et al. Probabilistic-KNN: A novel algorithm for passive indoor-localization scenario. In: *IEEE Vehicular Technology Conference*. Glasgow, United Kingdom: [s.n.], 2015. p. 1–5.

[Zhang e Yang 2015]ZHANG, Y.; YANG, Y. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, Elsevier, v. 187, n. 1, p. 95–112, 2015.

[Zhao et al. 2010]ZHAO, Z. et al. Advancing feature selection research. *ASU feature selection repository*, p. 1–28, 2010.