A novel probabilistic Jaccard distance measure for classification of sparse and uncertain data

I. Martire[1], P. N. da Silva[1], A. Plastino[1], F. Fabris[2], A. A. Freitas[2]

[1] Universidade Federal Fluminense, Brazil
igormartire@id.uff.br, {psilva, plastino}@ic.uff.br
[2] School of Computing, University of Kent, Canterbury, Kent, CT2 7NF, UK
{F.Fabris, A.A.Freitas}@kent.ac.uk

Abstract. Classification is one of the most important tasks in the data mining field, allowing patterns to be leveraged from data in order to try to properly classify unseen instances. Also, more and more often, the classification task has to be performed on datasets containing uncertain data. Although an increasing number of studies have been developed to handle uncertainty in classification in the last decade, there are still many underexplored scenarios — such as sparse data, usual in the bioinformatics field. Thus, in this work, we propose a novel distance measure for sparse and uncertain binary data based on the widely used Jaccard distance, testing its performance using the 1NN classifier. We evaluate the classification performance of our proposed method on 28 biological aging-related datasets with sparse and probabilistic binary features and compare it with a common technique to handle uncertainty by employing data transformation and traditional classification. The experimental results show that our proposed distance measure has both a smaller runtime and a better predictive performance than the traditional transformation approach.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.5.4 [**Pattern Recognition**]: Applications; J.3 [**Computer Applications**]: Life and Medical Sciences

Keywords: aging-related genes, classification, data mining, Jaccard distance, uncertain data

1. INTRODUCTION

The classification task is one of the most relevant tasks in the data mining field [Han et al. 2011]. Given a dataset of pre-labeled instances, the classification task comprises the induction of a classification model that is capable of predicting the class of an unseen instance based solely on its features. These features can have a numerical or categorical domain, with certain or uncertain values. Naturally, because of their higher prevalence, the majority of the techniques that have been developed so far focus on the handling of certain data [Aggarwal 2014].

In this work, we focus on the classification of uncertain data, specifically in sparse datasets. This kind of data can originate from many sources due to various factors, such as measurement precision limits, measurement errors, approximations or even lack of information. Even though the number of studies on classifying uncertain data has significantly increased in the last decade [Aggarwal 2014], there are still many underexplored areas, as is the case of sparse datasets.

We are particularly interested in the study of aging-related genes (represented as instances in our datasets) in order to identify the effect of genes on the longevity of an organism. These datasets commonly use binary features extracted from the Gene Ontology (GO) database [Ashburner et al. 2000], but another important type of feature are protein-protein interactions (PPIs) [Stojanova et al. 2013]. PPI features indicate whether or not an aging-related protein interacts with each of a set of

other proteins (which may or may not be aging-related proteins). For that purpose, we can use the STRING database [Szklarczyk et al. 2014], a popular source of PPI datasets in the bioinformatics literature. Note, however, that instead of providing binary values for the PPIs, the STRING database provides confidence scores for each interaction. This allows the dataset to present more PPI data, but adds uncertainty to it.

When not working with uncertain data, an approach to classify genes described by binary features in the aging literature is to use the k-Nearest Neighbors classifier with the Jaccard distance [Wan et al. 2015] [Wan and Freitas 2017]. Since the Jaccard distance is not able to directly handle uncertain binary values, a data transformation procedure would be required to "remove" the uncertainty, which, of course, could cause loss of valuable data. A simple and common transformation is applying a cut-off on the PPI values, so that when the confidence score is over (below) a certain value it is converted to 1 (0). The problem would then lie on how to choose an appropriate cut-off value. However, in the bioinformatics literature there is usually no concern on optimizing this value and not even an explanation about the reasons behind its choice.

Thus, the main contribution of this work is to provide an intuitive, fast and accurate method to handle uncertain PPI data in distance-based classification. For that purpose, we propose a novel Jaccard distance measure able to handle uncertain binary features, without requiring any data transformation procedure or parameter optimization. Also, it allows the algorithm to benefit from the uncertain information available, removing the need to rely on arbitrary cut-off values or to spend much time on optimizing its value.

The remainder of this article is organized as follows. Section 2 describes the related work. In Section 3, we introduce the novel distance measure for classification in sparse datasets with probabilistic binary features. Section 4 presents the datasets used in this work. Computational results are presented in Section 5. Lastly, in Section 6, we present the conclusions and future research directions.

2. RELATED WORK

The classification of uncertain data has been extensively studied 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 et al. 2010], Decision Trees [Tsang et al. 2011], k-Nearest Neighbors [Yang et al. 2015] and Support Vector Machines [Yang and Li 2009]. Most of them focus on uncertain numerical features, not specifically on 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].

As mentioned in the previous section, a lot of the research done so far in the bioinformatics field has simply ignored the uncertain information provided by the STRING database about PPIs. This has been done by applying *ad-hoc* cut-off values such as 0.4 [Shi et al. 2017], 0.7 [Gao et al. 2017], and 0.9 [Lin et al. 2016].

3. A NOVEL PROBABILISTIC JACCARD DISTANCE MEASURE

Distance-based classifiers use the intuitive idea that instances of the same class are more similar among themselves than among instances of other classes [Han et al. 2011]. Similarity (distance) measures, like the Jaccard index (distance), are functions that calculate how similar (distant) two objects are to (from) each other, and thus are the basis of supervised distance-based classification algorithms.

Next, we present the definitions of the traditional Jaccard measure (which cannot directly handle uncertainty, since a data transformation is needed to handle it) and of our proposed distance measure (which handles uncertainty directly).

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

$$\text{Jaccard}(s_j, s_{j'}) = \frac{|s_j \cap s_{j'}|}{|s_j \cup s_{j'}|} \tag{1}$$

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

$$\delta_{\text{Jaccard}}(j, j') = 1 - \text{Jaccard}(s_j, s_{j'}). \tag{2}$$

Note that Equation (1), and consequently Equation (2), 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 (3) 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.

$$\text{ProbJaccard}(s_j, s_{j'}) = \frac{\sum_{i=1}^{n}[p_i(s_j) \times p_i(s_{j'})]}{\sum_{i=1}^{n}[p_i(s_j) + p_i(s_{j'}) - p_i(s_j) \times p_i(s_{j'})]} \tag{3}$$

Like Equation (1), the numerator of Equation (3) 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 (3).

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

$$\delta_{\text{ProbJaccard}}(j, j') = 1 - \text{ProbJaccard}(s_j, s_{j'}). \tag{4}$$

Note that all these indexes and distances take values in the interval [0,1]. Also note that, when working with certain data, Equations (3) and (4) become equivalent to Equations (1) and (2), and, thus, they can be used in datasets with both certain and uncertain binary features.

4. EXPERIMENTAL 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) [de 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 when using only GO terms [Wan et al. 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 I 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 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).

5. EXPERIMENTS

In our datasets, as shown in Table I, the distribution of instances belonging to the two classes is imbalanced. 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 and 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 and 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 and Bland 1994]. The reported Gmean value for each dataset is the average of all the 10 Gmean values generated by the well-known stratified 10-fold cross-validation procedure [Witten et al. 2016].

In this work, we use the 1-Nearest Neighbor (1NN) classifier since, in previous work, it has been shown effective for classification in aging-related datasets [Wan et al. 2015][Wan and Freitas 2017].

We start by testing the improvement in predictive performance when the PPI features are added to the original database composed of GO terms only. Since this inserted data is uncertain and the Jaccard distance does not handle uncertain values, we decided to use, as a baseline, a 5-fold Internal

Table I: Statistics for each dataset.

Organism	Dataset	# features	# (%) GO features	# PPI features	avg. % GO = 0	avg. % 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

Cross-Validation (ICV) method (accessing the training set only) to automatically choose a cut-off value to discretize the feature (feature values greater or equal than the cut-off are set to 1 and set to 0 otherwise). This ICV is performed in each iteration of the external cross-validation procedure. This baseline method is here called Jaccard-ICV. Applying this cut-off on the uncertain data allows us to convert it to certain binary values and then use it with the 1NN classifier using the traditional Jaccard distance.

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. These values have also been extensively employed in the related literature [Lin et al. 2016] [Shi et al. 2017] [Gao et al. 2017]. For these two reasons, the ICV focused on choosing the best out of these four cut-off values.

The results are shown in Table II, where the boldface numbers denote the highest Gmean value obtained for each dataset. The first two columns are the same as in Table I, explained in the previous section. The third column shows the Gmean values obtained by the 1NN classifier in datasets with GO features only and using the traditional Jaccard distance metric. The fourth and fifth columns show the values obtained with the classification on the datasets composed of both GO and PPI data. While the fourth column shows the results with the internal cross-validation approach explained above, the fifth column shows the results obtained when using 1NN with our new proposed distance measure. Each row represents a different dataset in the same way as in Table I. Table II, however, has two additional rows. 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 3. So, the smaller the average rank of a method, the better its overall predictive performance. Finally, the last row, #Wins, shows the number of datasets where each method has obtained the best predictive performance. Again, the boldface numbers denote the best result in each of these two rows.

Table II: Comparison of predictive performance using Gmean as evaluation measure.

		GO	GO + PPI	
Group	Dataset	Jaccard	Jaccard-ICV	Prob-Jaccard
C. elegans	BP	55.91	**65.30**	64.13
	CC	59.73	61.01	**63.21**
	MF	53.47	64.86	**66.41**
	BP.CC	61.14	65.25	**66.44**
	BP.MF	58.07	**67.15**	65.19
	CC.MF	60.33	**63.12**	62.30
	BP.CC.MF	58.11	**68.49**	66.25
D. melanogaster	BP	**64.17**	52.39	61.13
	CC	70.44	**72.08**	68.03
	MF	50.65	**60.52**	58.13
	BP.CC	61.87	55.05	**65.19**
	BP.MF	62.88	63.24	**63.81**
	CC.MF	58.69	64.14	**64.93**
	BP.CC.MF	62.57	**65.49**	63.30
M. musculus	BP	62.98	**68.31**	63.07
	CC	50.74	56.27	**63.95**
	MF	53.94	65.64	**69.18**
	BP.CC	**61.84**	55.56	56.81
	BP.MF	63.81	**66.29**	65.30
	CC.MF	56.61	67.23	**68.89**
	BP.CC.MF	62.27	**63.49**	58.51
S. cerevisiae	BP	53.69	57.34	**58.26**
	CC	50.61	53.56	**61.45**
	MF	40.34	58.69	**58.99**
	BP.CC	58.32	55.88	**65.39**
	BP.MF	51.03	57.83	**58.29**
	CC.MF	41.56	**63.74**	60.73
	BP.CC.MF	53.60	57.32	**62.88**
Average Rank		2.71	1.75	**1.54**
# Wins		2	11	**15**

The results in Table II show that Prob-Jaccard, which uses our proposed distance measure, achieves the best predictive performance on 15 datasets, followed by Jaccard-ICV (best results on 11 datasets) and Jaccard (2 datasets). To determine whether the differences in performance are statistically significant, we ran the non-parametric Friedman test followed by the Nemenyi test [Japkowicz and Shah 2011]. Both tests were used at the 0.05 significance level. The Friedman test indicated that there was at least one pair of classifiers with a statistical difference in the predictive performance. Hence, we employed the post-hoc Nemenyi test to discover in which pairs this difference occurs. The Nemenyi test showed that both Prob-Jaccard and Jaccard-ICV are significantly superior to Jaccard-GO, which does not include PPI features. However, even though Prob-Jaccard achieves both a better average rank and a higher number of wins than Jaccard-ICV, the difference in the performance between Prob-Jaccard and Jaccard-ICV was not statistically significant.

One could think of using the Euclidean distance with the 1NN 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 two methods explored in this article. 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.

Based on the conducted experiments, we can notice a great improvement by simply adding the PPI features and optimizing the choice of cut-off value for each fold via internal cross-validation. However, this approach is slow, which could be a big problem when working with larger datasets. We then compared the runtime performance of the Jaccard-ICV method with our proposed Prob-Jaccard method to demonstrate how much faster this proposed method can be in comparison to the internal cross-validation one, without losing in overall predictive performance (and actually improving it most of the times). These results are presented in Table III. In this table, the first two columns are exactly the same as the ones in the previous tables. The third and fourth columns show the average time in seconds that was taken to classify a fold in the 10-fold cross-validation procedure. Notice that the reported times for the Jaccard-ICV method include the time spent in the selection of the cut-off value. The last column shows the ratio of the values in the third column to the values in the fourth column, which indicates how many times faster the Prob-Jaccard method is in comparison to Jaccard-ICV. These times were measured in a computer with 1.6 GHz Intel Core i5 processor and 4 GB 1600 MHz DDR3 memory.

Table III: Comparison of average CPU time in seconds per cross-validation fold.

		GO + PPI		$\frac{\text{Jaccard-ICV}}{\text{Prob-Jaccard}}$
Group	Dataset	Jaccard-ICV	Prob-Jaccard	
C. elegans	BP	18.048	0.835	21.614
	CC	9.577	0.399	24.003
	MF	10.802	0.438	24.662
	BP.CC	20.492	0.845	24.251
	BP.MF	20.316	0.861	23.596
	CC.MF	13.394	0.594	22.549
	BP.CC.MF	21.181	0.855	24.773
D. melanogaster	BP	0.639	0.023	28.684
	CC	0.492	0.018	24.200
	MF	0.469	0.016	25.059
	BP.CC	0.705	0.020	27.500
	BP.MF	0.676	0.021	29.905
	CC.MF	0.527	0.020	25.667
	BP.CC.MF	0.718	0.022	30.700
M. musculus	BP	0.639	0.023	27.783
	CC	0.492	0.018	27.333
	MF	0.469	0.016	29.313
	BP.CC	0.705	0.020	35.250
	BP.MF	0.676	0.021	32.190
	CC.MF	0.527	0.020	26.350
	BP.CC.MF	0.718	0.022	32.636
S. cerevisiae	BP	3.796	0.112	33.893
	CC	3.321	0.097	34.237
	MF	3.274	0.105	31.181
	BP.CC	4.008	0.116	34.552
	BP.MF	3.925	0.118	33.263
	CC.MF	3.488	0.108	32.296
	BP.CC.MF	4.190	0.126	33.254
Average		5.314	0.233	28.596

The last row of Table III shows that, on average, the Jaccard-ICV approach took 5.3 seconds to classify a single fold, while Prob-Jaccard took only 0.2 seconds. The last column in that row shows that the Prob-Jaccard approach was able to classify a single fold 28.6 times faster on average.

6. CONCLUSIONS

In this work, we presented a novel Jaccard distance measure for nearest-neighbor classification in sparse datasets with probabilistic binary features. We compared both the speed and the predictive performance of the 1NN classifier using both our novel distance measure and the traditional Jaccard distance (by applying an internal cross-validation to optimize the cut-off value).

The 1NN classifier using the proposed ProbJaccard distance measure is significantly faster than the Jaccard-ICV method. This is due to the fact that ProbJaccard handles the uncertainty from the data directly, so there is no need to perform an internal cross-validation to optimize a cut-off parameter. Additionally, the proposed ProbJaccard method has shown an overall improvement in the predictive performance of the 1NN classifier across 28 aging-related datasets, with a better average rank and higher number of wins when compared with the Jaccard-ICV method and a dataset with GO terms only, as shown in Table II; even though there was no statistically significant difference between the results of ProbJaccard and Jaccard-ICV.

Finally, this new distance measure can be extended to handle categorical features with more general types of uncertain values in sparse classification datasets. We leave this research for future work.

REFERENCES

Aggarwal, C. C. *Data classification: algorithms and applications.* CRC Press, 2014.

Altman, D. G. and Bland, J. M. Diagnostic tests 1: sensitivity and specificity. *British Medical Journal* 308 (6943): 1552, 1994.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., et al. Gene Ontology: tool for the unification of biology. *Nature genetics* 25 (1): 25–29, 2000.

de Magalhães, J. P., Budovsky, A., Lehmann, G., Costa, J., Li, Y., Fraifeld, V., and Church, G. M. The Human Ageing Genomic Resources: online databases and tools for biogerontologists. *Aging cell* 8 (1): 65–72, 2009.

Gao, Y., Xu, D., Zhao, L., and Sun, Y. 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* 795 (1): 15–26, 2017.

Ge, J., Xia, Y., and Nadungodage, C. UNN: a neural network for uncertain data classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Hyderabad, India, pp. 449–460, 2010.

Han, J., Pei, J., and Kamber, M. *Data mining: concepts and techniques.* Morgan Kaufmann, 2011.

Japkowicz, N. and Shah, M. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, 2011.

Lin, D., Zhang, J., Li, J., Xu, C., Deng, H.-W., and Wang, Y.-P. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics* 17 (1): 247, 2016.

Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., and Cheung, D. Naive Bayes Classification of Uncertain Data. In *IEEE International Conference on Data Mining*. Miami, United States of America, pp. 944–949, 2009.

Shi, J., Zhang, Y., Qi, S., Liu, G., Dong, X., Huang, N., Li, W., Chen, H., and Zhu, B. Identification of potential crucial gene network related to seasonal allergic rhinitis using microarray data. *European Archives of Oto-Rhino-Laryngology* 274 (1): 231–237, 2017.

Stojanova, D., Ceci, M., Malerba, D., and Dzeroski, S. Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC Bioinformatics* 14 (1): 285, 2013.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43 (D1): D447–D452, 2014.

Tsang, S., Kao, B., Yip, K. Y., Ho, W.-S., and Lee, S. D. Decision trees for uncertain data. *IEEE transactions on knowledge and data engineering* 23 (1): 64–78, 2011.

Wan, C. and Freitas, A. A. An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features. *Artificial Intelligence Review*, 2017.

Wan, C., Freitas, A. A., and de Magalhães, J. P. 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* 12 (2): 262–275, 2015.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

Xu, J., Li, N., Mao, X.-J., and Yang, Y.-B. Efficient probabilistic frequent itemset mining in big sparse uncertain data. In *Pacific Rim International Conference on Artificial Intelligence*. Gold Coast, Australia, pp. 235–247, 2014.

Yang, J.-L. and 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, pp. 163–168, 2009.

Yang, L., Chen, H., Cui, Q., Fu, X., and Zhang, Y. Probabilistic-KNN: A novel algorithm for passive indoor-localization scenario. In *IEEE Vehicular Technology Conference*. Glasgow, United Kingdom, pp. 1–5, 2015.