BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 24 2009, pages 3303–3309
doi:10.1093/bioinformatics/btp585

Data and text mining

Predicting citation count of *Bioinformatics* papers within four years of publication

Alfonso Ibáñez*, Pedro Larrañaga and Concha Bielza

Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, 28660 Madrid, Spain

Received on August 14, 2009; revised on September 29, 2009; accepted on October 3, 2009

Advance Access publication October 9, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Nowadays, publishers of scientific journals face the tough task of selecting high-quality articles that will attract as many readers as possible from a pool of articles. This is due to the growth of scientific output and literature. The possibility of a journal having a tool capable of predicting the citation count of an article within the first few years after publication would pave the way for new assessment systems.

Results: This article presents a new approach based on building several prediction models for the *Bioinformatics* journal. These models predict the citation count of an article within 4 years after publication (global models). To build these models, tokens found in the abstracts of *Bioinformatics* papers have been used as predictive features, along with other features like the journal sections and 2-week post-publication periods. To improve the accuracy of the global models, specific models have been built for each *Bioinformatics* journal section (*Data and Text Mining, Databases and Ontologies, Gene Expression, Genetics and Population Analysis, Genome Analysis, Phylogenetics, Sequence Analysis, Structural Bioinformatics and Systems Biology*). In these new models, the average success rate for predictions using the naive Bayes and logistic regression supervised classification methods was 89.4% and 91.5%, respectively, within the nine sections and for 4-year time horizon.

Availability: Supplementary material on this experimental survey is available at http://www.dia.fi.upm.es/∼concha/bioinformatics.html

Contact: aibanez@fi.upm.es

1 INTRODUCTION

Publishers nowadays face the problem of deciding which of the many papers they receive are of higher quality for publication in their journals. The current method used for article assessment is peer review. This process involves two or more authors reading and discussing different papers to determine the validity of the ideas and results, and their potential impact on the world of science.

Although if used properly peer review is assumed to be the most reliable system, it is slow, expensive and unwieldy (Cobo *et al.*, 2007; Mulligan, 2005; Scarpa, 2006). Other authors contest this appraisal (Hanks, 2005; Horrobin, 2001). This difference of opinion among authors has led to the development of several quantitative metrics associated with scientific production. One such metric is citation count. Citation count is the number of citations received by a paper in a period of time. Although citations are a measure of visibility, they can be considered as an indirect measure of article quality. The aim of this measure is to mirror the impact and quality of papers (Bornmann and Daniel, 2008).

Our work is based on the construction of predictive models to forecast the citation count of a paper within 4 years after publication. For this study we focus on papers published in *Bioinformatics* from January 1, 2005 to December 31, 2007. The supervised classification methods used in this article are Bayesian networks (naive Bayes and K2), logistic regression, decision trees and the K-nearest neighbor (K-NN) algorithm. These methods will be compared with each other.

2 RELATED WORK

In recent years, several researchers have investigated the prediction of citation count. Their work differs primarily as regards the prediction time horizon for the citation count and the predictive features used.

Several papers predict the number of citations using information gathered *after* publication. Brody *et al.* (2005) used download data within 6 months after publication as a predictive feature. However, the aim was to show the Open Access advantage. Castillo *et al.* (2007) used the number of citations, the authors' reputation and the source of the paper citations as predictive features. Lokker *et al.* (2008) used features related to the article and journal, like number of authors, pages, references and so on.

These three works used measures taken after the paper was published to predict its citation count in the future. The main disadvantage of using this feature is that the required values are not available until after publication.

On the other hand, others papers like Fu and Aliferis (2008) attempt to forecast citation count with the information available at the time of publication. Fu and Aliferis (2008) predict citation count within 10 years after publication with bibliometric information (number of articles for the first author, number of citations for the first author, number of authors, number of institutions and so on), the journal impact factor and the content of the article (title, abstract and MeSH terms). All these features are available at the time of publication. Support vector machine classification models were used as the learning algorithm. Predictions were made for a simple binary response variable that is defined by a set of citation thresholds to determine if an article is labeled positively or negatively. For a given threshold t, a positive label means that an article received at least

*To whom correspondence should be addressed.

t citations within 10 years after publication. These thresholds were 20 (mildly influential), 50 (relatively influential), 100 (influential) and 500 (extremely influential). Depending on the threshold used, the models output area under the receiver operating characteristic (ROC) curve (AUC) values ranging from 0.857 to 0.918.

As in Fu and Aliferis (2008), we also deal with the response variable as a discrete variable. Unlike Fu and Aliferis (2008), the variable that counts the number of citations is discrete rather binary but, taking three possible values (*few*, *some* and *many* citations). This leads to the use of classification methods rather than regression models to predict citation counts. Unlike Fu and Aliferis (2008) that use only support vector machines, we will take into account several classification methods and analyze which one provides better predictions for the problem. Moreover, our models will be constructed especially to predict annual time horizons (each of the first 4 years after publication) and for each *Bioinformatics* journal section. The information required from each article is its abstract content and the number of 2-week periods after publication. Hence, as opposed to other previous models described above that require information that is not available until after publication, our predictions will be available at publication time. Also, we will exploit the information output by the model, like the identification of key features (e.g. words in the abstract) that increase the chances of citation. This method can actually inform publishers about which articles will have a bigger impact in the future before they are published.

3 METHODS

3.1 Dataset

In the following, we illustrate the different phases for building the predictive models. In this article, we will build two different types of predictive models: global models and specific models. Global models attempt to predict the number of citations received by an article within each of the 4 years after publication, using information on all papers published in *Bioinformatics* over 3 years, from January 1, 2005 to December 31, 2007. Specific models have the same objective but, in this case, they use the information related to articles published within a specific *Bioinformatics* journal section.

The collection of abstracts published in *Bioinformatics* is the starting point for the construction of predictive models.

3.1.1 Collecting abstracts We selected *Bioinformatics* as the journal for this study. The basic elements of this work are the abstracts published in the *Bioinformatics* journal sections (*Data and Text Mining, Databases and Ontologies, Gene Expression, Genetics and Population Analysis, Genome Analysis, Phylogenetics, Sequence Analysis, Structural Bioinformatics* and *Systems Biology*) from 2005 to 2007. Before that date, no such sections existed. We accessed the *Bioinformatics* web site (http://bioinformatics.oxfordjournals.org/) to collect these abstracts. Once we had gathered this information, we stored the abstracts, the journal section and the number of 2-week periods from the beginning of the year to the publication date in a database designed for this purpose. This database is available at our web page.

3.1.2 Indexing abstracts The objective of this step was to use one of the Lucene library functions to build an index. Using this index, which references all abstracts in the corpus, we can more easily build datasets.

Lucene is an open source information retrieval library originally implemented in JAVA (http://lucene.apache.org/). It is used for programming search engines. Its main objectives are document indexation and retrieval.

3.1.3 Documenting citation count The next phase after collecting and indexing abstracts was to get the number of citations received by each article within each year after publication until December 31, 2008. For this purpose, we accessed the information available in the Web of Knowledge (http://www.isiknowledge.com/). The Web of Knowledge platform is composed of several databases. We chose the Web of Science (WoS) database as our citation count source. The information collected was stored in our database. This data will belong to the predictive models' training set.

3.1.4 Extracting tokens Different abstracts will be used depending on the model to be built (global models or specific models). In the case of global models, all the abstracts available in our database will be used, whereas abstracts belonging to the selected section will be used to build specific models.

The first step of this process is to output a ranking of tokens ordered by frequency of occurrence in the abstract set. This ranking is composed of one-, two- and three-word tokens.

The second step is to filter the ranking to reduce the large number of different tokens. The proposed filter is based on removing tokens that appear only occasionally in the abstract set. In this way, tokens that have a frequency of occurrence of less than three will be removed.

The next phase eliminates tokens that are repeated frequently and are irrelevant to the case study. For example, prepositions and articles are classic examples of stopwords. Generally, these tokens appear in all abstracts, and play no role in building the predictive model.

The last step is to associate tokens with their morphological root. We used the Porter algorithm provided by Lucene.

3.1.5 Building the dataset To construct the final dataset we need the information stored in our database, the tokens output by the above process and data from searches in the Lucene index.

In this step, we must design the dataset structure. The dataset structure will be different depending on the model to be built. The dataset structure of global models is made up of the *Section*, *Date*, *Token-1*, ..., *Token-n* features and *Citation* variable; whereas the specific models have the same structure except for the *Section* feature, which is constant.

In the case of global models, *Section* can take the values: *1-Data and Text Mining, 2-Databases and Ontologies, 3-Gene Expression, 4-Genetics and Population Analysis, 5-Genome Analysis, 6-Phylogenetics, 7-Sequence Analysis, 8-Structural Bioinformatics* and *9-Systems Biology*. These values correspond to the different *Bioinformatics* journal sections.

The feature *Date* refers to the number of 2-week periods from the beginning of the year to the publication date. It can take the values $\{1, 2, ..., 24\}$.

Token-i are the features that belong to the list of the tokens output in Section 3.1.4. These features are binary, and take the value 1 or 0 depending on whether or not the token is present in the selected abstract.

Finally, the *Citation* variable corresponds with the class label. It can take the values {*few*, *some*, *many*}. The first value, *few*, describes papers that receive at most one citation in a specific year according to the WoS. The value *some* applies to papers that receive 2, 3 or 4 citations in a year. And finally, the value *many* refers to papers that receive a number of citations equal to or greater than five.

3.2 Supervised classification methods

3.2.1 Selecting features To determine whether all dataset features are equally important or necessary to discriminate between the values {*few*, *some*, *many*}, we ran feature selection. The objective of feature selection is to build parsimonious models. Features that are irrelevant or redundant will not appear in these models. The benefits of applying feature selection include better classification performance, faster classification models, smaller databases and the ability to gain more insight into the process that is being modeled (Saeys *et al.*, 2007).

Table 1. Distribution of the data (papers), according to nine journal sections and citation count (*few*, *some* and *many*) across the 4-year time horizon

	Number of papers															
	First-year				*Second-year*				*Third-year*				*Fourth-year*			
	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*	Total	*f*	*s*	*m*
1-Data and Text Mining	**88**	81	7	0	**88**	39	31	18	**50**	10	17	23	**24**	8	7	9
2-Databases and Ontologies	**37**	32	3	2	**37**	13	14	10	**25**	6	8	11	**15**	6	3	6
3-Gene Expression	**283**	253	26	4	**283**	107	114	62	**192**	38	66	88	**107**	23	37	47
4-Genetics and Population Analysis	**46**	41	5	0	**46**	19	16	11	**27**	7	9	11	**21**	1	11	9
5-Genome Analysis	**103**	93	9	1	**103**	26	46	31	**82**	19	27	36	**53**	11	22	20
6-Phylogenetics	**28**	23	4	1	**28**	6	14	8	**20**	2	9	9	**11**	2	4	5
7-Sequence Analysis	**190**	170	16	4	**190**	73	77	40	**141**	49	46	46	**82**	22	31	29
8-Structural Bioinformatics	**150**	130	19	1	**150**	49	55	46	**103**	22	35	46	**54**	6	13	35
9-Systems Biology	**161**	140	20	1	**161**	56	65	40	**100**	14	32	54	**53**	3	12	38
All journal sections	**1086**	963	109	14	**1086**	388	432	266	**740**	167	249	324	**420**	82	140	198

f = *few*, *s* = *some*, *m* = *many*. Numbers in boldface represent the total number of papers belonging to a journal section in a particular year.

In this case, we used correlation-based feature selection (CFS) (Hall, 1999) as our feature selection algorithm. The basic idea behind this algorithm is to find a good set of features that are highly correlated with the class to be predicted (in our case *Citation*), but are not correlated with each other. CFS is a filter (Kohavi and John, 1997) that uses a correlation-based heuristic algorithm to evaluate each feature subset.

3.2.2 Naive Bayes This method (Minsky, 1961) is a Bayesian classifier. It is based on the Bayes' theorem under the assumption of conditional independence of predictors given the class.

3.2.3 K2 This algorithm greedily learns a Bayesian network from a dataset by using the marginal likelihood score (Cooper and Herskovits, 1992). Starting from the empty graph and a fixed order of the variables, this algorithm adds a variable as a parent to a given variable whenever its inclusion represents an improvement in the marginal likelihood score.

3.2.4 Logistic regresion The probability of an event is assumed to be a logistic function of certain variables that are considered potentially influential. The parameters of the model are estimated using the method of maximum likelihood and describe the size of the contribution of each variable to the model (Hosmer and Lemeshow, 2000).

3.2.5 C4.5 The C4.5 algorithm aims at inducing a decision tree that represents the knowledge of the problem with a tree structure by a recursive division of the predictors' space. This algorithm is an improvement of the ID3 algorithm (Quinlan, 1993).

3.2.6 K-NN The basic idea of the K-NN method is that a new case will be classified as the most frequent class among its K-NN. Euclidean distance is used to estimate the nearest neighbors of a given case (Hart, 1968).

3.3 Assessment procedure

We chose k-fold cross-validation as the procedure for estimating the probability of models classifying new cases according to the value of the predictive features. This method divides all cases from the dataset into k disjoint subsets of approximately equal size. Each subset is used to test a model that is learned from the other $k-1$ subsets. The k percentages of well-classified cases are averaged to output the estimated value of the model learned from all cases to classify new cases (Stone, 1974).

4 RESULTS

We used an open source machine learning package called Weka (Witten and Frank, 2005) to output the results shown below. In this research, we used the following Weka implementations: *NaiveBayesSimple* for naive Bayes, *BayesNet (K2)* for general Bayesian networks, *Logistic* for logistic regression, *J48* for decision trees and *IBK* for the K-NN algorithm.

4.1 Data distribution

Table 1 shows the distribution of the articles selected in this research. This table illustrates the number of papers belonging to a journal section in a particular year. Furthermore, it shows the distribution associated with each value of the class to be predicted.

The number of articles selected to build the predictive models varies depending on the year. To construct the models assigned to the *first-* and *second-year*, articles published in the years 2005, 2006 and 2007 were used (1086 papers). On the other hand, the models for the *third-year* used papers published in 2005 or 2006 (740 papers), and finally, the predictive models for the *fourth-year* used articles published in 2005 only (420 papers). Clearly, the longer the prediction horizon is the fewer papers are used to induce the models.

To give an understanding of the meaning of Table 1, some examples are explained below. The value (*All journal sections; Second-year; Total*) shows that 1086 articles are available to induce the global models in the *second-year*. According to the number of citations received, these articles are further divided into *few* (388), *some* (432) and *many* (226). Table 1 also lists the number of papers used in the specific models. For example, the models associated with *5-Genome Analysis* and *third-year* use 82 papers.

Analyzing Table 1, we find that the number of articles used in the *first-year* is 1086. Section *3-Gene Expression* accounts for 26.01% of all these articles. This is the section with most associated articles. At the other end of the spectrum, the sections with fewer papers in the *first-year* are *2-Databases and Ontologies* (3.41%), *4-Population Genetics and Analysis* (4.23%) and *6-Phylogenetics* (2.58%). The sections with more and fewer papers are the same across all years.

4.2 Global models

Several global models have been constructed for predicting the citation count of all the articles within 4 years after publication. Each model is associated with one of the 4 years to be predicted and one of the five supervised classification methods studied. Table 2 shows the results for each model.

These results could be better since apart from *first-year* models, model accuracy is <80%. There are some classification methods that provide better results than others. In this case, Bayesian classifiers have a higher average success rate within the 4 years (naive Bayes: 73.40% and K2: 70.37%), whereas logistic regression (65.85%), decision trees (60.15%) and K-NN (56.47%) yield the worst results.

Table 2. Accuracy and SD of global models

	All journal sections			
	First-year	*Second-year*	*Third-year*	*Fourth-year*
NB	91.4 ± 1.62	57.4 ± 6.08	68.9 ± 5.07	75.9 ± 6.39
K2	89.7 ± 2.54	57.4 ± 5.38	65.3 ± 4.48	69.1 ± 6.83
LR	84.7 ± 3.95	56.6 ± 2.75	59.3 ± 5.56	62.8 ± 7.20
C4.5	88.2 ± 0.47	48.8 ± 4.02	48.6 ± 4.69	55.0 ± 4.74
K-NN	88.5 ± 0.73	44.6 ± 4.72	38.5 ± 4.55	54.3 ± 4.95

NB = naive Bayes, K2 = K2 algorithm, LR = logistic regression, C4.5 = C4.5.

Although the *first-year* model has a much higher success rate than the models for the other years, the results are not satisfactory. This is because most cases belong to the *few* class (Table 1), and this is an obstacle to learn about the *some* and *many* classes since models avoid classifying cases into these classes. The C4.5 and K-NN methods especially tend to make this error for the *first-year* time horizon, whereas Bayesian classifiers and logistic regression are not prone to this error. The confusion matrices associated with these models are available at our web page.

4.3 Specific models

In response to accuracy concerns in the global models, new specific models were developed. Each model is associated with one of the nine journal sections, one of the four time horizons and one of the five supervised classification methods studied. Table 3 shows results for the new models.

Table 3 shows that the results depend of the journal section, the time horizon and the supervised classification method used. The highest percentage of correctly classified cases is 100%, which

Table 3. Accuracy and SD of specific models

	Section 1	Section 2	Section 3	Section 4	Section 5	Section 6	Section 7	Section 8	Section 9
First-year	7 features	7 features	30 features	5 features	11 features	18 features	33 features	54 features	48 features
NB	**96.6 ± 5.42**	91.9 ± 12.1	94.0 ± 4.42	**95.6 ± 8.45**	93.2 ± 4.73	**100 ± 0.00**	**95.8 ± 3.58**	**97.3 ± 3.42**	**96.9 ± 3.32**
K2	**95.6 ± 5.74**	92.5 ± 12.1	94.0 ± 3.26	**98.0 ± 6.32**	92.3 ± 4.09	**96.7 ± 10.5**	**96.3 ± 3.55**	**95.3 ± 4.50**	**96.3 ± 3.20**
LR	**98.9 ± 3.53**	86.5 ± 17.7	90.8 ± 2.55	**95.7 ± 8.46**	94.2 ± 5.06	92.9 ± 14.0	94.7 ± 4.36	92.0 ± 4.26	91.9 ± 4.81
C4.5	92.0 ± 5.42	86.5 ± 13.2	90.1 ± 2.22	89.1 ± 10.5	90.3 ± 0.48	82.1 ± 17.6	89.5 ± 2.53	86.7 ± 3.19	88.2 ± 3.59
K-NN	92.0 ± 5.42	86.5 ± 13.2	89.4 ± 0.22	89.1 ± 10.5	90.3 ± 0.48	82.1 ± 17.6	89.5 ± 0.00	86.7 ± 0.00	87.0 ± 1.65
Second-year	71 features	50 features	187 features	52 features	109 features	29 features	128 features	134 features	106 features
NB	82.9 ± 12.3	89.2 ± 21.9	86.2 ± 6.23	84.8 ± 14.1	86.4 ± 6.59	89.3 ± 16.1	85.3 ± 6.65	88.7 ± 7.20	87.0 ± 8.52
K2	69.3 ± 15.4	72.5 ± 23.3	75.3 ± 7.39	78.0 ± 15.3	70.6 ± 14.5	70.0 ± 24.6	80.0 ± 7.94	68.0 ± 12.1	80.8 ± 11.5
LR	**96.6 ± 5.43**	94.6 ± 12.4	90.1 ± 5.42	**95.7 ± 9.62**	**95.1 ± 6.93**	92.9 ± 14.0	**95.3 ± 5.87**	94.7 ± 7.60	93.2 ± 6.17
C4.5	56.8 ± 11.4	37.8 ± 11.3	53.7 ± 6.06	58.7 ± 22.8	61.2 ± 14.9	57.1 ± 30.6	58.9 ± 8.03	54.0 ± 6.90	56.5 ± 9.60
K-NN	53.4 ± 9.34	73.0 ± 20.1	62.2 ± 8.94	50.0 ± 17.1	63.1 ± 15.5	60.7 ± 24.6	57.4 ± 7.91	72.7 ± 15.5	60.2 ± 9.35
Third-year	58 features	37 features	143 features	37 features	87 features	18 features	149 features	109 features	83 features
NB	90.0 ± 14.1	80.0 ± 21.9	85.9 ± 6.81	88.9 ± 22.5	**95.1 ± 6.34**	85.0 ± 24.1	86.5 ± 7.14	89.3 ± 11.0	84.0 ± 11.7
K2	72.0 ± 21.5	73.3 ± 25.1	70.8 ± 7.23	63.3 ± 28.1	72.9 ± 14.4	90.0 ± 21.1	71.7 ± 14.2	77.0 ± 15.4	77.0 ± 14.9
LR	94.0 ± 9.78	**100 ± 0.00**	90.6 ± 7.74	92.6 ± 14.0	93.9 ± 6.66	**100 ± 0.00**	**95.7 ± 5.09**	93.2 ± 6.70	93.0 ± 8.20
C4.5	50.0 ± 14.1	48.0 ± 12.3	66.7 ± 7.86	48.1 ± 21.4	54.9 ± 9.78	70.0 ± 35.0	58.2 ± 9.23	58.3 ± 11.1	57.0 ± 11.6
K-NN	64.0 ± 15.8	52.0 ± 12.3	54.7 ± 8.70	92.6 ± 21.1	58.5 ± 14.2	75.0 ± 26.3	58.9 ± 9.12	65.1 ± 11.4	60.0 ± 6.70
Fourth-year	32 features	12 features	91 features	25 features	58 features	10 features	80 features	39 features	22 features
NB	91.7 ± 21.1	80.0 ± 24.1	86.0 ± 4.53	90.5 ± 18.0	90.6 ± 9.93	81.8 ± 42.2	93.9 ± 10.6	87.0 ± 12.0	90.6 ± 13.9
K2	73.3 ± 33.5	70.0 ± 34.9	71.3 ± 16.9	86.7 ± 21.9	77.7 ± 22.7	80.0 ± 42.1	75.6 ± 15.7	83.7 ± 9.90	88.3 ± 13.8
LR	91.7 ± 18.0	66.7 ± 45.9	93.4 ± 10.8	85.7 ± 24.9	88.7 ± 15.7	63.7 ± 51.6	**95.1 ± 12.1**	81.5 ± 19.8	90.6 ± 19.4
C4.5	33.3 ± 24.1	26.7 ± 35.4	55.1 ± 15.8	66.7 ± 27.7	41.6 ± 14.1	63.7 ± 47.4	59.8 ± 10.2	59.3 ± 18.6	69.8 ± 11.1
K-NN	83.3 ± 24.1	53.3 ± 40.8	72.9 ± 14.8	66.7 ± 27.7	58.5 ± 16.3	45.4 ± 49.7	63.4 ± 19.9	68.5 ± 9.30	71.7 ± 8.80

Numbers in boldface represent an average success rate better than 95%.

was achieved on three occasions by the naive Bayes and logistic regression methods. On the other hand, the results were poorest for the C4.5 and K-NN methods that output values of <50%.

Table 3 also shows the number of features accounted for the different predictive models. Fixing a specific journal section and analyzing the average number of features within the 4-year time horizon, we observe that sections with fewer features are *6-Phylogenetics* (18.75) and *2-Databases and Ontologies* (26.5), whereas the sections with most features are *3-Gene Expression* (112.75) and *7-Sequence Analysis* (97.5).

Looking at the behavior of the classifier for each value to be predicted {*few*, *some*, *many*}, Table 4 shows the confusion matrices of the models associated with sections *1-Data and Text Mining* and *2-Databases and Ontologies*, and with the logistic regression and decision trees methods, respectively. These models were chosen because they are the ones that are most and least accurate within each of the four time horizons, respectively (Table 3).

To check the good behavior of the logistic regression method, we focus, for example, on the confusion matrix of *1-Data and Text Mining* and the *second-year* model. This matrix shows that the total number of cases to be predicted is 88. Of these, 85 cases are well classified (96.6%) and three cases are wrongly classified (3.4%). Analyzing each value of the class, we find that the success rate for the values *few* and *some* is 100%, whereas three errors are made for the value *many*, where one is classified as *few* and two as *some*. On the other hand, the confusion matrix of the *2-Databases and Ontologies* and *fourth-year* model is an example of the poor behavior of C4.5. In this case, the model tries to predict 15 cases, of which four are well classified (26.7%) and the rest are wrongly classified (73.3 %). Analyzing the different values of the class, we find that the success rate for the values for *few* and *many* is 33%, whereas success for the value for *some* is 0%, where all instances of this value are classified as *many* rather than as *some*.

Figures 1 and 2 illustrate the results of these new specific models. In Figure 1, the height of the bars indicates the average percentages for the different classifiers within the four time horizons, with a fixed journal section. Taking the first bar as an example, the value displayed is 90.3%. This value is the mean accuracy for naive Bayes applied to *1-Data and Text Mining* averaged across the four time horizons.

Figure 1 shows that the journal section predicted with the highest success rate is method dependent. Logistic regression and naive Bayes achieve some notable results. Logistic regression predicts the *1-Data and Text Mining* journal section with a 95.30% success rate across the four time horizons, whereas naive Bayes predicts the *5-Genome Analysis* with an average accuracy of 91.32% across the four time horizons. On the other hand, the *4-Genetics and Population Analysis* journal section has the highest average percentage of cases well classified by all five algorithms (80.82%), whereas *2-Databases and Ontologies* is the journal section with the lowest percentage of well classified cases with an average accuracy of 73.05% for all the tested algorithms.

On the other hand, the height of the bars in Figure 2 indicates the average percentages scored by the different classifiers for the nine journal sections studied with a fixed year of publication. Taking the first bar as an example, the value displayed is 95.70%. This value is the mean accuracy of applying the naive Bayes classification method for the *first-year* of publication averaged over all journal sections.

The best average results are for the *first* time horizon at 92.06% across all classifiers. The second, third and fourth time horizons have many similarities with each other, where percentages range from 73% to 75%. Looking at the scores for each algorithm, note that naive Bayes, K2, C4.5 and K-NN predict the *first-year* more

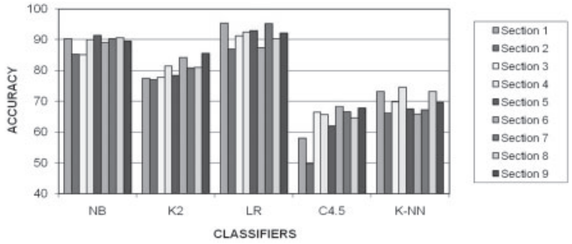

Fig. 1. Average accuracy within the four prediction years by each section and classification method.

Table 4. Confusion matrices of two specific models (logistic regression and decision trees models).

Section *1-Data and Text Mining* (logistic regression)

First-year (98.9 ± 3.53)				*Second-year* (96.6 ± 5.43)				*Third-year* (94.0 ± 9.78)				*Fourth-year* (91.7 ± 18.0)			
a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as
81	0	0	a = *few*	**39**	0	0	a = *few*	**10**	0	0	a = *few*	**8**	0	0	a = *few*
1	**6**	0	b = *some*	0	**31**	0	b = *some*	0	**17**	0	b = *some*	0	**7**	0	b = *some*
0	0	**0**	c = *many*	1	2	**15**	c = *many*	0	3	**20**	c = *many*	2	0	**7**	c = *many*

Section *2-Databases and Ontologies* (C4.5)

First-year (86.5 ± 13.2)				*Second-year* (37.8 ± 11.3)				*Third-year* (48.0 ± 12.3)				*Fourth-year* (26.7 ± 35.4)			
a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as	a	b	c	← Classified as
32	0	0	a = *few*	**7**	5	1	a = *few*	**1**	0	5	a = *few*	**2**	0	4	a = *few*
3	**0**	0	b = *some*	7	**6**	1	b = *some*	2	**0**	6	b = *some*	0	**0**	3	b = *some*
2	0	**0**	c = *many*	6	3	**1**	c = *many*	0	0	**11**	c = *many*	2	2	**2**	c = *many*

Numbers in boldface represent the total number of well-classified cases in each class.

accurately. However, logistic regression predicts the *third-year* more accurately, although this result is not significant compared with *first-* and *second-year* results (Fig. 2).

After analyzing all results, we can conclude that logistic regression and naive Bayes are the supervised classification methods that solve the problem more accurately. Comparing these methods, logistic regression achieves a higher success rate, scoring 91.55% on average across the nine sections within the four time horizons, whereas naive Bayes attains 89.38%. Additionally, these methods are the only ones that correctly classified 100% of cases for a specific year and section (Table 3). Regarding the journal sections and time horizons, logistic regression specializes in section *1-Data and Text Mining* (95.30%) and in the *third-year* (94.78%), whereas naive Bayes specializes in *5-Genome Analysis* (91.32%) and the *first-year* (95.70%).

4.4 Exploiting the best models

The purpose of this section is to find out whether there are any tokens that influence an article's citation counts within the journal sections and time horizons. This analysis shows the results of predicting the number of citations of a new article using the models of the journal section *6-Phylogenetics* in the *third-year* learned by naive Bayes and logistic regression models (18 features, see Table 3). Other models and more predictions are available at our web page.

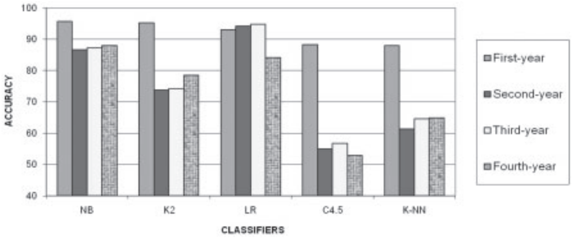

Fig. 2. Average accuracy within the nine journal sections by each prediction year and classification method.

Analyzing the probability distributions stored in the features of the naive Bayes model, we find that the fact that an article receives *few*, *some* or *many* citations determines the probability of occurrence of tokens in the article. Similarly, if some tokens appear in an article, they influence the citation count, and thus determine the value of the class to be predicted.

The three probability columns $P(X_i|f)$, $P(X_i|s)$ and $P(X_i|m)$ in Table 5 show the distributions of each token subject to our class values. These distributions show that there are some tokens like *linear*, *probability*, *discussed*, *automated*, *time* and *nucleotide*, which tend to appear more frequently in the papers with *few* citations. On the other hand, for papers that have *some* citations, these tokens are *time*, *nucleotide*, *dynamic*, *entire*, *independent*, *compared*, *interaction*, *clustering* and *protein*. Finally, *parameter*, *performance*, *analyze*, *researchers*, *likelihood based*, *linear* and *probability* are the tokens with a higher frequency of occurrence in articles that receive *many* citations.

The above probabilities and the marginal probability of each class value $P(C{=}c)$ with $c = f, s, m$ (Table 1), are the basic elements of the naive Bayes model used to predict the citation count of a specific paper ($\boldsymbol{x}$). This model is

$$P(C=c\,|\,\boldsymbol{x}) \propto P(C=c)\prod_{i=1}^{n} P(X_i=x_i\,|\,C=c).$$

On the other hand, the logistic regression model requires some coefficients (β_i) to calculate the class value with higher a posteriori

Table 5. Parameters that define naïve Bayes and logistic regression models

	$P(X_i=1\|C=c)$			*Coeff. LR*		*New article*
Feature (X_i)	$P(X_i\|f)$	$P(X_i\|s)$	$P(X_i\|m)$	β_i^f	β_i^s	$\boldsymbol{x}$
Parameter	0.25	0.09	**0.27**	−6.76	−10.93	
Performance	0.25	0.09	**0.27**	−6.76	−10.93	
Analyze	0.25	0.09	**0.36**	−12.35	−11.57	√
Researchers	0.25	0.09	**0.27**	−7.53	−10.93	√
Likelihood based	0.25	0.09	**0.27**	−8.13	−10.93	
Linear	**0.50**	0.09	**0.27**	−13.17	−11.57	
Probability	**0.50**	0.09	**0.27**	−3.18	−11.57	
Discussed	**0.75**	0.09	0.09	28.34	−10.93	
Automated	**0.50**	0.09	0.09	26.85	−10.35	√
Time	**0.50**	**0.36**	0.09	12.38	7.22	
Nucleotide	**0.50**	**0.36**	0.09	12.38	7.22	√
Dynamic	0.25	**0.27**	0.09	5.22	12.19	√
Entire	0.25	**0.27**	0.09	5.22	12.20	√
Independent	0.25	**0.36**	0.09	5.53	12.91	
Compared	0.25	**0.45**	0.09	5.88	13.72	√
Interaction	0.25	**0.27**	0.09	5.22	12.20	
Clustering	0.25	**0.27**	0.09	5.22	17.42	√
Protein	0.25	**0.45**	0.09	5.88	13.72	
Intercept (β_0)				−16.1833	−3.6972	

Naive Bayes and logistic regression models have been used for predicting the number of citations in the third year of a new article published in section *6-Phylogenetics*. Numbers in boldface represent the highest probability values in each class.

probability. These coefficients are shown in the middle columns (β_i^f and β_i^s) of Table 5. The models used for these predictions are

$$P(C=f\,|\,\boldsymbol{x})=\frac{e^{(\beta_0^f+\sum_{i=1}^n \beta_i^f x_i)}}{1+e^{(\beta_0^f+\sum_{i=1}^n \beta_i^f x_i)}+e^{(\beta_0^s+\sum_{i=1}^n \beta_i^s x_i)}}$$

$$P(C=s\,|\,\boldsymbol{x})=\frac{e^{(\beta_0^s+\sum_{i=1}^n \beta_i^s x_i)}}{1+e^{(\beta_0^f+\sum_{i=1}^n \beta_i^f x_i)}+e^{(\beta_0^s+\sum_{i=1}^n \beta_i^s x_i)}}$$

$$P(C=m\,|\,\boldsymbol{x})=1-P(C=f\,|\,\boldsymbol{x})-P(C=s\,|\,\boldsymbol{x}).$$

The new case to be predicted is shown in the last column of Table 5. This new case is a paper abstract. *Analyze*, *researchers*, *automated*, *nucleotide*, *dynamic*, *entire*, *compared* and *clustering* are the tokens that appear in the abstract. After propagating this evidence, the results predicted by naive Bayes are $P(f|\boldsymbol{x})=0.30$, $P(s|\boldsymbol{x})=0.67$ and $P(m|\boldsymbol{x})=0.03$. On the other hand, the results predicted by logistic regression are $P(f|\boldsymbol{x})=0.18$, $P(s|\boldsymbol{x})=0.81$ and $P(m|\boldsymbol{x})=0.01$.

The results of both models show that an abstract with the above tokens published in the journal section *6-Phylogenetics* will receive *some* citations (i.e. 2, 3 or 4 citations) in the *third-year* after publication.

5 CONCLUSIONS

The use of models capable of predicting the citations that an article will receive in the first few years after publication can be a useful tool for publishers' assessment process. For this reason, we focus on building models to predict the citation count of articles that are published in *Bioinformatics*. We predicted citation count in each of the first 4 years after publication. This time horizon was chosen considering that it can help to estimate the journal impact factor.

The construction of specific models for each section of *Bioinformatics* solved the problems associated with global models. As a whole, the results of specific models achieved a greater rate of success across the 4 years than the global models. Model specialization affects not only the *Bioinformatics* journal sections, but also each of the 4 years in the time horizon.

The logistic regression and naive Bayes classification methods output high average scores in the nine journal sections and across the four time horizons, achieving rates of 91.5% (AUC = 0.943) and 89.4% (AUC = 0.983), respectively.

We found that the appearance of certain words in the paper abstracts can influence the number of citations received. The probabilities assigned and the tokens selected depend on the journal section and chosen time horizon. The selected tokens could be used as a point of reference to identify the hot topics.

Unlike the models developed by Brody *et al.* (2005), Castillo *et al.* (2007) and Lokker *et al.* (2008), the predictions of our models are not based on information available after publication. Our models use the information content of the article abstract. In this way, predictions can be made at publication time, and it is not necessary to wait until the end of a data collection period to predict citation count.

It could be worthwhile comparing our models with models proposed by Fu and Aliferis (2008) because, although they use different features, datasets, response variable and prediction horizon, they both attempt to predict citations before publication with tokens contained in the article abstract. However, in our case the accuracy of the naive Bayes (AUC = 0.983) and logistic regression (AUC = 0.943) supervised classification methods were higher than the accuracy achieved by models developed by Fu and Aliferis (2008) (AUC = 0.918).

In the future, our target will be to build new models that incorporate other paper-based features (title, keywords, conclusions, etc.), new author-based features (h-index, number of papers, number of citations, etc.) and new journal-based features (impact factor, immediacy index, category, etc.). These models would be induced using different machine learning methods. The way citation count is handled influences the results. It could be modeled as a continuous variable using other methods like regression, regularized regression, or local regression. Finally, the number of citations could vary depending on the source consulted (Google Scholar, Scopus, ISI WoS, etc.) (Bar-Ilan, 2008; Meho and Yang, 2007), which is a point to be taken into account.

Funding: Spanish Ministry of Science and Innovation, Project TIN2008-04528-E. Spanish Ministry of Science and Innovation, (grants TIN2007-62626, TIN2008-06815-C02-02 and Consolider Ingenio 2010-CSD2007-00018).

Conflict of Interest: none declared.

REFERENCES

Bar-Ilan,J. (2008) Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, **74**, 257–271.

Bornmann,L. and Daniel,H. (2008) What do citation counts measure? *J. Doc.*, **64**, 45–80.

Brody,T. *et al.* (2005) Earlier web usage statistics as predictors of later citation impact. *J. Am. Assoc. Inf. Sci. Technol. (JASIST)*, **57**, 1060–1072.

Castillo,C. *et al.* (2007) Estimating the number of citations using author reputation. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, Vol. 4726, Springer, Santiago, Chile, pp. 107–117.

Cobo,E. *et al.* (2007) Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PLoS ONE*, **2**, e332.

Cooper,G. and Herskovits,E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, **9**, 309–347.

Fu,L. and Aliferis,C. (2008) Models for predicting and explaining citation count of biomedical articles. In *AMIA Annual Symposium Proceedings*, Vol. 2008, pp. 222–226.

Hall,M. (1999) *Correlation-based Feature Selection for Machine Learning*. PhD Thesis, Department of Computer Science, Waikato University, New Zealand.

Hanks,G. (2005) Peer review in action: the contribution of referees to advancing reliable knowledge. *Palliat. Med.*, **19**, 359–370.

Hart,P.E. (1968) The condensed nearest neighbour rule. *Trans. Inf. Theory*, **14**, 515–516.

Horrobin,D. (2001) Something rotten at the core of science. *Trends Pharmacol. Sci.*, **22**, 51–52.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*, 2nd edn. Wiley, New York.

Kohavi,R. and John,G. (1997) Wrappers for feature subset selection. *Artif. Intelli.*, **97**, 273–324.

Lokker,C. *et al.* (2008) Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *Br. Med. J.*, **336**, 655–657.

Meho,L. and Yang,K. (2007) Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *J. Am. Soc. Inf. Sci. Technol.*, **58**, 2105–2115.

Minsky,M. (1961) Steps toward artificial intelligence. *IRE*, **49**, 8–30.

Mulligan,A. (2005) Is peer review in crisis? *Oral Oncology*, **41**, 135–141.

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, USA.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Scarpa,T. (2006) Peer review at NIH. *Science*, **311**, 41.

Stone,M. (1974) Cross-validation choice and assesment of statistical predictions. *J. R. Stat. Soc.*, **36**, 111–147.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco, USA.