Assigning Gene Ontology terms to biotext by classification methods

Theodosios Theodosiou[1], Morfoula Fragkopoulou[2], Lefteris Angelis[3], Athena Vakali[4],

[1,2,3,4]Department of Informatics, Aristotle University of Thessaloniki,
54124, Thessaloniki, Greece.
Email: {theodos,mfragkop,lef,avakali}@csd.auth.gr

Abstract

Biomedical literature databases constitute valuable repositories of up to date scientific knowledge. The development of efficient classification methods in order to facilitate the organization of these databases and the extraction of novel biomedical knowledge is becoming increasingly important. Several of these methods use bio-ontologies, like Gene Ontology to concisely describe and classify biological documents. The purpose of this paper is to compare two classical statistical classification methods, namely multinomial logistic regression (MLR) and linear discriminant analysis (LDA), to a machine learning classification method, called support vector machines (SVM). Although all the methods have been used with success for classifying texts, there is not a direct comparison between them for classifying biological text to specific Gene Ontology terms. The results from the study show that LDA performs better (accuracy 80.32%) than SVM (77.18%) and MLR (57.4%). LDA not only performs well in the assignment of Gene Ontology terms to documents, but also reduces the dimensions of the original data, making them easier to manage.

Keywords: biotext, classification, linear discriminant analysis, multinomial logistic regression, support vector machines

1. Introduction

The use of high-throughput experimental methods, such as microarrays, and the results from large-scale computational experiments produce huge amounts of biological information [Rubin D.L. et al. (2005)], which is usually stored in data repositories, like sequence databases. One of the biggest repositories of biomedical information, used extensively by researchers, is the PubMed [PubMed] database.of published biomedical articles containing bibliographic citations and abstract articles from more than 4800 biomedical journals [Ashburner M. (2000)].

As the amount of biological information and the number of biomedical documents increase, their management, categorization and exploitation through manual means becomes a difficult task since the size of the data is huge [Rubin D.L. et al. (2005)],

[Spasic I. et. al. (2005)]. Several machine learning and text mining methods [Spasic I. et. al. (2005)] try to facilitate and automate the organization of biological information described in documents. Also, some research efforts involve natural language techniques [Andrade M.A. et. al. (1997)], [Eisenhaber F. et. al. (1999)]. Other efforts [Raychaudhuri S. et. al. (2002)], [Kazawa H. et. al. (2004)], [Theodosiou T. et. al. (2006)] have utilized ontologies, like Gene Ontology (GO), and applied data mining and machine learning methods to published biological literature stored in the PubMed database. Ontologies are considered to be prerequisites for advanced text mining and very useful for concisely describing the information inside biomedical documents [Spasic I. et. al. (2005)].

Gene Ontology (GO) [Ashburner M. et. al. (2000)], is a standard, controlled and structured vocabulary mainly for describing a gene's functions, but has also been used for document description. Specifically, in [Raychaudhuri S. et. al. (2002)] 21 GO codes were used in order to compare the performance of three different classification models (maximum entropy, naive Bayes and nearest- neighborhood) concluding that the maximum entropy method outperforms. In [Kazawa H. et. al. (2004)], a set of 12 GO codes was used in order to evaluate Support Vector Machines (SVM) showing that SVM outperform the maximum entropy classification. In [Theodosiou T. et. al. (2006)] a methodology based on linear discriminant analysis (LDA) was proposed for classifying 12 GO terms to a set of biomedical documents.

The purpose of the present work is to compare the performance of classical statistical classification methods, namely multinomial logistic regression (MLR) and LDA to the performance of SVM, which is a typical machine learning method. The choice of the methods was based on the fact that all of them have been used with success to the field of biological text mining. It should be noted that biological text has specific properties, like term variation and term ambiguity, etc. compared to other knowledge domains [Spasic I. et. al. (2005)] which makes the task of text mining and classification quite demanding. The paper focuses on measuring the performance of the methods in classifying the documents to 12 GO terms. The use of GO instead of another Ontology is based on the fact that it is popular in literature and can be used for every organism [Camon E. et. al. (2003)]. The effectiveness of the methods is evaluated by specific performance metrics (such as recall, precision, F-measure, cross-validation and hold-out sample validation).

The remainder of the paper is structured as follows: Section 2 describes the methodology we used. Section 3 presents the results of experimentation. Conclusions and future work are given in Section 4.

2. Methodology

2.1 Data Representation

The source of published biomedical literature is the PubMed database. The documents[1] in this study are represented by the use of Vector Space Model (VSM). The steps of transforming each document into a vector are described briefly below [Manning C.D. et. al. (1999)]:

1. **Tokenization**: extraction of all words appearing in an entire set of documents;

2. **Stopword removal**: elimination of non-informative words (stopwords) such as ”a”, ”and”, ”the”, etc.;

3. **Stemming**: use only the root of each word;

4. **Frequency counting**: counting of the number of occurrences of each word in each document;

5. **Filtering**: elimination of non-content-bearing high-frequency and low-frequency words;

6. **Vector creation**: construction of a weight vector. In our study we used the binary weighting scheme, which is the simplest. According to it, the weights are binary numbers (0 denoting the absence of a word and 1 denoting the presence). We have to emphasize however, that the statistical methodology we used can be applied as well to other weighting schemes appearing in the literature [Manning C.D. et. al. (1999)].

The aforementioned procedure results in a keyword dataset where each document is represented by a binary vector of size p (total number of keywords).

2.2 Multinomial Logistic Regression (MLR)

Multinomial logistic regression (MLR) is the generalization of the logistic regression and it is used when the categorical dependent variable has more than two classes while simple logistic regression is used when the classes are only two. MLR builds a classification model based on any type (numerical or categorical) of independent variables (keywords). In our case the dependent categorical variable is the one containing the different GO groups.

MLR can be used to predict the GO code of a document, based on the keywords it contains. It applies maximum likelihood estimation after transforming the GO groups into a logit variable (the natural log of the odds of the GO group occurring). In this

[1] In our context a document signifies the title and the abstract only and not the whole document.

way, MLR estimates the probability of a certain document to belong to a specific GO group based on the keywords it contains [Tabachnick B. G. et. al. (2001)].

2.3 Linear Discriminant Analysis (LDA)

As in MLR, the purpose of LDA is to model the relationship between a dependent categorical variable (the GO codes) with a set of independent variables (the keywords).

The LDA procedure produces a number of statistical results leading to the estimation of the probability that a vector (of word weights) belongs in a particular group. The original idea of LDA was the projection of the $n \times p$ data matrix $\mathbf{X}$ (n is the number of PubMed documents and p the number of words) onto a single dimension using the *Fisher's linear discrimination function* $\mathbf{z} = \mathbf{Xa}$. The vector of coefficients $\mathbf{a}$ should be chosen in such a way that an optimal discrimination is achieved via the maximization of the ratio of the *between-group-sum of square* to the *within-group-sum of squares* [Hair J.F. et. al. (1998)].

In the case where the categories of the dependent variable (the 12 GO categories in our context) are $k > 2$, this idea can be generalized and the method computes

$$m = \min(k - 1, p) \quad \textbf{(1)}$$

new variables, the canonical discriminant functions which can be used to exhibit the differences among the categories. For detailed description of LDA we refer to [Hair J.F. et. al. (1998)].

2.4 Support Vector Machines (SVM)

SVM is a typical binary classifier, which may only solve classifications problems of two groups. In case of (more than two) multi-groups classification needs the standard approach is to reduce the multi-class problem into several binary sub-problems [Baldi P. et. al. (2003)].

SVM method has already been used for similar classification problems as in our case (such as in [Kazawa H. et. al. (2004)]). The basic idea in SVM is to define an optimal separating hyperplane that could separate the observations into two classes. This hyperplane is constructed based on certain vectors of the dataset (the so called support vectors) and a proper (so called) kernel function (to project the original data to this hyperplane). Typically, this kernel function requires several parameters which should be appropriately tuned. Examples of such parameters are the degree of the polynomial (to be set in polynomial kernel), or the value of the gamma parameter (to be set in radial kernel). The simplest form of the SVM has a linear kernel, and it is used as a basis for comparisons with LDA and MLR. It must be noted that kernel

methods can also be used for LDA [Kocsor A. (2001)], but the aim of this paper was to use the simplest forms of LDA and SVM.

2.5 Performance metrics

The validation of the models sensitivity (changes in the accuracy) can be achieved by cross validation. We use the hold-out [Hair J.F. et. al. (1998)] and the 10-fold cross validation methods. The hold-out method randomly splits the initial dataset into a subset containing most of the observations used for training the model and another smaller (hold-out sample) for validating the classification performance. The size of the hold out samples is determined empirically (there is no standard formula) and usually is not larger than 30% of the original dataset [Hair J.F. et. al. (1998)]. The 10-fold cross validation partitions the initial dataset into 10 different subsets. Of the 10 subsets a single one is retained for validating the classification model and the remaining 9 subsets are used for training the model. This is repeated 10 times with each of the 10 subsets used exactly once as the validation data.

The basis for the computation of various performance measures is the confusion or classification matrix [Hair J.F. et. al. (1998)]. Specifically, the most common performance measures are:

1. **Accuracy** [Hair J.F. et. al. (1998)] of the classification is the ratio of the correct classified documents to the total number of documents in the dataset.

2. **Precision** [Hair J.F. et. al. (1998)] is the percentage of the correct classifications of a specific group to all the documents assigned to that group.

3. **Recall** [Hair J.F. et. al. (1998)] is the percentage of the correct classifications of a specific group to all the documents of the group contained in the dataset.

4. **F-measure** [Manning C.D. et. al. (1999)] is the harmonic mean between precision and recall.

3. Experimentation

The proposed method was tested under a dataset of 12 GO terms used in earlier works, e.g. [Theodosiou T. et. al. (2006)]. For practical purposes, each GO code was represented by a group number ranging from 1 to 12. The assignment of the numbers was random and is listed in Table 1. Due to the space limitations, we outline the most important results of this study.

As a dataset we use a set of $n = 9009$ documents. The number of keywords extracted from the documents is $p = 1642$. In order to reduce the number of dimensions (keywords) we applied Principal Component Analysis (PCA) [Hair J.F. et. al. (1998)] to the dataset. PCA reduces the dimensions of the data and retains as much as possible of the information contained in the original data. After applying

PCA the $p = 1642$ keywords were reduced to $q = 622$ new real numbered variables. Thus, the data matrix used in our experiments consists of $n = 9009$ vector representations of real valued variables $q = 622$, accompanied by one of $k = 12$ GO codes.

The hold-out validation involved for each classification method the split of the dataset to 10 different hold-out samples. The samples were randomly created using 30% of the total number of documents. Every sample was used for validating the classification model constructed by the remaining training dataset. The accuracies of the ten different models were averaged in order to have more statistically reliable estimation for the overall performance.

Table 1. *GO codes, Terms and corresponding group number*

GO code	**GO term**	**Group no.**
GO:0006914	Autophagy	1
GO:0007049	Cell cycle	2
GO:0008283	Cell proliferation	3
GO:0007267	Cell cell signalling	4
GO:0006943	Chemimechanical coupling	5
GO:0007126	Meiosis	6
GO:0008152	Metabolism	7
GO:0007048	Oncogenesis	8
GO:0006950	Stress response	9
GO:0006810	Trasnport	10
GO:0008219	Cell death	11
GO:0007165	Signal transduction	12

It is evident from Table 2 that overall the performance of LDA is much better (80.32%) than those of MLR (57.4%) and SVM (77.18%). Although, MLR has the worst predictive accuracy of the three methods, it has the highest fitting accuracy (99.9%) which means that the model is overfitted to the training data. This means that MLR can not predict, as accurately as the other two methods, new data which have not been used for the training of the model. The 10-fold cross-validation results also show that LDA (81%) and SVM (79.38%) perform similarly and much better than MLR (60.49%).

Table 2. *Average fitting and predictive accuracy for hold-out experiments.*

Classification method	**Fitting Accuracy**	**Predictive accuracy**
MLR	99.9%	57.4%
LDA	93.63%	80.32%
SVM	96.44%	77.18%

Table 3 show the precision, recall and F-measure of each GO group, which are calculated from the classification matrices (not shown here due to space limitations).

In studying the accuracy results (Table 3 & Figure 1), we notice that for MLR group 1 (autophagy) has the lowest F-measure (31.8%), whereas for LDA and SVM the lowest F-measure corresponds to group 3 (cell proliferation), 57.14% and 26.8% respectively. This indicates that LDA can handle better data that are difficult to classify. The training dataset is informative about 'cell proliferation', but the information is highly correlated to other GO categories. For example, from the definition of 'cell proliferation' [Ashburner M. et. al. (2000)], we can see that the correlation with 'oncogenesis' is high and so the two terms are referenced together very often. A similar, but negative, correlation exists with 'cell death'.

Table 3. *Performance metrics (in %) for MLR, LDA & SVM*

	MLR			**LDA**			**SVM**		
GO	**Recall**	**Precision**	**F-meas.**	**Recall**	**Precision**	**F-meas.**	**Recall**	**Precision**	**F-meas.**
1	71.7	20.43	31.8	84.51	100	91.6	71.7	100	83.52
2	45.49	39.85	42.48	67.35	70.21	68.75	52.1	73.1	60.83
3	44.66	30.67	36.36	65.98	50.39	57.14	16.5	73.91	26.98
4	50.82	28.97	36.90	79.63	82.69	81.13	42.86	96	59.26
5	60.32	66.9	63.44	83.23	88.16	85.62	83.27	75.33	79.1
6	69	76.95	72.76	88.21	96.86	92.34	90.84	91.85	91.34
7	58.19	59.22	58.7	76.37	76.37	76.37	74.84	80.2	77.43
8	50.29	57.14	53.5	78.90	74.42	76.6	89.11	60.53	72.09
9	53.96	64.11	58.6	78.83	77.32	78.06	77.31	77.54	77.43
10	63.66	74.82	68.79	87.63	89.12	88.36	87.65	93.11	90.3
11	65.19	68.21	66.67	81.76	90.55	85.94	78.35	83.71	80.94
12	55.15	61.68	58.24	82.18	72.15	76.84	81.98	69.46	75.2

It must be also noted that an important advantage of LDA compared to MLR and SVM is that it reduces the number of dimensions of the original data. Since in our data the number of independent variables is quite large ($q = 622$) and the number of GO codes is $k = 12$, LDA results in only 11 variables ($m = 11$), according to Formula 1. The 11 new variables are ordered in the sense that the first is the most important for the discrimination between groups, and so on [Venables W. N. et. al. (2002)] and can be used for further analysis and interpretation of the data. For example, it is possible to graphically explore the discriminative power of the first two variables of LDA as in Figure 2. Figure 2 shows that by plotting the centroids (means) of the 12 groups with respect to the first two variables, GO categories 1, 8, and 11 are clearly discriminated. Also, GO groups 4 and 5 are further apart from the other groups. Of course, in order to classify all the 12 groups it is necessary to combine all the 11 new discriminant variables.

4. Conclusions

This work demonstrates that LDA and SVM can be used to facilitate the process of automatically assigning a GO category to a biomedical document with very good results. MLR, on the other hand, performs significantly less than the aforementioned methods. Although LDA is a classical statistical classification method it competes very well with SVM, a machine learning method. Furthermore, LDA achieves the reduction of the dimensions of the original data, which can help manage and understand better the information inside the documents. In the future we would like to present extended results on our comparisons and to include more methods, classical statistical and machine learning, in our tests.

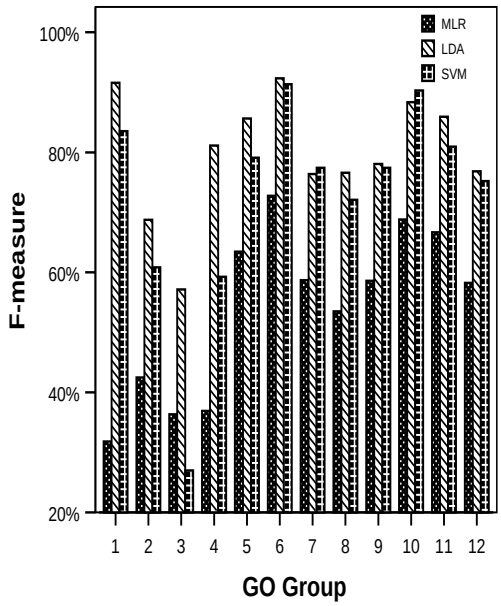

Figure 1. *F-measure for each GO group and classification method*

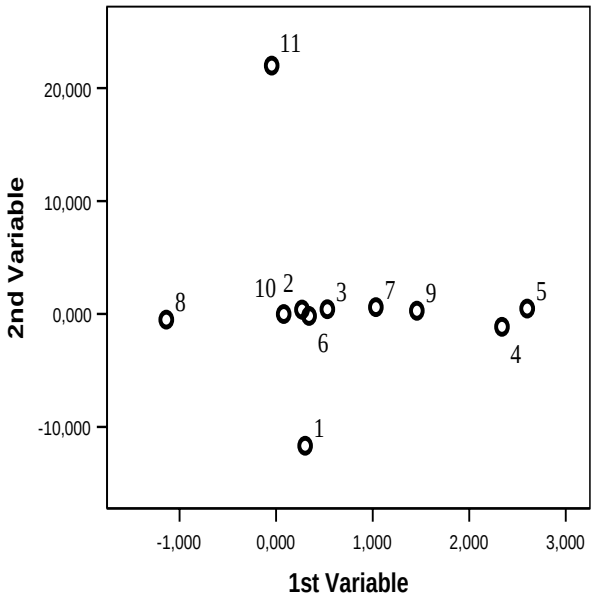

Figure 2. *The discriminative ability of the first 2 variables of LDA*

References

Andrade M.A., Valencia A., "Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system.", Proc Int Conf Intell Syst Mol Biol, 1997, pp.25–32.

Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., Sherlock G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nat Genet 2000;25(1):25–9.

Baldi P., Frasconi P., Smyth P., Modeling the Internet and the Web, John Wiley and Sons Ltd, 2003.

Camon E., Magrane M., Barrell D., Binns D., Fleischmann W., Kersey P., Mulder N., Oinn T., Maslen J., Cox A., Apweiler R., The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res. 13 (4) (2003) 662–672, http://dx.doi.org/10.1101/gr.461403.

Eisenhaber F., Bork P., "Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries.", Bioinformatics 15(7-8), 1999, pp. 528–35.

Hair J.F., Tatham R. L., Anderson R. E., Black W., Mutlivariate Data Analysis Prentice Hall, PTR., 5th edition, 1998.

Kazawa H., Izumitani T., Taira H., "Assigning gene ontology categories (go) to yeast genes using text-based supervised learning methods.", In Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004), 2004.

Kocsor A., Tóth L., Paczolay D., A Nonlinearized Discriminant Analysis and its Application to Speech Impediment Therapy, in: V. Matousek, P. Mautner, R. Moucek, K. Tauser (Eds.): Proceedings of Text, Speech and Dialogue: 4th International Conference, TSD 2001, LNAI 2166, pp. 249-257, Springer Verlag, 2001.

Manning C.D., Schutze H., Foundations of Statistical Natural Language Processing. The MIT Press; 1999.

PubMed. National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institute of Health (NIH). http://www.pubmed.com

Raychaudhuri S., Chang J.T., Sutphin P.D., Altman R.B., "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature.", Genome Res 12(1), Jan2002, pp. 203–14.

Rubin D.L., Thorn C.F., Klein T.E., Altman R.B., A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge. J Am Med Inform Assoc 2005 Mar–Apr;12(2): 121–129.

Spasic I., Ananiadou S., McNaught J., Kumar A., Text mining and ontologies in biomedicine: making sense of raw text., Brief Bioinform. 2005 Sep;6(3):239-51.

Tabachnick B. G., Fidell L. S., Using Multivariate Statistics, Allyn, Bacon, 2001, ISBN: 0-321-05677-9.

Theodosiou T., Angelis L., Vakali A., Thomopoulos G.N., Gene functional annotation by statistical analysis of biomedical articles. International Journal of Medical Informatics Accepted 26 April 2006, In Press

Venables W. N., Ripley B. D., Modern Applied Statistics with S. Fourth Edition. Springer, 4th edition, 2002.