Evaluating the effect of unbalanced data in biomedical document classification

Rosalía Laza[*], Reyes Pavón, Miguel Reboiro-Jato and Florentino Fdez-Riverola

ESEI, Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico,
Campus Universitario As Lagoas s/n, 32004, Ourense, Spain

Summary

Nowadays, document classification has become an interesting research field. Partly, this is due to the increasing availability of biomedical information in digital form which is necessary to catalogue and organize. In this context, machine learning techniques are usually applied to text classification by using a general inductive process that automatically builds a text classifier from a set of pre-classified documents. Related with this domain, imbalanced data is a well-known problem in many practical applications of knowledge discovery and its effects on the performance of standard classifiers are remarkable. In this paper, we investigate the application of a Bayesian Network (BN) model for the triage of documents, which are represented by the association of different MeSH terms. Our results show that BNs are adequate for describing conditional independencies between MeSH terms and that MeSH ontology is a valuable resource for representing Medline documents at different abstraction levels. Moreover, we perform an extensive experimental evaluation to investigate if the classification of Medline documents using a BN classifier poses additional challenges when dealing with class-imbalanced prediction. The evaluation involves two methods, under-sampling and cost-sensitive learning. We conclude that BN classifier is sensitive to both balancing strategies and existing techniques can improve its overall performance.

1 Introduction

The automated classification of texts into predefined categories has experimented an increasing interest given the fact that the number of on-line biomedical documents is constantly growing and it is necessary to organize them. With more than one thousand specialized biological databases in use today, the task of automatically identifying novel relevant data from such databases is increasingly important [1]. As a well-known example, the National Library of Medicine (NLM) uses Medical Subject Headings[1] (MeSH ontology) [2] to index articles from its Medline database [3].

MESH represents a taxonomic hierarchy of medical and biological terms suggested by the U.S. NLM since 1960. A new version is released every year, supplying a controlled vocabulary that represents biomedical concepts to be used for indexing publications included in the Medline database. All terms in MeSH are hierarchically organized with most general terms higher in the taxonomy than most specific ones.

In this context, one of the current challenges motivated by the explosive growth of biomedical literature is to help biologists in identifying relevant information from the huge amount of

[*] To whom correspondence should be addressed. Email: rlaza@uvigo.es

[1] http://www.nlm.nih.gov/mesh

doi:10.2390/biecoll-jib-2011-177

existing Medline documents. For this task, the dominant approach is based on the application of machine learning techniques, where a general inductive process automatically builds a classifier by learning the characteristics of the underlying predictive class. Related with this challenge, the Text Retrieval Conference (TREC) includes a Genomics track since 2003 [4]. One of the target tasks of the 2004 and 2005 Genomics track editions was a biomedical document triage task, which aimed to identify relevant articles from different areas of interest in an automated way [5].

Class-imbalanced data are usual in the field of text categorization, mainly characterized by a lot of irrelevant documents but very few articles belonging to the interesting category. In this context, BN models are commonly applied as standard classifiers given their accurate results and their ability for representing relationships among variables, but frequently ignoring the underlying class-imbalanced problem [6, 7]. With the goal of improving the accuracy of standard classification methods working in a class-imbalanced scenario, several strategies have been previously developed. These techniques mainly include: *sampling algorithms*, *cost-sensitive methods*, *recognition-based strategies* and *active learning approaches*.

Sampling strategies have been used to overcome the class imbalance problem by either eliminating some data from the majority class (*under-sampling*) or adding some artificially generated or duplicated data to the minority class (*over-sampling*) [8, 9, 10, 11, 12]. Cost sensitive learning uses a cost-matrix for different types of errors or instances in order to facilitate learning from imbalanced data sets (i.e., it uses different cost-matrices describing the penalty for misclassifying any particular data sample). This approximation has a similar effect to over-sampling the minority class and may end up with over specific rules or rules overfitting training [13, 14, 15, 16]. Recognition-based learning approaches learn rules from the minority class with or without using the examples of the majority class, guaranteeing that some rules are learned for the minority class [17, 18, 19]. Active learning techniques are conventionally used to solve problems related to unlabeled training data. Instead of searching the entire training data space, these methods can effectively select informative instances from a random set of training populations, therefore significantly reducing the computational cost when dealing with large imbalanced data sets [20, 21].

Based on our previous work dealing with text classification of biomedical literature [22], the aim of the current study is to investigate how class imbalance affects the accurate triage of Medline manuscripts represented by different combinations of MeSH terms that are classified using a BN. We devoted special attention to the representational capabilities of MeSH vocabulary and to the effectiveness of some strategies that were previously proposed to deal with class imbalance. To our knowledge, the joint effect of both aspects influencing BN classifier has not been thoroughly investigated. Specifically, in this paper we deal with the application of *random under-sampling* with different spreads between the minority and the majority class up to 1:1, where a full balance is reached. Additionally, and motivated by the different importance that classification errors have for the end-user (i.e. false negative errors may lead to ignore interesting papers while false positive errors only results in unnecessary reading) we used a *cost-sensitive* BN classifier.

The rest of the paper is structured as follows: in Section 2 we explain how we construct the proposed BN model taking into consideration the different levels existing in the MeSH ontology. Section 3 introduces the corpus used for the experimentation and presents the results obtained by the proposed BN model. Section 4 comments and discuss on the results obtained and finally,

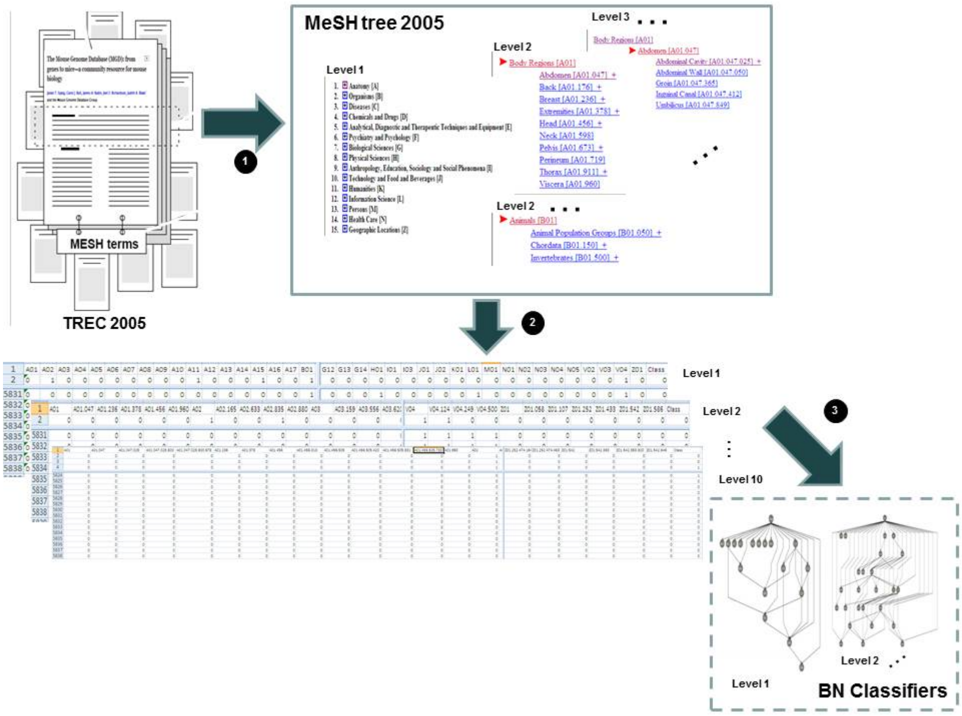

Figure 1: MeSH document representation for BN classification.

Section 5 concludes the work.

2 Representation and Classification of Documents using a BN Model

In the present study we will use a BN model as a probabilistic framework for the automatic classification of Medline documents represented by different vector descriptions constructed from their original set of MeSH terms. The main goal is to gain a deeper knowledge about how the hierarchical structure of MeSH thesaurus can influence the classification process. In our work, we consider that when a document is labeled with a specific MeSH term (positioned in one level of the eleven-level hierarchy) it is also related with all the ancestors of that term in the hierarchy. For example, a document initially indexed with the term A01.047.025.600.451 will be also represented by the terms A01, A01.047, A01.047.025 and A01.047.025.600. The purpose of this assumption is to extend the ontology-based document representation initially provided by Medline records. Moreover, for this study we are interested in representing documents only taking into consideration those MeSH terms belonging to a given level, so we define ten levels of representation for each document by applying our previously successful extension procedure [23]. Figure 1 exemplifies how this method is applied over a given document.

In order to represent each document D_i using our extension approach, a vector $< t_1^i, t_2^i, ...,$ $t_m^i, c^i >$ is defined in which t_j^i stands for the value of presence or absence of the MeSH term t_j in the document *i*, and c^i represents the value of the class variable C={*relevant*, *irrelevant*}. For this MeSH-based document representation we use binary vectors containing 22.996 elements (i.e., 22.995 MeSH descriptors plus the class attribute). After representing the documents belonging to the training and testing corpora (steps 1 and 2 in Figure 1), we can induce different BN models from the training data (one for each level) owning specific generalization capabilities (step 3 in Figure 1). The implementation of the BN classifier used in our experiments was provided by Weka[2] environment, and the learning strategy applied for inducing the Bayes network was the K2 algorithm [24] with a bayes score to judge the quality of the network structure. Conditional probability tables comprising the BN classifier were estimated directly from data. In order to reduce the high dimensionality of input matrices we used the *CfsSubsetEval* [25]feature selection method available in Weka with *GreedyStepwise* parameter for searching through the space of attribute subsets.

Once the BN models have been created, it is possible to carry out the classification of new instances. Thus, given an unseen document, D_{n+1}, we are able to compute the posterior probabilities of the class attribute (*relevant/irrelevant*). In order to perform these calculations, all the evidences (augmented MeSH terms belonging to the new document) need to be instantiated in the network and propagated through its internal structure. The category having the maximum value for the posterior probability will indicate the class of the document.

3 Identifying the Most Representative Mesh Level

In order to test how the different levels comprising MeSH thesaurus influence the accurate triage of biomedical documents, our evaluation framework uses a Medline triage task organized by the Genomics track of the TREC 2005, which is based on the simplified versions of the MGI (*Mouse Genome Informatics*) triage process. It consists on the triage subtask from the TREC 2004 Genomics track, which aims to identify articles for *Gene Ontology annotation (G)*, as well as three other major topics of interest to MGI: *Alleles of mutant types (A), Embryologic gene expression (E) and Tumor biology (T)*. For TREC 2004, full text articles published in 2002 and 2003 by three major journals (*Journal of Biological Chemistry, Journal of Cell and Proceedings of the National Academy of Science*) were obtained. Those articles containing the terms 'mouse', 'mice' or 'murine' were identified and separated into a training corpora (5.837 documents from 2002) and a test corpora (6.043 documents from 2003). The same data was used in the TREC 2005 triage task [26], for which Table 1 shows the number of relevant and irrelevant documents for the training and testing corpora.

Table 1: TREC 2005 Genomic track corpora description.

	A Alleles of mutant types		G Gene Ontology annotation		E Embryologic gene expression		T Tumor biology	
	Relevant	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
Train	338	5499	462	5375	81	5756	36	5801
Test	332	5711	518	5525	105	5938	20	66023

From Table 1 it can be observed that these corpora are heavily skewed. In category A only 5,79% of documents are *relevant*, while in category G only the 7,91%. This fact is augmented in categories E (1,38%) and T (0,61%).

In order to assess the accuracy of the classifier for each representation level belonging to each category, *F-score* was selected as the main evaluation criteria for combining *recall* and *precision* measures: $F - score = \frac{2*P*R}{P+R}$ where *P* stands for *precision* (i.e., $\frac{TP}{(TP+FP)}$) and *R* stands

[2] Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/ml/weka/

for *recall* (i.e., $\frac{TP}{(TP+FN)}$). TP is the number of relevant documents correctly labeled as relevant, FP represents the number of irrelevant documents incorrectly labeled as relevant and finally, FN stands for the number of relevant documents incorrectly labeled as irrelevant. Tables 2, 3, 4 and 5 show the accuracy of the BN classifier working at different MeSH ontology levels without taking into consideration the class imbalance problem.

Table 2: Results achieved by the BN classifier for each MeSH level in the category A.

	Precision		Recall		F-score	
Level	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
1	0,330	0,954	0,115	0,978	0,170	0,970
2	0,437	0,961	0,261	0,982	0,327	0,971
3	0,524	0,972	0,470	0,977	0,496	0,974
4	0,519	0,967	0,372	0,982	0,433	0,974
5	0,506	0,975	0,538	0,972	0,521	0,973
6	0,544	0,973	0,486	0,979	0,518	0,976
7	0,544	0,973	0,486	0,979	0,518	0,976
8	0,537	0,977	0,569	0,974	**0,553**	0,975
9	0,534	0,976	0,561	0,974	0,547	0,975
10	0,567	0,973	0,486	0,980	0,523	0,976

Table 3: Results achieved by the BN classifier for each MeSH level in the category G.

	Precision		Recall		F-score	
Level	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
1	0,000	0,919	0,000	1,000	0,000	0,958
2	0,429	0,921	0,030	0,996	0,056	0,957
3	0,321	0,921	0,023	0,996	0,042	0,957
4	0,346	0,921	0,023	0,996	0,042	0,957
5	0,370	0,922	0,043	0,994	0,076	0,956
6	0,278	0,923	0,055	0,987	0,092	0,954
7	0,250	0,922	0,045	0,988	0,076	0,954
8	0,309	0,922	0,053	0,990	**0,090**	0,955
9	0,309	0,922	0,053	0,990	0,090	0,955
10	0,309	0,922	0,053	0,990	0,090	0,955

As it can be globally seen from Table 2, the best MeSH level for representing documents in category A is level 8. Table 3 shows that better MeSH levels for representing documents in category G are 6 and 8. Table 4 indicates that levels 5 and 8 are the best whilst in Table 3 are 3 and 5 for categories E and T respectively.

Taking into consideration the differences related with balance/unbalanced data (see Table 1) and the results obtained in this preliminary study level 8 is selected for categories A and G and level 5 is selected for categories E and T. We will use this document representation for further experimentation.

Table 4: Results achieved by the BN classifier for each MeSH level in the category E.

	Precision		Recall		F-score	
Level	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
1	0,000	0,984	0,000	1,000	0,000	0,992
2	0,500	0,984	0,025	1,000	0,048	0,992
3	0,125	0,984	0,025	0,997	0,042	0,991
4	0,261	0,985	0,075	0,997	0,117	0,991
5	0,367	0,986	0,138	0,996	**0,200**	0,991
6	0,333	0,985	0,100	0,997	0,154	0,991
7	0,360	0,986	0,113	0,997	0,171	0,991
8	0,345	0,986	0,125	0,996	0,183	0,991
9	0,345	0,986	0,125	0,996	0,183	0,991
10	0,345	0,986	0,125	0,996	0,183	0,991

Table 5: Results achieved by the BN classifier for each MeSH level in the category T.

	Precision		Recall		F-score	
Level	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
1	0,000	0,997	0,000	0,999	0,000	0,998
2	0,167	0,998	0,231	0,997	0,194	0,997
3	0,333	0,998	0,308	0,998	0,320	0,998
4	0,400	0,998	0,308	0,999	0,348	0,998
5	0,400	0,998	0,308	0,999	**0,348**	0,998
6	0,400	0,998	0,308	0,999	0,348	0,998
7	0,400	0,998	0,308	0,999	0,348	0,998
8	0,400	0,998	0,308	0,999	0,348	0,998
9	0,400	0,998	0,308	0,999	0,348	0,998
10	0,400	0,998	0,308	0,999	0,348	0,998

4 Assessing the Impact of Class-Imbalance Data

As discussed previously, unbalanced data set represents a common problem in many real applications of knowledge discovery and its effects on the performance of standard classifiers are remarkable. Therefore, one objective of our current research has been to further advance the previous study and to investigate if the classification of Medline documents using MeSH controlled vocabulary poses additional challenges when dealing with class-imbalanced prediction (see Table 1 for the proportion of relevant/irrelevant examples from the original data sets).

In particular, this section presents an experimental study that applies existing strategies able to cope with the unbalanced data problem in order to show the impact of unbalanced data in the performance of a BN classifier. The selected strategies were *random under sampling*, that consists of randomly eliminating elements of the over-sized class until it matches the size of the other class, and *cost-sensitive learning*, that consists of modifying the relative cost associated to misclassifying the positive and negative class so that it compensates for the imbalance ratio of the two classes. Results obtained with both strategies will be compared against the performance obtained without balancing.

As a baseline reference, Tables 2, 3, 4 and 5 show the accuracy of the BN classifier working at different MeSH ontology levels without taking into consideration the class imbalance problem. From Table 2 and Table 3 we observe that the best value obtained by the BN classifier for the F-score corresponding to relevant documents is provided by level 8 of MeSH ontology. In the same way, from Table 4 and Table 5 we observe that the best value is provided by level 5 of MeSH ontology. Therefore, we have selected these levels of each category for our subsequent analysis. Moreover, with the aim of understanding the influence of the degree of imbalance on document classification, we show the results taking into consideration the degree of imbalance between existing classes. Consequently, Section 4.1 illustrates the impact of imbalance class for categories A and G where the percentage of relevant documents presented in the corpora are over 6%, whereas Section 4.2 shows the results of balancing at categories E and T where classes are heavily skewed.

4.1 Experimental Results with Categories A and G

The first technique we have tested has been *random under-sampling*. We have used the filter *SpreadSubsample* available in Weka, that allows us specifying the maximum spread between the minority and the majority class up to 1:1, where a full balance is reached. Tables 6 and 7 show the accuracy of the BN classifier under these conditions working with documents from category A and G respectively. These documents are represented by MeSH terms belonging to the 8th level of the ontology (best values are highlighted in Tables 2 and 3).

From Table 6 we can observe that recall values of relevant documents increase as the problem of imbalance is corrected, but unfortunately precision also diminishes. The highest F-score value for both classes is achieved without balancing the data set. However, the FP rate of the irrelevant class (those interesting documents classified as not relevant) decreases when a full balance of the data is forced.

Table 6: Results of BN classifier for category A when applying different spread for random under-sampling.

	Precision		Recall		F-score		FP rate	
Spread	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
12:1	**0,523**	0,977	0,573	**0,972**	**0,547**	**0,974**	**0,028**	0,427
6:1	0,411	0,981	0.664	0,949	0,508	0,965	0,051	0,336
3:1	0,373	0,988	0.787	0,929	0,506	0,958	0,071	0,213
1,5:1	0,327	0,991	0.854	0,906	0,473	0,947	0,094	0,146
1:1	0,296	**0,992**	**0,886**	0,889	0,441	0,938	0,111	**0,134**

Similarly to category A, from Table 7 we can observe that recall values of relevant documents at category G increase as the balance between classes augment. Conversely, as the imbalance decreases FP rate of the irrelevant class diminishes. In both cases, the best values are not achieved in a state of full balance, probably motivated by a corpus originally less unbalanced. However, contrary to category A, the highest F-score value for the relevant class is achieved with a ratio between the classes 1,5:1.

Motivated by the different importance that classification errors have for the end-user, the second experiment consisted in the application of a cost-sensitive BN classifier. In particular, we have used the *CostSensitiveClassifier* available in Weka.

Table 7: Results of BN classifier for category G when applying different spread for random undersampling.

	Precision		Recall		F-score		FP rate	
Spread	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
10:1	**0,268**	0,924	0,075	**0,982**	0,117	**0,952**	**0,018**	0,925
6:1	0,265	0,928	0,143	0,965	0,185	0,946	0,035	0,858
3:1	0,254	0,941	0,353	0,909	0,295	0,925	0,091	0,648
1,5:1	0,181	**0,983**	**0,868**	0,655	**0,299**	0,786	0,345	**0,133**
1:1	0,181	0,983	0,868	0,655	0,299	0,786	0,345	0,133

In this context, in which category A presents a 1:16 class-imbalanced ratio in favor of the negative class, we have assigned different cost factors (i.e., 2, 3, 5, 8 and 15) to misclassify a positive example (false negative) using a matrix cost. This proposal is based on the idea that the cost of miss-classifying a positive instance is correlated with the ratio between true and false classes. Table 8 summarizes the results obtained from the BN classifier taking into consideration this scenario. From Table 8 we can observe a situation very similar to the first experiment, confirming previous findings.

Table 8: Results of a cost-sensitive BN classifier using different cost values for category A.

	Precision		Recall		F-score		FP rate	
Cost	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
2	**0,492**	0,980	0,636	**0,965**	**0,555**	**0,972**	**0,035**	0,364
3	0,433	0,983	0,696	0,951	0,534	0,967	0,049	0,304
5	0,408	0,984	0,723	0,944	0,522	0,964	0,056	0,277
8	0,351	0,989	0,810	0,919	0,490	0,953	0,081	0,190
15	0,309	**0.992**	**0,862**	0,896	0,455	0,942	0,104	**0,138**

For category G, since the class-imbalanced ratio in favor of the negative class is 1:11 we have replaced a cost factor of 15 by a cost factor of 10. Table 9 summarizes the results obtained from the BN classifier taking into consideration these cost values. Again, from Table 9 we can observe a situation very similar to the random under-sampling experiment, confirming previous findings.

Table 9: Results of a cost-sensitive BN classifier using different cost values for category G.

	Precision		Recall		F-score		FP rate	
Cost	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
2	**0,28**	0,922	0,053	**0,988**	0,088	**0,954**	**0,012**	0,948
3	0,272	0,938	0,3	0,93	0,285	0,934	0,07	0,7
5	0,271	0,938	0,3	0,929	0,285	0,934	0,071	0,7
8	0,184	**0,979**	**0,835**	0,675	**0,302**	0,799	0,325	**0,165**
10	0,184	0,979	0,835	0,675	0,302	0,799	0,325	0,165

4.2 Experimental Results with Categories E and T

With respect to categories E and T, also the first experiment was the application of *random under-sampling* to documents represented by those MeSH terms belonging to the 5th level of the ontology. Tables 10 and 11 show the accuracy of the BN classifier under these conditions.

In this case, we have experimented with more different spreads between the minority and the majority classes due to the greater imbalance.

From Table 10 we can observe that recall values belonging to relevant documents suffer for a considerable increment as the problem of imbalance is corrected. This rate is much higher than the loss of precision. The best value of recall is obtained when a full balance is achieved. However, the highest F-score value for the relevant class is achieved nor with the original data or with a complete balance between the classes, but when we use a class-imbalance ratio of 6:1. This means that if the imbalance between classes is very large, a full balance with random under-sampling is not the best strategy.

Table 10: Results of BN classifier for category E when applying different spread for random undersampling.

Spread	Precision		Recall		F-score		FP rate	
	Relevant	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
60:1	**0,276**	0,985	0,1	**0,996**	0,147	**0,991**	**0,004**	0,9
48:1	0,25	0,986	0,138	0,993	0,177	0,99	0,007	0,863
24:1	0,203	0,987	0,188	0,988	0,195	0,987	0,012	0,813
12:1	0,204	0,987	0,238	0,985	0,22	0,986	0,015	0,763
6:1	0,15	0,992	0,55	0,949	**0,235**	0,97	0,051	0,45
3:1	0,083	0,995	0,691	0,892	0,148	0,941	0,108	0,309
1,5:1	0,078	0,994	0,675	0,869	0,14	0,927	0,131	0,325
1:1	0,03	**0,997**	**0,9**	0,517	0,057	0,681	0,483	**0,1**

Again, from Table 11 we can observe that the highest recall value belonging to relevant documents is achieved when classes are full balanced. However, unlike what happened so far, as the problem of imbalance is corrected the recall value not always increases. Similar fluctuations occur in the precision values of relevant documents. The reason of this behavior is due to the fact that the original corpus is heavily unbalanced and removing too many examples from the negative class may cause the classifier to miss important concepts pertaining to the majority class. As a consequence, the highest F-score value for both classes is not achieved with the original data set, but when we specify that there are at most a 24:1 difference in class frequencies. Finally, the FP rate of the irrelevant class decreases when a full balance of the data is forced.

Table 11: Results of BN classifier for category T when applying different spread for random undersampling.

Spread	Precision		Recall		F-score		FP rate	
	Relevant	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
96:1	**0,357**	0,998	0,385	**0,998**	0,37	0,998	**0,002**	0,615
48:1	0,211	0,988	0,308	0,997	0,25	0,998	0,003	0,692
24:1	0,296	0,999	0,615	0,996	**0,4**	**0,998**	0,004	0,385
12:1	0,189	0,999	0,769	0,991	0,303	0,995	0,009	0,231
6:1	0,113	0,999	0,615	0,987	0,19	0,993	0,013	0,385
3:1	0,019	0,999	0,692	0,907	0,038	0,951	0,093	0,308
1,5:1	0,016	0,999	0,769	0,876	0,031	0,934	0,124	0,231
1:1	0,018	**1,0**	**0,846**	0,877	0,035	0,934	0,123	**0,154**

As in the previous section, the second experiment carried out with categories E and T was the application of a cost-sensitive BN classifier. In this context, in which category E presents a 1:71 class-imbalanced ratio, we have used several cost matrices. Table 12 summarizes the results obtained from the BN classifier taking into consideration this scenario. We can observe a performance very similar to the one obtained with the under-sampling strategy.

Table 12: Results of a cost-sensitive BN classifier using different cost values for category E.

	Precision		Recall		F-score		FP rate	
Cost	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
2	**0,298**	**0,987**	0,175	**0,993**	0,22	**0,99**	**0,007**	0,825
3	0,264	0,987	0,175	0,992	0,211	0,989	0,008	0,825
5	0,218	0,987	0,238	0,986	0,228	0,987	0,014	0,763
8	0,217	0,99	0,375	0,978	0,275	0,984	0,022	0,625
15	0,175	0,993	0,563	0,957	**0,267**	0,974	0,043	0,438
30	0,152	0,992	0,563	0,948	0,239	0,97	0,052	0,438
60	0,083	0,994	**0,688**	0,875	0,148	0,931	0,125	**0,313**

Finally, Table 13 shows the results obtained by the cost sensitive BN classifier from documents of category T. Taking into consideration the severe class-imbalanced ratio of this corpus, we can observe that the lower costs assigned to misclassify a positive example (i.e cost 2 and 3) hardly affects the results. Moreover, contrary to the under-sampling strategy, no fluctuations in the recall values of relevant documents were observed. The rest of the measures show a similar behavior to the sampling strategy.

Table 13: Results of a cost-sensitive BN classifier using different cost values for category T.

	Precision		Recall		F-score		FP rate	
Cost	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*	*Relevant*	*Irrelevant*
2	0,286	0,998	0,308	0,998	0,296	0,998	0,002	0,692
3	**0,286**	0,998	0,308	**0,998**	0,296	**0,998**	**0,002**	0692
5	0,267	0,999	0.615	0,996	**0,372**	0,997	0,004	0,385
8	0,25	0,999	0.692	0,995	0,367	0,997	0,005	0,308
15	0,231	0,999	0.692	0,994	0,346	0,997	0,006	0,308
30	0,213	0,999	0.769	0,993	0,333	0,996	0,007	0,231
60	0,182	**1**	**0,923**	0,989	0,304	0,994	0,011	**0,077**

5 Conclusions

The purpose of this paper was to carry out an empirically study in order to assess both (i) the suitability of MeSH ontology for classifying Medline documents using a BN classifier and (ii) the impact of class imbalance distribution over the performance of this classifier.

Concerning the adequacy of MeSH ontology it was demonstrated that incrementing the number of MeSH terms used for representing Medline document can improve classification accuracy. For categories A and G best results are obtained with level 8. In categories E and T, which present severe imbalance distribution, best results are obtained from level 5.

Moreover, our experiments allowed us to conclude that BN classifiers are sensitive to imbalanced class distributions and that under-sampling as cost-sensitive learning strategies are effective to deal with this situation. Both balancing strategies provide similar improvements. More specifically, from the results obtained we can conclude that the number of relevant documents correctly identified (recall) increases with both strategies. However, incrementing the correct classification percentage of relevant documents also implies to slightly decrease the classification accuracy of irrelevant documents, but this situation is acceptable in the current domain.

However, taking into consideration the F-score measure, the effectiveness of cost-sensitive learning strategy is higher than the random under-sampling. We can observe that the F-score of relevant documents increases in the four categories when the cost-sensitive strategy is applied, whilst it decreases in categories A and T when under sampling method is selected.

The results obtained seem to contradict the idea that the class imbalance problem depends on the degree of class imbalance. So categories G and E, with different degree of imbalance, significantly improve the F-score of relevant documents compared with categories A and T. Therefore, we conclude that the degree of class imbalances does not seem to systematically cause performance degradation, and we are interesting in testing if other factors (e.g. the overlapping between classes) have more influence.

In addition, future work includes (i) the comparison of several methods for dealing with the class imbalance problem: under sampling and over sampling focused, learning from the positive class, tomek links, Condensed Nearest Neighbor (CNN) Rule, etc. and (ii) testing the selected strategies with other classifiers such as SVM, Naïve Bayes, IB3 and/or C4.5 in order to assess the sensibility of other classifiers to the imbalance problem. Moreover, we will use ROC (*Receiver Operating Characteristic*) curves to evaluate the accuracy of the different models. This will provide us with a deep understanding about the behavior of balancing and cleaning methods.

Acknowledgements

This work is supported in part by the project *MEDICAL-BENCH: Platform for the development and integration of knowledge-based data mining techniques and their application to the clinical domain* (TIN2009-14057-C03-02) from Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union.

References

[1] A.K. Sehgal, M.H. Saier, and C. Elkan. Identifying relevant data for a biological database: Handcrafted rules versus machine learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):851–857, 2011.

[2] S. J. Nelson, D. Johnston, and B. L. Humphreys. Relationships in medical subject headings. In C. A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 171–184. Kluwer Academic Publishers, 2001.

[3] Aurelie Neveol, Sonya E. Shooshan, Susanne M. Humphrey, James G. Mork, and Alan R. Aronson. A recent advance in the automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, 42(5):814–823, 2009.

[4] Patrick Ruch, Christine Chichester, Gilles Cohen, Frédéric Ehrler, Paul Fabry, Johan Marty, Henning Müller, and Antoine Geissbühler. Report on the TREC 2004 experiment: Genomics track. In *Proceedings of the Text Retrieval Conference*, 2004.

[5] William Hersh and Ellen Voorhees. TREC genomics special issue overview. *Inf. Retr.*, 12:1–15, 2009.

[6] Susan Dumais, John Platt, Mehran Sahami, and David Heckerman. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 1998 ACM Conference on Information and Knowledge Management*, pages 148–155. ACM Press, 1998.

[7] Wai Lam, Kon Fan Low, and Chao Yang Ho. Using a bayesian network induction approach for text categorization. In *Proceedings of 15th International Joint Conference on Artificial Intelligence*, pages 745–750, 1997.

[8] Willie Ng and Manoranjan Dash. An evaluation of progressive sampling for imbalanced data sets. In *Proceedings of the 6th IEEE International Conference on Data Mining*, pages 657–661, 2006.

[9] S.-J. Yen, Y.-S. Lee, C.-H. Lin, and J.-C. Ying. Investigating the effect of sampling methods for imbalanced data distributions. In *IEEE International Conference on Systems and Man and Cybernetics*, pages 4163–4168, 2006.

[10] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[11] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997.

[12] Chris Drummond and Robert C. Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the International Conference on Machine Learning*, 2003.

[13] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[14] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 155–164. ACM Press, 1999.

[15] Zhi hua Zhou and Xu ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18:63–77, 2006.

[16] Xu-Ying Liu and Zhi-Hua Zhou. The influence of class imbalance on cost-sensitive learning: An empirical study. In *Proceedings of the 6th IEEE International Conference on Data Mining*, pages 970–974, 2006.

[17] Nathalie Japkowicz, Catherine Myers, and Mark Gluck. A novelty detection approach to classification. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, pages 518–523, 1995.

[18] Miroslav Kubat, Robert Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 195–215, 1998.

[19] M. Molinara, M.T. Ricamato, and F. Tortorella. Facing imbalanced classes through aggregation of classifiers. *Image Analysis and Processing*, pages 43–48, 2007.

[20] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 127–136, 2007.

[21] Seyda Ertekin, Jian Huang, and C. Lee Giles. Active learning for class imbalance problem. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 823–824, 2007.

[22] Reyes Pavón, Rosalía Laza, Miguel Reboiro-Jato, and Florentino Fdez-Riverola. Assessing the impact of class-imbalanced data for classifying relevant/irrelevant medline documents. In *Proceedings of Advances in Intelligent and Soft Computing*, volume 93, pages 345–353, 2011.

[23] Rosalía Laza, Reyes Pavón, Miguel Reboiro-Jato, and Florentino Fdez-Riverola. Assessing the suitability of mesh ontology for classifying medline documents. In *Proceedings of Advances in Intelligent and Soft Computing*, volume 93, pages 337–344, 2011.

[24] G. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[25] M.A Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, 1998.

[26] Aynur A. Dayanik, Alexander Genkin, Paul B. Kantor, David D. Lewis, and David Madigan. DIMACS at the TREC 2005 genomics track. In *Proceedings of the Text Retrieval Conference*, 2005.