MEDLINE MeSH Indexing: Lessons Learned from Machine Learning and Future Directions

Antonio Jimeno-Yepes
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
antonio.jimeno@gmail.com

James G. Mork
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
mork@nlm.nih.gov

Bartłomiej Wilkowski
Technical University of Denmark
DTU Informatics
Richard Petersens Plads
B321, DK-2800, Kongens Lyngby, Denmark
wilkowskib@gmail.com

Dina Demner Fushman
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
ddemner@nlm.nih.gov

Alan R. Aronson
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
alan@nlm.nih.gov

ABSTRACT

Due to the large yearly growth of MEDLINE, MeSH indexing is becoming a more difficult task for a relatively small group of highly qualified indexing staff at the US National Library of Medicine (NLM). The Medical Text Indexer (MTI) is a support tool for assisting indexers; this tool relies on MetaMap and a k-NN approach called PubMed Related Citations (PRC). Our motivation is to improve the quality of MTI based on machine learning. Typical machine learning approaches fit this indexing task into text categorization. In this work, we have studied some Medical Subject Headings (MeSH) recommended by MTI and analyzed the issues when using standard machine learning algorithms. We show that in some cases machine learning can improve the annotations already recommended by MTI, that machine learning based on low variance methods achieves better performance and that each MeSH heading presents a different behavior. In addition, there are several factors which make this task difficult (e.g. limited access to the full-text of the citations) which provide direction for future work.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.3.1 [**Content Analysis and Indexing**]: Indexing methods

General Terms

Algorithms

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Keywords

Indexing methods, Text categorization, MeSH, MEDLINE

1. INTRODUCTION

MEDLINE® citations are indexed using the Medical Subject Headings (MeSH)® controlled vocabulary. This indexing is performed by a relatively small group of highly qualified indexing staff at the US National Library of Medicine (NLM). Their task is becoming more difficult due to the ever increasing size of MEDLINE, currently around 700k articles per year[1]. We hope that the situation can be eased through improvements to the recommendations made by NLM's indexing tool, the Medical Text Indexer (MTI) [2, 4].

MTI is a support tool for assisting indexers as they add MeSH indexing to MEDLINE. MTI has two main components: MetaMap [3] and the PubMed® Related Citations (PRC) algorithm. MetaMap performs an analysis of the citations and annotates them with Unified Medical Language System (UMLS)® concepts. Then, the mapping from UMLS to MeSH follows the *Restrict-to-MeSH* [6] approach which is based primarily on the semantic relationships among UMLS concepts. The PRC [11] algorithm is a modified k-NN algorithm which relies on document similarity to assign MeSH headings (MHs). This method attempts to increase the recall of MetaMap by proposing indexing candidates for MeSH headings which are not explicitly present in the citation but which are used in similar context.

Our motivation is to improve MTI's recommendations using machine learning because there is a large number of MeSH headings, around 26k, and previously indexed citations are available as training data. On the other hand, indexers have access to the full-text. Automatic indexing has no access to this due to license restrictions. We encounter issues, some of which are common to text categorization:

1. Imbalance between the number of positive and negative instances where the negative class usually overwhelms the positive one. Some machine learning algo-

[1]http://www.nlm.nih.gov/bsd/bsd_key.html

rithms have difficulty with this imbalance. We tested several approaches to deal with this issue to balance the datasets and to use a method based on the optimization of a multivariate measure instead of relying on accuracy. Joachims [10] proposed an adaptation of SVM to optimize measures like F-measure or the area under the ROC-curve instead of accuracy, being an alternative to balancing the positive and negative instances.

2. Even if a MeSH heading is correctly identified with a citation, it might not be significant enough to be included in the indexing.

3. Inconsistencies in the annotations might appear due to:

   (a) Inconsistency between MeSH indexers [7].

   (b) Changes in indexing policy over time can introduce inconsistencies with previously-indexed citations. This can even apply to routine changes to the structure of MeSH. In the selection of our set we carefully avoided this issue by selecting MHs which were already in MeSH during the current indexing period.

In this paper, we study the use of machine learning algorithms in the task of MeSH indexing for some MeSH headings and present several characteristics of the task. We show that the citation text has limited prediction capability and that other sources of information (e.g. fulltext) or representations of the citations other than unigrams and bigrams could still be explored. In the discussion, we point to future work and, based on statistics about MeSH indexing and MTI's performance.

2. RELATED WORK

Previous work has seen the indexing task as a text categorization task. The large body of related work provides valuable insights with respect to classification of MEDLINE citations and feature selection methods.

We find that most of the methods fit either into pattern matching methods which are based on a reference terminology (like UMLS or MeSH) and machine learning approaches which learn a model from examples of previously indexed citations.

Among the pattern matching methods we find the first component of MTI, as mentioned above, and an information retrieval approach by Ruch [13]. Ruch's system is a combination of information retrieval and boosting based on pattern matching. In his approach, the categories are the documents and the query is the text to be indexed. Pattern matching considers only the inner structure of the terms but not the terms with which they co-occur. This means that if an article is related to a MeSH heading but does not appear in the reference source (usually restricted to abstract text and title due to availability of full-text), it will not be suggested.

This problem has been approached in several ways from a machine learning point of view. Machine learning methods tend to be ineffective with many categories; i.e. turn the multi-class problem into a binary classification problem. Small scale studies with machine learning approaches already exist [1, 15]. But the presence of a large number of categories has forced machine learning approaches to be combined with information retrieval methods designed to reduce the size of the problem. For instance, PRC and a k-NN approach by Trieschnigg et al. [14] look for similar citations in MEDLINE and predict MeSH headings by a voting mechanism on the top-scoring citations. Experience with MTI shows that k-NN methods produce high recall but low precision indexing. Other machine learning algorithms have been evaluated which rely on a more complex representation of the citations which do not rely only on unigrams or bigrams, e.g., learning based on ILP (Inductive Logic Programming) [12].

3. MACHINE LEARNING ANALYSIS

Experiments have been performed on the MTI experiment set for the 2009 MeSH indexing. This set-up allows avoiding any interference provided by policy change in the indexing. We have selected candidate MHs highly represented in MEDLINE but with poor recall performance by MTI. The list of selected MHs is found in Table 1 along with their MeSH identifiers and tree code[2]. MTI performance for each MH is available in this table and as well in Table 4.

Considering that the total number of citations in the training set is 409279, we can see that the number of mentions (Positives) of these MeSH headings is very low. We find a very imbalance data set in which the negative examples exceed by far the number of positive ones. MTI identifies correctly a small amount of the positives (MTITP) but on the other hand, a large set of false positives is incorrectly predicted (MTIFP).

MeSH Heading	Unique ID	Potives	MTITP	MTIFP
Acute Disease	D000208	2739	526	1857
Gene Expression	D015870	3442	841	4225
Health Services	D006296	967	301	1963
Hormones	D006728	291	108	2094
Infection	D007239	437	182	3113
RTPCR	D020133	6953	3428	13711

Table 1: Selected MeSH headings based on 2010 MeSH and MTI performance on the training set

This selection has been previously used in [9]. In the current work a two stage approach to the problem is presented, in which the first step attempts to improve recall while the latter to increase precision. Compared to this previous work, we focus on a deeper analysis of the second step, in which a previously selected subset of documents is further analyzed according to the methods and the representation of the documents.

In the first step, the idea is to reduce the whole dataset to ease the work with machine learning algorithms. This implies identifying a set of classification rules with high recall, which might have a low precision performance. This reduction is performed by doing feature selection using Latent Dirichlet Allocation (LDA) [5] to extract the most salient terms in the groups and selecting the terms with a higher prediction performance based on the combination of decision trees (DT) common branches of the trees among cross-

[2] *RTPCR* stands for *Reverse Transcriptase Polymerase Chain Reaction* and *Health Services* stands for *Health Services Needs and Demands*

validation sets and decision trees. The DT derived rules (recall rules) reduce the total set of citations to be considered by the false positive filtering study, see Table 4. We can see that in almost all the cases we can reduce the size of the set, keeping recall high for each MeSH heading but still with low precision.

In Table 2, we show several terms which appeared in the LDA analysis for *Gene Expression*. We find that terms like *expression* have high coverage but low precision, since there are terms which can be used in different situations. On the other hand, we find the term *gene expression* which has lower recall, but surprisingly the precision is still very low. This means that there are cases in which the term *gene expression* appears in the citation but does not qualify to be included as a candidate MH. Machine learning will not only have to ensure that the term is used in the proper sense but that it is significant enough to qualify, showing further the complexity of the task.

Term	Rec	Prec	F1
gene expression	0.2543	0.1668	0.2014
mrna	0.2965	0.1243	0.1752
expression	0.7704	0.0933	0.1664
gene	0.5492	0.0725	0.1281
expressed	0.3033	0.0771	0.1230

Table 2: Gene expression feature prediction study

Some of the MeSH headings in our study are parents of more specific headings in the MeSH taxonomy (e.g. *Hormones*). These more specific headings (e.g. *thyroid hormones*) might be used for indexing instead of the MHs we are considering. To evaluate the impact of this phenomenon we have identified the children of the MHs under study. This includes the immediate and all their descendants.

In Table 3 we show that some MHs like *Hormones* and *Infection* have a large number of children and seem to overlap with the indexing performed for these MHs (FP+Children). In the case of *Hormones*, half of the false positives (FP) are indexed with a hormone type. Methods based on pattern matching might avoid this issue selecting the MH matching the largest span of text. Examples of these methods are MetaMap and Ruch's approach.

MeSH Heading	Children	FP+Children	Total FP
Gene Expression	3	984	24978
Health Services	2	76	27475
Hormones	212	2290	4181
Infection	148	3408	49796

Table 3: Overlap of FPs and annotation of more specific MeSH Headings

In the second step, to the reduced set produced by the recall rules, we have applied the following machine learning algorithms. Each algorithm relies on different learning bias which would allow closer examination of the results for each one of the cases.

1. Traditional classifiers (SVM, Naïve Bayes, decision trees, k-NN and AdaBoost).
2. Multivariate SVM [10], the training is done to optimize $F1$-measure.
3. We have performed class noise removal based on the algorithms by Zhu et al.[16].

False positive filtering experiments (Filtering) have been performed for each one of the learning algorithms listed above. Unigrams and bigrams are used in the representation of the documents. Results are presented in Table 4, considering Filtering results, only the results for the best performing method are shown. We show the MTI results, MTI with machine learning filtering (MTI+Filtering), the outcome of the recall rules and the recall rules with machine learning filtering (RecRul+Filtering). The data sets for (MTI+Filtering) are derived from the MTI results while the data sets for (RecRul+Filtering) are derived from the recall rules presented above. Further experiments are performed balancing the data sets.

Considering the MTI+Filtering results, as observed already in [9], is that machine learning improves the precision of the MeSH heading recommendation but at the cost of recall. AdaBoost performs better for *Acute Disease* and *Gene Expression*. Class noise reduction improves *Health Services* and *Infection*, while the method used from this reduced set are decision tree and Naïve Bayes respectively. Multivariate SVM is the preferred method for *Hormones* and *RTPCR*. In all the methods but multivariate SVM, balancing the positive and negative examples increases the performance of the classifiers.

Considering the RecRul+Filtering, AdaBoost is the best performing method. As in the previous set, balancing the positive and negative examples improves the performance of the classifiers. Only in the case of *Acute Disease*, the best performing method is multivariate SVM.

We also show results of the children analysis in Table 4 for *Hormones* and *Infection*. We can see that children analysis improves the performance of the recommendations, meaning that the MeSH structure should be further studied in order to improve the recommendations. In both cases, AdaBoost is the best performing method.

From the machine learning algorithms used in the experiments, AdaBoost and multivariate SVM achieve the best performance in many of the filtering results, meaning that low variance methods achieve a better performance. On the other hand, decision trees achieve the lowest performance which correlates with previous studies on text categorization.

4. DISCUSSION

In our study, we have used a data set from 2009 MTI experiments, and we have analyzed some of the characteristics of the results obtained by applying machine learning on them. We have presented the issues which machine learning algorithms face when dealing with MeSH indexing.

As we have noted above, each MH seems to have a different behavior according to the method used. Since there are 26k MHs, to train and maintain up-to-date a system which can manage the different MHs, it might be possible to place the effort on highly represented MHs. Systems based on k-NN [11, 14] or matching strategies like MetaMap and Ruch's approach [13] manage the size problem efficiently. In this section, we present different statistics on the MeSH indexing which could help deciding on focusing the effort on a specific set of MHs.

Table 5 shows the micro/macro-average performance of MTI evaluated for all 26k MHs. We can see that while recall is almost the same, precision is much lower for micro-average. This might mean that there are MHs which are

Acute Disease	Prec	Rec	F1	F2
MTI	0.2664	0.1580	0.1984	0.1720
MTI+Filtering	**0.4272**	0.1395	0.2103	0.1612
Recall rules	0.1176	**0.8562**	0.2068	0.3795
RecRul+Filtering	0.1941	0.6611	**0.3001**	**0.4463**
Gene Expression	Prec	Rec	F1	F2
MTI	0.1958	0.2712	**0.2274**	0.2518
MTI+Filtering	**0.2642**	0.1389	0.1896	0.1805
Recall rules	0.0645	**0.8165**	0.1195	0.2450
RecRul+Filtering	0.1130	0.5220	0.1858	**0.3029**
Health Services	Prec	Rec	F1	F2
MTI	0.1810	0.3533	0.2394	**0.2968**
MTI+Filtering	**0.2636**	0.2387	**0.2505**	0.2433
Recall rules	0.0169	**0.6293**	0.0329	0.0763
RecRul+Filtering	0.0723	0.3547	0.1201	0.1992
Hormones	Prec	Rec	F1	F2
MTI	0.0726	0.4000	0.1229	0.2103
MTI+Filtering	**0.1310**	0.2800	**0.1785**	**0.2281**
Recall rules	0.0328	**0.6311**	0.0624	0.1359
RecRul+Filtering	0.0839	0.3600	0.1361	0.2172
Recall no children	0.0698	0.6311	0.1258	0.2421
Recall nc filter	0.1845	0.3911	0.2507	0.3195
Infection	Prec	Rec	F1	F2
MTI	0.0649	0.4013	0.1117	0.1970
MTI+Filtering	**0.1568**	0.2492	**0.1925**	**0.2229**
Recall rules	0.0048	**0.7767**	0.0095	0.0234
RecRul+Filtering	0.0216	0.4660	0.0412	0.0910
Recall no children	0.0051	0.7767	0.0102	0.0251
Recall nc filter	0.0276	0.4854	0.0523	0.1126
RTPCR	Prec	Rec	F1	F2
MTI	0.2790	0.3738	0.3213	0.3535
MTI+Filtering	**0.5316**	0.3038	**0.3879**	0.3188
Recall rules	0.0931	**0.7191**	0.1648	0.3066
RecRul+Filtering	0.2048	0.4863	0.2883	**0.3815**

Table 4: Results of different methods on selected MeSH headings

highly represented in MeSH indexing (e.g. Female) for which MTI achieves a result with low precision.

	Precision	Recall	F-measure
Macro-average	0.4164	0.5111	0.4589
Micro-average	0.3268	0.5118	0.3989

Table 5: MTI macro and micro averaging based on ln $frequency$

Table 6 shows the distribution of MHs according to their occurrence frequency in MEDLINE. In order to properly distribute the MHs, we have placed them into bins according to the logarithm of the frequency. MHs indicate the number of individual MHs, the total is the actual total mention of MHs, and precision, recall and F-measure is the average performance in each one of these categories. MTI's performance seems to decrease slightly as the total number of citations indexed by the MHs increases. The exception is the last category with only the single MH *Humans*. We can see that the last five categories have a low number of MHs but the total number of occurrences in MEDLINE is quite high. The most popular terms in our dataset are *Humans* with 471,467 occurrences, *Female* with 233,499 and *Male* with 227,052.

There are MHs with very low number of mentions in MEDLINE. We can assume that these MHs are rare, but even if you find the term it does not mean that it is significant enough to be added to the indexing.

We find as well that there are 1,314 MHs which are never considered for indexing[3]. Some MHs are used to specify the *Publication Characteristics* (Tree V), which in some cases allow the identification of funding support for the article[4]. Other MHs are used to organize the MeSH taxonomy.

ln($freq$)	MHs	Total	Prec	Rec	F1
0	833	833	0.2878	0.4898	0.3626
1	1933	5704	0.4448	0.5108	0.4755
2	3375	27296	0.4910	0.5363	0.5126
3	4393	94692	0.4834	0.5430	0.5115
4	4795	273297	0.4671	0.5456	0.5033
5	4313	650906	0.4230	0.5399	0.4743
6	2698	1091380	0.3860	0.5454	0.4520
7	1319	1392237	0.3500	0.5602	0.4309
8	465	1303683	0.3321	0.5574	0.4162
9	115	898067	0.3263	0.5208	0.4012
10	22	429109	0.4074	0.4413	0.4237
11	7	369217	0.4735	0.3472	0.4007
12	5	874276	0.5817	0.2964	0.3927
13	1	471467	0.9155	0.6914	0.7878

Table 6: MTI macro averaging based on ln $frequency$

Table 7 shows the macro average performance of MTI according to each one of the MeSH trees. A detailed list of the current tree codes is available from[5]. We can see that there are trees which contain a low number of MeSH headings but embody a large number of indexed citations like CT (Check Tags), G (Analytical, Diagnostic and Therapeutic Techniques and Equipment) and E (Phenomena and Processes).

One possible next step would consist of focusing on these sets of MeSH headings and try, in addition, to identify commonalities among the MHs.

Tree	MHs	Total	Prec	Rec	F1
A	1614	480326	0.3723	0.5404	0.4641
B	3546	248804	0.5459	0.6465	0.5989
C	4394	757400	0.4600	0.5682	0.5107
CT	34	1804516	0.4393	0.3007	0.3041
D	8805	1287185	0.4323	0.5327	0.4740
E	2396	1412951	0.3515	0.4146	0.3590
F	739	281784	0.3208	0.3944	0.3352
G	1360	822398	0.2970	0.4253	0.3491
H	292	91239	0.2796	0.3122	0.2656
I	410	133224	0.3068	0.3389	0.2977
J	193	46574	0.3123	0.4165	0.3557
K	145	13341	0.3135	0.2505	0.2404
L	246	86219	0.2171	0.2519	0.2066
M	154	48106	0.3371	0.3564	0.3106
N	737	233104	0.2528	0.2536	0.2194
V	146	0	0.0000	0.0000	0.0000
Z	377	134993	0.5006	0.5391	0.5125

Table 7: MTI macro averaging based on MeSH Tree code

5. CONCLUSION

Experiments show that machine learning can be used to improve the results of MTI, but the results are still low for production purposes. From the results, we can see that low variance machine learning methods provide better results.

[3]From the MEDLINE Baseline http://mbr.nlm.nih.gov/index.shtml
[4]http://www.nlm.nih.gov/bsd/funding_support.html
[5]http://www.nlm.nih.gov/mesh/trees.html

This implies that noise resilient methods are preferred, even though it is still difficult to know how much noise is derived from attribute noise or class noise. Further work might be devoted to understand both noise types and devise approaches to deal with them. In addition, indexing methods exhibit different behavior depending on the MH. This might be taken into account when training a system for all the MHs in MeSH.

We have presented results on a limited number of examples. Extending the work to more MeSH headings would provide better insights in the comparison of machine learning approaches. For instance, improvement the results for the Check Tag MHs, given the low number of them and the large number of citations, would provide a boosting in the performance of the MTI. Very frequent MHs like *Humans*, *Male* and *Female* belong as well to this category of MHs.

Balancing the number of positives and negatives by removing instances from the negatives (subsampling) has improved the performance of many classifiers. On the other hand, subsampling has removed negative instances with features that should be considered. We plan to consider other sampling approaches including synthetic sampling.

We have performed experiments on the text provided by the abstract and title of the citations. The results point out that the citations might not provide enough information to index the citations, e.g. for around 15% of the citations only the title is present. Only the title is not enough to decide on the MeSH headings to be used to index the documents. In addition, we have seen in Table 2 that there is a limited number of terms related to *Gene Expression* with high F-measure performance. The analysis performed in this paper indicate that AdaBoost performs reasonably well. AdaBoost in this study uses a decision tree as base learner, this means that capturing relations between features will increase performance. A larger set of features, available in full text, might increase the performance of the classifiers. Further studies on full-text might be required, but only 15% of the PMIDs in our dataset could be matched to full-text identifiers in PubMed Central®.

Another possibility to extend the feature set is to consider existing meta-data already available in the citations. One way of doing this might be correlating the MeSH headings with the journals in which the citations appears. This might be approached using the Journal Descriptor indexing which has already been proposed in the literature [8].

Acknowledgment

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. The third author also gratefully acknowledges funding from the Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (Cimbi.org), Otto Mønsteds Fond, Kaj og Hermilla Ostenfelds Fond, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Fond.

6. REFERENCES

[1] Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C.F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association*, 12(2):207–216, 2005.

[2] A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, JG Mork, SJ Nelson, TC Rindflesch, and WJ Wilbur. The NLM Indexing Initiative. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2000.

[3] A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010.

[4] A.R. Aronson, J.G. Mork, C.W. Gay, S.M. Humphrey, and W.J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. In *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*, page 268. OCSL Press, 2004.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] K.W. Fung and O. Bodenreider. Utilizing the UMLS for semantic mapping between terminologies. American Medical Informatics Association, 2005.

[7] M.E. Funk and C.A. Reid. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176, 1983.

[8] S.M. Humphrey. Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science*, 50(8):661–674, 1999.

[9] A. Jimeno-Yepes, B. Wilkowski, J.G. Mork, E. Van Lenten, D. Demner Fushman, and A.R. Aronson. A bottom-up approach to MEDLINE indexing recommendations. In *AMIA Symposium*. American Medical Informatics Association, 2011.

[10] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.

[11] J. Lin and W.J. Wilbur. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423, 2007.

[12] A. Névéol, S. Shooshan, and V. Claveau. Automatic inference of indexing rules for MEDLINE. *BMC bioinformatics*, 9(Suppl 11):S11, 2008.

[13] P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658, 2006.

[14] D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412, 2009.

[15] M. Yetisgen-Yildiz and W. Pratt. The effect of feature representation on MEDLINE document classification. In *AMIA Annual Symposium Proceedings*, volume 2005, page 849. American Medical Informatics Association, 2005.

[16] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proceedings of the 20th International Conference on Machine Learning*, 2010.