MIS Quarterly

RESEARCH ARTICLE

EXPLAINING DATA-DRIVEN DOCUMENT CLASSIFICATIONS

David Martens
Department of Engineering Management, Faculty of Applied Economics, University of Antwerp, Prinsstraat 13,
2018 Antwerp, BELGIUM {david.martens@uantwerpen.be}

Foster Provost
Department of Information, Operations and Management Sciences, Stern School of Business, New York University,
44 West 4th Street, New York, NY 10012-1126 U.S.A. {fprovost@stern.nyu.edu}

Appendix A

News Item Categorization

Twenty Newsgroups Data Set

To demonstrate generality and to illustrate some additional properties of the method, we also apply the explanation method to a second domain: classifying news stories. The 20 newsgroups data set is a benchmark data set used in document classification research. It contains about 20,000 news items from 20 newsgroups representing different topics, and has a vocabulary of 26,214 different words (after stemming) (Lang 1995). The 20 topics can be categorized into seven top-level usenet categories with related news items: alternative (alt), computers (comp), miscellaneous (misc), recreation (rec), science (sci), society (soc), and talk (talk). One typical problem studied with this data set is to build classifiers to identify stories from these seven high-level news categories, which for our purposes gives a wide variety of different topics across which to provide document classification explanations. Looking at the seven high-level categories also provides realistic richness to the task: in many real document classification tasks, the class of interest is actually a collection (disjunction) of related concepts (consider, for example, "hate speech" in the safe-advertising domain).

We build a classifier system to distinguish the seven top-level categories using all words in the vocabulary. This permits us to examine a wide variety of explanations of different combinations of true class and predicted class, in a complicated domain, but one where we have at least a high-level intuitive understanding of the classes. The examination shows that even for news items grouped within the same top-level category, the explanations for their classifications can vary greatly and are intuitively related to their true lower-level newsgroup.

Results

The classifier system for distinguishing the seven top-level newsgroups (alt, comp, misc, rec, sci, soc, talk) operates in a one-versus-others setup (i.e., seven classifiers are built, each distinguishing one newsgroup from the rest). For training (on 60% of the data) and for prediction (remaining 40% as test data), if a news item is (predicted to be) from the given newsgroup, the class variable is set to one; if not, the class variable is set to zero. To demonstrate the method with different types of model, here we build both linear and nonlinear SVM classifiers.

In Table A1, each cell shows at least one explanation (where possible) of an example from one of the 20 low-level categories (specified in the row header) being classified into one of the top-level categories (specified in the column header). If no explanation is given in a cell, either no misclassified instances exist, which occurs most frequently, or no explanation was found with a maximum 10 words. The shaded cells on

Table A1. Explanations for Twenty Newsgroups Dataset (showing why for any cell, documents from the newsgroup at the beginning of the row are classified as the newsgroup at the top of the column)

	Classification models in one-versus-others setup: "newsgroup" versus "not newsgroup." Explanations why news items are classified as "newsgroup."			
	alt vs. not alt	**comp vs. not comp**	**misc vs. not misc**	**rec vs. not rec**
alt.atheism	ico bibl moral god believ	unm	wustl distribut	com
	ico bibl moral god read	carina screen	wustl 5	univers
	ico bibl moral accept god	carina join	wustl origin	distribut
comp.graphics	umd	quicktim 3do centris resolut card program	bigwpi wpi distribut	nb canada ca
	wam	quicktim 3do centris resolut ac card	bigwpi wpi pleas	nb luck canada
	mistak cant	quicktim 3do centris resolut fax card	bigwpi wpi email	nb archiv canada
comp.os. ms-windows.misc		mous microsoft cant	distribut	6
		mous microsoft solution	look	tom
		mous microsoft switch	pleas	archiv com
comp.sys.ibm.pc. hardware		hardwar thank	distribut	cornel buffalo
		hardwar appreci	repli	buffalo cc wonder
		adam hardwar	call	ubvmsb buffalo cc
comp.sys.mac. hardware	kmr4po read	vga monitor mac advenc card am	offer sale distribut	univers
	kmr4po follow	vga monitor mac advenc card repli	offer sale card	recent
	kmr4po note	vga monitor mac advenc card thank	jame offer sale	price
comp.windows.x		enterpoop lcs fax	pleas	street final list
		enterpoop lcs mit	includ	2154 street final com
		enterpoop xpertexpo lcs inc	send	2154 street final pleas
misc.forsale		driver program	sale	insur
		driver card	2190	gasket massachusett ser
		pc driver	pc mention	gasket jacket massachusett
rec.autos		window call	distribut	geico insur distribut
		window email	3	geico insur ca
		window 4	compani	geico insur usa
rec.motorcycles		greyscal color	mile	dod
		greyscal pictur	pad	ottawa ca
		greyscal directori	rosevil deal	ottawa canada
rec.sport.baseball			offer	miller brave gatech nl seri team technologi game
			game 3	miller brave gatech nl seri team institut game
			game 5	miller brave gatech nl seri team plai game
rec.sport.hockey		michel comput	susan	buffalo ny team
		michel 4	game call	bruin buffalo team
		co michel	buffalo game	sabr buffalo team
sci.crypt	mathew	42 print messag	ohio	usa
	rusnew mantis umd consult couldnt agre	42 print seen	cincinnati	list
	rusnew mantis umd consult couldnt stop	42 print net	victor	free

Table A1. Explanations for Twenty Newsgroups Dataset (Continued)

	Classification models in one-versus-others setup: "newsgroup" versus "not newsgroup." Explanations why news items are classified as "newsgroup."			
	alt vs. not alt	**comp vs. not comp**	**misc vs. not misc**	**rec vs. not rec**
sci.electronics		softwar	sell price email pleas	univers
		prefer	sell price game email	distribut
		appl	ncsu sell price email	ca
sci.med	atheist	lcs mit address thank	nyx	canada cc bad pleas univers
	god believ	lcs laboratori mit address	denver du	canada cc bad pleas thank
	god start	lcs mit address email am	denver dept distribut	canada cc bad i'v pleas
sci.space		michel help	internet	riversid due
		site help	servic	riversid ucr
		help thank am	institut	riversid prbaccess com
soc.religion. christian	atheist	wrote	call	chanc
		technologi	person	dave
		9	includ	princeton
talk.politics.guns		richard drive	holonet norton internet	sfasu
		richard fax	holonet norton modem	arlen thank
		bryan richard	holonet norton pete	arlen pleas
talk.politics.mid-east	wrote	ai repli	hous	cc
	evid	ai mit	amherst	columbia
	religion	ai cant 3	pl7	lion
talk.politics.misc	religi god	cwru	ohio	car
	religi religion	jone	jone	watch
	islam religi	cleveland western	hela ins cleveland reserv western usa 2	jm
talk.religion.misc	bill	site	institut	refer
	explain	ca system	gold	mike
	cration	usa system	polytechn	univ

	Classification models in one-versus-others setup: "newsgroup" versus "not newsgroup." Explanations why news items are classified as "newsgroup."		
	sci vs. not sci	**soc vs. not soc**	**talk vs. not talk**
alt.atheism	latech	translat	ha atom 2000 moral object evid
	scisur	familiar	ha overwhelm atom 2000 moral object
	rayengr help	translat god	microscop ha atom 2000 moral object
comp.graphics	map	scott pleas	david
	pub inc	scott read	happen
	pub ftp	scott answer	list
comp.os.ms-windows.misc	public	book	speak
	date	pa	limit
	std	steven	stand
comp.sys.ibm.pc. hardware	nz mark		address
	nz 1.1		student
	nz network		utexa
comp.sys.mac. hardware	bounc suppli		purdu
	bounc circuit		cc center
	sync bounc happen		pure cc

Table A1. Explanations for Twenty Newsgroups Dataset (Continued)

	Classification models in one-versus-others setup: "newsgroup" versus "not newsgroup." Explanations why news items are classified as "newsgroup."		
	sci vs. not sci	**soc vs. not soc**	**talk vs. not talk**
	nz	scienc	re
comp.windows.x	aukuni time	sorc	time
	aukuni scienc	upenn	name
	tube	pa	usa
misc.forsale	catalog	sex accept	21
	umb etc	sex hell	gun
	max low fone	chuck	utexa call
rec.autos	max cycl fone	discuss pleas	utexa center
	max pl9 effect fone	discuss read	utexa care
	ibm		righteous racist stupid mean
rec.motorcycles	week fone		righteous racist stupid own
	rochest fone 10		righteous racist stupid opinion
	list 10	dt	buffalo love cc
rec.sport.baseball	list scienc	nswc	buffalo stand cc
	std list	carderock	buffalo stori cc
	ericsson inc	oppos	john
rec.sport.hockey	ericsson commun	csd	boulder center
	ericsson user	chuck	boulder depart
	inform		congress law john
sci.crypt	commun		preced congress john
	offic		nagl congress john
	adcom	god	re
sci.electronics	preamp chip sound	accept	david
	preamp network chip	recent	citi
	handed rsilverworld sight domin eye commun	sex	perot
sci.med	handed rsilverworld sight domin eye indic	grade fysic	16 happen
	handed rsilverworld sight domin guest eye look	fysic speak reason	edward happen
	space	book	terror moral govern
sci.space	nasa follow	discuss	terror moral law
	nasa scienc	fysic	terror moral major
	greet marie angel	religion pleas	homosexu
soc.religion. christian	gabriel greet mari 12	religion question	abus behavior love
	gabriel greet mari various	religion follow	abus sexual love peopl
	chip	marri christ life	batf waco clinton question
talk.politics.guns	explode	marri christ view	batf waco clinton law
	medic understand	marri christ religion	batf waco clinton evid
	ai	ab4zvirginia beyer	holocaust arab militari plan evid kill
talk.politics. mideast	amend lab	ab4zvirginia beyer andi	holocaust arab militari attack evid kill
	amend messag 10	blanket ab4zvirginia beyer andi	holocaust arab militari reach evid kill
	acid scienc	serbian	homosexu moral law
talk.politics.misc	acid commun	bomb york 2	homosexu moral stop
	acid sorc	bomb york position	homosexu moral pass
	messag	pa christian	malcolm weapon jew christian
talk.religion.misc	institut	mormon faith christian 2	malcolm weapon jew kill
	apr	mormon faith hous christian	malcolm weapon jew hous

the diagonal are the explanations for correct classifications; the rest are explanations for errors. For example, the first explanation in the upper-left cell (excluding the header rows) shows that this correct classification of a news story in the alt.atheism category is explained by the inclusion of the terms *ico*, *bibl*, *moral*, *god*, and *believ*: if these words alone are removed, the classifier would no longer place this story correctly into the alt category.

Several cells below, we see explanations for why a sci.med story was misclassified as belonging to alt: because of the occurrence of the word *atheist* (first explanation), or the words *god* and *believe* (second explanation). Further investigation of this news story reveals it concerns organ donation. In general, the explanations shown in Table A1—the correctly classified test instances (grayed cells on the diagonal)—usually are indeed intuitively related to the topic.

The categories themselves often occur as words in the explanations, such as *hardwar*, *microsoft*, *mac*, and *space*. Importantly, the different subcategories of the newsgroups show different explanations, which motivates using instance- rather than global-level explanations. For example, for the computer newsgroup (shown in the second column), the terms used to explain classifications from the different subgroups are quite different and intuitively related to the specific subgroups.

The misclassified explanations (outside of the shaded cells) often show the ambiguity of certain words as reason for the misclassification. For example *window* is a word that can be related to computers, but also can be related to automobiles. The explanations for the misc.forsale news items indicate they are most often misclassified because the item that is being sold comes from or is related to the category in which it is misclassified. With this individual-instance approach, similar ambiguities as well as intuitive explanations for each of the subgroups also can be found for the other categories. The results also demonstrate how the explanations can hone in on possible overfitting, such as with "unm" and "umd" in the cells adjacent to the upper-left cell we discussed above.

The test accuracy (in terms of percentage correctly classified instances, PCC) and explainability metrics when allowing a maximum of 10 words in an explanation are shown in Table A2, for the positive classifications. Although most of the test instances are explained (PE around 90–95% for all models) some instances still remain unexplained. If we allow up to 30 words in an explanation, all instances are explained for each of the models. Of particular note is that for this widely used benchmark with a vocabulary of 26,214 words, on average only a small fraction of a second (ADF of 0.02–0.08 second for the linear models) is needed to find a first explanation. As previously mentioned, this is because our SEDC explanation algorithm is independent of the vocabulary size. Explaining the nonlinear model requires more time, since backtracking occurs and the model evaluation takes longer than for a linear model. Nevertheless, on average still less than a second is needed to find an explanation.

These results in a second domain, with a wide range of document topics, provide support that our type of instance-level document classification is capable of providing better understanding of the functioning of text classifiers, and that the SEDC algorithm is generally effective and fast as well. Further, this second study provides an additional demonstration of the futility of global explanations in domains such as this. Specifically, there are very many different reasons for different classifications; at best they would be muddled in any global explanation, and likely they would simply be incomprehensible.

Table A2. Explanation Performance on the Test Set of the 20 Newsgroups Data Set for a Linear (left) and Nonlinear (right) SVM Model, Limiting Explanations to 10 Words (Maximum)

	Linear SVM							Nonlinear RBF SVM						
Model	**PCC**	**PE**	**AWS**	**ANS**	**ANT**	**ADF**	**ADA**	**PCC**	**PE**	**AWS**	**ANS**	**ANT**	**ADF**	**ADA**
alt	81.5%	96.1%	2.7	6.1	18.5	0.05	0.16	76.8%	95.7%	2.5	7.2	30.1	0.62	1.35
comp	93.7%	89.1%	3.1	6.1	13.3	0.05	0.12	94.9%	81.7%	3.3	5.4	12.4	0.54	0.88
misc	92.8%	98.1%	1.9	4.9	12.9	0.02	0.12	90.5%	96.6%	1.8	6.0	17.0	0.14	0.38
rec	94.2%	94.8%	2.4	5.7	13.7	0.04	0.11	93.6%	92.9%	2.4	7.0	16.7	0.40	0.79
sci	85.4%	93.5%	2.7	8.0	19.6	0.06	0.15	83.1%	90.4%	2.7	9.7	23.2	1.01	1.62
soc	94.2%	94.4%	1.8	6.5	16.9	0.03	0.15	90.2%	91.5%	2.4	10.0	29.5	0.39	0.78
talk	88.5%	92.1%	2.5	7.8	23.8	0.08	0.21	86.8%	90.0%	2.0	10.5	28.5	1.30	2.90

Appendix B

A Word on Scaling Up

Let us first consider a linear model. For a document with m_D unique words, SEDC evaluates sequentially m_D "documents" (the original document with one word removed), then iteratively works on the best of these, leading to the evaluation of $m_D - 1$ documents (the original with two words removed); next $m_D - 2$ documents are evaluated, and so on. When an explanation of size s is found a total of $O(s \times m_D)$ evaluations have occurred. The computational complexity depends, therefore, on (1) the time needed for model evaluation (sometimes very fast, sometimes not so), (2) the number of words needed for an explanation s, which in our case study went to about 50, and (3) the number of unique words in the document m_D, which is generally very small as compared to the overall vocabulary. Most importantly, the computational complexity is independent of the overall size of the vocabulary, unlike previous instance-level explanation approaches. This complexity could be lowered further for linear models to $O(s)$ by incrementally evaluating the word combinations with the next most highly ranked word removed (recall Lemma 1 and Theorem 1). Our implementation does not include this speed-up mechanism in order to present a technique applicable to all models and not just to linear ones.

For a nonlinear model, the heuristic search will likely backtrack; a better local improvement may be found elsewhere. The extent to which this occurs depends on the shape of the model's decision boundary. In the worst case scenario, backtracking over all words occurs, leading to $m_D + {m_D}^{m_D}$ evaluations. Thus, for nonlinear models the worst-case complexity grows exponentially with the depth of the search tree.

Appendix C

Some Additional Related Work

The goal of the present approach seems similar to that of inverse classification (Mannino and Koushik 2000). However, the definition of an explanation, the specific optimization problem, and the search algorithms are all quite different. First, for document classification, we should only consider reducing the values for the corresponding variables. Increasing the value of variables does not make sense. Second, we don't need to decide on step sizes for changes in the values, as removing the occurrences of a word corresponds to setting the value to zero. In the optimization routine of inverse classification, the search problem is exactly to find the minimal distance for each dimension. The optimization is completely different for explanations of documents' classifications, as we will discuss next. Third, applying inverse classification approaches to document classification generally is not feasible, due to the huge dimensionality of these data sets. Our approach takes advantage of the sparseness of document representations, and only needs to consider those words actually present in the document. Fourth, we provide a general framework to obtain explanations independent of the classification technique.

Finally, note the link with *K*- (different from the *k* above) nearest neighbor (KNN) approaches. If such a technique is used as classification method (see D'Silva et al. 2011; Han et al. 2001), showing these *K*-nearest neighbors and their classes "explains" why the model chose that classification. This technical "explanation" notwithstanding, the comprehensibility of such classification models is disputable. What is it exactly about the present document that makes it most similar to a set of documents that yield the predicted class? The KNN technique does not tell us. If the document had been slightly different would it simply be closer to a different set of documents that yields the same predicted class? In "Hyper-Explanations Are Necessary," we discuss how showing the nearest neighbor(s) as an explanation for the classification made by *any* type of model can be used as secondary support for an explanation, for example, showing training data that may have been mislabeled and led a model to make erroneous classifications (see hyper-explanation 3 in the article). This can help us to improve a model if the explanation reveals an error.

References

D'Silva, S., Joshi, N., Rao, S., Venkatraman, S., and Shrawne, S. 2011. "Improved Algorithms for Document Classification and Query-Based Multi-Document Summarization," Journal of Engineering and Technology (3:4), pp. 404-409.

Han, E-H., Karypis, G., and Kumar, V. 2001. "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," in *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, London: Springer-Verlag, pp. 53-65.

Lang, K. 1995. "Newsweeder: Learning to Filter Netnews," in *Proceedings of the 12th International Conference on Machine Learning*, A. Prieditis and S. J. Russell (eds.), San Francisco: Morgan Kaufmann, pp. 331-339.

Mannino, M., and Koushik, M. 2000. "The Cost-Minimizing Inverse Classification Problem: A Genetic Algorithm Approach," *Decision Support Systems* (29:3), pp. 283-300.