



Machine Learning in Practice Project Advice

Carolyn Penstein Rosé

*Language Technologies Institute/
Human-Computer Interaction
Institute*



Yingbo et al., paper



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis



What Makes a Good Paper

- **Argue why the problem is important, what your big result is, and what you learned about your application area**
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis

Application: Staff Assignment



<http://www.lewrockwell.com/rogers/line.jpg>

- Staff assignment normally performed manually
- Poor assignment of human resources causes unnecessary expenses



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- **Summarize your approach**
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis

Background on Task

- **Workflow model**: a plan that can be executed
 - Composed of **activities** and **tasks**
- Each activity is performed by one or more **actors**
- An **event entry** records that the activity was accomplished
- Event entries are recorded in an **event log**
- One path through the plan is called a **case** or **workflow instance**



<http://ec.europa.eu/enterprise/regulation/goods/images/cars.jpg>

	Start Date	End Date	Workflow Models	Activities	Workflow Instances	Event Entries	Actors	Related Organizations
A	May-30-2005	Jul-26-2006	24	399	4,005	42,099	244	29
B	Oct-31-2003	Jun-06-2006	49	922	8,612	99,765	147	28

Setting Up the Problem

- Raw data: an event log that shows over time who executes which activities
 - Each event is an instance
 - Class attribute: nominal, names of people who are potential assignees
 - Features: “event entries of those completed activities in the same workflow instance”
 - Representation is like a word vector where each activity is like a word
- Suitable activities are ones that are executed many times and have many potential actors
 - 22 activities from enterprise A and 35 from B

Table 2 Statistics about activities

		Actor Count			Activity Log Entry Count		
Range		1-3	4-10	10+	1-300	301-1000	1000+
Activity Count	A	306	59	34	368	19	12
	B	857	50	15	857	54	11



More on Feature Space Design

- Information in event log “actor’s identity, data contained in the workflow and structure information”
 - Too vague to see what they are doing here
 - Referring to documents that people access while executing the tasks
 - Later they mention social network information and expertise



More on Feature Space Design

- Did separate experiments for each workflow model
 - Each instance is a vector representing the set of activities that were executed by an actor to achieve a goal as part of a case
 - Ordering information must be lost (“bag of activities”)
 - Basically, you are identifying difference characteristic styles of executing a plan that are associated with different actors



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- **Justify your evaluation metrics, corpus, gold standard, and baselines**
- **Present your results**
- Discuss your error analysis

Why did they pick SVM?

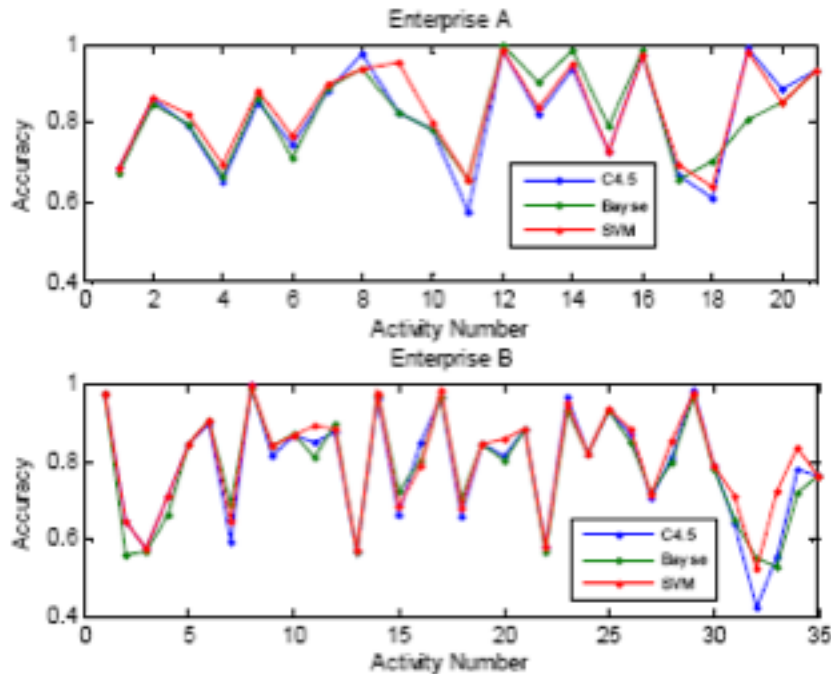


Table 3 Average prediction accuracy

	C4.5	Naïve Bayes	SVM
Enterprise A	85.41%	86.24%	85.82%
Enterprise B	77.04%	76.64%	80.06%

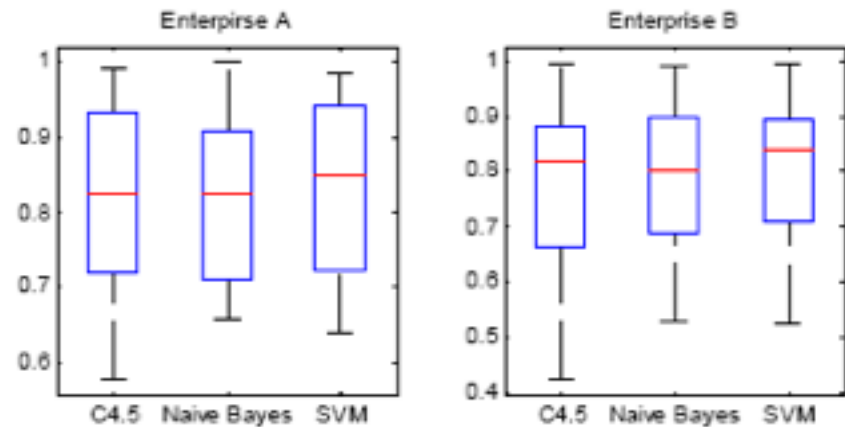


Figure 2 Box plot of accuracy in both enterprises

- No significant difference on average
- Decided to go with the most consistent algorithm

Application: Staff Assignment



<http://www.asia-pacific-connections.com/images/Car-manufacturing.jpg>

- Results for car manufacturing are 85.5% for one company and 80.1% for another
 - But what is the kappa?
 - How hard is the problem?
 - What kinds of mistakes does it make?
 - What does accuracy mean here?

What Makes a Good Paper

- 😊 Argue why the problem is important, what your big result is, and what you learned about your application area
- 😊 Explain prior work, where previous approaches “went wrong”, and where your approach fits
- 😐 Summarize your approach
- 😐 Justify your evaluation metrics, corpus, gold standard, and baselines
- 😐 Present your results
- 😞 Discuss your error analysis


Take Home Message

■ Positives

- Real world application of machine learning
- Evaluation on realistic dataset
- Related work discussion points out what is the unique contribution of this work
 - But the latter portion of the discussion sounds like a “laundry list”
- Lots of consideration of feature space design
 - Not sure feature space design makes sense for the task though
 - *is it learning the right thing?*

■ Negatives

- Evaluation seems weak – and no error analysis
- Writing is not clear
 - You should try to explain things more clearly than this



How would you have
approached the
problem differently?



Stumpf et al., paper



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis



What Makes a Good Paper

- **Argue why the problem is important, what your big result is, and what you learned about your application area**
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis



Research Goals

- Email classification: hard classification problem
- Collaboration between machine learning and humans
 - Getting advice from users
 - Making machine learning transparent
 - Increasing trust
- User centered evaluation methodology
 - Use think aloud protocols to evaluate alternative approaches to offering explanations to users
- Users are able to offer “valuable” feedback
 - Machine learning technology needs to be extended in order to take advantage of it



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- **Explain prior work, where previous approaches “went wrong”, and where your approach fits**
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis



Related Work

- Other related work on “collaborative” machine learning
 - Offering explanation of predictions
 - Getting feedback
 - Lack of work integrating the two
- Making predictions more transparent to users increases preference and trust

Related Work

- Obtaining user feedback
 - **Typical:** users indicate whether a prediction is right or what the correct prediction would have been
 - **New:**
 - Algorithm explains how it made its prediction
 - User offers detailed feedback on how performance could be improved



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- **Summarize your approach**
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- Discuss your error analysis

Think Aloud Study

- External validity: use real world data, but not the user's own data
- Machine learning algorithms sorted emails from the ENRON corpus into folders and “explained their reasoning”
 - Three different explanation approaches: rule based, keyword based, similarity based
 - Users were free to give feedback in any form
- Evaluate user feedback: willingness, accuracy, demonstrated understanding of computer explanations, usability by computer algorithms



Procedure

- Offline preparation for study:
 - Classify all emails
 - Generate three forms of explanation for each email
- Within subject manipulation – every subject exposed to every form of explanation
 - Approach selected randomly for each message
- Subjects were graduate students (6 males and 7 females)
- Native English speakers with computer experience but not computer science majors
- Paper prototypes, so they would not give the impression of a finished system
- Users could rearrange piles of “classified” email
- Questionnaire data every 15 minutes and at the end to assess effort and preference

Explanation Generation

- Used JRIP and Naïve Bayes
- Popular for email classification
- Decent performance, though not great
 - Rule based feedback generated by highlighting the relevant JRIP rule
 - Keywords from Naïve Bayes model –words with strongest positive and negative evidence
 - Similarity based by finding the message that would have had the biggest impact on the prediction of the message if it was removed from the training set

Rule Based Approach

Resume

From: toni.graham@enron.com
To: daren.farmer@enron.com
Subject: re: job posting

Daren, is this position budgeted and who does it report to?
Thanks,
Toni Graham

The reason the system thinks that this email message belongs to folder "Resume" is because the highest priority rule that fits this email message was:

- Put the email in folder "Resume" if:
It's from toni.graham@enron.com.

The other rules in the system are:

- Put the email in folder "Personal" if:
The message does not contain the word "Enron" and
The message does not contain the word "process" and
The message does not contain the word "term" and
The message does not contain the word "link".
- Put the email in folder "Enron News" if:
No other rule applies.

Figure 3: (Top): Email.
(Bottom): Rule-based explanation excerpt.

Keyword Based Approach

Personal

From: buylow@houston.rr.com
To: j.farmer@enron.com
Subject: life in general

Good **god** -- where do you find time for all of that? You should w...

By the way, what is your new address? I may want to come by ...
your work sounds **better** than anything on TV.

You will make a good trader. Good relationships and flexible pri...
a few **zillion** other intangibles you will run into. It beats the hell o...
other **things**.

I'll let you be for now, but do keep those stories coming we **love**...

The reason the system thinks that this email message belongs to folder "Personal" is because it found the following top 5 words in the email message:

1.	ll
2.	love
3.	better
4.	things
5.	god

But if the following words were not in the message, it would be more sure the email message really goes here.

1.	keep
2.	find
3.	trader
4.	book
5.	general

Figure 4: (Top): Excerpt from email.
(Bottom): Keyword-based explanation, supplementing the highlights in the email.

Similarity Based Approach

Resume

Message #2
From: 4Denron@enron.com
To: All ENW employees
Subject: enron net works t&e policy
From: Greg Piper and Mark Pickering

Please print and become familiar with the updated ENW T&E P...
business-first travel, with supervisor approval, for International fl...
Mexico). Supervisors will be responsible for making the decision...

If you have any questions about the policy or an expense not co...
Costello.

Wow! The message is really similar to the message #3 in "Resume"
because #2 and #3 have important words in common.

Message #3
From: toni.graham@enron.com
To: lisa.csikos@enron.com, rita.wynne@enron.com,
daren.farmer@enron.com
CC: renda.herod@enron.com
Subject: confirming requisitions

Confirming the open requisitions for your group. If your records
indicate otherwise, please let me know.

Lisa Csikos 104355, 104001
Rita Wynne 104354
Daren Farmer 104210
Mike Elben 104323
Pat Clynes 104285

The posting dates have all been **updated** to reflect a current
posting date.
Thanks for your support!!
Toni

Figure 5: (Top): Excerpt from email.
(Bottom): Its Similarity-based explanation.



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- **Justify your evaluation metrics, corpus, gold standard, and baselines**
- Present your results
- Discuss your error analysis

Qualitative Analysis of User Comments

- Designed two coding schemes
 - First identify positive and negative comments
 - Then focus on finer grained distinctions in negative comments – investigating how these can be used by the algorithms
- Assessed inter-rater reliability

Code	Description	Example from data	Count (% of total)
Breakdown	Expressing confusion or lack of understanding with the explanation of the algorithm.	I don't understand why there is a second email.	41 (8%)
Understand	Explicitly showing evidence of understanding the explanation of the algorithm.	I see why it used "Houston" as negative	85 (17%)
Emotion	Expressing emotions.	It's funny to me.	15 (3%)
Trust	Stating that he or she trusted the system.	I would probably trust it if I was doing email.	1 (<1%)
Expectation	Expressing an expectation for the system to behave in a certain way.	I hope that eventually the intelligent assistant would learn to give more reasons.	2 (<1%)
Suggest change	Correcting the explanations or otherwise suggesting changes to the system's reasoning.	Different words could have been found in common, like "Agreement," "Ken Lay."	161 (32%)
Negative comment	Making negative comments about the explanation (without suggesting an improvement).	...arbitrary words: "energy" especially bad.	100 (20%)
Positive comment	Making positive comments about the explanation.	The Resume rules are good.	94 (19%)

Table 1: The main coding scheme.



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- **Present your results**
- Discuss your error analysis



Summary of Findings

- Users understood and preferred the rule based approach most, based on the questionnaire data
 - But enough liked the other approaches that sticking with just one does not seem advisable
 - Users liked the “conversational style” of similarity based feedback
 - Users expected the explanations to seem sound and reasonable
- Task based explanation: User’s generate a classification based on explanation
 - Most trouble with similarity based
- When asked to provide an explanation, users typically used a rule based approach



Counterintuitive finding!

- Some users were slower when the prediction was correct!
- The better the accuracy, the harder it may be for users to spot the flaws
- This is why it's important to user test!!!

Most Common Suggestions for Keyword Based Explanations

- Adjust weights
 - Flip polarity
 - Consider different words
 - Generalization across words
 - Syntactic and semantic variants
 - Feature combinations
 - Meta data: from line, email threads
-
- ***Discussion Question:*** How would you do this?
And what effect do you think it would have?

User Feedback

	KB-English	KB-common-sense	KB-domain	KB-other	Total	%
1. Adjust weight	11	11	4	13	39	12%
2. Select different features (words)	70	64	25	16	175	53%
3. Parse or extract in a different way	7	17	10	0	34	10%
4. Employ feature combinations	9	5	2	1	17	5%
5. Relational features	0	9	5	0	14	4%
6. Other	3	12	4	33	52	16%
Total	100	118	50	63	331	
%	30%	36%	15%	19%		

Table 4: Types of participants' changes (in rows) that required various background knowledge (in columns).



What Makes a Good Paper

- Argue why the problem is important, what your big result is, and what you learned about your application area
- Explain prior work, where previous approaches “went wrong”, and where your approach fits
- Summarize your approach
- Justify your evaluation metrics, corpus, gold standard, and baselines
- Present your results
- **Discuss your error analysis**

Error Analysis



What Makes a Good Paper

- 👤 Argue why the problem is important, what your big result is, and what you learned about your application area
- 😊 Explain prior work, where previous approaches “went wrong”, and where your approach fits
- 😊 Summarize your approach
- 😊 Justify your evaluation metrics, corpus, gold standard, and baselines
- 😊 Present your results
- 😞 Discuss your error analysis



Take Home Message

- Interesting use of user study data in connection with machine learning
 - Recent focus on how people use machine learning models and interact with machine learning algorithms
- Error analysis is most commonly lacking in published work
 - But I want to see that in your work!